WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at
http://www.mrw.interscience.wiley.com/eot

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor.
p. cm.
includes index.
ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G.
TK5102 .W55 2002
621.382′03 — dc21

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

I am pleased to welcome the readers to the *Wiley Encyclopedia of Telecommunications*. The Encyclopedia contains 275 tutorial articles focused on modern telecommunications topics. The contents include articles on communication networks, source coding and decoding, channel coding and decoding, modulation and demodulation, optical communications, satellite communications, underwater acoustic communications, radio propagation, antennas, multiuser communications, magnetic storage systems, and a variety of standards. Additional details on these topics are given below. The authors of these articles were selected for their expertise and leadership in their respective areas in the field of telecommunications. All of the authors possess advanced graduate degrees and have published widely in prestigious international journals and conferences.

COMMUNICATION NETWORKS

There are approximately 60 articles on the subject of communication networks, including several articles on protocols, such as the wireless application protocol (WAP) and MAC protocols; network flow; congestion control; admission control broadband integrated digital networks; local area networks and standards; satellite networks; network reliability and fault tolerance; DWDM ring networks, wireless ad hoc networks; and multi-protocol label switching (MPLS).

MODULATION AND DEMODULATION

There are over 40 articles covering various basic modulation and demodulation techniques, including analog amplitude modulation (AM), frequency modulation (FM) and phase modulation; digital modulation techniques, namely, pulse amplitude modulation (PAM); phase-shift keying (PSK), quadrature amplitude modulation (QAM), continuous-phase modulation (CPM); continuous-phase frequency shift-keying (CPFSK); partial response signals; spread spectrum modulation; adaptive equalization; turbo equalization; and orthogonal frequency-division multiplexing (OFDM).

OPTICAL COMMUNICATIONS

There are approximately 30 articles about optical communication, including articles on optical modulation; optical detectors; optical amplifiers; optical correlators; optical filters; photonic A/D conversion; optical transport; optical multiplexers and demultiplexers, optical switching; and characterization of optical fibers.

ANTENNAS

Various types of antennas and antenna arrays are described in 10 articles, including parabolic antennas; microstrip antennas; waveguide antennas; television antennas; loop antennas; horn antennas; leaky wave antennas; and helical and spiral antennas.

PROPAGATION

Six articles are devoted to electromagnetic radio signal propagation, including propagation at very low frequencies (VLF), low frequencies (LF), medium frequencies (MF), high frequencies (HF), very high frequencies (VHF), microwave frequencies, and millimeter wave frequencies.

CHANNEL CODING AND DECODING

Approximately 35 articles cover various channel codes and decoding algorithms, including BCH codes; convolutional codes; concatenated codes; trellis codes; space-time codes; turbo codes; Gold codes; Kasami codes; Golay codes; finite geometry codes; codes for magnetic recording channels; Viterbi decoding algorithm; and sequential decoding algorithm.

SOURCE CODING AND DECODING

Eight articles cover various data compression and source coding and decoding methods, including waveform coding techniques such as pulse code modulation (PCM) and differential PCM (DPCM); linear predictive coding (LPC); Huffman coding; and high definition television (HDTV).

MULTIUSER COMMUNICATION

There are 12 articles focused on multiuser communications, including multiple access techniques such as code-division multiple access (CDMA), frequency-division multiple access (FDMA), time-division multiple access (TDMA), and carrier-sense multiple access (CSMA); Ethernet technology; multiuser detection algorithms; and third-generation (3G) digital cellular communication systems.

ACOUSTIC COMMUNICATIONS

There are 5 articles on acoustic communications dealing with acoustic transducers; underwater acoustic communications and telemetry; underwater acoustic moderns, and acoustic echo cancellation.

SATELLITE COMMUNICATIONS

Two articles focus on geosynchronous satellite communications and on low-earth-orbit (LEO) and medium-earth-orbit (MEO) satellite communication systems.

John G. Proakis, Editor
Northeastern University

CONTRIBUTORS

Behnaam Aazhang, *Rice University, Houston, Texas,* Multiuser Wireless Communication Systems

Ali N. Akansu, *New Jersey Institute of Technology, Newark, New Jersey,* Orthogonal Transmultiplexers: A Time-Frequency Perspective

Nail Akar, *Bilkent University, Ankara, Turkey,* BISDN (Broadband Integrated Services Digital Network)

Arda Aksu, *North Carolina State University, Raleigh, North Carolina,* Unequal Error Protection Codes

Naofal Al-Dhahir, *AT&T Shannon Laboratory, Florham Park, New Jersey,* Space–Time Codes for Wireless Communications

Edward E. Altshuler, *Electromagnetics Technology Division, Hanscom AFB, Massachusetts,* Millimeter Wave Propagation

Abeer Alwan, *University of California at Los Angeles, Los Angeles, California,* Speech Coding: Fundamentals and Applications

Moeness G. Amin, *Villanova University, Villanova, Pennsylvania,* Interference Suppression in Spread-Spectrum Communication Systems

John B. Anderson, *Lund University, Lund, Sweden,* Continuous-Phase-Coded Modulation

Alessandro Andreadis, *University of Siena, Siena, Italy,* Wireless Application Protocol (WAP)

Peter Andrekson, *Chalmers University of Technology, Göthenburg, Sweden,* Optical Solitons

Oreste Andrisano, *University of Bologna, DEIS, Italy,* Communications for Intelligent Transportation Systems

A. Annamalai, *Virginia Tech, Blacksburg, Virginia,* Wireless Communications System Design

Cenk Argon, *Georgia Institute of Technology, Atlanta, Georgia,* Turbo Product Codes for Optical CDMA Systems

Hüseyin Arslan, *Ericsson Inc., Research Triangle Park, North Carolina,* Channel Tracking in Wireless Communication Systems

Tor Aulin, *Chalmers University of Technology, Göteborg, Sweden,* Serially Concatenated Continuous-Phase Modulation with Iterative Decoding

James Aweya, *Nortel Networks, Ottawa, Ontario, Canada,* Transmission Control Protocol

Ender Ayanoglu, *University of California, Irivne, California,* BISDN (Broadband Integrated Services Digital Network)

Krishna Balachandran, *Lucent Technologies Bells Labs, Holmdel, New Jersey,* Wireless Packet Data

Constantine A. Balanis, *Arizona State University, Tempe, Arizona,* Antennas

Stella N. Batalama, *State University of New York at Buffalo, Buffalo, New York,* Packet-Rate Adaptive Receivers for Mobile Communications

Rainer Bauer, *Munich University of Technology (TUM), Munich, Germany,* Digital Audiobroadcasting

S. Benedetto, *Politecnico di Torino, Torino (Turin), Italy,* Serially Concatenated Codes and Iterative Algorithms

Toby Berger, *Cornell University, Ithaca, New York,* Rate-Distortion Theory

Steven Bernstein, *MIT Lincoln Laboratory, Lexington, Massachusetts,* Communication Satellite Onboard Processing

Claude Berrou, *ENST Bretagne, Brest, France,* Turbo Codes

Randall Berry, *Northwestern University, Evanston, Illinois,* Information Theory

H. L. Bertoni, *Polytechnic University, Brooklyn, New York,* Path Loss Prediction Models in Cellular Communication Channels

Christian Bettstetter, *Technische Universität München, Institute of Communication Networks, Munich, Germany,* General Packet Radio Service (GPRS); GSM Digital Cellular Communication System

Ravi Bhagavathula, *Wichita State University, Wichita, Kansas,* Modems

Andrea Bianco, *Politecnico di Torino, Torino (Turin), Italy,* Multimedia Networking

Jayadev Billa, *BBN Technologies, Cambridge, Massachusetts,* Speech Recognition

Bjørn A. Bjerke, *Qualcomm, Inc., Concord, Massachusetts,* Pulse Amplitude Modulation

Fletcher A. Blackmon, *Naval Undersea Warfare Center Division Newport, Newport, Rhode Island,* Acoustic Telemetry

Ian F. Blake, *University of Toronto, Ontario, Toronto, Canada,* Cryptography

Martin Bossert, *University of Ulm, Ulm, Germany,* Hadamard Matrices and Codes

Gregory E. Bottomley, *Ericsson Inc., Research Triangle Park, North Carolina,* Channel Tracking in Wireless Communication Systems

Torsten Braun, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Madhukar Budagavi, *Texas Instruments, Incorporated, Dallas, Texas,* Wireless MPEG-4 Videocommunications

Kenneth Budka, *Lucent Technologies Bells Labs, Holmdel, New Jersey,* Wireless Packet Data

R. Michael Buehrer, *Virginia Tech, Blacksburg, Virginia,* Mobile Radio Communications

Julian J. Bussgang, *Signatron Technology Corporation, Concord, Massachusetts,* HF Communications

Jens Buus, *Gayton Photonics, Gayton, Northants, United Kingdom,* Optical Sources

Søren Buus, *Northeastern University, Boston, Massachusetts,* Speech Perception

Maja Bystrom, *Drexel University, Philadelphia, Pennsylvania,* Image Processing

Henning Bülow, *Optical Systems, Stuttgart, Germany,* Polarization Mode Dispersion Mitigation

A. R. Calderbank, *AT&T Shannon Laboratory, Florham Park, New Jersey,* Space–Time Codes for Wireless Communications

Gilberto M. Camilo, *OmniGuide Communications, Cambridge, Massachusetts,* Characterization of Optical Fibers

G. Cariolaro, *Università di Padova, Padova, Italy,* Pulse Position Modulation

Jeffrey B. Carruthers, *Boston University, Boston, Massachusetts,* Wireless Infrared Communications

John H. Carson, *George Washington University, Washington, District of Columbia,* Local Area Networks

Anne Cerboni, *France Télécom R&D, Issy Moulineaux, France,* IMT-2000 3G Mobile Systems

Kavitha Chandra, *Center for Advanced Computation and Telecommunications, University of Massachusetts Lowell, Lowell, Massachusetts,* Statistical Multiplexing

Sekchin Chang, *University of Texas at Austin, Austin, Texas,* Compensation of Nonlinear Distortion in RF Power Amplifiers

Matthew Chapman Caesar, *University of California at Berkeley, Berkeley, California,* IP Telephony

Jean-Pierre Charles, *France Télécom R&D, Issy Moulineaux, France,* IMT-2000 3G Mobile Systems

Chi-Chung Chen, *University of California, Los Angeles, California,* Chaos in Communications

Po-Ning Chen, *National Chi Tung University, Taiwan,* Sequential Decoding of Convolutional Codes

Thomas M. Chen, *Southern Methodist University, Dallas, Texas,* ATM Switching

Zhizhang (David) Chen, *Dalhousie University, Halifax, Nova Scotia, Canada,* Millimeter-Wave Antennas

Andrew R. Chraplyvy, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Nonlinear Effects in Optical Fibers

Christos G. Christodoulou, *University of New Mexico, Albuquerque, New Mexico,* Antennas for Mobile Communications

Michael T. Chryssomallis, *Democritus University of Thrace, Xanthi, Greece,* Antennas for Mobile Communications

Keith M. Chugg, *University of Southern California, Los Angeles, California,* Iterative Detection Algorithms in Communications

Habong Chung, *Hongik University, Seoul, Korea,* Gold Sequences

Leonard J. Cimini Jr., *AT&T Labs-Research, Middletown, New Jersey,* Orthogonal Frequency-Division Multiplexing

J. Cioffi, *Stanford University, Stanford, California,* Very High-Speed Digital Subscriber Lines (VDSLs)

Wim M. J. Coene, *Philips Research Laboratories, Eindhoven, The Netherlands,* Constrained Coding Techniques for Data Storage

Robert A. Cohen, *Troy, New York,* Streaming Video

Giovanni Emanuele Corazza, *University of Bologna, Bologna, Italy,* cdma2000

Steven Cummer, *Duke University, Durham, North Carolina,* Extremely Low Frequency (ELF) Electromagnetic Wave Propagation

Milorad Cvijetic, *NEC America, Herndon, Virginia,* Optical Transport System Engineering

Nelson L. S. da Fonseca, *Institute of Computing, State University of Campinas Brazil,* Bandwidth Reduction Techniques for Video Services; Network Traffic Modeling

Dirk Dahlhaus, *Communication Technology Laboratory, Zurich, Switzerland,* Chirp Modulation

Roger Dalke, *Institute for Telecommunication Sciences, Boulder, Colorado,* Local Multipoint Distribution Services (LMDS)

Marc Danzeisen, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Pankaj K. Das, *University of California, San Diego, La Jolla, California,* Surface Acoustic Wave Filters

Héctor J. De Los Santos, *Coventor, Inc., Irvine, California,* MEMS for RF/Wireless Applications

Filip De Turck, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Piet Demeester, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Jing Deng, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

Michael Devetsikiotis, *North Carolina State University, Raleigh, North Carolina,* Network Traffic Modeling

Olufemi Dosunmu, *Boston University, Boston, Massachusetts,* High-Speed Photodetectors for Optical Communications

Alexandra Duel-Hallen, *North Carolina State University, Raleigh, North Carolina,* Fading Channels

Tolga M. Duman, *Arizona State University, Tempe, Arizona,* Interleavers for Serial and Parallel Concatenated (Turbo) Codes

K. L. Eddie Law, *University of Toronto, Toronto, Canada,* Optical Switches

Thomas F. Eibert, *T-Systems Nova GmbH, Technologiezentrum, Darmstadt, Germany,* Antenna Modeling Techniques

Evangelos S. Eleftheriou, *IBM Zurich Research Laboratory, Rueschlikon, Switzerland,* Signal Processing for Magnetic Recording Channels

Amro El-Jaroudi, *University of Pittsburgh, Pittsburgh, Pennsylvania,* Linear Predictive Coding

Matthew Emsley, *Boston University, Boston, Massachusetts,* High-Speed Photodetectors for Optical Communications

T. Erseghe, *Università di Padova, Padova, Italy,* Pulse Position Modulation

Sonia Fahmy, *Purdue University, West Lafayette, Indiana,* Network Traffic Management

David R. Famolari, *Telecordia Technologies, Morristown, New Jersey,* Wireless IP Telephony

Li Fan, *OMM, Inc., San Diego, California,* Optical Crossconnects

András Faragó, *University of Texas at Dallas, Richardson, Texas,* Medium Access Control (MAC) Protocols

Aiguo Fei, *University of California at Los Angeles, Los Angeles, California,* Multicast Algorithms

Robert J. Filkins, *University of California, San Diego, La Jolla, California,* Surface Acoustic Wave Filters

John P. Fonseka, *University of Texas at Dallas, Richardson, Texas,* Quadrature Amplitude Modulation

M. Fossorier, *University of Hawaii at Manoa, Honolulu, Hawaii,* Finite-Geometry Codes

Roger Freeman, *Independent Consultant, Scottsdale, Arizona,* Community Antenna Television (CATV) (Cable Television); Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH)

Fabrizio Frezza, *"La Sapienza" University of Rome, Roma, Italy,* Leaky-Wave Antennas

Thomas E. Fuja, *University of Notre Dame, Notre Dame, Indiana,* Automatic Repeat Request

Alessandro Galli, *"La Sapienza" University of Rome, Roma, Italy,* Leaky-Wave Antennas

Costas N. Georghiades, *Texas A&M University, College Station, Texas,* EM Algorithm in Telecommunications

Leonidas Georgiadis, *Aristotle University of Thessaloniki, Thessaloniki, Greece,* Carrier-Sense Multiple Access (CSMA) Protocols

Mario Gerla, *University of California at Los Angeles, Los Angeles, California,* Multicast Algorithms

Pierre Ghandour, *France Télécom R&D, South San Franciso, California,* IMT-2000 3G Mobile Systems

K. Ghorbani, *RMIT University, Melbourne, Australia,* Microstrip Patch Arrays

Dipak Ghosal, *University of California at Davis, Davis, California,* IP Telephony

Giovanni Giambene, *University of Siena, Siena, Italy,* Wireless Application Protocol (WAP)

Arthur A. Giordano, *AG Consulting Inc., LLC, Burlington, Massachusetts,* Statistical Characterization of Impulsive Noise

Stefano Giordano, *University of Pisa, Pisa, Italy,* Multimedia Networking

Alain Glavieux, *ENST Bretagne, Brest, France,* Turbo Codes

Savo G. Glisic, *University of Oulu, Oulu, Finland,* Cochannel Interference in Digital Cellular TDMA Networks

Dennis L. Goeckel, *University of Massachusetts, Amherst, Massachusetts,* Bit-Interleaved Coded Modulation

Virgilio E. Gonzalez-Lozano, *Department of Electrical and Computer Engineering, University of Texas at El Paso, El Paso, Texas,* Optical Fiber Local Area Networks

Vivek Goyal, *Digital Fountain Inc., Fremont, California,* Transform Coding

Larry J. Greenstein, *AT&T Labs-Research, Middletown, New Jersey,* Orthogonal Frequency-Division Multiplexing

Marcus Greferath, *San Diego State University, San Diego, California,* Golay Codes

Manuel Günter, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Jaap C. Haartsen, *Ericsson Technology Licensing AB, Emmen, The Netherlands,* Bluetooth Radio System

Zygmunt J. Haas, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

David Haccoun, *Ecole Polytechnique de Montréal, Montréal, Quebec, Canada,* High-Rate Punctured Convolutional Codes

Abdelfatteh Haidine, *Dresden University of Technology, Dresden, Germany,* Powerline Communications

M. Hajian, *Delft University of Technology, Delft, The Netherlands,* Microwave Waveguides

Mounir Hamdi, *Hong Kong University of Science and Technology, Hong Kong,* Multimedia Medium Access Control Protocols for WDM Optical Networks

Yunghsiang S. Han, *National Chi Yan University, Taiwan,* Sequential Decoding of Convolutional Codes

Marc Handlery, *Lund University, Lund, Sweden,* Tailbiting Convolutional Codes

Eberhard Hänsler, *Darmstadt University of Technology, Darmstadt, Germany,* Acoustic Echo Cancellation

Fred Harris, *San Diego State University, San Diego, California,* Sigma-Delta Converters in Communication Systems

Christian Hartmann, *Technische Universität München, Institute of Communication Networks, Munich, Germany,* General Packet Radio Service (GPRS); GSM Digital Cellular Communication System

Mark Hasegawa-Johnson, *University of Illinois at Urbana-Champaign, Urbana, Illinois,* Speech Coding: Fundamentals and Applications

Homayoun Hashemi, *Sharif University of Technology, Teheran, Iran,* Wireless Local Loop Standards and Systems

Dimitrios Hatzinakos, *University of Toronto, Toronto, Ontario, Canada,* Spatiotemporal Signal Processing in Wireless Communications

Michelle C. Hauer, *University of Southern California, Optical Communications Laboratory, Los Angeles, California,* Digital Optical Correlation for Fiberoptic Communication Systems

Simon Haykin, *McMaster University, Hamilton, Ontario, Canada,* Maximum-Likelihood Estimation

Da-ke He, *University of Waterloo, Waterloo, Ontario, Canada,* Huffman Coding

Juergen Heiles, *Siemens Information & Communication Networks, Munich, Germany,* DWDM Ring Networks

Tor Helleseth, *University of Bergen, Bergen, Norway,* Ternary Sequences

Thomas R. Henderson, *Boeing Phantom Works, Seattle, Washington,* Leo Satellite Networks

Naftali Herscovici, *Anteg, Inc., Framingham, Massachusetts,* Microstrip Antennas

Pin-Han Ho, *Queen's University at Kingston, Ontario, Canada,* Survivable Optical Internet

Henk D. L. Hollmann, *Philips Research Laboratories, Eindhoven, The Netherlands,* Constrained Coding Techniques for Data Storage

R. Hoppe, *Institut Fuer Hochfrequenztechnik, University of Stuttgart, Stuttgart, Germany,* Propagation Models for Indoor Communications

Jiongkuan Hou, *New Jersey Institute of Technology, University Heights, Newark, New Jersey,* Admission Control in Wireless Networks

Halid Hrasnica, *Dresden University of Technology, Dresden, Germany,* Powerline Communications

Laura L. Huckabee, *Time Domain Corporation, Huntsville, Alabama,* Ultrawideband Radio

Abbas Jamalipour, *University of Sydney, Sydney, Australia,* Satellites in IP Networks

Pertti Järvensivu, *VTT Electronics, Oulu, Finland,* Cellular Communications Channels

Bahram Javidi, *University of Connecticut, Storrs, Connecticut,* Secure Ultrafast Data Communication and Processing Interfaced with Optical Storage

Rolf Johannesson, *Lund University, Lund, Sweden,* Tailbiting Convolutional Codes

Thomas Johansson, *Lund University, Lund, Sweden,* Authentication Codes

Sarah J. Johnson, *University of Newcastle, Callaghan, Australia,* Low-Density Parity-Check Codes: Design and Decoding

Douglas L. Jones, *University of Illinois at Urbana — Champaign, Berkeley, California,* Shell Mapping

Biing-Hwang Juang, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Hidden Markov Models

Edward V. Jull, *University of British Columbia, Vancouver, British Columbia,, Canada,* Horn Antennas

Peter Jung, *Gerhard-Mercator-Universität Duisburg,, Duisburg, Germany,* Time Division Multiple Access (TDMA)

Apostolos K. Kakaes, *Cosmos Communications Consulting Corporation, Centreville, Virginia,* Communication System Traffic Engineering

Dimitris N. Kalofonos, *Northeastern University, Boston, Massachusetts,* Multicarrier CDMA

Güneş Karabulut, *University of Ottawa, School of Information Technology and Engineering, Ottawa, Ontario, Canada,* Waveform Coding

Khalid Karimullah, *Hughes Network Systems, Germantown, Maryland,* Geosynchronous Satellite Communications

Magnus Karlsson, *Chalmers University of Technology, Göthenburg, Sweden,* Optical Solitons

Tadao Kasami, *Hiroshima City University, Hiroshima, Japan,* Kasami Sequences

Timo Kaukoranta, *Turku Centre for Computer Science (TUCS), University of Turku, Turku, Finland,* Scalar and Vector Quantization

Mohsen Kavehrad, *Pennsylvania State University, University Park, Pennsylvania,* Diversity in Communications

Haruo Kawakami, *Antenna Giken Corp., Laboratory, Saitama City, Japan,* Television and FM Broadcasting Antennas

Jürgen Kehrbeck, *Head, Division of e-Commerce and Mobile Communications, LStelcom Lichtenau, Germany,* Cell Planning in Wireless Networks

Gerd Keiser, *PhotonicsComm Solutions, Inc., Newton Center, Massachusetts,* Optical Couplers; Optical Fiber Communications

John Kieffer, *University of Minnesota, Minneapolis, Minnesota,* Data Compression

Dennis Killinger, *University of South Florida, Tampa, Florida,* Optical Wireless Laser Communications: Free-Space Optics

Kyungjung Kim, *Syracuse University, Syracuse, New York,* Adaptive Antenna Arrays

Ryuji Kohno, *Hiroshima City University, Hiroshima, Japan,* Kasami Sequences

Israel Korn, *University of New South Wales, Sydney, Australia,* Quadrature Amplitude Modulation

Sastri Kota, *Loral Skynet, Palo Alto, California,* Trends in Broadband Communication Networks

Hamid Krim, *ECE Department, North Carolina State University Centennial Campus, Raleigh, North Carolina,* Wavelets: A Multiscale Analysis Tool

Frank R. Kschischang, *University of Toronto, Toronto, Canada,* Product Codes

Erozan M. Kurtas, *Seagate Technology, Pittsburgh, Pennsylvania,* Design and Analysis of Low-Density Parity-Check Codes for Applications to Perpendicular Recording Channels

Alexander V. Kuznetsov, *Seagate Technology, Pittsburgh, Pennsylvania,* Design and Analysis of Low-Density Parity-Check Codes for Applications to Perpendicular Recording Channels

Henry K. Kwok, *University of Illinois at Urbana — Champaign, Berkeley, California,* Shell Mapping

Hyuck Kwon, *Wichita State University, Wichita, Kansas,* Modems

Cedric F. Lam, *Opvista Inc., Irvine, California,* Modern Ethernet Technologies

Paolo Lampariello, *"La Sapienza" University of Rome, Roma, Italy,* Leaky-Wave Antennas

F. Landstorfer, *Institut Fuer Hochfrequenztechnik, University of Stuttgart, Stuttgart, Germany,* Propagation Models for Indoor Communications

Greg D. LeCheminant, *Agilent Technologies, Santa Rosa, California,* Test and Measurement of Optically Based High-Speed Digital Communications Systems and Components

Frederick K. H. Lee, *Queen's University, Kingston, Ontario, Canada,* Nonuniformly Spaced Tapped-Delay-Line Equalizers for Sparse Multipath Channels

Jhong Sam Lee, *J.S. Lee Associates, Inc., Rockville, Maryland,* CDMA/IS95

Lin-Nan Lee, *Hughes Network Systems, Germantown, Maryland,* Geosynchronous Satellite Communications

Ralf Lehnert, *Dresden University of Technology, Dresden, Germany,* Powerline Communications

Hanoch Lev-Ari, *Northeastern University, Boston, Massachusetts,* Digital Filters

Allen H. Levesque, *Marlborough, Massachusetts,* BCH Codes — Binary; BCH Codes — Nonbinary and Reed-Solomon

Shipeng Li, *Microsoft Research Asia, Beijing, P.R. China,* Image and Video Coding

Weiping Li, *WebCast Technologies, Inc., Sunnyvale, California,* Image and Video Coding

Ben Liang, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

L. P. Ligthart, *Delft University of Technology, Delft, The Netherlands,* Microwave Waveguides

Jae S. Lim, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* High-Definition Television

Dave Lindbergh, *Polycom, Inc., Andover, Massachusetts,* H.324: Videotelephony and Multimedia for Circuit-Switched and Wireless Networks

K. J. Ray Liu, *University of Maryland, College Park, Maryland,* Multimedia Over Digital Subscriber Lines

Stephen S. Liu, *Verizon Laboratories, Waltham, Massachusetts,* ATM Switching

Alfio Lombardo, *University of Catania, Catania, Italy,* Multimedia Networking

Steven H. Low, *California Institute of Technology, Pasadena, California,* Network Flow Control

M. Luise, *University of Pisa, Dipartimento Ingegneria Informazione, Pisa, Italy,* Synchronization in Digital Communication Systems

Steven S. Lumetta, *University of Illinois Urbana — Champaign, Urbana, Illinois,* Network Reliability and Fault Tolerance

Wei Luo, *Lucent Technologies Bells Labs, Holmdel, New Jersey,* Wireless Packet Data

Maode Ma, *Nanyang Technological University, Singapore,* Multimedia Medium Access Control Protocols for WDM Optical Networks

Rangaraj Madabhushi, *Agere Systems, Optical Core Networks Dividion, Breinigsville, Pennsylvania,* Optical Modulators — Lithium Niobate

Aarne Mämmelä, *VTT Electronics, Oulu, Finland,* Cellular Communications Channels

Elias S. Manolakos, *Northeastern University, Boston, Massachusetts,* Neural Networks and Applications to Communications

Jon W. Mark, *University of Waterloo, Waterloo, Ontario, Canada,* Wideband CDMA in Third-Generation Cellular Communication Systems

Donald P. Massa, *Massa Products Corporation, Hingham, Massachusetts,* Acoustic Transducers

James L. Massey, *Consultare Technology Group, Bethesda, Denmark,* Threshld Decoding

Osamu Matoba, *University of Tokyo, Tokyo, Japan,* Secure Ultrafast Data Communication and Processing Interfaced with Optical Storage

John E. McGeehan, *University of Southern California, Optical Communications Laboratory, Los Angeles, California,* Digital Optical Correlation for Fiberoptic Communication Systems

Peter J. McLane, *Queen's University, Kingston, Ontario, Canada,* Nonuniformly Spaced Tapped-Delay-Line Equalizers for Sparse Multipath Channels

Steven W. McLaughlin, *Georgia Institute of Technology, Atlanta, Georgia,* Turbo Product Codes for Optical CDMA Systems

Donald G. McMullin, *Broadcom Corporation, Irvine, California,* Cable Modems

Muriel Médard, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* Network Reliability and Fault Tolerance

Seapahn Megerian, *University of California at Los Angeles, West Hills, California,* Wireless Sensor Networks

U. Mengali, *University of Pisa, Dipartimento Ingegneria Informazione, Pisa, Italy,* Synchronization in Digital Communication Systems

Lazaros Merakos, *University of Athens, Panepistimiopolis, Athens Greece,* Wireless ATM

Jan De Merlier, *Ghent University — IMEC, Ghent, Belgium,* Optical Signal Regeneration

Alfred Mertins, *University of Wollongong, Wollongong, Australia,* Image Compression

John J. Metzner, *Pennsylvania State University, University Park, Pennsylvania,* Aloha Protocols

Alan R. Mickelson, *University of Colorado, Boulder, Colorado,* Active Antennas

Arnold M. Michelson, *Marlborough, Massachusetts,* BCH Codes — Binary; BCH Codes — Nonbinary and Reed-Solomon

Leonard E. Miller, *Wireless Communications Technologies Group, NIST, Gaithersburg, Maryland,* CDMA/IS95

Mario Minami, *University of São Paulo, São Paulo, Brazil,* Low-Bit-Rate Speech Coding

Joseph Mitola III, *Consulting Scientist, Tampa, Florida,* Software Radio

Urbashi Mitra, *Communication Sciences Institute, Los Angeles, California,* Adaptive Receivers for Spread-Spectrum Systems

Eytan Modiano, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* Wavelength-Division Multiplexing Optical Networks

Peter Monsen, *P.M. Associates, Stowe, Vermont,* Tropospheric Scatter Communication

G. Montorsi, *Politecnico di Torino, Torino (Turin), Italy,* Serially Concatenated Codes and Iterative Algorithms

Tim Moors, *University of New South Wales, Sydney, Australia,* Transport Protocols for Optical Networks

Pär Moqvist, *Chalmers University of Technology, Göteborg, Sweden,* Serially Concatenated Continuous-Phase Modulation with Iterative Decoding

M. Morelli, *University of Pisa, Dipartimento Ingegneria Informazione, Pisa, Italy,* Synchronization in Digital Communication Systems

Geert Morthier, *Ghent University — IMEC, Ghent, Belgium,* Optical Signal Regeneration

Hussein T. Mouftah, *Queen's University at Kingston, Ontario, Canada,* Survivable Optical Internet

Biswanath Mukherjee, *University of California, Davis, Davis, California,* Design and Analysis of a WDM Client/Server Network Architecture

Hannes Müsch, *GN ReSound Corporation, Redwood City, California,* Speech Perception

Rohit U. Nabar, *Stanford University, Stanford, California,* Mimo Communication Systems

Ayman F. Naguib, *Morphics Technology Inc., Campbell, California,* Space–Time Codes for Wireless Communications

Masao Nakagawa, *Keio University, Japan,* Communications for Intelligent Transportation Systems

Hisamatsu Nakano, *Hosei University, Koganei, Tokyo, Japan,* Helical and Spiral Antennas

A. L. Narasimha Reddy, *Texas A&M University, College Station, Texas,* Differentiated Services

Krishna R. Narayanan, *Texas A&M University, College Station, Texas,* Turbo Equalization

Tomoaki Ohtsuki, *Tokyo University of Science, Noda, Chiba, Japan,* Optical Synchronous CDMA Systems

Yasushi Ojiro, *Antenna Giken Corp., Laboratory, Saitama City, Japan,* Television and FM Broadcasting Antennas

Rolf Oppliger, *eSECURITY Technologies Rolf Oppliger, Bern, Switzerland,* Network Security

Alessandro Orfei, *CNR, Istituto di Radioastronomia, Bologna, Italy,* Parabolic Antennas

Douglas O'Shaughnessy, *INRS-Telecommunications, Montreal, Quebec, Canada,* Speech Processing

Tony Ottosson, *Chalmers University of Technology, Goteborg, Sweden,* Signature Sequences for CDMA Communications

Sebnem Ozer, *MeshNetworks, Inc., Orlando, Florida,* Admission Control in Wireless Networks

Ryan A. Pacheco, *University of Toronto, Toronto, Ontario, Canada,* Spatiotemporal Signal Processing in Wireless Communications

K. Pahlavan, *Center for Wireless Information Network Studies Worcester Polytechnic Institute, Worcester, Massachusetts,* Trends in Wireless Indoor Networks

Algirdas Pakštas, *London Metropolitan University, London, England,* Intranets and Extranets

Constantinos B. Papadias, *Global Wireless Systems Research, Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Multiple Antenna Transceivers for Wireless Communications: A Capacity Perspective

Panagiotis Papadimitratos, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

Symeon Papavassiliou, *New Jersey Institute of Technology, University Heights, Newark, New Jersey,* Admission Control in Wired Networks; Admission Control in Wireless Networks

Peter Papazian, *Institute for Telecommunication Sciences, Boulder, Colorado,* Local Multipoint Distribution Services (LMDS)

Matthew G. Parker, *University of Bergen, Bergen, Norway,* Golay Complementary Sequences; Peak-to-Average Power Ratio of Orthogonal Frequency-Division Multiplexing

So Ryoung Park, *Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea,* Polyphase Sequences

Steen A. Parl, *Signatron Technology Corporation, Concord, Massachusetts,* HF Communications

Gianni Pasolini, *University of Bologna, DEIS, Italy,* Communications for Intelligent Transportation Systems

Nikos Passas, *University of Athens, Panepistimiopolis, Athens Greece,* Wireless ATM

Kenneth G. Paterson, *University of London, Egham, Surrey,* Golay Complementary Sequences

Arogyaswami J. Paulraj, *Stanford University, Stanford, California,* Mimo Communication Systems

Fotini-Niovi Pavlidou, *Aristotle University of Thessaloniki, Thessaloniki, Greece,* Frequency-Division Multiple Access (FDMA): Overview and Performance Evaluation

Menelaos K. Perdikeas, *National Technical University of Athens, Athens, Greece,* Distributed Intelligent Networks

Lance C. Perez, *University of Nebraska, Lincoln, Omaha,* Soft Output Decoding Algorithms

Athina P. Petropulu, *Drexel University, Philadelphia, Pennsylvania,* Interference Modeling in Wireless Communications

Stephan Pfletschinger, *Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany,* DMT Modulation

Raymond L. Pickholtz, *George Washington University, Washington, District of Columbia,* Code-Division Multiple Access

Pekka Pirinen, *University of Oulu, Oulu, Finland,* Cochannel Interference in Digital Cellular TDMA Networks

Leon Poladian, *University of Sydney, Eveleigh, Australia,* Optical Filters

Anastasis C. Polycarpou, *Arizona State University, Tempe, Arizona,* Antennas

Dimitrie C. Popescu, *Rutgers WINLAB, Piscataway, New Jersey,* Interference Avoidance for Wireless Systems

Miodrag Potkonjak, *University of California at Los Angeles, West Hills, California,* Wireless Sensor Networks

Edward J. Powers, *University of Texas at Austin, Austin, Texas,* Compensation of Nonlinear Distortion in RF Power Amplifiers

John G. Proakis, *Northeastern University, Boston, Massachusetts,* Amplitude Modulation; Companders; Intersymbol Interference in Digital Communication Systems; Matched Filters in Signal Demodulation; Power Spectra of Digitally Modulated Signals; Sampling of Analog Signals; Shallow-Water Acoustic Networks; Spread Spectrum Signals for Digital Communications

Chunming Qiao, *SUNY at Buffalo, Buffalo, New York,* Optical Switching Techniques in WDM Optical Networks

Hayder Radha, *Troy, New York,* Streaming Video

Harold Raemer, *Northeastern University, Boston, Massachusetts,* Atmospheric Radiowave Propagation

Daniel Ralph, *BTexact Technologies, Ipswich, Suffolk, United Kingdom,* Services Via Mobility Portals

Miguel Arjona Ramírez, *University of São Paulo, São Paulo, Brazil,* Low-Bit-Rate Speech Coding

Carey Rappaport, *Northeastern University, Boston, Massachusetts,* Reflector Antennas

Theodore S. Rappaport, *The University of Texas at Austin, Austin, Texas,* Mobile Radio Communications

Lars K. Rasmussen, *University of South Australia, Mawson Lakes, Australia,* Iterative Detection Methods for Multiuser Direct-Sequence CDMA Systems

Joseph A. Rice, *Northeastern University, Boston, Massachusetts,* Shallow-Water Acoustic Networks

Matti Rintamäki, *Helsinki University of Technology, Helsinki, Finland,* Power Control in CDMA Cellular Communication Systems

Apostolos Rizos, *AWARE, Inc., Bedford, Massachusetts,* Partial-Response Signals for Communications

Patrick Robertson, *Institute for Communications Technology, German Aerospace Center (DLR), Wessling, Germany,* Turbo Trellis-Coded Modulation (TTCM) Employing Parity Bit Puncturing and Parallel Concatenation

Ulrich L. Rohde, *Synergy Microwave Corporation, Paterson, New Jersey,* Frequency Synthesizers

Kai Rohrbacher, *Head, Department of Mobile Communication Software, LStelcom Lichtenau, Germany,* Cell Planning in Wireless Networks

Christopher Rose, *Rutgers WINLAB, Piscataway, New Jersey,* Interference Avoidance for Wireless Systems; Paging and Registration in Mobile Networks

George N. Rouskas, *North Carolina State University, Raleigh, North Carolina,* Routing and Wavelength Assignment in Optical WDM Networks

Michael Ruane, *Boston University, Boston, Massachusetts,* Optical Memories

William E. Ryan, *University of Arizona, Tucson, Arizona,* Concatenated Convolutional Codes and Iterative Decoding

Ashutosh Sabharwal, *Rice University, Houston, Texas,* Multiuser Wireless Communication Systems

John N. Sahalos, *Radiocommunications Laboratory, Aristotle University of Thessaloniki, Thessaloniki, Greece,* Antenna Arrays

S. Sajama, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

Magdalena Salazar Palma, *Universidad Politecnica de Madrid, Madrid, Spain,* Adaptive Antenna Arrays

Masoud Salehi, *Northeastern University, Boston, Massachusetts,* Frequency and Phase Modulation

Burton R. Saltzberg, *Middletown, New Jersey,* Carrierless Amplitude–Phase Modulation

Sheldon S. Sandler, *Lexington, Massachusetts,* Linear Antennas

Hikmet Sari, *Juniper Networks, Paris, France,* Broadband Wireless Access

Tapan K. Sarkar, *Syracuse University, Syracuse, New York,* Adaptive Antenna Arrays

Iwao Sasase, *Keio University, Yokohama, Japan,* Optical Synchronous CDMA Systems

Ali H. Sayed, *University of California, Los Angeles, California,* Wireless Location

Christian Schlegel, *University of Alberta, Edmonton, Alberta, Canada,* Trellis Coding

Jeffrey B. Schodorf, *MIT Lincoln Laboratory, Lexington, Massachusetts,* Land-Mobile Satellite Communications

Thomas A. Schonhoff, *Titan Systems Corporation, Shrewsbury, Massachusetts,* Continuous Phase Frequency Shift Keying (CPFSK); Statistical Characterization of Impulsive Noise

Henning Schulzrinne, *Columbia University, New York, New York,* Session Initiation Protocol (SIP)

Romed Schur, *Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany,* DMT Modulation

Kenneth Scussel, *Benthos, Inc., North Falmouth, Massachusetts,* Acoustic Modems for Underwater Communication

Randall G. Seed, *MIT Lincoln Laboratory, Lexington, Massachusetts,* Multibeam Phased Arrays

M. Selim Ünlü, *Boston University, Boston, Massachusetts,* High-Speed Photodetectors for Optical Communications

Husrev Sencar, *New Jersey Institute of Technology, Newark, New Jersey,* Orthogonal Transmultiplexers: A Time-Frequency Perspective

Mehdi Shadaram, *Department of Electrical and Computer Engineering, University of Texas at El Paso, El Paso, Texas,* Optical Fiber Local Area Networks

Ippei Shake, *NTT Network Innovation Laboratories, Kanagawa, Japan,* Signal Quality Monitoring in Optical Networks

K. Sam Shanmugan, *University of Kansas, Lawrence, Kansas,* Simulation of Communication Systems

John M. Shea, *University of Florida, Gainesville, Florida,* Multidimensional Codes

Chris Shephard, *BTexact Technologies, Ipswich, Suffolk, United Kingdom,* Services Via Mobility Portals

Barry L. Shoop, *United States Military Academy, West Point, New York,* Photonic Analog-to-Digital Converters

Mark Shtaif, *Tel-Aviv University, Tel-Aviv, Israel,* Modeling and Analysis of Digital Optical Communications Systems

Marvin K. Simon, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California,* Minimum-Shift-Keying

Kazimierz (Kai) Siwiak, *Time Domain Corporation, Huntsville, Alabama,* Loop Antennas; Ultrawideband Radio

David R. Smith, *George Washington University, Ashburn, Virgina,* Terrestrial Microwave Communications

Josep Sole i Tresserres, *France Télécom R&D, South San Franciso, California,* IMT-2000 3G Mobile Systems

Hong-Yeop Song, *Yonsei University, Seoul, South Korea,* Feedback Shift Register Sequences

Iickho Song, *Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea,* Polyphase Sequences

Ethem M. Sozer, *Northeastern University, Boston, Massachusetts,* Shallow-Water Acoustic Networks

Andreas Spanias, *Arizona State University, Tempe, Arizona,* Vocoders

Predrag Spasojević, *Rutgers, The State University of New Jersey, Piscataway, New Jersey,* EM Algorithm in Telecommunications

Joachim Speidel, *Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany,* DMT Modulation

Per Ståhl, *Lund University, Lund, Sweden,* Tailbiting Convolutional Codes

Alexandros Stavdas, *National Technical University of Athens, Athens, Greece,* Optical Multiplexing and Demultiplexing

Marc-Alain Steinemann, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Dimitrios Stiliadis, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Packet-Switched Networks

Milica Stojanovic, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* Acoustic (Underwater) Communications; Shallow-Water Acoustic Networks

Detlef Stoll, *Siemens ICN, Optisphere Networks, Boca Raton, Florida,* DWDM Ring Networks

Erik Strom, *Chalmers University of Technology, Goteborg, Sweden,* Signature Sequences for CDMA Communications

Carl-Erik W. Sundberg, *iBiquity Digital Corp., Warren, New Jersey,* Continuous-Phase-Coded Modulation

Arne Svensson, *Chalmers University of Technology, Goteborg, Sweden,* Signature Sequences for CDMA Communications

Violet R. Syrotiuk, *University of Texas at Dallas, Richardson, Texas,* Medium Access Control (MAC) Protocols

Chintha Tellambura, *University of Alberta, Edmonton, Alberta,* Golay Complementary Sequences; Peak-to-Average Power Ratio of Orthogonal Frequency-Division Multiplexing; Wireless Communications System Design

Hemant K. Thapar, *LSI Logic Corporation, San Jose, California,* Magnetic Storage Systems

Ioannis Tomkos, *Athens Information Technology, Peania, Greece,* WDM Metropolitan-Area Optical Networks

Lang Tong, *Cornell University, Ithaca, New York,* Channel Modeling and Estimation

Brent Townshend, *Townshend Computer Tools, Menlo Park, California,* V.90 MODEM

William H. Tranter, *Virginia Tech, Blacksburg, Virginia,* Mobile Radio Communications

H. J. Trussell, *North Carolina State University, Raleigh, North Carolina,* Image Sampling and Reconstruction

Jitendra K. Tugnait, *Auburn University, Auburn, Alabama,* Blind Equalization Techniques

B. E. Usevitch, *University of Texas at El Paso, El Paso, Texas,* JPEG2000 Image Coding Standard

Steven Van Den Berghe, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Pim Van Heuven, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Richard van Nee, *Woodside Networks, Breukelen, The Netherlands,* Wireless LAN Standards

Alessandro Vanelli-Coralli, *University of Bologna, Bologna, Italy,* cdma2000

Emmanouel Varvarigos, *University of Patras, Patras, Greece,* Computer Communications Protocols

Theodora Varvarigou, *National Technical University, Patras, Greece,* Computer Communications Protocols

Bane Vasic, *University of Arizona, Tucson, Arizona,* Design and Analysis of Low-Density Parity-Check Codes for Applications to Perpendicular Recording Channels

Iakovos S. Venieris, *National Technical University of Athens, Athens, Greece,* Distributed Intelligent Networks

Roberto Verdone, *University of Bologna, DEIS, Italy,* Communications for Intelligent Transportation Systems

A. J. Viterbi, *Viterbi Group, San Diego, California,* Viterbi Algorithm

Emanuele Viterbo, *Politechnico di Torino, Torino* (*Turin*)*, Italy,* Permutation Codes

Branimir R. Vojčić, *George Washington University, Washington, District of Columbia,* Code-Division Multiple Access

John L. Volakis, *University of Michigan, Ann Arbor, Michigan,* Antenna Modeling Techniques

John C. H. Wang, *Federal Communications Commission, Washington, District of Columbia,* Radio Propagation AT LF, MF, and HF

Xiaodong Wang, *Columbia University, New York, New York,* Blind Multiuser Detection

R. B. Waterhouse, *RMIT University, Melbourne, Australia,* Microstrip Patch Arrays

Steven R. Weller, *University of Newcastle, Callaghan, Australia,* Low-Density Parity-Check Codes: Design and Decoding

Wushao Wen, *University of California, Davis, Davis, California,* Design and Analysis of a WDM Client/Server Network Architecture

Lih-Jyh Weng, *Maxtor Corporation, Shrewsbury, Massachusetts,* Coding for Magnetic Recording Channels

Richard D. Wesel, *University of California at Los Angeles, Los Angeles, California,* Convolutional Codes

Krzysztof Wesołowski, *Poznań University of Technology, Poznań, Poland,* Adaptive Equalizers

Stephen B. Wicker, *Cornell University, Ithaca, New York,* Cyclic Codes

Werner Wiesbeck, *Director, Institute for High Frequency Technology and Electronics Karlsruhe University, Germany,* Cell Planning in Wireless Networks

Alan E. Willner, *University of Southern California, Optical Communications Laboratory, Los Angeles, California,* Digital Optical Correlation for Fiberoptic Communication Systems

Stephen G. Wilson, *University of Virginia, Charlottesvaille, Virginia,* Trellis-Coded Modulation

Bernhard Wimmer, *Siemens AG, Munich, Germany,* H.324: Videotelephony and Multimedia for Circuit-Switched and Wireless Networks

Peter J. Winzer, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Optical Transmitters, Receivers, and Noise

G. Woelfle, *Institut Fuer Hochfrequenztechnik, University of Stuttgart, Stuttgart, Germany,* Propagation Models for Indoor Communications

Tan F. Wong, *University of Florida, Gainesville, Florida,* Multidimensional Codes

Thomas Wörz, *Audens ACT Consulting GmbH, Wessling, Germany,* Turbo Trellis-Coded Modulation (TTCM) Employing Parity Bit Puncturing and Parallel Concatenation

William W. Wu, *Consultare Technology Group, Bethesda, Denmark,* Threshld Decoding

Yiyan Wu, *Communications Research Centre Canada, Ottawa, Ontario, Canada,* Terrestrial Digital Television

Jimin Xie, *Siemens ICN, Optisphere Networks, Boca Raton, Florida,* DWDM Ring Networks

Fuqin Xiong, *Cleveland State University, Cleveland, Ohio,* Digital Phase Modulation and Demodulation

En-hui Yang, *University of Waterloo, Waterloo, Ontario, Canada,* Huffman Coding

Jie Yang, *New Jersey Institute of Technology, University Heights, Newark, New Jersey,* Admission Control in Wired Networks

Xueshi Yang, *Drexel University, Philadelphia, Pennsylvania,* Interference Modeling in Wireless Communications

Kung Yao, *University of California, Los Angeles, California,* Chaos in Communications

Bülent Yener, *Rensselaer Polytechnic University, Troy, New York,* Internet Security

Ikjun Yeom, *Korea Advanced Institute of Science and Technology, Seoul, South Korea,* Differentiated Services

Abbas Yongaçoğlu, *University of Ottawa, School of Information Technology and Engineering, Ottawa, Ontario, Canada,* Waveform Coding

Myungsik Yoo, *Soongsil University, Seoul, Korea,* Optical Switching Techniques in WDM Optical Networks

Nabil R. Yousef, *Adaptive Systems Laboratory, Department of Electrical Engineering, University of California, Los Angeles, California,* Wireless Location

Jens Zander, *Royal Institute of Technology, Stockholm, Sweden,* Radio Resouce Management in Future Wireless Networks

Yimin Zhang, *Villanova University, Villanova, Pennsylvania,* Interference Suppression in Spread-Spectrum Communication Systems

Haitao Zheng, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Multimedia Over Digital Subscriber Lines

Shihua Zhu, *Xian Jiaotong University, Xian, Shaanxi, People's Republic of China,* Wideband CDMA in Third-Generation Cellular Communication Systems

Rodger E. Ziemer, *University of Colorado, Colorado Springs, Colorado,* Mobile Radio Communications

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

A

ACOUSTIC ECHO CANCELLATION

EBERHARD HÄNSLER
Darmstadt University of Technology
Darmstadt, Germany

1. INTRODUCTION

In 1877 the front page of *Scientific American* showed a picture of a man using "the new Bell telephone" [1]. He held a microphone in front of his mouth and an identical-looking device — the loudspeaker — close to one of his ears. So, at the beginning of telecommunications both hands were busy while making a telephone call. This troublesome way of operation was due to the lack of efficient electroacoustic converters and amplifiers. The inconvenience, however, guaranteed optimal conditions: a high signal-to-(environmental) noise ratio at the microphone input, a perfect coupling between loudspeaker and the ear of the listener, and — last but not least — a high attenuation between the loudspeaker and microphone. The designers of modern speech communication systems still dream of getting back those conditions.

It did not take long until the microphone and the loudspeaker of a telephone were mounted in a handset. Thus, one hand had been freed. To provide a fully natural communication between two users of a speech communication system — to allow both to speak at the same time and to interrupt each other, with both hands free — is still a problem that keeps hundreds of researchers and industrial developers busy.

This article is meant to explain the problem of acoustical echoes and their cancellation. It will focus on the hands-free telephone as one of the applications mostly asked for. The statements, however, hold for other applications such as hearing aids, voice input systems, and public-address systems as well.

The problem of acoustic echo cancellation arises wherever a loudspeaker and a microphone are placed such that the microphone picks up the signal radiated by the loudspeaker and its reflections at the borders of the enclosure. As a result, the electroacoustic circuit may become unstable and produce howling. In addition, the users of telecommunication systems are annoyed by listening to their own speech delayed by the round-trip time of the system. To avoid these problems, the attenuation of the acoustic path between loudspeaker and microphone has to be sufficiently high.

In general, acoustic echo cancellation units as used in hands-free communication systems consist of three subunits: (1) a loss control circuit (LC), (2) a filter parallel to the loudspeaker–enclosure–microphone system (LEMS) — the echo cancellation filter (ECF), and (3) a second filter — the residual echo-suppressing filter (RESF) — within the path of the output signal (see Fig. 1).

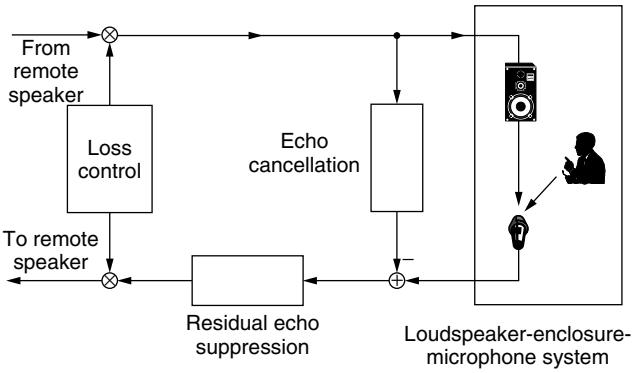


Figure 1. General structure of an acoustic echo cancellation system.

Their functions are obvious — the loss control circuit can attenuate the input and/or output signal such that the communication loop always remains stable. The ECF in parallel to the LEMS is able to cancel echoes picked up by the microphone according to the degree to which this filter is matched to the LEMS. The RESF within the path of the output signal can be used to suppress residual echoes and background noise. In the early days of acoustic echo control a so-called center clipper (see Fig. 2) took the place of this filter.

Of these subunits, the loss control circuit has the longest history in hands-free communication systems. In its simplest form it reduces the usually full-duplex communication system to a half-duplex one by alternatively switching the input and output lines on and off. Apart from preventing howling and suppressing echoes, any natural conversation was prevented, too. Only the ECF in parallel to the LEMS can help to provide full-duplex (i.e., fully natural) communication.

A device for hands-free telephone conversation using voice switching was presented in 1957 [2]. The introduction of a center clipper in 1974 [3] resulted in a noticeable improvement. Laboratory experiments applying an adaptive filter for acoustic echo cancellation were reported in 1975 [4]. At that time, however, an economically feasible implementation of such a filter for acoustic echo cancellation was far out of sight.

Because of the high interest in providing hands–free speech communication, an enormous number of papers has been published since the early 1980s. Among those are a number of bibliographies [5–8], overview papers [9–11], and books [12,13].

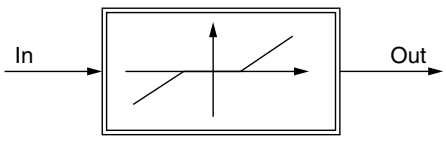


Figure 2. Center clipper.

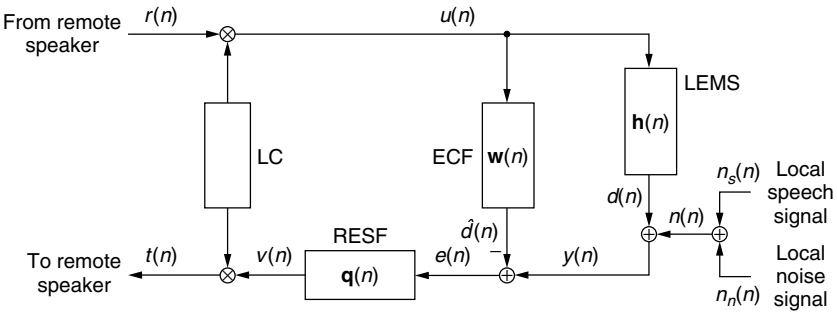


Figure 3. Notation used in this contribution: LC = loss control circuit, LEMS = loudspeaker–enclosure–microphone system, ECF = echo-canceling filter, RESF = residual echo suppressing filter.

For the following considerations, the notation as given in Fig. 3 will be used. Lowercase boldface letters will indicate column vectors; uppercase boldface letters will denote matrices.

2. SCENARIO

2.1. Systems

2.1.1. Loudspeaker–Enclosure–Microphone System (LEMS). In a LEMS the loudspeaker and the microphone are connected by an acoustical path formed by a direct connection (if both can "see" each other) and in general a large number of reflections at the boundaries of the enclosure. For low sound pressure and no overload of the converters, this system may be modeled with sufficient accuracy as a linear system. The impulse response can be described by a sequence of delta impulses delayed proportionally to the geometric length of the related path and the inverse of the sound velocity. The amplitudes of the impulses depend on the reflection coefficients of the boundaries and on the inverse of the pathlengths. As a first-order approximation one can assume that the impulse response decays exponentionally. A measure for the degree of this decay is the *reverberation time* T_{60}, which specifies the time necessary for the sound energy to drop by 60 dB after the sound source has been switched off [14]. Depending on the application, it may be possible to design the boundaries of the enclosure such that the reverberation time is small, resulting in a short impulse response. Examples are telecommunication studios. For ordinary offices, the reverberation time T_{60} is typically in the order of a few hundred milliseconds. For the interior of a passenger car, this quantity is a few tens of milliseconds long. Figure 4 shows the impulse responses of LEMSs measured in an office (left) and in a passenger car (right). The microphone signals have been sampled at 8 kHz according to the standards for telephone signals. It becomes obvious that the impulse response of an office exhibits amplitudes noticeably different from zero even after 1000 samples, that is to say, after 125 ms. In comparison, the impulse response of the interior of a car decays faster because of the smaller volume of this enclosure.

The impulse response of a LEMS is highly sensitive to any changes such as the movement of a person within it. This is explained by the fact that, assuming a sound velocity of 343 m/s and 8 kHz sampling frequency, the distance traveled between two sampling instants is 4.3 cm. Therefore, a 4.3-cm change in the length of an echo path, the move of a person by only a few centimeters, shifts the related impulse by one sampling interval. Thus, the echo cancellation filter (ECF) that has to mimic the LEMS must be an adaptive filter.

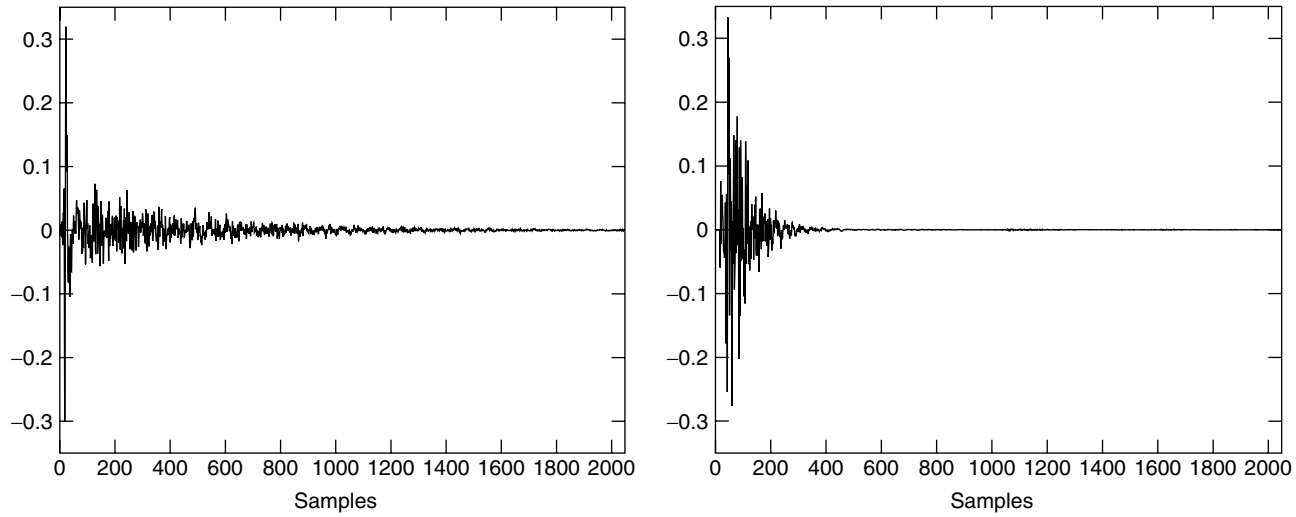


Figure 4. Impulse responses measured in an office (left) and in a car (right) (sampling frequency = 8 kHz).

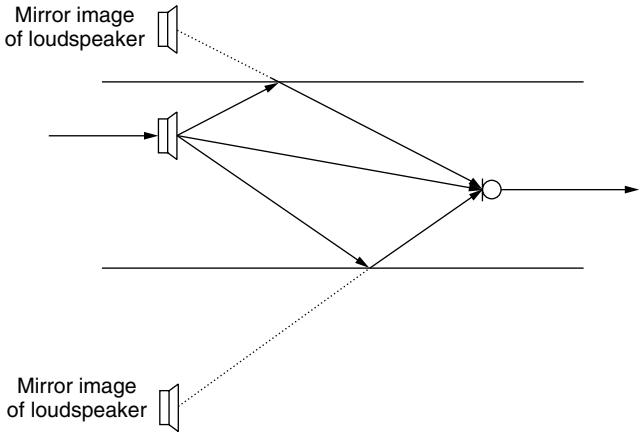


Figure 5. Simulation of an impulse response: direct path and two reflections.

2.1.2. Simulation of a LEMS. The impulse response of a LEMS can be simulated by using the principle of an *image source* [15] just as used in geometric optics. A reflected sound ray can be viewed as originating from a mirror image of the original source (see Fig. 5). At the reflection the sound pressure is attenuated according to a *reflection coefficient* that depends on the material at the reflection point. Typical reflection coefficients are in the order of 0.4 for "soft" material and close to but smaller than one for rigid walls [16]. Reflections occur at all boundaries of the enclosure, and sound rays may be reflected several times before reaching the microphone. In addition to (multiple) reflections, the sound pressure is also attenuated proportionally to the inverse of the length of its path.

2.1.3. Electronic Replica of LEMSs. From a control engineering point of view, acoustic echo cancellation is a system identification problem. However, the system to be identified — the LEMS — is highly complex; its impulse response exhibits up to several thousand sample values noticeably different from zero and it is time-varying at a speed mainly according to human movements. The question of the optimal structure of the ECF has been discussed intensively. Since a long impulse response has to be modeled by the ECF, a recursive (IIR) filter seems best suited at first glance. At second glance, however, the impulse response exhibits a highly detailed and irregular shape. To achieve a sufficiently good match, the replica must offer a large number of adjustable parameters. Therefore, an IIR filter does not have an advantage over a nonrecursive (FIR) filter [17,18]. The even more important argument in favor of an FIR filter is its guaranteed stability during adaptation.

Figure 6 shows an FIR filter of length N. The N values of the input signal $u(n)$ can be combined in a column vector $\mathbf{u}(n)$:

$$\mathbf{u}(n) = [u(n), u(n-1), u(n-2), \ldots, u(n-N+2), \times u(n-N+1)]^T \quad (1)$$

If we also combine the filter coefficients, the tap weights, in a column vector $\mathbf{w}(n)$

$$\mathbf{w}(n) = [w_0(n), w_1(n), w_2(n), \ldots, w_{N-2}(n), w_{N-1}(n)]^T \quad (2)$$

the output signal $\hat{d}(n)$ can be written as an inner product:

$$\hat{d}(n) = \sum_{k=0}^{N-1} w_k(n)\, u(n-k) = \mathbf{w}^T(n)\, \mathbf{u}(n) = \mathbf{u}^T(n)\, \mathbf{w}(n) \quad (3)$$

A measure to express the effect of an ECF is the *echo return loss enhancement* (*ERLE*):

$$ERLE = 10 \log \frac{E[d^2(n)]}{E[(d(n) - \hat{d}(n))^2]} \text{ dB} \quad (4)$$

where the echo $d(n)$ is equal to the microphone output signal $y(n)$ in case the loudspeaker is the only signal source within the LEMS [i.e., the local speech signal $n_s(n)$ and the local noise $n_n(n)$ are zero], and $\hat{d}(n)$ describes the ECF output. Denoting the (assumed to be time invariant for the moment) impulse responses of the LEMS by $h_i, i = 0, \ldots, \infty$, and the ECF by $\mathbf{w}$ respectively, it follows that

$$d(n) = \sum_{i=0}^{\infty} h_i u(n-i) \quad (5)$$

and

$$\hat{d}(n) = \sum_{i=0}^{N-1} w_i u(n-i) \quad (6)$$

where $N-1$ is the degree of the nonrecursive ECF. Assuming, also for simplicity, a stationary white input signal $u(n)$, the *ERLE* can be expressed as

$$ERLE = 10 \log \frac{E[u^2(n)] \sum_{i=0}^{\infty} h_i^2}{E[u^2(n)] \left(\sum_{i=0}^{\infty} h_i^2 - 2 \sum_{i=0}^{N-1} h_i w_i + \sum_{i=0}^{N-1} w_i^2 \right)} \text{ dB} \quad (7)$$

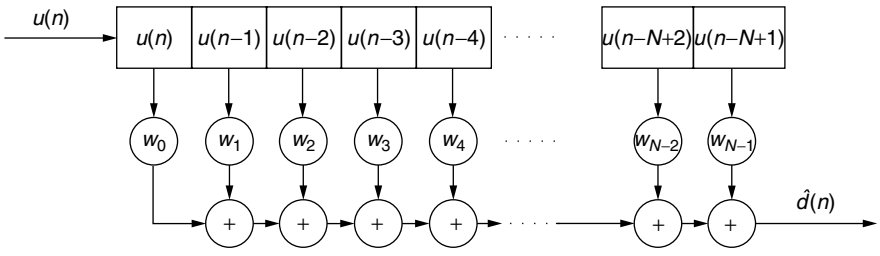


Figure 6. Transversal (FIR) filter as used to model the LEMS.

An upper bound for the effect of an ECF of degree $N-1$ can be calculated by assuming a perfect match of the first N coefficients of the ECF with the LEMS:

$$w_i = h_i \quad \text{for} \quad 0 \le i < N \tag{8}$$

In this case Eq. (7) reduces to

$$ERLE_{\max}(N) = 10 \log \frac{\sum_{i=0}^{\infty} h_i^2}{\sum_{n=N}^{\infty} h_i^2} \text{ dB} \tag{9}$$

Figure 7 shows the upper bounds of the *ERLE* achievable with transversal ECFs of length N for an office and a car with impulse responses as given in Fig. 4. An attenuation of only 25 dB needs filter lengths of about 1100 for the office and about 250 for the car.

2.2. Signals

2.2.1. Speech Signals. Acoustic echo cancellation requires the adaptation of FIR filters. In general, the performance of adaptation algorithms crucially depends on the properties of the signals involved. In the application considered here, one has to deal primarily with speech signals additively disturbed by other speech signals (in the case of doubletalk, i.e., if both communication partners talk simultaneously) and by noise. Performing signal processing with this type of signals turns out to be very difficult.

Speech is characterized by nearly periodic segments, by noiselike segments, and by pauses. The signal envelope fluctuates enormously. In speech processing it is widely accepted that parameters derived from a speech signal have to be updated after intervals of about 20 ms. Short-time variances may differ by more than 40 dB [19]. Sampling frequencies range from 8 kHz in telephone systems up to about 40 kHz in high-fidelity systems. Even in the case of 8 kHz sampling frequency, consecutive samples are highly correlated. The normalized autocorrelation coefficient $s_{uu}(1)/s_{uu}(0)$ of neighboring samples assumes values in the range of 0.8–0.95. Short-time autocorrelation matrices very often become singular. Thus, special precautions are necessary to prevent instability of algorithms that use — directly or indirectly — the inverse of the autocorrelation matrix. To summarize, speech signals are *nonpersistent*. Figure 8 shows a segment of a speech signal (left) and an estimate of a power spectral density of speech signals (right). The spectrum clearly indicates that the statistical properties of speech are quite different from those of a white process.

2.2.2. Noise. The noise signals involved in echo-canceling applications are typically those existing in offices or in moving cars. Especially the noise in a passenger car moving at constant speed exhibits a power density spectrum that is decaying slightly faster than the one of speech (Fig. 9). In both cases the major part of the energy is concentrated at low frequencies.

2.3. Side Constraints

Acoustic echo cancelers used in telephones have to comply with a number of standards issued by the international standardization organizations like the International Telecommunication Union (ITU) or the European Telecommunications Standards Institute (ETSI). Especially important are requirements concerning delay and echo attenuation [20–22]. For ordinary telephones the ITU allows only 2 ms additional delay for front-end processing. In case of mobile telephones 39 ms are permitted. The maximum of 2 ms prohibits the application of efficient frequency domain or block processing methods. Concerning the overall echo attenuation a minimum of 45 dB is necessary in singletalk situations. In case of doubletalk, the attenuation can be reduced to 30 dB taking into consideration that the echo of the far-end signal is masked by the local speech signal. This high echo attenuation has

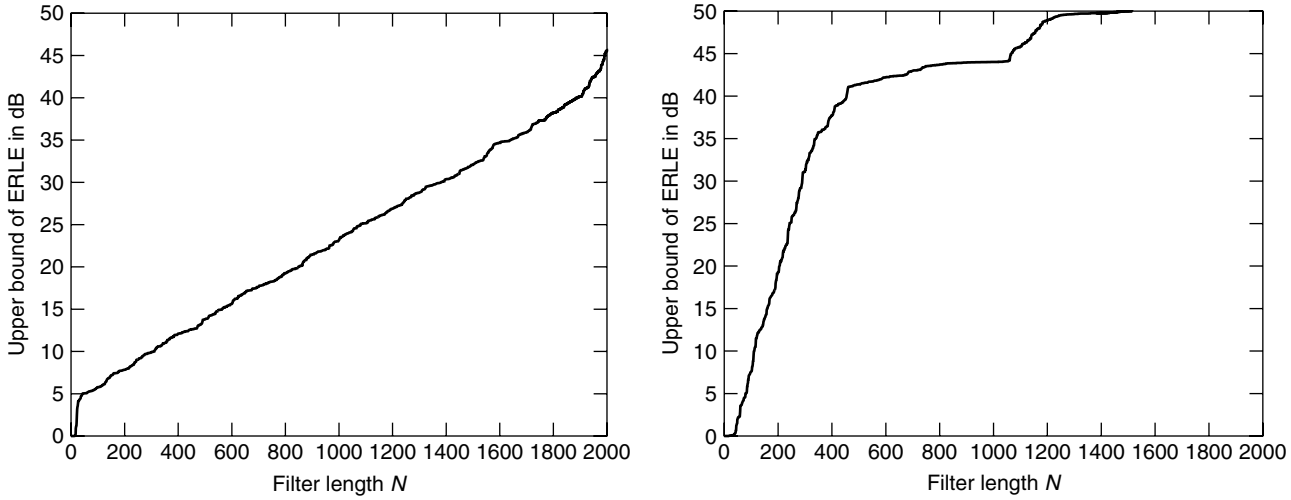


Figure 7. Maximal attenuations achievable with a transversal filter of length N in an office (left) and in a car (right) (sampling frequency = 8 kHz).

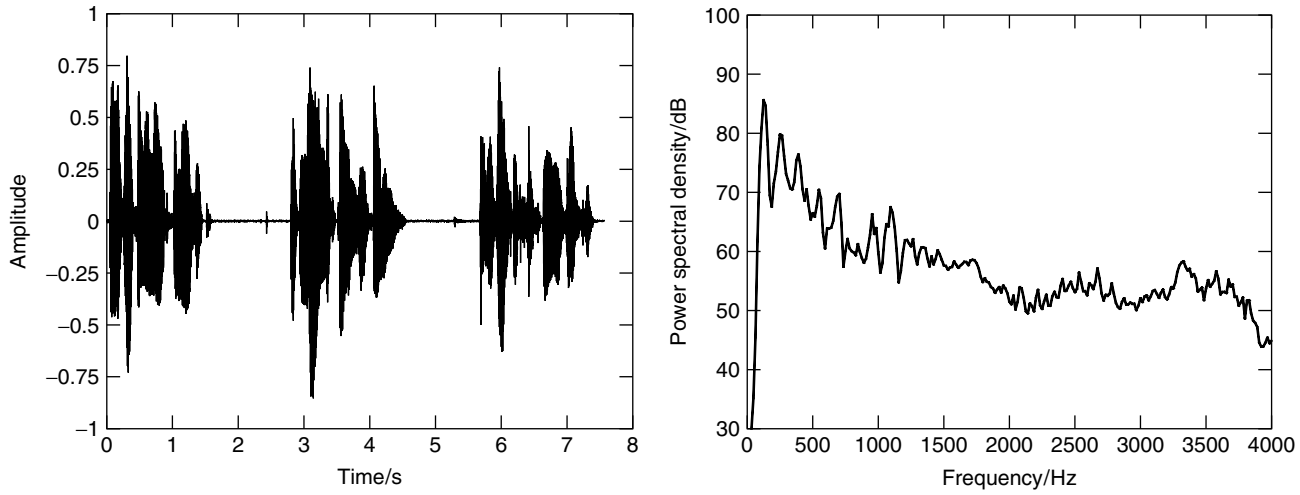


Figure 8. Section of a speech signal and estimate of the power spectral density of speech.

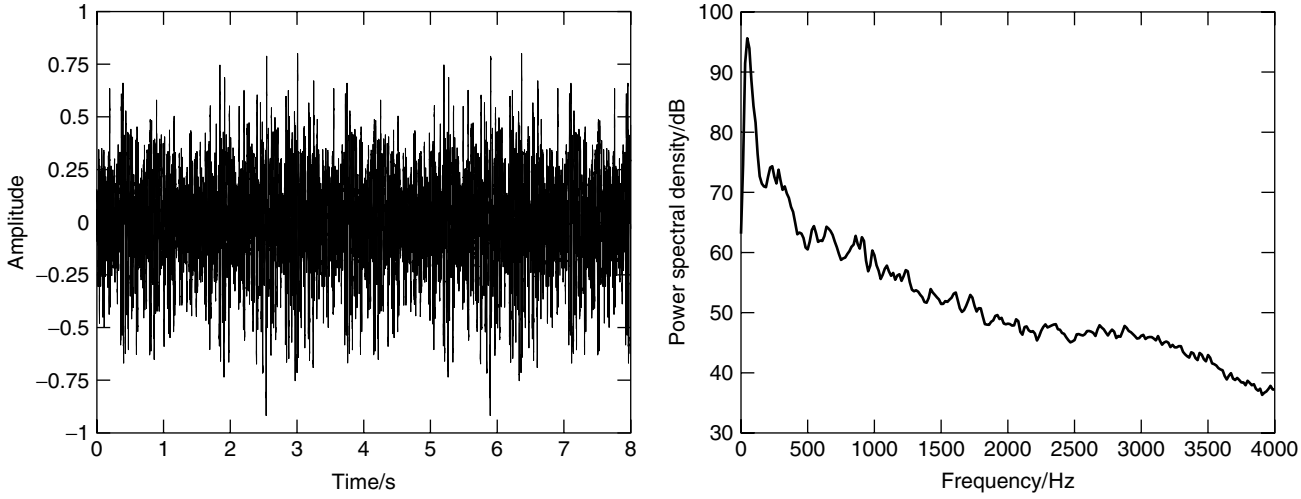


Figure 9. Section of a car noise signal and estimate of the power spectral density.

to be provided at all times of the connection and at all situations. Since adaptive filters (the ECF and the RESF; see Fig. 3) may not (yet) have converged to their optimal settings the attenuation required can be guaranteed only with the help of an additional loss control device.

3. METHODS TO STABILIZE THE ELECTROACOUSTIC LOOP

3.1. Traditional Methods

Most voice communication systems operate as a closed loop. In the case of the telephone system with customers using handsets, the attenuations between loudspeakers and microphones crucially contribute to the stabilization of this loop. Therefore, with the introduction of hands-free devices the provision of a stable electroacoustic loop became a major problem.

The simplest way to achieve stability is reducing the full-duplex connection to half-duplex. This can be done manually — as it is still done by astronauts — or by voice-controlled switches. The problem related to those switches is that they do not distinguish between speech and noise signals. Therefore, a strong noise can mislead the switching circuit. Consequently an active speaker may be switched off in favor of a noise at the remote location.

A moderate form of switching off one direction of a connection consists of artificially increasing its attenuation. This results in shifting an attenuation to the incoming or the outgoing circuit depending on where the loss control device detects the lower speech activity. In case no activity is detected on both circuits, the additional attenuation can be distributed equally between both directions. As mentioned before, even modern echo-canceling devices cannot fulfill the ITU requirements without loss insertion. However, the attenuation already provided by echo-canceling (and other) circuits can be estimated and only the lacking attenuation has to be inserted. In the case of a well-adapted ECF, this may only be a few decibels that do not disturb the speakers.

A method to stabilize an electroacoustic loop has been proposed [23]. It is especially designed for systems like public-address systems where the loudspeaker output signal feeds back into the talker microphone directly. It consists of a frequency shift of a few hertz, implemented by a single-sideband modulation — within the microphone–loudspeaker circuit. Thus, stationary howling cannot build up. Echoes are not canceled. They are, however, shifted to higher or lower frequencies — depending on whether the modulation frequency is positive or negative — until they "fall" into a minimum of the transfer function of the LEMS and become inaudible. In speech communication systems frequency shifts of ~3–5 Hz are scarcely audible. The stability gain achievable with this method depends on the signal and the acoustical properties of the enclosure. For speech signals and rooms with short reverberation times the gain is in the order of 3–5 dB; for rooms with long reverberation times it can go up to ~10 dB.

3.2. Adaptive Filters

With the availability of powerful digital signal processors, the application of an adaptive filter to cancel acoustic echoes, the ECF, and a second adaptive filter, the RESF, to suppress residual echoes not canceled by the ECF became feasible (see Fig. 1). As explained in the previous section, a transversal filter of high order is used for the ECF. The RESF, typically, is implemented in the frequency domain by an adaptive filter as well.

3.2.1. The Echo-Canceling Filter (ECF). For the following considerations we assume that the impulse responses of the ECF and of the LEMS both have the same length N. In reality, the impulse response of the LEMS may be much longer than that of the ECF. Nevertheless, this assumption means no restriction because the shorter impulse response can always be extended by zeros. Equivalent to Eqs. (2) and (3), one can write the impulse response of the LEMS at time n as a vector $\mathbf{h}(n)$

$$\mathbf{h}(n) = [h_0(n), h_1(n), h_2(n), \ldots, h_{N-2}(n), h_{N-1}(n)]^T \quad (10)$$

and the output signal $d(n)$ as an inner product:

$$d(n) = \sum_{k=0}^{N-1} h_k(n)\, u(n-k) = \mathbf{h}^T(n)\, \mathbf{u}(n) = \mathbf{u}^T(n)\, \mathbf{h}(n) \quad (11)$$

The mismatch between LEMS and ECF can be expressed by a *mismatch vector* $\boldsymbol{\varepsilon}(n)$:

$$\boldsymbol{\varepsilon}(n) = \mathbf{h}(n) - \mathbf{w}(n) \quad (12)$$

Later, the squared L_2-norm $\boldsymbol{\varepsilon}^T(n)\, \boldsymbol{\varepsilon}(n)$ of the system mismatch vector will be called the *system distance*:

$$\Delta(n) = \boldsymbol{\varepsilon}^T(n)\, \boldsymbol{\varepsilon}(n) = \|\boldsymbol{\varepsilon}(n)\|^2 \quad (13)$$

The quantity $e_u(n)$

$$e_u(n) = d(n) - \hat{d}(n) = \boldsymbol{\varepsilon}^T(n)\, \mathbf{u}(n) \quad (14)$$

represents the *undisturbed error*, that is, the error signal when the locally generated signals $n_s(n)$ and $n_n(n)$ are zero. Finally, the error signal $e(n)$ is given by

$$e(n) = y(n) - \hat{d}(n) = e_u(n) + n(n) = e_u(n) + n_s(n) + n_n(n) \quad (15)$$

This error will enter the equation used to update the coefficients of the ECF. Obviously, only the fraction expressed by the undisturbed error $e_u(n)$ contains "useful" information. The locally generated signal $n(n)$, however, causes the filter to diverge and thus to increase the system distance. Therefore, independent of the used adaptive algorithm a control procedure is necessary to switch off or slow down the adaptation when $n(n)$ is large compared to the echo $d(n)$.

3.2.2. The Residual Echo-Suppressing Filter (RESF). The impact of the ECF on the acoustical echo is limited by — at least — two facts: (1) only echoes due to the linear part of the transfer function of the LEMS can be canceled and (2) the order of the ECF typically is much smaller than the order of the LEMS (see Section 2.1.3). Therefore, a second filter — the RESF — is used to reduce the echo further. The transfer function of this filter is given by the well-known Wiener equation [24,25]:

$$Q(\Omega, n) = \frac{S_{en}(\Omega, n)}{S_{ee}(\Omega, n)} \quad (16)$$

In this equation Ω is a normalized frequency, $S_{ee}(\Omega, n)$ is the short-term auto-power spectral density of the error signal $e(n)$, and $S_{en}(\Omega, n)$ is the short-term cross-power spectral density of $e(n)$ and the locally generated signal $n(n)$. In good agreement with reality one can assume that the undisturbed error $e_u(n)$ and $n(n)$ [see Eq. (15)] are orthogonal. Then the power spectral density $S_{ee}(\Omega, n)$ reduces to

$$S_{ee}(\Omega, n) = S_{e_u e_u}(\Omega, n) + S_{nn}(\Omega, n) \quad (17)$$

Furthermore, the cross-power spectral density $S_{en}(\Omega, n)$ is given by

$$S_{en}(\Omega, n) = S_{nn}(\Omega, n) \quad (18)$$

Then, it follows from Eq. (16) and after some manipulations for the transfer function of the RESF that

$$Q(\Omega, n) = 1 - \frac{S_{e_u e_u}(\Omega, n)}{S_{ee}(\Omega, n)} \quad (19)$$

The impulse response of the RESF is found by an inverse Fourier transformation.

Since the signals involved are highly nonstationary, the short-term power spectral densities have to be estimated for time intervals no longer than 20 ms. The overwriting problem, however, is that the locally generated signal $n(n)$ is observable only during the absence of the remote excitation signal $u(n)$. Since $n(n)$ is composed of local speech and local noise the RESF suppresses local noise, as well. It should be noted, however, that any impact of the RESF on residual echoes also impacts the local speech

signal $n_s(n)$ and, thus, reduces the quality of the speech output of the echo-canceling unit.

When applying Eq. (19), the power spectral densities $S_{ee}(\Omega, n)$ and $S_{e_u e_u}(\Omega, n)$ have to be replaced by their estimates $\hat{S}_{ee}(\Omega, n)$ and $\hat{S}_{e_u e_u}(\Omega, n)$. Therefore, it is possible that the quotient becomes larger than one. Consequently, the filter exhibits a phase shift of π. To prevent that, Eq. (19) of the filter transfer function is (heuristically) modified to

$$Q(\Omega, n) = 1 - \min\left[\frac{\hat{S}_{e_u e_u}(\Omega, n)}{\hat{S}_{ee}(\Omega, n)}, Q_{\min}\right] \tag{20}$$

where $Q_{\min}$ determines the maximal attenuation of the filter. Details can be found in, for example, Quatieri's book [26]. The problem of residual echo suppression is very similar to the problem of noise suppression and both are treated simultaneously [27,28].

4. ADAPTIVE ALGORITHMS

4.1. Normalized Least-Mean-Square (NLMS) Algorithm

The majority of implementations of acoustic echo-canceling systems use the NLMS algorithm to update the ECF. This gradient type algorithm minimizes the mean-square error [24]. The update equation is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\mu(n)}{\mathbf{u}^T(n)\,\mathbf{u}(n)}\,\mathbf{u}(n)\,e(n) \tag{21}$$

The term $\mathbf{u}^T(n)\,\mathbf{u}(n)$ in the denominator represents a normalization according to the energy of the input vector $\mathbf{u}(n)$. This is necessary because of the high variance of this quantity for speech signals. The step size of the update is controlled by the *step-size factor* $\mu(n)$. In general, the algorithm is stable (in the mean square) for $0 < \mu < 2$. Reducing the step size is necessary to prevent divergence of the filter coefficients in case of strong local signals $n_s(n)$ and/or $n_n(n)$.

The NLMS algorithm has no memory; that is, it uses only signals that are available at the time of update. This is advantageous for tracking changes of the LEMS. The update is performed in the direction of the input signal vector $\mathbf{u}(n)$ (see Fig. 10). For speech signals, consecutive vectors may be highly correlated, meaning that their directions differ only slightly. This is the reason for the low speed of convergence of the NLMS algorithm in case of speech excitation. Additional measures such as decorrelation of the input signal $u(n)$ and/or controlling the step-size parameter $\mu(n)$ (see Section 5) are necessary to speed up convergence.

The motivation for using the NLMS algorithm in the application discussed here is its robustness and its low computational complexity that is only in the order of $2N$ operation per coefficient update.

Decorrelating the input signal offers a computationally inexpensive method to improve the convergence of the filter coefficients. To achieve this two (identical) decorrelation filters and an inverse filter have to be added to the echo-canceling system (see Fig. 11). The decorrelation filter has to be duplicated since the loudspeaker needs the original signal $u(n)$. Simulations show that even filters of first order approximately double the speed of convergence in acoustic echo-canceling applications [11]. Further improvements require adaptive decorrelation filters because of the nonstationarity of speech signals. Also, in case of higher-order filters the necessary interchange of the decorrelation filter and the time-varying LEMS causes additional problems. Therefore, only the use of a first-order decorrelation filter can be recommended. The curves in Fig. 11(a) are averages over several speech signals with pauses removed.

4.2. Affine Projection (AP) Algorithm

The AP algorithm [29] overcomes the weakness of the NLMS algorithm concerning correlated input signals by updating the filter coefficients not just in the direction of the current input vector but also within a hyperplane spanned by the current input vector and its $M-1$ immediate predecessors (see Fig. 10). To accomplish this an *input signal matrix* $\mathbf{U}(n)$ is formed

$$\mathbf{U}(n) = [\mathbf{u}(n), \mathbf{u}(n-1), \mathbf{u}(n-2), \ldots, \mathbf{u}(n-M+2) \times \mathbf{u}(n-M+1)] \tag{22}$$

and an error vector is calculated

$$\mathbf{e}(n) = [y(n), y(n-1), \ldots, y(n-M+1)]^T - \mathbf{U}^T(n)\,\mathbf{w}(n) \tag{23}$$

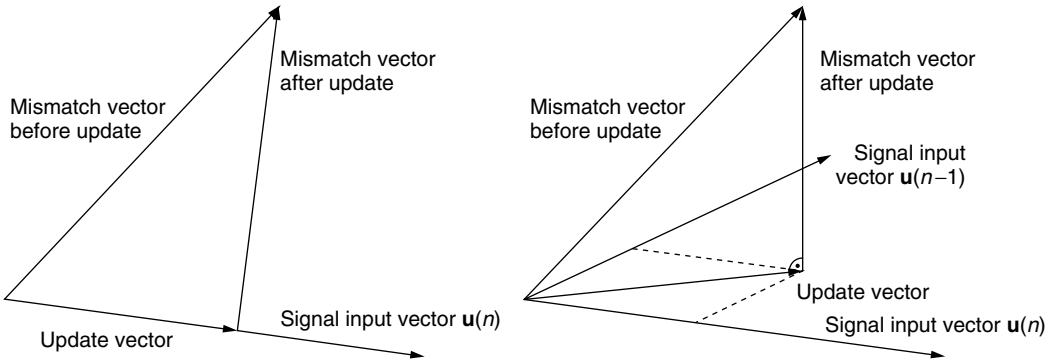


Figure 10. Updates of the system mismatch vector according to the NLMS algorithm (left) and to the AP algorithm (right).

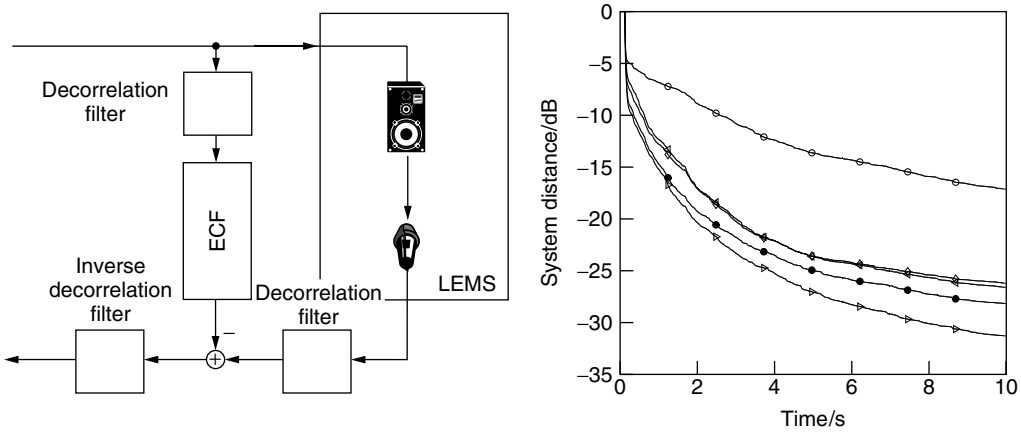


Figure 11. Insertion of decorrelation filters (**a**) and improved concergence of NLMS algorithm utilizing decorrelation filters (**b**): ○: none, ◇: fixed first order, ◁: fixed second order, ●: adaptive tenth order, ▷: adaptive 18th order (sampling frequency = 8 kHz).

collecting the errors for the current and the $M-1$ past input signal vectors applied to the ECF with the *current* coefficient setting. The price to be paid for the improved convergence is the increased computational complexity caused by the inversion of an $M \times M$ matrix required at each coefficient update. Fast versions of this algorithm are available [30,31].

Finally, the update equation is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(n)\,\mathbf{U}(n)\,(\mathbf{U}^T(n)\,\mathbf{U}(n))^{-1}\,\mathbf{e}(n) \qquad (24)$$

Numerical problems arising during the inversion of the matrix $\mathbf{U}^T(n)\,\mathbf{U}(n)$ can be overcome by using $\mathbf{U}^T(n)\,\mathbf{U}(n) + \delta\,\mathbf{1}$ instead of $\mathbf{U}^T(n)\,\mathbf{U}(n)$, where $\mathbf{1}$ is the unit matrix and δ is a small positive constant. For $M = 1$ the AP algorithm is equal to the NLMS procedure. For speech input signals even $M = 2$ leads to a considerably faster convergence of the filter coefficients (see Fig. 12). Suggested values for M are between 2 and 5 for the ECF update.

It should be noted, however, that faster convergence of the ECF coefficients also means faster divergence in case of strong local signals. Therefore, faster control of the step size is required as well. Its optimal value is based on estimated quantities (see Section 5). Since their reliabilities depend on the lengths of the data records usable for the estimation, a very high speed of convergence may not be desirable. Nevertheless, the AP algorithm seems to be a good candidate to replace the NLMS algorithm in acoustic echo-canceling applications.

4.3. Recursive Least-Squares (RLS) Algorithm

The RLS algorithm minimizes the sum of the squared error

$$\overline{e^2(n)} = \sum_{k=0}^{n} \lambda^{n-k}\, e^2(k) \qquad (25)$$

where $e(n)$ is given by Eqs. (15) and (3). It calculates an estimate $\hat{\mathbf{S}}_{uu}(n)$ of the autocorrelation matrix of the input

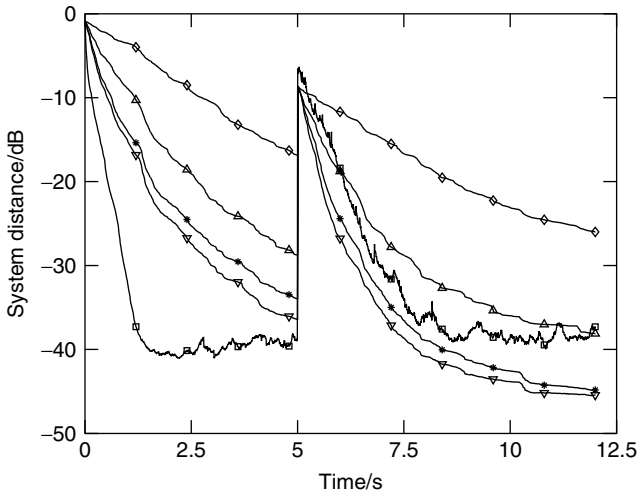


Figure 12. Convergence of the filter coefficients for different adaptation algorithms (filter length = 1024, sampling frequency = 8 kHz): ◇: NLMS ($\mu = 1$), △: AP of order two, ∗: AP of order 5, ▽: AP of order 10 (all AP algorithms with $\mu = 1$), □: RLS ($\lambda = 0.9999$). The impulse response of the LEMS is changed at $t = 5$ s.

signal vector

$$\hat{\mathbf{S}}_{uu}(n) = \sum_{k=0}^{n} \lambda^{n-k}\, \mathbf{u}(k)\, \mathbf{u}^T(k) \qquad (26)$$

and uses the inverse of this $N \times N$ matrix to decorrelate the input signal in the update equation. The factor λ is called the *forgetting factor*. It is chosen close to but smaller than one and assigns decreasing weights to the input signal vectors the further they are in the past. In addition, the *a priori error* $\tilde{e}(n)$, defined by

$$\tilde{e}(n+1) = d(n+1) - \mathbf{u}^T(n+1)\,\mathbf{w}(n) \qquad (27)$$

is calculated. This is the error calculated with the new input vector but with the not yet updated filter coefficients.

Finally, the update equation of the RLS algorithm is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(n)\,\hat{\mathbf{S}}_{uu}^{-1}(n+1)\,\mathbf{u}(n+1)\,\tilde{e}(n+1) \tag{28}$$

In contrast to the AP algorithm, now an $N \times N$ matrix has to be inverted at each coefficient update. This can be done recursively, and fast RLS algorithms are available [32] with numerical complexity in the order of $7N$.

The RLS algorithm exhibits considerable stability problems. One problem is caused by finite wordlength errors, especially when the procedure is executed in 16 bit fixed-point arithmetic. The second problem arises from the properties of speech signals (see Section 2.2.1), temporarily causing the estimate of the (short-term) autocorrelation matrix to become singular. A forgetting factor λ very close to one helps to overcome this problem. On the other hand, however, the long memory caused by a λ close to one slows down the convergence of the coefficients of the ECF after a change of the LEMS.

In spite of the speed of convergence achievable with the RLS algorithm, the numerical complexity and the stability problems so far prevented the use of this algorithm in acoustical echo-cancelation applications.

A unified analysis of least squares adaptive algorithms can be found in the paper by Glentis et al. [33].

5. STEP-SIZE CONTROL OF THE NLMS ALGORITHM

Independent of the specific algorithm used, the update of the coefficients of the ECF strongly depends on the error signal $e(n)$. This signal is composed of the undisturbed error $e_u(n)$ and the locally generated signal $n(n)$ [see Eq. (15)]. Only $e_u(n)$ steers the coefficients toward their optimal values. The step-size factor $\mu(n)$ is used to control the update according to the ratio of both contributions. Assuming the filter has converged, the error signal $e(n)$ has assumed a certain value. Suddenly the amplitude of $e(n)$ is increased. This may have two reasons that require different actions: (1) a local speaker became active or a local noise started — in this case the step size has to be reduced to prevent losing the degree of convergence achieved before, and (2) the impulse response of the LEMS has changed, for example, by the movement of the local talker. Now, the step size has to be increased to its maximal possible value in order to adapt the ECF to its new impulse response as fast as possible.

A major problem becomes visible with this consideration — the not-directly-observable undisturbed error signal $e_u(n)$ needs to be known in order to control the adaptation process. Another leading point should be mentioned here. The first situation requires immediate action. In the second case, a delayed action causes an audible echo but no divergence of the ECF.

5.1. Optimal Step Size for the NLMS Algorithm

In this section an optimal step size for the NLMS algorithm will be derived assuming that all required quantities are available. The following sections explain how to estimate them from measurable signals.

Using Eq. (21) and assuming that the impulse response of the LEMS does not change

$$\mathbf{h}(n+1) = \mathbf{h}(n) \tag{29}$$

the mismatch [see Eq. (12)] is given by

$$\begin{aligned}\boldsymbol{\varepsilon}(n+1) &= \mathbf{h}(n+1) - \mathbf{w}(n+1) \\ &= \mathbf{h}(n) - \mathbf{w}(n) - \frac{\mu(n)}{\|\mathbf{u}(n)\|^2}\,\mathbf{u}(n)\,e(n) \qquad (30)\\ &= \boldsymbol{\varepsilon}(n) - \frac{\mu(n)}{\|\mathbf{u}(n)\|^2}\,\mathbf{u}(n)\,e(n) \qquad (31)\end{aligned}$$

Using Eqs. (13) and (14), the expectation of the system distance can be expressed as

$$\begin{aligned}E\{\Delta(n+1)\} = E\{\Delta(n)\} &- 2\,\mu(n)\,E\left\{\frac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\} \\ &+ \mu^2(n)\,E\left\{\frac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\}\end{aligned} \tag{32}$$

For an optimal step size, it is required that

$$E\{\Delta(n+1)\} - E\{\Delta(n)\} \le 0 \tag{33}$$

Inserting Eq. (32) leads to

$$\mu^2(n)\,E\left\{\frac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\} - 2\,\mu(n)\,E\left\{\frac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\} \le 0 \tag{34}$$

Thus, the condition for the optimal step size is given by

$$0 \le \mu(n) \le 2\,\frac{E\left\{\dfrac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\}}{E\left\{\dfrac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\}} \tag{35}$$

A step size in the middle of this range achieves the fastest decrease of the system distance. The optimal step size μ_{opt} therefore is

$$\mu_{\text{opt}}(n) = \frac{E\left\{\dfrac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\}}{E\left\{\dfrac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\}} \tag{36}$$

To simplify this result, one can assume that the L_2 norm of the input signal vector $\mathbf{u}(n)$ is approximately constant. This can be justified by the fact that in echo-canceling applications the length of this vector typically is in the order of 512–2048. Then, the optimal step size is given by

$$\mu_{\text{opt}}(n) \approx \frac{E\{e(n)\,e_u(n)\}}{E\{e^2(n)\}} \tag{37}$$

Since the undisturbed error $e_u(n)$ and the locally generated signal $n(n)$ are uncorrelated, this expression further simplifies to

$$\mu_{\text{opt}}(n) \approx \frac{E\{e_u^2(n)\}}{E\{e^2(n)\}} \tag{38}$$

Finally, the denominator may be extended using Eq. (15), and again the property that $e_u(n)$ and $n(n)$ are orthogonal:

$$\mu_{\text{opt}}(n) \approx \frac{E\{e_u^2(n)\}}{E\{e_u^2(n)\} + E\{n^2(n)\}} \tag{39}$$

Equation (39) emphasizes the importance of the undisturbed error $e_u(n)$, specifically, if there is a good match between the LEMS and the ECF, this term is small. If at the same time the local signal $n(n)$ is large, the optimal step size approaches zero; the adaptation freezes.

We have discussed here only a scalar step-size factor $\mu(n)$. This means that the same factor is applied to all filter coefficients. Numerous suggestions have been made to replace the scalar step size factor by a diagonal matrix in order to apply distinct factors to different elements of the coefficient vector. An example is an exponentially weighted step size taking into account the exponential decay of the impulse response of LEMS [34,35].

The implementation of the optimal step size needs the solution of a number of problems that will be discussed in the following section.

5.2. Implementation of the Optimal Step Size

The implementation of the optimal step size derived in the previous section requires the estimation of several quantities, including

- The expectations of signal powers have to be approximated by short-term estimates.
- An estimation method for the not-directly-observable undisturbed error $e_u(n)$ has to be derived.

5.2.1. Estimation of Short-Term Signal Power. Short-term signal power can be easily estimated by squaring the signal amplitude and smoothing this value by an IIR filter. A filter of first order proved to be sufficient [36]. If a rising signal amplitude should be detected faster than a falling one, a shorter time constant for a rising edge can be used than for a falling one. Typically both constants are chosen out of [0.9, 0.999]. Applying different time constants gives rise to a (small) bias that can be neglected in this application. Where squaring the signal amplitude causes a problem because of fixed-point arithmetic, the square can be replaced by the magnitude of the amplitude. Both square and magnitude are related by a factor depending on the probability density function of the signal amplitude. If two short-term estimates of signal powers are compared with each other, as is done in most cases in controlling the ECF, this factor cancels out.

5.2.2. Estimation of the Undisturbed Error. Two methods will be described to estimate the undisturbed error. The first one will use so-called delay coefficients. The second procedure compares signal powers at the input and the output of the LEMS. Both need supporting measures in order to distinguish between local activities and alterations of the impulse response of the LEMS.

5.2.2.1. Estimation via Delay Coefficients. Estimating the undisturbed error needs an estimate of the mismatch vector $\varepsilon(n)$ [see Eq. (12)]. Obviously, the impulse response vector $\mathbf{h}(n)$ of the LEMS is not known. However, if an artificial delay of N_D samples is inserted before the loudspeaker [37], the ECF also models this part of the unknown impulse response. The impulse response coefficients related to this delay are zero:

$$h_i(n) = 0 \quad \text{for } i = 0, \ldots, N_D - 1 \tag{40}$$

The NLMS algorithm has the property to distribute coefficient errors equally over all coefficients. Therefore, from the mismatch of the first N_D coefficients, one can estimate the system distance [see Eq. (13)]:

$$\hat{\Delta}(n) = \frac{N}{N_D} \sum_{i=0}^{N_D-1} w_i^2(n) \tag{41}$$

Assuming statistical independence of the input signal $u(n)$ and the filter coefficients, the optimal step size according to Eq. (38) is approximately given by

$$\mu_{\text{opt}}(n) \approx \frac{E\{u^2(n)\}\,\hat{\Delta}(n)}{E\{e^2(n)\}} \tag{42}$$

The performance of this method proves to be quite reliable. It has one deficiency, however. The update of the ECF freezes in case of a change of the impulse response of the LEMS. The reason for this behavior is that the coefficients related to the artificial delay remain equal to zero in that case. Therefore, applying this method requires an additional detector for changes of the LEMS.

Several methods are known [36]. A reliable indicator is based on a so-called *shadow filter*. This is a short adaptive filter in parallel to the ECF. Its step size is controlled only by the excitation signal $u(n)$. Consequently, it does not stop in case of a change of the LEMS. Since the shadow filter has far fewer coefficients than the ECF, it converges (but also diverges) much faster. During normal operation periods, the output signal of the shadow filter is inferior to the output signal of the ECF due to the nonoptimal step size control and the insufficient degree compared to the degree of the LEMS. Only immediately after a system change, the error calculated from the output of the shadow filter and the microphone output is smaller than the error based on the output signal of the ECF. If this situation is detected, the step size of the ECF adaptation is increased artificially in order to restart adaptation.

5.2.2.2. Estimation via Power Comparison. A second estimation method of the short-term power of the undisturbed error $E\{e_u^2(n)\}$ is based on the estimation of a so-called power transfer factor $\beta(n)$. If there is sufficient remote excitation and if there are no local activities $[n(n) = 0]$ this factor is given by

$$\beta(n) = \frac{E\{e^2(n)\}}{E\{u^2(n)\}} \tag{43}$$

For the expectations, short-term estimates (see Section 5.2.1) can be used. The estimation of the power transfer factor requires intervals of remote singletalk. These have

to be determined by a *doubletalk detector*. One method is based on a correlation measure ρ between the microphone output signal $y(n)$ and the output signal $\hat{d}(n)$ of the ECF. In case of a sufficiently converged ECF $\hat{d}(n)$ is a good estimate of the echo signal $d(n)$. Also, it is synchronized with $d(n)$. Therefore, the correlation must be evaluated only for a delay of zero. In contrast, correlation of $u(n)$ and $y(n)$ requires to search for the maximum of the measure. Further, it is advisable to normalize the correlation measure $\rho(n)$ defined as

$$\rho(n) = \frac{\left| \sum_{k=0}^{N_c-1} \hat{d}(n-k)\, y(n-k) \right|}{\sum_{k=0}^{N_c-1} |\hat{d}(n-k)\, y(n-k)|} \tag{44}$$

The value for N_c has to be determined by a compromise between a reliable correlation measure and the delay required for its calculation. Remote singletalk is assumed if $\rho(n)$ is larger than a given threshold. Since even in the case of singletalk, low local noise $n_n(n)$ may be present in the microphone output $y(n)$. Consequently, the estimate for the power coupling factor β may be too large. Therefore, the optimal step size $\mu_{\text{opt}}(n)$ should not exceed a given upper bound. For a reliable operation it is necessary to smooth the result of Eq. (44).

During intervals with no remote singletalk condition the power coupling factor $\beta(n)$ has to be frozen. Let the most recent doubletalk interval start at time n_1. Then the power transfer factor $\beta(n)$ is set to $\beta(n_1 - 1)$ during this interval.

With this modifications the factor $\beta(n)$ can replace $\hat{\Delta}(n)$ in Eq. (42):

$$\mu_{\text{opt}}(n) \approx \begin{cases} \dfrac{E\{u^2(n)\}\,\overline{\beta(n)}}{E\{e^2(n)\}} & \text{during remote singletalk} \\ \dfrac{E\{u^2(n)\}\,\overline{\beta(n_1-1)}}{E\{e^2(n)\}} & \text{during doubletalk} \end{cases} \tag{45}$$

where $\overline{\beta(n)}$ is a smoothed value of $\beta(n)$.

In general, doubletalk detectors can be based on other measures such as the cepstral distance, coherence, or the likelihood ratio [38–41].

6. ECHO CANCELLATION FOR STEREOPHONIC SYSTEMS

In stereophonic telecommunication systems there are four ECFs (per location) necessary to model the four echo paths between left and right loudspeakers and left and right microphones (see Fig. 13). In addition to providing the increased processing power a problem specific to stereophonic systems has to be solved — the signals on the left and the right channel originate from the same signal source and are separated only by the convolution with the impulse responses $\mathbf{g}_R(n)$ and $\mathbf{g}_L(n)$ of the transmission paths between signal source and left and right microphone at the remote location. Typically, both impulse responses exhibit minimal phase components that are invertible. Therefore, the impulse responses of the ECFs do not necessarily converge to the impulse responses of the related echo paths. Preprocessing of the loudspeaker input signals is required in order to decorrelate both signals. A large number of methods to achieve this goal have been suggested [42–48].

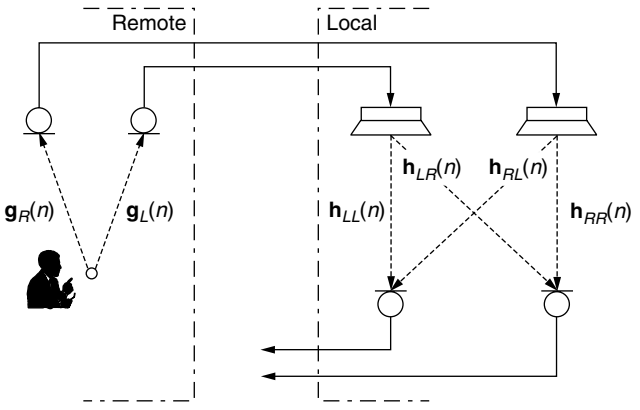


Figure 13. Signal paths in a stereophonic telecommunication system.

One method for decorrelation of left and right channels consists of introducing a nonlinearity into one or both of the channels. Up to a certain degree that depends on the quality of the audio equipment, this distortion proved to be not audible. A special suggestion is adding signals to the inputs of the loudspeakers that are nonlinear functions, such as half-wave rectifications of these signals [49]. Another way to achieve decorrelation is obtained by the insertion of a periodically varying delay [50,51]. The tolerable amount of delay has to be limited such that the spatial information is maintained. Other proposals consist of adding noise such that it is masked by the speech signal [52] and using audio coding methods [53].

7. SUBBAND ECHO CANCELLATION

Cancelling acoustic echoes in subbands (see Fig. 14) leads to a reduced processing load and increases the degrees of freedom for optimizing the echo cancelling system [54–57].

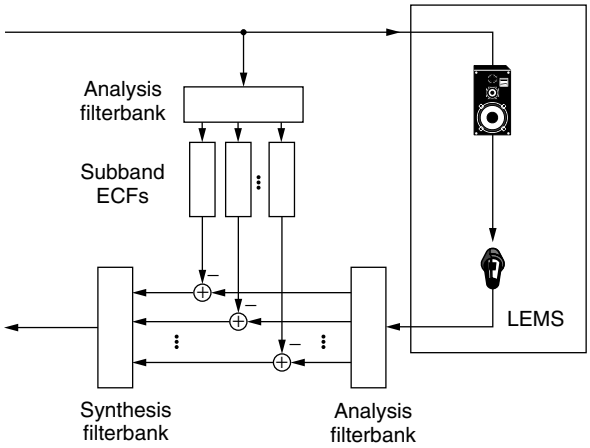


Figure 14. General structure of subband echo cancellation filters.

Prohibitive in many applications, however, is the enlarged signal delay (see Section 2.3) that is inevitably connected with the insertion of filterbanks.

Assume that the signal $u(n)$ will be split into K subbands. Then K filters have to be adapted. The subband signals can be decimated by a factor of K (critical decimation according to Nyquist's law) and the lengths of the subband ECFs can be reduced to N/K, assuming for simplicity that N is a multiple of K. Adaptation takes place after every K sampling intervals. Compared to full-band processing, the number of operations is given by

$$K(\text{filters}) \times \frac{1}{K}(\text{reduced filter lengths}) \times \frac{1}{K}(\text{extended updating interval}) \quad (46)$$

resulting in a reduction by the factor K. In addition, however, two analysis and one synthesis filterbanks are necessary. If the signals are split into equally wide subbands polyphase filters may be applied that can be efficiently implemented [58–62]. Critical decimation of subband signals may give rise to crosstalk problems between adjacent frequency bands. A decimation factor slightly smaller than K eases these problems [63–65]. Using the NLMS algorithm, the speed of convergence of the filter coefficients is inversely proportional to the filter length. For subband ECFs this length is N/K and the extended adaptation interval is compensated by the faster convergence. Splitting into subbands partially whitens the signals also causing a faster convergence (using the NLMS algorithm).

Additional savings in processing power are possible by using the degrees of freedom offered by subband processing for a better tuning of the echo-canceling system. The subband filters need not have the same lengths. Since the major part of the energy of speech signals is concentrated within the lower subbands (see Fig. 8) — and so is the echo — the ECFs of the upper bands can be shorter than those of the lower-frequency bands. In addition, the sound absorption of materials used in typical enclosures such as offices increases with frequency. Thus, echoes in higher-frequency bands are attenuated faster. A design criterion can be the contribution of the energy of the noncanceled echo tails to the energy of the total error $e(n)$ [66]. Finally, routines such as the doubletalk detector need to be implemented only in one subband. Choosing the band in which the speech signal has its highest energy enhances the reliability of the detection.

8. BLOCK PROCESSING SOLUTIONS

The adaptation of the coefficients of the ECF requires a huge amount of computational power. It can be reduced by applying block processing algorithms [67]. In this case an update takes place only after intervals of length $BT, B > 1$, where T is the sampling time. Between updates the data are assembled in blocks of length B. At each update, a block of B samples of the filter output signal is generated. The efficiency of block processing algorithms increases with the blocklength. On the other hand, however, collecting signal samples in blocks introduces a delay corresponding to the blocklength. Therefore, in time-critical applications, such as hands-free telephones, only short blocks can be tolerated.

Since adaptation takes place only once in B sampling intervals, the speed of convergence of the filter coefficients is reduced accordingly. To speed up convergence, it is possible to correct the error signal (i.e., a vector of length B in block processing procedures) such that it is identical to the error signal generated by adaptation at each sampling instant [68–70]. In this case the NLMS algorithm based on block processing behaves exactly such as the ordinary NLMS algorithm. The filter adaptation may be performed in the time or in the frequency domain [71].

An desirable choice of the blocklength B would be the length N of the ECF [72]. In typical applications, however, this filter has to cover echoes that are in the order of 32–125 ms long. A blocklength equal to N, therefore, would introduce a nontolerable signal delay. To overcome this problem, the ECF has to be partitioned into subfilters of smaller length [73]. Thus the blocklength and the delay related to it can be tailored to the specific needs of the application. An efficient algorithm is available that combines the error signal correction and the partitioning of the filter impulse response with an overlap-save implementation calculating the subfilter updates and the filter output in the frequency domain [11,12].

9. CONCLUSIONS AND OUTLOOK

Powerful and affordable acoustical echo-canceling systems are available. Their performance is satisfactory, especially if compared to solutions in other voice processing areas such as speech recognition or speech-to-text translation. The fact that echo control systems have not yet entered the market on a large scale seems not to be a technical but a marketing problem — a customer who buys a high-quality echo suppressing system pays for the comfort of his her communication partner. Using a poor system only affects the partner at the far end, who usually is too polite to complain.

Future research and development in the area of acoustic echo cancellation certainly will not have to take into account processing power restrictions. This has a number of consequences; the implementation of even sophisticated procedures on ordinary (office) PCs will be possible. This will make it easier to test modifications of existing procedures or completely new ideas in real time and in real environments. The performance of future systems will approach limits given only by the environment they have to work in. It will no longer be limited by the restricted capabilities of affordable hardware. It will depend only on the quality of the algorithms implemented.

This does not necessarily mean that future systems will be perfectly reliable in all situations. The reliability of estimation procedures used to detect system states such as a change of the impulse response of the LEMS or the beginning of doubletalk depends on the length of the usable data record. Since, however, the working environment is highly time-varying and nonstationary the usage of too long records can cause the loss of the real-time capability.

Up to now the NLMS algorithm plays the role of the "workhorse" for acoustic echo cancelling. The AP algorithm offers improved performance at modest additional implementation and processing cost. It does not cause stability problems that are difficult to solve. Rules for step-size control used for the NLMS algorithm, however, have to be reconsidered.

Customer demands are increasing with time. Using available systems, customers will certainly ask for better performance. Therefore, the need for new and better ideas will remain. Acoustic echo canceling will continue to be one of the most interesting problems in digital signal processing.

BIOGRAPHY

Eberhard Hänsler received his degrees (Dipl.-Ing., 1961, Dr.-Ing., 1968) in Electrical Engineering from Darmstadt University of Technology, Darmstadt, Germany. He worked with the Research Institute of the German PTT (1961–1963), the Electrical Engineering Department of Darmstadt University of Technology (1963–1968), and with the IBM Research Division at Zurich and Yorktown Heights (1968–1974). Since 1974 he has been Full Professor for Signal Theory at Darmstadt University of Technology.

His research interests are signal and system theory, adaptive systems, digital signal processing, echo cancellation, and noise reduction. He has been working on the hands-free telephone problem for several years. He is cofounder of the biennial International Workshop on Acoustic Echo and Noise Control (IWAENC), and has organized sessions on this topic at several international conferences and acted as guest editor of special issues on acoustic echo and noise control of *signal processing* (January 1998) and of the *European Transactions on Telecommunications* (March/April 2002).

Together with his group, he received the Annual European Group Technical Achievement Award in 2000 for "major contributions in the design and implementation of acoustic echo and noise control system."

BIBLIOGRAPHY

1. The new Bell telephone, *Sci. Am.* **37**: 1 (1877).
2. W. F. Clemency, F. F. Romanow, and A. F. Rose, The Bell system speakerphone, *AIEE. Trans.* **76**(I): 148–153 (1957).
3. D. A. Berkley and O. M. M. Mitchell, Seeking the ideal in "hands-free" telephony, *Bell Lab. Rec.* **52**: 318–325 (1974).
4. G. Pays and J. M. Person, Modèle de laboratoire d'un poste téléphonique à haut-parleur, *FASE* **75**: 88–102 (Paris) (1975).
5. E. Hänsler, The hands-free telephone problem — an annotated bibliography, *Signal Process.* **27**: 259–271 (1992).
6. E. Hänsler, The hands-free telephone problem — an annotated bibliography update, *Annales des Télécommunications* **49**: 360–367 (1994).
7. E. Hänsler, The hands-free telephone problem — a second annotated bibliography update, *Proc. 4th Int. Workshop on Acoustic Echo and Noise Control*, 1995, pp. 107–114.
8. A. Gilloire et al., Innovative speech processing for mobile terminals: An annotated bibliography, *Signal Process.* **80**(7): 1149–1166 (2000).
9. A. Gilloire, State of the art in acoustic echo cancellation, in A. R. Figueiras and D. Docampo, eds., *Adaptive Algorithms: Applications and Non Classical Schemes*, Univ. Vigo, 1991, pp. 20–31.
10. A. Gilloire, E. Moulines, D. Slock, and P. Duhamel, State of the art in acoustic echo cancellation, in A. R.Figueiras-Vidal, ed., *Digital Signal Processing in Telecommunications*, Springer, London, 1996, pp. 45–91.
11. C. Breining et al., Acoustic echo control, *IEEE Signal Process. Mag.* **16**(4): 42–69 (1999).
12. S. L.Gay and J.Benesty, eds., *Acoustic Signal Processing for Telecommunication*, Kluwer, Boston, 2000.
13. J. Benesty et al., *Advances in Network and Acoustic Echo Cancellation*, Springer, Berlin, 2001.
14. H. Kuttruff, Sound in enclosures, in M. J. Crocker, ed., *Encyclopedia of Acoustics*, Wiley, New York, 1997, pp. 1101–1114.
15. J. B. Allen and D. A. Berkley, Image method for efficiently simulating small-room acoustics, *J. Acoust. Soc. Am.* **65**: 943–950 (1975).
16. M. Zollner and E. Zwicker, *Elektroakustik*, Springer, Berlin, 1993.
17. M. Mboup and M. Bonnet, On the adequateness of IIR adaptive filtering for acoustic echo cancellation, *Proc. EUSIPCO-92*, Brussels, Belgium, 1992, pp. 111–114.
18. A. P. Liavas and P. A. Regalia, Acoustic echo cancellation: Do IIR filters offer better modelling capabilities than their FIR counterparts? *IEEE Trans. Signal Process.* **46**(9): 2499–2504 (1998).
19. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
20. International Telecommunication Union, *Acoustic Echo Controllers*, ITU-T Recommendation G.167, 1993.
21. International Telecommunication Union, *Control of Talker Echo*, ITU-T Recommendation G.131, 1996.
22. International Telecommunication Union, *Relation Between Echo Disturbances under Single Talk and Double Talk Conditions* (*Evaluated for One-Way Transmission Time of 100 ms*), ITU-T Recommendation G.131(App. II), 1999.
23. M. R. Schroeder, Improvement of acoustic-feedback stability by frequency shifting, *J. Acoust. Soc. Am.* **36**: 1718–1724 (1964).
24. S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice-Hall, Englewood Cliffs, NJ, 2002.
25. E. Hänsler and G. U. Schmidt, Hands-free telephones — joint control of echo cancellation and post filtering, *Signal Process.* **80**: 2295–2305 (2000).
26. T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice-Hall, Upper Saddle River, NJ, 2002.
27. R. Martin and P. Vary, Combined acoustic echo control and noise reduction for hands-free telephony — state of the art and perspectives, *Proc. EUSIPCO-96*, Triest, Italy, 1996, pp. 1107–1110.
28. S. Gustafsson, R. Martin, and P. Vary, Combined acoustic echo control and noise reduction for hands-free telephony, *Signal Process.* **64**: 21–32 (1998).

29. K. Ozeki and T. Umeda, An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, *Electron. Commun. Jpn.* **67-A**(5): 19–27 (1984).
30. S. Gay and S. Travathia, The fast affine projection algorithm, *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 3023–3027.
31. V. Myllylä, Robust fast affine projection algorithm for acoustic echo cancellation, *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, 2001, pp. 143–146.
32. D. Slock and T. Kailath, Fast transversal RLS algorithms, in N. Kalouptsidis and S. Theodoridis, eds., *Adaptive System Identification and Signal Processing Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
33. G.-O. Glentis, K. Berberidis, and S. Theodoridis, Efficient least squares adaptive algorithms for FIR transversal filtering: A unified view, *IEEE Signal Process. Mag.* **16**(4): 13–41 (1999).
34. S. Makino and Y. Kaneda, Exponentially weighted step-size projection algorithm for acoustic echo cancellers, *IECE Trans. Fund.* **E75-A**: 1500–1507 (1992).
35. S. Makino, Y. Kaneda, and N. Koizumi, Exponentially weighted step-size NLMS adaptive filter based on the statistics of a room impulse response, *IEEE Trans. Speech Audio Process.* **1**: 101–108 (1993).
36. A. Mader, H. Puder, and G. Schmidt, Step-size control for acoustic echo cancellation filters — an overview, *Signal Process.* **80**: 1697–1719 (2000).
37. S. Yamamoto and S. Kitayama, An adaptive echo canceller with variable step gain method, *Trans. IECE Jpn.* **E65**: 1–8 (1982).
38. T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, Double-talk detector based on coherence, *IEEE Trans. Commun.* **COM-44**: 1421–1427 (1996).
39. K. Ghose and V. U. Redd, A double-talk detector for acoustic echo cancellation applications, *Signal Process.* **80**: 1459–1467 (2000).
40. H. Ye and B. Wu, A new double-talk detection algorithm based on the orthogonality theorem, *IEEE Trans. Commun.* **COM-39**: 1542–1545 (1991).
41. A. H. Gray and J. D. Markel, Distance measures for speech processing, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-24**: 380–391 (1976).
42. J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, Adaptive filtering algorithms for stereophonic acoustic echo cancellation, *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 3099–3102.
43. S. Shimauchi and S. Makino, Stereo projection echo canceller with true echo path estimation, *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 3059–3062.
44. F. Amand, J. Benesty, A. Gilloire, and Y. Grenier, A fast two-channel projection algorithm for stereophonic acoustic echo cancellation, *Proc. ICASSP-95*, Atlanta, GA, 1996, pp. 949–952.
45. S. Shimauchi, Y. Haneda, S. Makino, and Y. Kaneda, New configuration for a stereo echo canceller with nonlinear pre-processing, *Proc. ICASSP-98*, Seattle, OR, 1998, pp. 3685–3688.
46. S. Shimauchi et al., A stereo echo canceller implemented using a stereo shaker and a duo-filter control system, *Proc. ICASSP-99*, Phoenix, AZ, 1999, pp. 857–860.
47. T. Gänsler and J. Benesty, New insights to the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution, *IEEE Trans. Speech Audio Process.* **9**: 686–696 (1998).
48. A. Sugiyama, Y. Joncour, and A. Hirano, A stereo echo canceler with correct echo-path identification based on an input-sliding technique, *IEEE Trans. Signal Process.* **49**: 2577–2587 (2001).
49. J. Benesty, D. R. Morgan, and M. M. Sondhi, A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation, *IEEE Trans. Speech Audio Process.* **6**: 156–165 (1998).
50. Y. Joncour and A. Sugiyama, A stereo echo canceler with preprocessing for correct echo-path identification, *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 3677–3680.
51. M. Ali, Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation, *Proc. ICASSP-98*, Seattle, OR, 1998, pp. 3689–3692.
52. A. Gilloire and V. Turbin, Using auditory properties to improve the behavior of stereophonic acoustic echo cancellers, *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 3681–3684.
53. T. Gänsler and P. Eneroth, Influence of audio coding on stereophonic acoustic echo cancellation, *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 3649–3652.
54. I. Furukawa, A design of canceller of broad band acoustic echo, *Int. Teleconf. Symp.*, Tokyo, Japan, Jan. 8–Aug. 8, 1984.
55. A. Gilloire, Adaptive filtering in sub-bands, *Proc. ICASSP-88*, New York, 1988, pp. 1572–1576.
56. W. Kellermann, Analysis and design of multirate systems for cancellation of acoustical echoes, *Proc. ICASSP-88*, New York, 1988, pp. 2570–2573.
57. W. Kellermann, Zur Nachbildung physikalischer Systeme durch parallelisierte digitale Ersatzsysteme im Hinblick auf die Kompensation akustischer Echos, *Fortschr.-Ber. VDI Reihe* 10(102), VDI Verlag, Düsseldorf, Germany, 1989.
58. R. E. Chrochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
59. P. Vary and G. Wackersreuther, A unified approach to digital polyphase filter banks, *AEÜ Int. J. Electron. Commun.* **37**: 29–34 (1983).
60. G. Wackersreuther, On the design of filters for ideal QMF and polyphase filter banks, *AEÜ Int. J. Electron. Commun.* **39**: 123–130 (1985).
61. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
62. P. P. Vaidyanathan, Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial, *Proc. IEEE* **78**: 56–93 (1990).
63. S. Weiss, R. W. Stewart, A. Stenger, and R. Rabenstein, Performance limitations of subband adaptive filters, *Proc. EUSIPCO-98*, Rhodos, Greece, 1998, pp. 1245–1248.
64. S. Weiss, *On Adaptive Filtering in Oversampled Subbands*, Ph.D. dissertation, Dept. Electronic and Electrical Engineering, Univ. Strathclyde, May 1998.
65. G. U. Schmidt, Entwurf und Realisierung eines Multiratensystems zum Freisprechen, *Fortschr.-Ber. VDI Reihe* 10(674), VDI Verlag, Düsseldorf, Germany, 2001.
66. G. U. Schmidt, Acoustic echo control in subbands — an application of multirate systems, *Proc. EUSIPCO-98*, Rhodos, Greece, 1998, pp. 1961–1964.

67. J. Shynk, Frequency-domain and multirate adaptive filtering, *IEEE Signal Process. Mag.* **9**(1): 14–37 (1992).

68. J. Benesty and P. Duhamel, A fast exact least mean square adaptive algorithm, *IEEE Trans. Signal Process.* **40**: 2904–2920 (1992).

69. B. Nitsch, The partitioned exact frequency domain block NLMS algorithm, a mathematically exact version of the NLMS algorithm working in the frequency domain, *AEÜ Int. J. Electron. Commun.* **52**: 293–301 (1998).

70. B. Nitsch, Real-time implementation of the exact block NLMS algorithm for acoustic echo control in hands-free telephone systems, in S. L. Gay and J. Benesty, eds., *Acoustic Signal Processing for Telecommunication*, Kluwer, Boston, 2000.

71. B. Nitsch, A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain, *Signal Process.* **80**: 1733–1745 (2000).

72. A. O. Ogunfumni and A. M. Peterson, Fast direct implementation of block adaptive filters, *Proc. ICASSP-89, Glasgow*, UK, 1989, pp. 920–923.

73. P. Estermann and A. Kaelin, A hands-free phone system based on partitioned frequency domain adaptive echo canceller, *Proc. EUSIPCO-96*, Trieste, Italy, 1996, pp. 1131–1134.

ACOUSTIC MODEMS FOR UNDERWATER COMMUNICATION

KENNETH SCUSSEL
Benthos, Inc.
North Falmouth, Massachusetts

1. INTRODUCTION

Conventional in-air wireless communications typical rely on RF or electromagnetic means to convey information. Because such signals do not propagate well under water, sound waves, or acoustic signals, are the obvious choice for wireless underwater communication. This article discusses the development of modems for acoustic communications (acomms) and provides an overview of illustrative applications of such modems. The acomms channel is considerably more difficult than those encountered in typical terrestrial applications. The acoustic modem has to overcome several major obstacles in using the underwater channel: slow propagation, limited bandwidth, nonuniform propagation conditions, multipath, and low signal-to-noise ratio (SNR). The speed of sound in water is approximately 1500 m/s, while electromagnetic signals travel at nearly the speed of light. Thus, at practical distances (>2 km), the propagation delay for acoustic signals is measured in seconds compared to the nearly instantaneous propagation of RF signals. Another problem is that high-frequency sound waves are severely attenuated, so operation at any practical distance requires that the acoustic signal be less than approximately 30 kHz. At these frequencies, the current state of the art in transducer development limits the available bandwidth, which is a major obstacle to achieving high data rates. A third major obstacle is that the speed of sound in water varies with temperature and salinity. Thus, if an acoustic signal encounters a large temperature or salinity gradient, its path is bent or refracted, leaving "holes" in the channel where acoustic energy may be greatly reduced. This can result in a shadow zone where no communication is possible. The next major obstacle is multipath reflections from the seabed, surface, or other boundary. This can cause destructive interference resulting in frequency-dependent fading or intersymbol interference (ISI). Finally, acoustic modems may be required to operate at very low SNR caused by a combination of ambient noise and interference, limited transmitter source level, and transmission loss. At short ranges, the transmitted signal spreads spherically from its source. This transmission loss (TL) can be expressed as a function of range, $\mathrm{TL} = 20 * \log_{10}$ (range). In addition, there may be high levels of ambient noise present because of weather conditions at the sea surface and shipping noise. Despite these challenges, several companies have developed commercially available acoustic modems. There is a small but growing demand for wireless underwater telemetry in applications where cables are impractical or simply too expensive. These include command and control for autonomous undersea vehicles (AUVs), instrumentation around oil fields, any application in areas that are frequently fished by bottom trawlers, and deep-water instrumentation, to name a few. This article discusses the architecture of acoustic communication system, the hardware and software implementation of commercially available acoustic modems, and some examples of real-world applications.

2. ARCHITECTURE OF ACOUSTIC COMMUNICATION SYSTEMS

Acoustic modems generally segment an incoming stream of data or information bits into smaller blocks called *packets*, each of which is transmitted as an individual waveform over the physical channel. The acoustic communication system can be divided into multiple layers similar to the Open System Interconnection (OSI) reference model. The lowest layer is the physical layer, which, on the transmitter side, applies error correction coding, modulates the message into a passband waveform and passes it into the channel. On the receive side, the physical layer consists of those functions that acquire and align the waveform, demodulate the message, and finally decode the message and present it to the next layer. The higher levels are dependent on the application. For point-to-point communications, where only two modems are involved, the link makes use only of the protocol layer, which is similar to the media access control (MAC) layer in the OSI model. This layer handles any packet retransmissions and is responsible for presenting the data to the users. In the case of networks of acoustic modems, there is a MAC layer and a network layer. The implementation of the acoustic modem networks uses a communication protocol that contains many elements similar to the IEEE 802.11 protocol. This section focuses on the physical layer.

The first portion of a packet processed by the physical layer is the acquisition. The acquisition signal is prepended to the beginning of every packet and can

be any waveform that is used to detect the presence of the packet and to synchronize the receiver. This is the most important part of the receive processing, as without properly detecting and aligning the receiver to the start of a packet, it is impossible to receive any of the data that follow. Therefore it is important that this waveform be as robust as possible, to assure detection of the packet. The most robust signal is one that uses all the available bandwidth and is as long as practical. The temporal duration of the waveform is limited by additional receiver complexity associated with processing longer signals. In addition, acquisition-related portions of the waveform do not convey any data and thus are overhead, which reduces the actual data throughput. Therefore, the selection of the waveform is a tradeoff between reliability, overhead, and receiver complexity. A couple of examples of possible acquisition waveforms are a linear frequency-modulated (LFM) chirp, or a pseudorandom broadband signal. Both types of signals are processed with a replica correlator (an envelope-detected matched filter). The replica correlator transforms a T second waveform with an input SNR of S_{in} to a compressed peak with an output SNR of $S_{\text{out}} \sim 2TWS_{\text{in}}$, where W is the effective signal bandwidth. The effective duration of the compressed peak is approximately $1/W$ (second). Given a priori knowledge of the temporal offset from the edge of the chirp to the edge of the message portion, one judges the start of the received message by the same distance relative to the correlator peak location.

Following the acquisition are the modulated data. Modulation is a technique used to transmit information or data using an "elemental" signal occupying a specific portion of the time–frequency band. Historically, the elemental waveform is usually called a "chip." The (information) data are usually obtained as a binary string of 1s and 0s. Generally, modulation of digital data is accomplished by varying the amplitude, frequency, or phase (or a combination thereof) of a sinusoidal chip. Amplitude modulation is problematic in the ocean environment and requires a high SNR, so it is seldom used in underwater acoustic modems. Frequency-shift-keyed (FSK) modulation is the predominant method used in low-data-rate, noncoherent modems.

The simplest form of FSK techniques is binary FSK, or BFSK. BFSK uses two discrete frequency "slots" within an allocated time–frequency block, where a logic 1 is represented by the first frequency and a logic 0 is represented by the second frequency. By switching between the two frequencies, a stream of digital data can be sent. There are several variations of frequency modulation: (1) *multiple frequency shift keying* (MFSK), using multiple frequency slots within a block; and (2) *frequency hopping*, in which blocks are hopped about the available signal band, so that only a few (generally one) tones are transmitted at one baud interval. To achieve the densest mapping of frequency slots, the width of the slots should be equal to 1/(chip duration). This orthogonal signaling will prevent adjacent frequencies from interfering with each other and will make the maximum use of the available bandwidth. All the FSK methods can be received simply by measuring the signal energy and ignoring the phase of the signal, and thus are usually referred to as "noncoherent." Typically, the signal energy in each of the M slots is compared, and the slot with the most energy is selected. If there is broadband noise, it will affect each frequency slot equally and the correct decision will still be the frequency bin with the most energy. However, in a multipath environment the transmitted tone could be lost as a result of frequency-dependent fading, resulting in selection of the wrong slot, thereby causing multiple bit errors.

Another way to map the data to the available spectrum is to divide the entire band into N slots without regard to blocks. We map sequential clusters of 5 data bits into one group of 20 slots, which is drawn from $2^5 = 32$ possible combinations of 20 slots. These 20 slot codewords are derived from a family of Hadamard codes. Each codeword consists of 10 ones and 10 zeros, and has the property that each has a minimum distance of 10. The advantage is that if a tone is lost, the receiver employs a soft decision algorithm to pick the codeword that is the closest match to one of the 32 possible codewords. This method is referred to as *Hadamard MFSK* and is effective in both the presence of noise and multipath, at the expense of bandwidth efficiency. This method provides for coding gain, which means that the modem can operate at a lower SNR, with fewer errors, all at the expense of a lower data rate.

Among the noncoherent techniques, those that provide the most transmitted energy to a single tone, and those that can provide some immunity to frequency-dependent fading are generally more reliable in a given channel. Thus, for a given chip duration, frequency hopping will be more reliable at low SNRs than will the other techniques, with Hadamard MFSK performing midway between the other two. In all cases, this conclusion assumes that the electronic and channel spectral response is flat across the signal band. With substantial frequency-dependent attenuation, Hadamard signaling will be degraded more than the other techniques.

Yet another technique for message modulation is to vary the phase of the carrier. The simplest form is binary phase shift keying (BPSK), where a 1 is represented by one phase and a 0 is represented by a 180° phase shift. Information bits can be clustered and transformed into an "alphabet," permitting modulation that is more compact. For example, there are precisely four ways to combine 2 bits. We can therefore modulate the phase in 90° increments to reflect any of the four combinations. This is referred to as *quadrature PSK*, *QPSK*, or *4PSK*. The phase can be broken up into even more divisions. The process of receiving phase-shifted data is referred to as "coherent" processing. Coherent techniques require much more sophisticated processing as the acoustic channel can severely affect the phase of the received signal. The state-of-the-art receiver uses a decision feedback equalizer (DFE) to remove the effects of multipath (both amplitude and phase distortion), in an attempt to convert the received signal into one similar to what was transmitted. This method is usually successful, but requires relatively high SNR, and cannot tolerate very rapid variation in the channel multipath.

3. HARDWARE IMPLEMENTATION

Designing acoustic modems requires overcoming two challenges: (1) an intensive amount of processing is required to overcome the obstacles presented by the underwater acoustic channel and (2) most acoustic modems are battery-powered and thus must operate with a minimum of power. Advancements in low-power digital signal processors (DSPs) have made commercially available acoustic modems possible. Although there are several vendors producing commercially available acoustic modems Benthos (formally Datasonics) was one of the first and is the leading supplier of acoustic modems. This article discusses the hardware and signal processing software of a typical Benthos acoustic modem.

Figure 1 is a block diagram of a typical acoustic modem. The DSP is the main component of the modem and implements the signal processing software required to generate the transmit signals and the processing required to make sense of the received signals. A fixed-point DSP is used, since it provides the required processing power with less energy and memory demand than a floating-point DSP.

The DSP generates the transmit signal and sends it to the D/A converter. The analog output of the D/A is then put into a power amplifier, which boosts the level of the transmit signal to generate sufficient acoustic energy or source level at the transducer. The transmit signal level is adjustable, allowing power control to deliver sufficient but not excessive SNR at the receiver. Power control provides transmission security, power conservation, and improved multiple-access networking. The power amplifier drives the transmitter/receiver (T/R) network and matching network. The T/R network prevents the sensitive receiver from being damaged by the large transmitted waveform, and the matching network is used to match the output impedance of the power amplifier to the transducer's impedance. The transducer is a piezoelectric ceramic, which converts the electrical transmit signal to an acoustic signal.

Received acoustic signals are generally very small and pass through the T/R network to the preamplifier. The output of the preamplifier goes into either the receiver or a wideband amp, depending on the mode of the DSP. The DSP operates in either a low-power mode or an active mode. In the low-power mode, the DSP runs off the slow clock and all functions are shut down except the wakeup logic. In this mode, the DSP is capable of processing only coded wakeup signals. This allows the modem to operate in a standby mode with a very low current drain. In active mode, the DSP runs off the fast clock, allowing all incoming acoustic signals to be processed. The received signal can have a large dynamic range. Thus, automatic gain control (AGC) is used to maintain a nearly constant signal level at the input of the A/D. The AGC is controlled by the DSP, which measures the signal level at the A/D. Large signals are attenuated and gain is applied

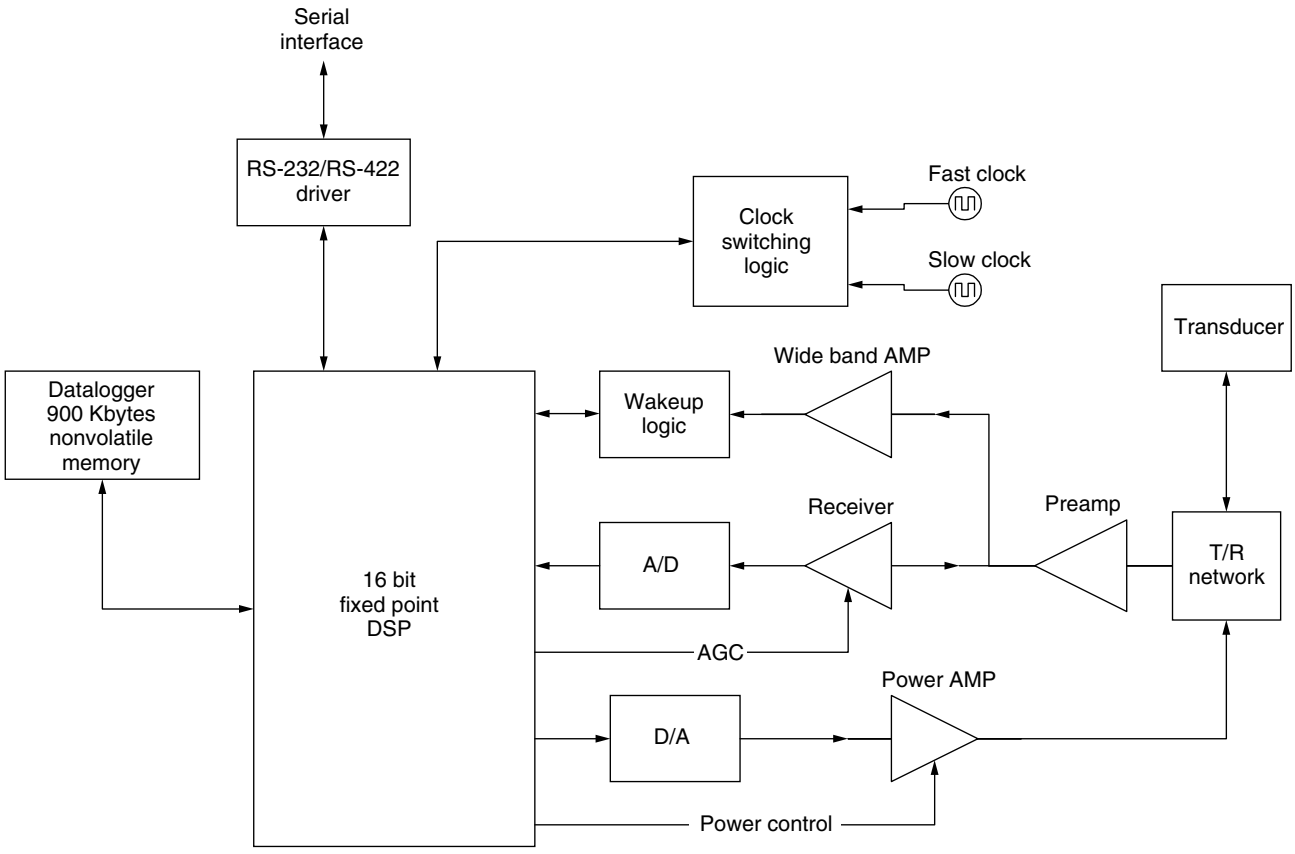


Figure 1. Top-level block diagram of a typical Benthos acoustic modem.

to small signals. The sampling rate of the A/D is set to greatly oversample incoming signals to reduce the need for expensive antialiasing filters.

Other peripherals include a serial interface and a datalogging capability. The serial interface allows control of the modem by a host PC or other processor through a standard RS-232 or RS-422 interface. RS-232 allows connection to most PCs and instruments at rates of $\leq$9600 baud. RS-422 allows the modem electronics to be placed closer to the transducer, allowing the host PC or instrument to be several kilometers from the electronics. This provides an alternative to noisy, lossy analog connections to the transducer via long cables. A datalogging capability using 900 kB (kilobytes) of nonvolatile memory is available for buffering and storage of incoming data. In many applications, it is desirable to store data for some time before acoustic transmission.

The modem electronics can be packaged in a variety of configurations. The simplest configuration is the modem board set shown in Fig. 2. In this configuration, the printed-circuit boards are mounted to a chassis, and are externally powered. Usually this configuration is used by original equipment manufacture (OEM) applications, where the modem is integrated with the instrumentation of another manufacturer. Another configuration is to package the modem with batteries in a self-contained pressure housing, for deployment underwater. The required water depth or pressure determines the type of housing. Figure 2 shows an ATM-885 acoustic modem packaged in a hardcoat anodized aluminum housing with a maximum depth rating of 2000 m. Note that the housing contains a connector for external power and the serial interface for connection to host equipment. For shipboard operation the electronics can also be installed in an AC-powered shipboard deckbox or 19-in. rack. The shipboard equipment makes use of an over the side or remote transducer. Figure 2 shows photographs of the AC-powered shipboard deckbox and a remote transducer. The final option is for the modem electronics to be packaged in a buoy containing a RF modem and a GPS receiver. The modem's transducer hangs below the buoy, and any data received by the modem are relayed via the RF link to a shore station, and vice versa.

4. SIGNAL PROCESSING IMPLEMENTATION

The transmit signal processing is as is shown in Fig. 3. The data or information bits are first convolutionally encoded for error correction. The output of the data encoder is mapped to MFSK frequency tones. Tones for Doppler tracking are also added. The phase of the frequency-domain signal is randomized to avoid large peak : average power ratio signals at the output. The spectrum is inverse Fast Fourier–transformed to obtain a time-domain baseband signal sampled at 10,240 Hz. The baseband signal is then interpolated to 163 kHz rate and quadrature-mixed to a passband carrier prior to digital-to-analog conversion.

The receiver signal processing is shown in Figure 4. Acoustic data sampled at 163 kHz are obtained from the A/D converter and resampled at a slightly different sample rate depending on the Doppler shift present in the communication channel. Automatic gain control detects the signal level and adjusts it as necessary via external hardware. A quadrature mixer converts the signal to complex baseband. The baseband signal is decimated to a 10,240-Hz sample rate. During the acquisition period, matched-filter processing is performed to look for the acquisition signal. Once the acquisition signal is detected, it is used to synchronize to the incoming data. Adjustments

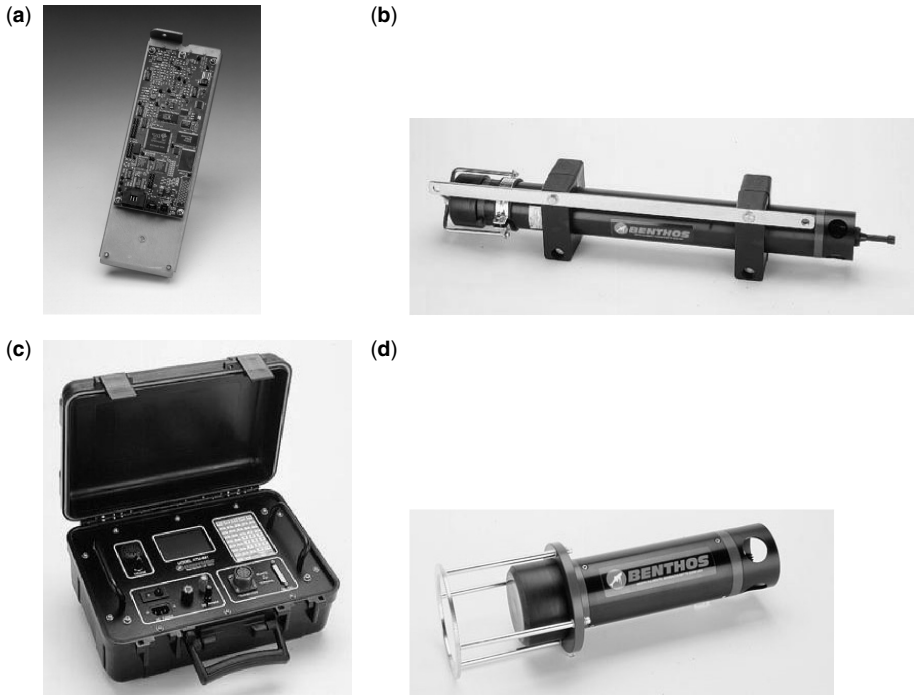


Figure 2. ATM-88x modem components: (**a**) OEM board set; (**b**) ATM-885; (**c**) ATM-881 deckbox; (**d**) remote transducer.

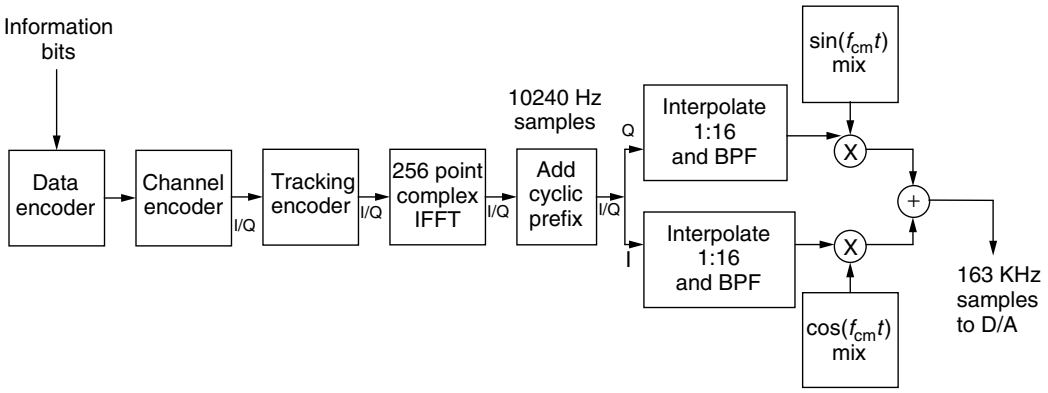


Figure 3. Transmit signal processing.

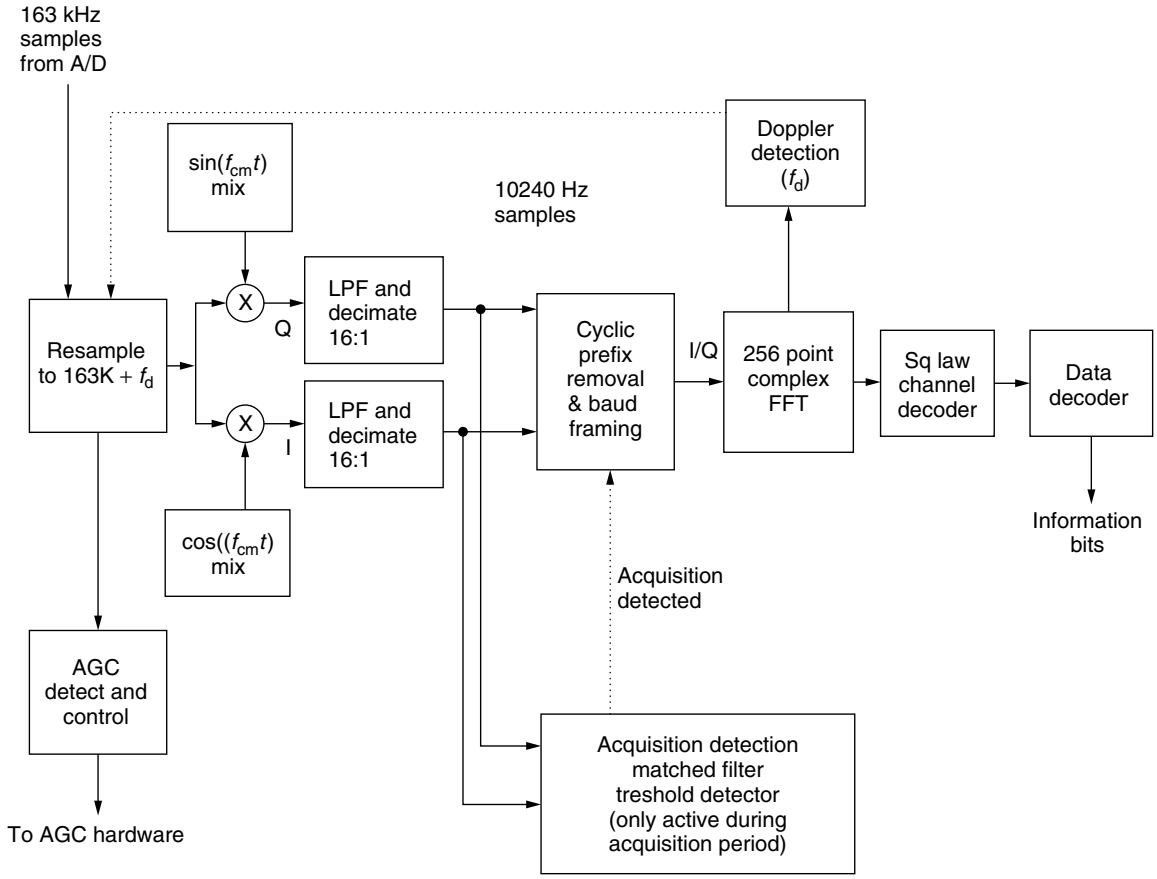


Figure 4. Receive signal processing.

are made for timing errors and to eliminate the prepended signal used to guard against multipath. A complex 256-point FFT converts the signal to the frequency domain, which is used for MFSK decoding and Doppler tracking. If a Doppler shift is detected, the next incoming samples are resampled to adjust for the shift. The frequency-domain data are then run through a square-law channel decoder to obtain the magnitude of the signal. This decodes the MFSK data. Finally, a Viterbi decoder interprets the convolutionally encoded data. The final information bits are sent to the user.

5. MODEM APPLICATIONS

The are numerous diverse applications for underwater acoustic modems. A few sample applications are described below.

5.1. Cone Penetrometer

In one modem application, wireless control and real-time data recovery are effected using a deep-water seabed cone penetrometer (CPT). Guardline Surveys, located in Great Yarmouth, England, uses the CPT for penetrating the

seabed in water depths of ≤2000 meters. In operation, the instrument is lowered over the side from a ship and to the seabed. The instrument is fitted with pressure and temperature sensors as well as inclinometers to assure proper attitude and stability. In the past, the CPT used an expensive electromechanical cable both to lower the instrument and for communication purposes. Guardline removed the constraints imposed by the electrical umbilical cable, replacing it with Benthos acoustic modems. With the modems the operator can communicate with the CPT all the way down during deployment to the seabed, sending commands, and receiving status information. During penetration, data from the sensors are sent to the operator in real time. With real-time remote recovery of data, the CPT can be lifted just off the bottom and maneuvered to another nearby site. Figure 5 is an illustration of the CPT in operation.

5.2. Pipeline Bending

Another acoustic modem application is the remote acquisition of pipeline bending stresses and vortex-induced vibration (VIV) from an offshore oil–gas platform. A monitoring project was established by the Petrobras R&D center to collect data vibrations and tensions on the mooring lines and risers. Petrobras contracted with Scientific Marine Services (SMS) to provide the instrumentation. It was determined that cables connecting the subsea instrumentation to the surface would have a minimum survival probability during the difficult pipe laying operations; therefore acoustic communication was selected. Benthos acoustic modems provide both the downlink command and control signaling and uplink data acquisition. The acoustic modems provide the means for controlling the VIV bottle synchronization, data uplink repetition rate, as well as other operating parameters. The data rate used is 600 bps (bits per second) at a range of 1300 m. Figure 6 is an illustration of the deployment.

5.3. Imaging and Telemetry

Command and control of an autonomous underwater vehicle is a growing application for acoustic modems. During one experiment,[1] a robotic crawler[2] carrying an acoustic modem, camera, and a digital signal processing unit was used to search autonomously for an underwater object. When the object was found, the robot informed the user using acomms that "something" had been found. The robot was then acoustically commanded to take a still-frame picture, compress it, and transmit using the acoustic modem. The grayscale images shown in Fig. 7 each consist

[1] The U.S. Navy's "AUVfest 2001" held off of Gulfport, Mississippi in October 2001.
[2] Developed by Foster-Miller Inc. for the U.S. Navy's Coastal Systems Station (CSS).

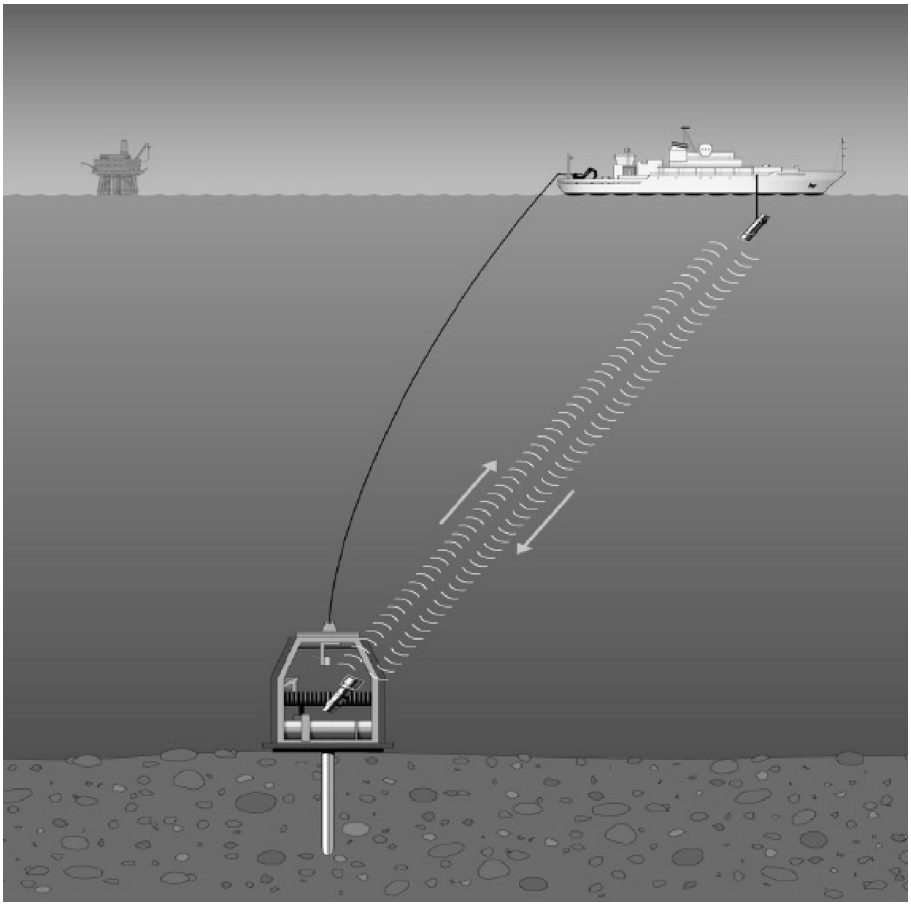


Figure 5. Illustration of penetrometer operation incorporating the telesonar acoustic modems for transmission of real-time command, control, and penetrometer data.

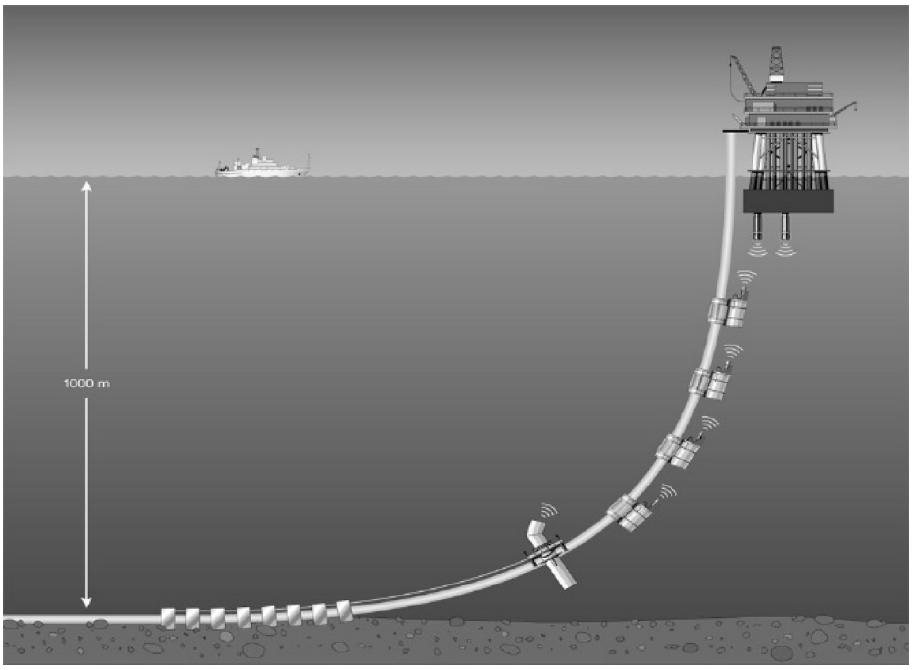


Figure 6. Illustration of instrumentation deployment scenario incorporating the Telesonar Acoustic Modems.

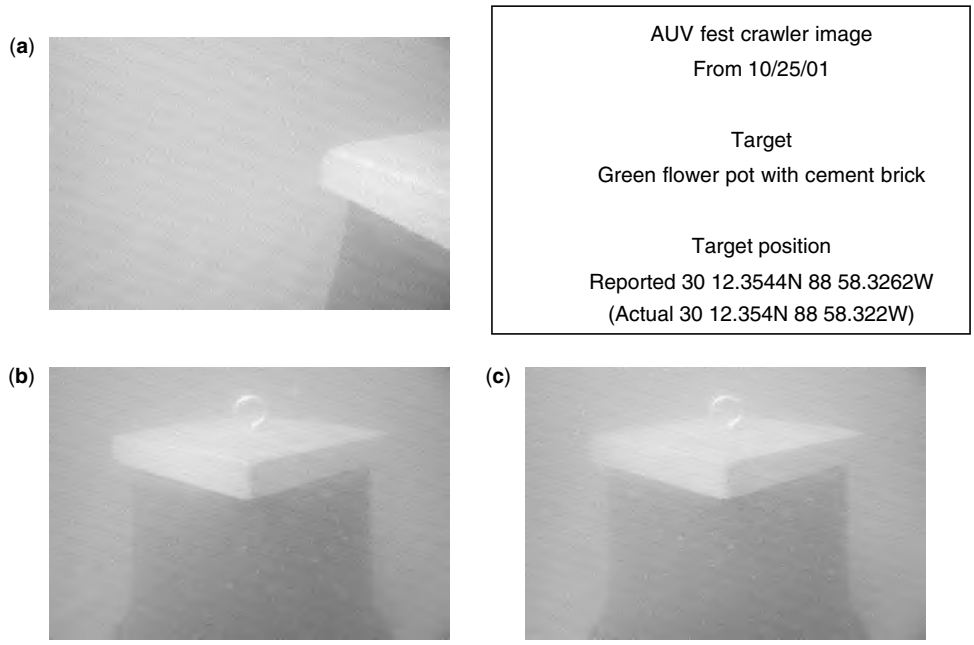


Figure 7. Automated detection, commanded acomms telemetry of compressed images: (**a**) acquisition image from raster pattern, acomm 600 bps, 50/1 compression; (**b**,**c**) ID image after vehicle reposition, acomm 600 bps, 50/1 compression (b) and acomm 1200 bps, 50/1 compression (c).

of 100 kB. With the 50 : 1 compression ratio used for this experiment, the transmitted image was only 2 kB, which, at 1200 bps transmission rate, required only 13 s to transmit. We note that the revolutionary wavelet-based compression technology used here[3] easily supports 200 : 1 and higher compression for such images.

[3] Developed by Professor Truong Nguyen of the University of California at San Diego.

We observe that, using coherent communications at a nominal bit rate of 4000 bps, and an image compression ratio of 200 : 1, we can maintain a real-time rate of one frame every 1.5 s. Thus, "slow scan" video is distinctly possible in those channels that support high-rate acomms.

BIOGRAPHY

Kenneth Scussel received a B.S. degree in Electrical Engineering in 1988 from the University of Connecticut,

and an M.S. degree in Electrical Engineering from Rensselaer Polytechnic Institute in 1992. In 1988 he joined General Dynamics, Electric Boat Division, where he developed embedded software for submarine launched cruise missile and torpedo tube control console simulators. Following General Dynamics, Mr. Scussel joined Datamarine International in 1993. At Datamarine he was a Project Engineer in charge of the development of consumer marine electronics. His accomplishments include putting a system that integrated depth, wind, boatspeed, and navigation sensors into production. This work included the development of depth sounder algorithms, which dramatically improved the performance of the depth sensor. Since 1995, he has been with Benthos (formerly Datasonics) as a Staff Software Engineer responsible for all aspects of software development in the acoustic modem product line. Mr. Scussel was part of the team that developed the original Datasonics ATM-87X series of acoustic modems. He has made significant enhancements to the signal processing, and enhanced the networking protocol algorithms that lead to the release of the ATM-88X series, a new generation of acoustic modems. His areas of interest are digital signal processing and developing software for real-time embedded processors.

FURTHER READING

R. Edwards, SMS deploys first successful acoustically coupled VIV measurement system in Campos Basin, published in the *Marine Analyst*, OTC-2000. SMS newsletter article available on their website www.SCIMAR.com.

M. Green and J. Rice, Channel-tolerant FH-MFSK acoustic signaling for undersea communications and networks, *IEEE J. Ocean. Eng.* **25**(1): 28–39 (Jan. 2000).

J. Preisig, Underwater acoustic communications, *IEEE Signal Process. Mag.* (July 1998).

J. Proakis, *Digital Communications*, McGraw-Hill, 1989.

K. Scussel, J. Rice, and S. Merriam, (1997). A new MFSK acoustic modem for operation in adverse underwater channels, *Oceans'97 MTS/IEEE Conf. Proc.*, 1997, Vol. 1, pp. 247–254.

R. Urick, *Principles of Underwater Sound*, McGraw-Hill, 1983.

ACOUSTIC TELEMETRY

FLETCHER A. BLACKMON
Naval Undersea Warfare Center
Division Newport
Newport, Rhode Island

1. INTRODUCTION

Humankind has always felt the need to communicate. It is a basic human desire that began most primitively as oral tradition and cave writings. The forms of communication were made more elaborate in the writings and hieroglyphics of ancient cultures such as the Egyptians and the Greeks. More modern cultures have refined the art of communication through language and pictures. The focus then became one of how to communicate over exceedingly larger distances. This human desire for the transfer of the printed word and audio and visual effects led to the creation of the telegraph, the U.S. Postal Service, and the telephone, radio, and television. More recently, the desire for information — the reason one communicates — has extended to wireless forms of communication using mobile and cellular phone technology, satellite communications, and communications from deep-space probes. It is no wonder that electronic mail (email) and the Internet, which currently provide a worldwide communications/information network, has literally taken over the world by involving everyone in the dialog of humankind. It is therefore natural and a vital step in this dialog to communicate and transfer information into, within, and out of the underwater environment, which covers more than 70% of the earth's surface.

2. BACKGROUND

The underwater environment has provided and is still providing one of the most interesting and challenging mediums for communication. These challenges include limited bandwidth, multipath induced time and spectral dispersion, and channel time variability. The available bandwidth is limited because of the frequency-dependent absorption characteristics of the underwater environment since higher frequencies are attenuated more strongly than lower frequencies and also as a function of range from the transmitter to the receiver. Another consideration relating to bandwidth is the receive signal strength, which decreases as a function of range as well as the noise present at the receiver due to ambient and human-made (synthetic) noise components. Bandwidth, signal transmission loss due to spreading and absorption, and noise are parameters that are used to determine the signal-to-noise ratio (SNR) at the receiver. Complicating this underwater picture is the presence of unwanted multipath in addition to the desired direct path signal. *Multipath* can be defined as one or more delayed signal replicas arriving at the receiver with time-varying amplitude and phase characteristics. These multipaths are responsible for temporal spreading and spectral spreading of the transmitted signal. These distorted signal replicas are produced by boundary reflection from the surface, bottom, and other objects as well as from acoustic ray bending in response to sound speed variation in the underwater environment. Add to these complications the time variability of these phenomena and a very interesting medium for acoustic telemetry results.

One of the first acoustic communication systems that was employed was the underwater telephone or UQC. This is a low-bandwidth (8–11 kHz) voice link that was developed by the United States government in 1945. It was used to communicate to and from submarines at speed and depth over a range of several kilometers. This method of acoustic communication is still today the standard for submarines and surface vessels. However,

modern technology with the advent of miniaturized, low power, digital signal processor (DSP) electronics and portable personal computers that can implement and support complex signal processing/communications algorithms has provided the capability to improve the quality, increase the data throughput, and increase the number of military and commercial telemetry applications that are possible.

Next, a brief telemetry system overview will be presented as a framework for the discussion to follow, which will focus on commercially available acoustic telemetry modems, past and present acoustic telemetry applications, the navy's range-based telemetry modems, several specific range-based telemetry applications, and finally current and future research in underwater acoustic telemetry.

3. TELEMETRY SYSTEM OVERVIEW

Telemetry data can take many different forms. Current speed data, sound velocity speed data, salinity data, pressure data, accelerometer data, temperature data, system health data, instrument command data, videos and images, data bearing files, digital voice, and interactive text are all clear examples of the various types of data that a telemetry system can be used to convey from one point to another point. In the underwater acoustic telemetry case, the data are telemetered from one point underwater to another point underwater that may be more accessible to those interested in the data. A telemetry system is comprised of a transmitter and a receiver. A typical transmission format is shown in Fig. 1, and a typical transmitter block diagram is shown in Fig. 2. The first portion of the transmission is a synchronization signal that is used to detect the presence of a telemetry signal and the location of data within the telemetry stream. The synchronization signal may be a large time bandwidth "chirp," or a differential or binary phase-shift keyed (DPSK or BPSK) signal with good auto- and cross-correlation properties. Guard time intervals are used after the synchronization signal and following the data prior to the next synchronization signal to mitigate the acoustic channel's multipath effects. A short training sequence follows the first guard time interval and is used to train the receiver's adaptive equalizer, specifically, the channel compensation capability for the modulated data that are to follow. Alternatively, the data may be transmitted by frequency shift-keyed (FSK) or multifrequency shift-keyed (MFSK) or spread spectrum modulation waveforms. A Doppler tracking tone may be superimposed over the telemetry packet for Doppler compensation at the receiver side of the link or may be derived from the synchronization signal itself. An optional channel probe in some telemetry systems can be used periodically prior to sending the telemetry packet or in an integral fashion to measure the channel characteristics so that the receiving system may take advantage of it. The transmitter block diagram in Fig. 2 shows the generation of the synchronization signal, coding and interleaving of the binary data, and the modulation of the binary data into transmission symbols as well as the power amplifier and electrical to acoustic transducer.

The acoustic telemetry receiver block diagram is shown in Fig. 3. The task of the receiver is to mitigate, compensate, and/or undo the effects of the underwater channel on the transmitted telemetry signal. The receiver incorporates acoustic to electric conversion via hydrophone, filtering, amplification, detection, synchronization, and Doppler compensation. The filtered, amplified, synchronized, and Doppler-corrected signal is then presented to a demodulator to provide baseband processing. The baseband signal is then passed to a symbol detector that in the case of coherent signals takes the form of an adaptive equalizer or a bank of filters for frequency-based incoherent systems or a set of correlators for direct-sequence spread-spectrum (DSSS)-based signals. Following the bit or symbol detection process, a decoding step can be performed if the data were coded at the transmitter. This decoding step seeks to correct errors that may still be present after the symbol/bit detection stage.

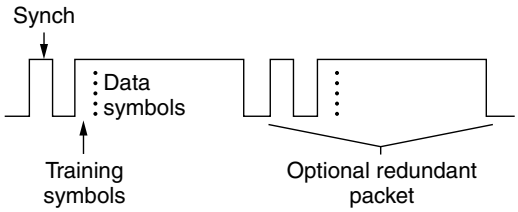


Figure 1. Typical telemetry transmit packet.

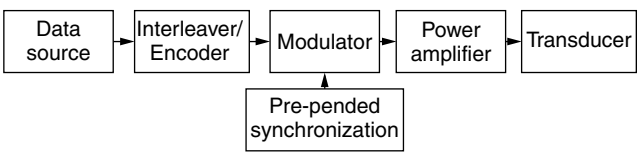


Figure 2. Acoustic telemetry transmitter block diagram.

4. ACOUSTIC TELEMETRY MODEMS

There are a number of commercially available acoustic telemetry modems at present. These modems span the gamut of communications techniques such as FSK, MFSK, spread-spectrum, and coherent signaling schemes such as BPSK and QPSK. In addition, these modems provide varying capabilities, chief among these are the bandwidth and data rate. Table 1 shows a brief comparison of these telemetry modems and their salient features.

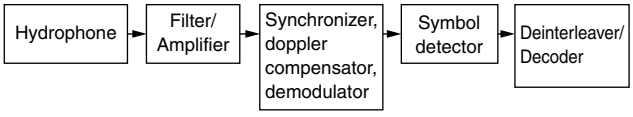


Figure 3. Acoustic telemetry receiver block diagram.

Table 1. Commercially Available Telemetry Modem Systems

Modem Manufacturer	Application	Modulation Format	Data Rate (bps)	Link Type
Benthos	Oil, environmental, military	FSK, HFSK, MFSK, BPSK, QPSK, TCM8PSK	10–10,000	Half -duplex
Linkquest	Oil, military	Spread-spectrum	200–2000	Half-duplex

5. TELEMETRY APPLICATIONS

A number of telemetry applications can be found in the fairly recent history of acoustic telemetry that chiefly began late in the 1980s after feasibility research was conducted in the area of incoherent FSK systems in the early 1980s [1]. The DATS system [1] was capable of operating at data rates up to 1200 bits per second (bps). As early as 1984, the concept of using acoustic telemetry to monitor and control oil wellheads was being tested and used in 100-m water depths and at horizontal ranges of 2 nautical miles [2]. Oil monitoring and control since this time and at present has been employing acoustic telemetry to reduce maintenance and operations costs. In 1989, it was demonstrated that a vertical acoustic telemetry link could be established to an undersea robot that was used to replace divers in the maintenance of submerged platforms [3]. In 1991, the Woods Hole Oceanographic Institute (WHOI) conducted a long-term acoustic telemetry experiment using their utility acoustic modem (UAM) to collect sensor data during a 6-month moored deployment [4]. This telemetry system, still used today, is based on TMS30C44 DSP technology and represents a versatile, configurable device for autonomous and moored telemetry applications. Again in 1991, acoustic data links employing telemetry were being explored for unmanned undersea vehicle (UUV) applications [5] using a 1250-bps data link with a range of 2 nautical miles using a bandwidth of 10 kHz. In 1992, a submersible was used to telemeter high-quality video imagery from a deep ocean trench to a surface vessel [6]. Throughout the 1990s, acoustic telemetry technology employed noncoherent modems. One of the more advanced of these systems, called the Telesonar system developed by Datasonics, Inc., and the U.S. Naval Command, Control and Ocean Surveillance Center, employed MFSK techniques using Hadamard coding as well as convolutional coding and frequency-hopping patterns [7]. This system is currently being used for a variety of applications and is commercially available from Benthos Inc.

First attempts were made in 1989 using phase-coherent signaling and minimal equalization techniques to telemeter images and commands between a surface vessel and a subsea robot in a very benign channel. Quadrature amplitude modulation (QAM) signaling was used with a transmission rate of 500 kbps with a bandwidth of 125 kHz centered at 1 MHz [8]. New and innovative ground breaking research in coherent acoustic communication techniques that allowed for tracking channel time variability in the early 1990s [9,10] made it possible to robustly communicate successfully in horizontal as well as vertical channels with higher data rates than were previously possible through the use of bandwidth-efficient MPSK and MQAM modulation schemes. In 1993, long-range, low-error-rate ($<10^{-4}$) telemetry over horizontal distances in excess of 200 km was shown to be possible [9]. In 1994, a prototype digital, acoustic underwater phone was described and demonstrated that compressed the data prior to transmission [11]. This type of system has not received much attention but still represents an important and much needed upgrade to the much older UQC voice system used aboard surface and subsurface vessels. In 1998, a practical coherent telemetry system for use onboard an autonomous underwater vehicle (AUV) was demonstrated using a bandwidth of 3 kHz and a bandwidth of 25 kHz with data rates of 2500 and 10,000 bps, respectively, in shallow water depths of 10–30 m [12]. Also in 1998, the WHOI UAM telemetry system was deployed on the MIT Odyssey and the Florida Atlantic University (FAU) Ocean Explorer UUVs [13]. The Odyssey UUV was integrated as part of an autonomous ocean sampling network (AOSN). Sensor data were transmitted from the vehicle to another telemetry modem. The data were then sent up a mooring cable to a surface buoy and then relayed via RF link to a support vessel for real time monitoring. In 2000, a surf-zone acoustic telemetry experiment (SZATE) was conducted alongside the pier at the Scripps Institution of Oceanography to demonstrate the ability to coherently telemeter data in the challenging very-shallow-water surf-zone environment [14]. It is envisioned that commercial and military applications of small size and/or miniature lemmings or crawlers with video and other sensors will be used in the surf zone in the near future and will have a need to telemeter this data to one or more remote undersea sites.

One of the most advanced acoustic communication/telemetry systems has been developed as part of the Acoustic Communications Advanced Technology Demonstration (ACOMMS ATD) funded by the U.S. Navy Advanced System Technology Office (ASTO). This tactical system can employ noncoherent as well as coherent modulation/demodulation techniques with multiple array sensor inputs and has been used for a multitude of naval demonstrations and applications involving transmission of voice, text, and video between UUVs, UUV and surface vessels, UUV and submarine, surface vessel and submarine, two submarines, and surface buoys. These telemetry links have been established at various data rates of $\leq$20 kbps in some cases and has operated within a

number of low-, medium-, and high-frequency bands over ranges of 2 km in shallow water, 3.7–5.6 km at high frequency, and 37–124 km at medium frequency [15] in deep water.

More recently, there has been growing interest in the application of acoustic telemetry to remote undersea networks given the success of point-to-point telemetry links. One such example of an undersea network is the Autonomous Oceanographic Surveillance Network (AOSN), which has been developed through funds from the Office of Naval Research (ONR) to network surface buoys and autonomous AUVs in order to sample the underwater environment [16]. The U.S. Navy is also developing sensor networks that employ acoustic telemetry with power control [17], protocol layers [18], and the ability for the sensor nodes to adaptively learn the network parameters such as node numbers, link quality, and connectivity tables [19]. This particular network has been demonstrated showing its surveillance capability as well as RF and satellite gateway capability to offload collected sensor data. A pictoral representation of an undersea acoustic telemetry network is shown in Fig. 4. Typically, these systems employ one or more of the following schemes for multiuser access: time-division multiple access (TDMA), frequency-division multiple access (FDMA), and code-division multiple access (CDMA) with a higher-level protocol layer that frequently uses ARQ and handshaking methodologies.

An excellent tutorial on acoustic communications presenting the history as well as an eye to the future has been presented in a review article [20]. Another excellent and also very readable acoustic underwater communications tutorial has been presented by Milica Stojanovic in this very same edition of the Wiley Encyclopedia of Telecommunications for the interested reader.

Next, we will take a more detailed look into a specific NUWC range-based modem telemetry system. Following this telemetry modem discussion, a number of range-based telemetry applications will be presented in detail as specific examples.

6. NUWC RANGE-BASED MODEM

The NUWC range-based telemetry system is a set of underwater acoustic telemetry modems developed by the Engineering, Test and Evaluation Department at the Naval Undersea Warfare Center Division, Newport (NUWCDIVNPT). The modems were developed as part of the Submarine Underwater Telemetry project that was tasked to provide robust bidirectional, full-duplex underwater acoustic communication subsystems to undersea test and evaluation and training ranges now under development. The goal was to produce acoustic modems that can reliably communicate with subsurface range participants at a throughput data rate of approximately 1 kbps at ranges out to 2 nautical miles in shallow and deep water while maintaining vehicle track.

The fact that navy ranges are designed for tracking exercise participants and not specifically for telemetry placed numerous constraints on the modem's design. The system's bandwidth and center frequencies were constrained, as was the choice of transducers to be used. The receiving hydrophones available on a range are typically widely spaced and omnidirectional, and available transmitters are also often omnidirectional. The benefits of beamforming, spatial diversity combining, and other directive techniques developed by researchers [23–25] in both the United Kingdom and the United States are not practical options under these conditions. Instead, the range-based modems must rely on adaptive equalization, paired with time redundancy and error correction schemes to achieve robust underwater communication. The cost of these techniques is reduced data rate. Although the system transmits at 4000 bps, the maximum sustained throughput data rates are approximately 900 bps with half-rate coding and 1800 bps without coding. The performance of this telemetry system has been documented in the literature [21].

Physically, each modem is a VME chassis with a notebook computer which serves as a user interface. The VME chassis is controlled by a FORCE-5CE CPU running SUN OS 4.1.3-u1. The chassis also contains a hard-disk drive, a VME clock and a timing board (VCAT) that serves as a master external clock, an optional GPS timing board for synchronizing transmissions with a GPS time reference, and two octal TMS320C40 digital signal processor boards as shown in Fig. 5.

Each modem has a transmitter and a receiver. The transmitter resides on two TMS320C40 processors. It reads binary data from a buffer on the UNIX host, then

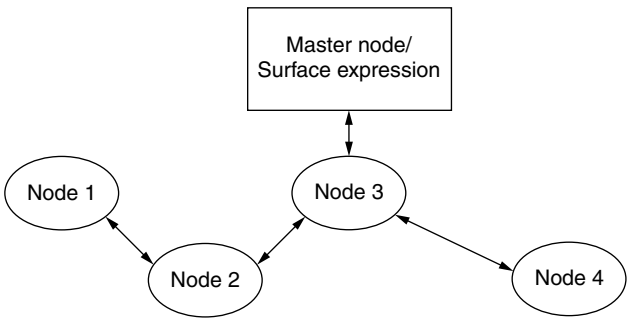


Figure 4. Underwater acoustic telemetry network diagram.

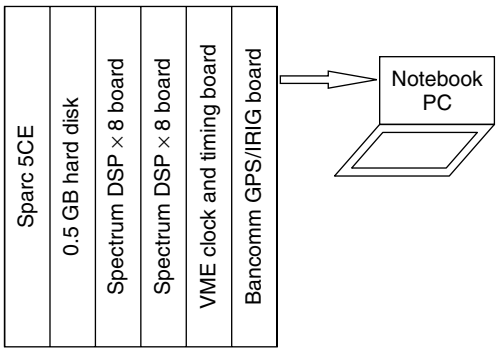


Figure 5. Block diagram of a NUWC range-based VME telemetry modem chassis.

packetizes and (optionally) convolutionally encodes the data. The transmission waveform is digitally synthesized, digital to analog converted and broadcast through the water. Figure 1 shows the format for the transmission. Each one second data packet is sent redundantly in a time diversity scheme to improve the robustness of the receiver.

The receiver for the system is a jointly adaptive decision feedback equalizer with a digital phase-locked loop (DFE-DPLL). The DFE-DPLL is patterned after a receiver first proposed by Stojanovic, Catipovic, and Proakis [24] with modifications to allow for efficient real-time implementation [25]. The receiver is implemented using eight TMS320C40 digital signal processors and consists of four functional blocks: packet detection and synchronization, Doppler estimation and compensation, complex demodulation, and equalization. Viterbi decoding (if required) is performed on the UNIX host.

The pairs of redundant data packets are jointly equalized in a manner similar to the spatial diversity combining technique presented in Ref. 24. Spatial diversity combining assumes that multiple spatially separated sensors are available to receive a given telemetry packet, and that the transmission travels through independent paths (or channels) to arrive at each sensor. The output of the sensors can then be jointly equalized to recover more of the signal than a single sensor. Since these telemetry modems cannot rely on the availability of multiple sensors, it employs multiple, time-separated transmissions of the same signal packets (i.e., time diversity). The ocean is a highly nonstationary, time-varying environment. Therefore, two transmissions of the same data packet, spaced sufficiently in time, travel through independent channels to arrive at the single sensor. Once the redundant data packets have arrived at the sensor, the net effect is essentially equivalent to spatial diversity. The adaptive equalizer algorithm has multiple inputs with each input containing the same signal information as received across an independent path. The cost of time diversity is a reduction in the throughput data rate, but diversity is often essential for low-error-rate telemetry.

7. RANGE-BASED TELEMETRY APPLICATIONS

A number of range-based acoustic telemetry applications will now be discussed. These include the mobile deep-range (MDR) application, the synthetic environment tactical integration virtual torpedo program (SETI VTP) application, and the underwater range data communication (URDC) application.

7.1. MDR

In November 1997, the NUWC range-based modems were used by personnel from the Naval Surface Warfare Center (NSWC), Detachment Annapolis to acoustically transfer data files collected during their mobile deep-range (MDR) exercises from a submerged submarine to a moored surface vessel as shown in Fig. 6. The MDR trial represents one of the initial uses of this modem system. One modem chassis with its notebook PC interface was set up in a data analysis station aboard the moored ship.

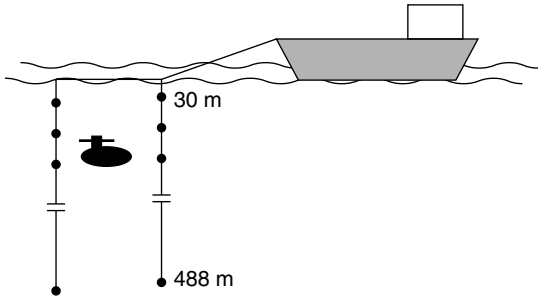


Figure 6. Configuration of the mobile deep-range telemetry.

At the analysis station, there was access to the outputs of 18 omnidirectional hydrophones contained in the two vertical legs of the MDR array shown in Fig. 6. A number of ASCII data files containing position and test configuration data were transmitted from the submarine under test and were received on the vertical hydrophone arrays. The acoustic telemetry modem receiver was successful in decoding the data files with few or no errors. These test configuration data files were then used by analysts aboard the measurement vessel to reconstruct and better analyze the data recorded with the arrays' acoustic and electromagnetic sensors in near real time.

7.2. SETI VTP

The submarine virtual torpedo program is the initial implementation of the synthetic environment tactical integration (SETI) project. The SETI project promotes the use of advanced distributed simulation (ADS) capabilities by creating high-fidelity antisubmarine warfare (ASW) training opportunities using live targets, synthetic torpedoes, and onboard submarine tactical and training systems in realistic external environments. The goal of the VTP project is to enable the real-time interaction of live submarines with high-fidelity simulated torpedoes. The SETI VTP conceptual drawings are shown in Figs. 7 and 8. In June 1998, the SETI VTP test conducted at the Atlantic Undersea Test and Evaluation Center (AUTEC) demonstrated the full-duplex capabilities of the NUWC range-based modems. The demonstrated capabilities include (1) encrypted, bidirectional, full-duplex data exchange between a submerged submarine operating on an instrumented navy range and remote modeling and simulation facilities via a wide-area network (WAN), which includes the range-based modems and a satellite link; (2) submarine launch control of simulated torpedoes from the remote modeling and simulation facilities and reception of tactical weapon data from the launched weapon; and (3) use of two alternate communication methods, distributed interactive simulation (DIS) protocols and high level architecture (HLA) runtime infrastructure (RTI) to transfer both weapon and positional data between the submarine and the remote modeling and simulation facilities.

7.3. URDC

The underwater range data communications (URDC) project is a currently ongoing U.S. Navy project that

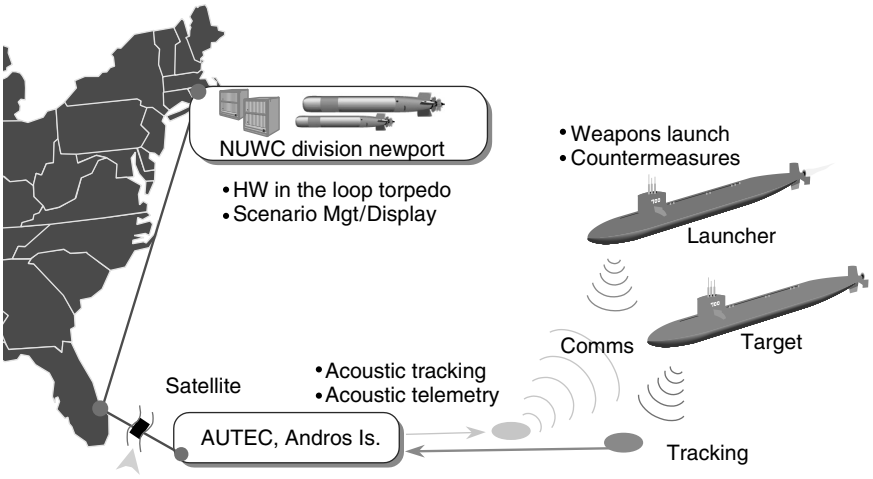


Figure 7. Illustration of the SETI VTP telemetry link concept.

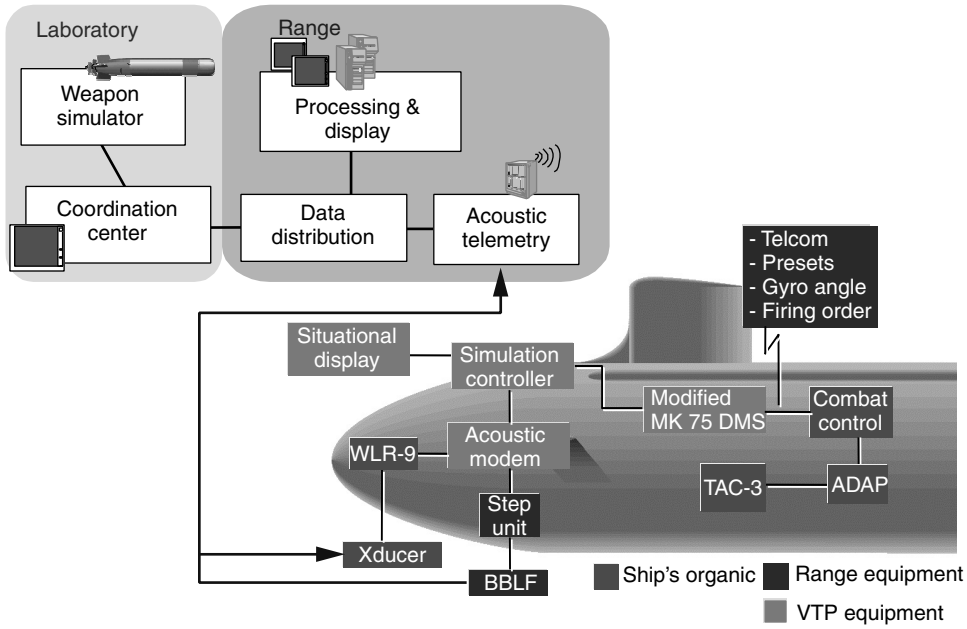


Figure 8. Ship/shore SETI VTP telemetry configuration.

is funded to provide acoustic telemetry capability on shore at the Command and Control (CC) building at AUTEC as well as a roll on/roll off utility on submarines engaged in Test and Evaluation exercises on range. This system will expand and modernize the older NUWC range-based telemetry modems discussed earlier. The URDC system will interface to range receive and transmit node infrastructure on shore as well as organic submarine sensors. In addition, this system will integrate submarine tracking capability with telemetry capability. This new telemetry system will integrate state-of-the-art SHARC-based DSP technology with other telemetry components in a small user-friendly package for navy sailor and range use. The system will incorporate multiple telemetry modes of operation including high-, medium-, and low-data-rate transfer in half-duplex and full-duplex FDMA modes. These modes will employ coherent signaling techniques coupled with strong error correction techniques and selective time diversity automatic repeat request (ARQ) capability. Initial prototype performance results have been reported previously in the literature [22]. In addition, for low probability of intercept (LPI) and potentially multiuser applications, variants of direct-sequence spread-spectrum (DSSS) techniques will be used. The URDC applications that are envisioned include rapid mission/exercise debrief, interactive text (chat) capability, digital voice, and file transfer as well as test and evaluation specific applications such as transferring ground truth track of ship's position from shore and exercise coordination information.

8. CURRENT AND FUTURE RESEARCH

Current and future acoustic telemetry research is focused on a number of aspects of the telemetry problem. One area of active research is the development of iterative and

integral equalization and decoding methodologies to obtain performance at or better than that possible by conventional equalization techniques [26,27]. These procedures require more complex processing by virtue of the feedback nature and number of iterations inherent in these schemes. Future work in this area will involve novel methods of Turbo equalization employing Turbo codes and decoders while reducing overall complexity while maintaining improved performance. Another active area of current and future research is that of multiuser and networked node telemetry, associated protocols, routing, and multiuser techniques. The goal of this daunting task is to include as many underwater nodes, such as environmental devices, UUVs, remotely operated vehicles (ROVs), and submarines in increasingly larger communication nets while demanding performance approaching that of single-point links as well as conserving precious bandwidth. Finally, another active area of telemetry research that this author with other researchers is currently involved in is the remote optoacoustic and acoustooptic telemetry technology that connects submerged platforms at speed and depth with in-air platforms.

9. FINAL REMARKS

Once again, the dialog of humankind demands the search, collection, and transfer of knowledge and information to and from the underwater world. Our need to communicate will ultimately drive the technology and applications that solve the problems of the future that in the past were thought impossible to solve.

BIOGRAPHY

Fletcher A. Blackmon received his B.S. degree in electrical engineering in 1988 from Southeastern Massachusetts University (now known as University of Massachusetts at Dartmouth), his M.S. degree in electrical engineering in 1991 from the University of Massachusetts at Dartmouth, and is currently enrolled as a Ph.D. student in electrical engineering at the University of Massachusetts at Dartmouth. He joined the Naval Undersea Warfare Center in 1989 as an electronics engineer. At NUWC he worked on the research, design, and development of underwater acoustic communications and telemetry modem systems for Navy ranges. Since 1995, he has been involved in research at NUWC, where he has been working on opto-acoustic systems and applications. Mr. Blackmon holds a number of patents in the area of signal generation as well as patents pending in the areas of iterative and integral equalization and coding/decoding for underwater acoustic communication systems as well as the areas of opto-acoustic methods for communication and sonar applications. His areas of interest are the design and performance of equalization and decoding algorithms for underwater acoustic communications and opto-acoustic/acousto-optic methods for sonar and communications applications.

BIBLIOGRAPHY

1. A. Baggeroer, D. E. Koelsch, K. V. Der Heydt, and J. Catipovic, DATS — a digital acoustic telemetry system for underwater communications, *Proc. Oceans '81*, Boston, MA, 1981.
2. F. C. Jarvis, Description of a secure reliable acoustic system for use in offshore oil blowout (BOP) or wellhead control, *IEEE. J. Ocean. Eng.* **OE-9**: 253–258 (1984).
3. A. Kaya and S. Yauci, An acoustic communication system for subsea robot, *Proc. Oceans '89*, 1989, pp. 765–770.
4. L. Freitag, S. Meriam, D. Frye, and J. Catipovic, A long term deep water acoustic experiment, *Proc. Oceans '91*, Honolulu, Hawaii, 1991.
5. G. Mackelburg, Acoustic data links for UUVs, *Proc. Oceans '91*, Honolulu, Hawaii, 1991.
6. M. Suzuli and T. Sasaki, Digital acoustic image transmission system for deep sea research submersible, *Proc. Oceans '92*, 1992, pp. 567–570.
7. K. Scussel, J. Rice, and S. Meriam, A new MFSK acoustic modem for operation in adverse underwater channels, *Proc. Oceans '97*, Halifax, Nova Scotia, Canada, 1997.
8. A. Kaya and S. Yauchi, An acoustic communication system for subsea robot, *Proc. Oceans '89*, Seattle, WA, 1989.
9. M. Stojanovic, J. Catipovic, and J. G. Proakis, Adaptive multichannel combining and equalization for underwater acoustic communications, Part 1, *J. Acoust. Soc. Am.* **94**(3): 1621–1631 (Sept. 1993).
10. M. Stojanovic, J. Catipovic, and J. G. Proakis, Phase-coherent digital communications for underwater acoustic channels, *IEEE. J. Ocean. Eng.* **19**: 100–111 (1994).
11. A. Goalic et al., Toward a digital acoustic underwater phone, *Proc. Oceans '94*, Brest, France, 1994.
12. L. Freitag et al., A bidirectional coherent acoustic communication system for underwater vehicles, *Proc. Oceans '98*, Nice, France, 1998.
13. L. Freitag, M. Johnson, and J. Preisig, Acoustic communications for UUVs, *Sea Technol.* **40**(5): (May 1999).
14. D. Green and F. Blackmon, Performance of channel-equalized acoustic communications in the surf zone, *Proc. Oceans '01*, Honolulu, Hawaii, 2001.
15. T. Curtin and R. Benson, ONR program in underwater acoustic communications, *Sea Technol.* **4**(5): (May 1999).
16. T. Curtin, J. Bellingham, J. Catipovic, and D. Webb, Autonomous oceanographic sampling networks, *Oceanography* **6**: 86–94 (1993).
17. J. Proakis, M. Stojanovic, and J. Rice, Design of a communication network for shallow water acoustic modems, *Proc. Ocean Community Conf. '98*, Baltimore, MD, 1998.
18. M. Green and J. Rice, Handshake protocols and adaptive modulation for underwater communication networks, *Proc. Oceans '98*, Nice, France, 1998.
19. E. Soizer, M. Stojanovic, and J. Proakis, Underwater acoustic networks, *IEEE. J. Ocean. Eng.* **25**: 72–83 (2000).
20. D. Kilfoyle and A. Baggeroer, The state of the art in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **25**: 4–27 (2000).
21. S. M. Jarvis, F. A. Blackmon, K. Fitzpatrick, and R. Morrissey, Results from recent sea trials of the underwater digital acoustic telemetry system, *Proc. Oceans '97*, Oct. 1997.
22. F. Blackmon and W. Canto, Performance comparison of several contemporary equalizer structures applied to selected field test data, *Proc. Oceans '00*, Sept. 2000.

23. D. Thompson et al., Performance of coherent PSK receivers using adaptive combining, beamforming, and equalisation in 50 km underwater acoustic channels, *Proc. Oceans '96*, Sept. 1996.

24. J. A. Catipovic and L. E. Freitag, Spatial diversity processing for underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **16**(1): 86–97 (Jan. 1991).

25. S. M. Jarvis and N. A. Pendergrass, Implementation of a multichannel decision feedback equalizer for shallow water acoustic telemetry using a stabilized fast transversal filters algorithm, *Proc. Oceans '95*, Oct. 1995.

26. F. Blackmon et al., Performance comparison of iterative/integral equalizer/decoder structures for underwater acoustic channels, *Proc. Oceans '01*, Honolulu, Hawaii, Nov. 2001.

27. E. Sozer, J. Proakis, and F. Blackmon, Iterative equalization and decoding techniques for shallow water acoustic channels, *Proc. Oceans '01*, Honolulu, Hawaii, Nov. 2001.

ACOUSTIC TRANSDUCERS

DONALD P. MASSA
Massa Products Corporation
Hingham, Massachusetts

1. INTRODUCTION AND HISTORICAL OVERVIEW

The purpose of this article is to provide a brief overview of the very extensive topic of acoustic transducers. Transducers are devices that transform one form of energy into another. A few acoustic transducers, such as whistles or musical instruments, transform mechanical energy into sound, but the following discussion is concerned primarily with electroacoustic transducers. They are classified as either transmitters that convert electricity to sound, or receivers that change acoustic energy into electrical signals.

The invention of the telephone in the late 1800s resulted in the first widespread use of electroacoustic transducers. The microphone in the telephone converted the acoustical energy of the human voice into electrical signals. The earpiece in the telephone converted the electrical signals back into acoustic energy so the voice of the person at the other end of the line can be heard.

New requirements for different types of electroacoustic transducers were created by the development of the phonograph at the turn of the last century, followed by increased consumer use of radio in the 1920s and the advent of sound motion pictures in the 1930s. Improved loudspeakers and microphones were required to meet the demands of these new industries, and the science of sound was transformed into the applied science of electroacoustics.

During the 1920s, electrical engineers began applying the concepts of "equivalent circuits" to characterize acoustic transducers. The mechanical and acoustical portions of the transducer were modeled by converting them to equivalent electric circuit components of inductors, capacitors, and resistors. These equivalent-circuit elements of the acoustic portions were coupled to the pure electrical portions of the transducer by means of an electromechanical transformer. This modeling allowed the pioneering generation of electroacoustic engineers to not only better understand how transducers operated but also to optimize transducer designs by using the well-known methods of electric circuit analysis. In 1929 the Acoustical Society of America was formed, and in 1934 the first engineering-based textbook on transducers, entitled *Applied Acoustics*, was published by Olson and Massa.

While significant improvement in the design of electroacoustic transducers for use in the audible frequency band in air were achieved during 1900–1940, a new requirement for electroacoustic transducers to operate underwater for sonar applications was only in its infancy. However, the military threat of submarines during World War II caused sonar transducer development to rapidly advance during the 1940s.

Following World War II, new types of electroacoustic transducers designed to operate in the ultrasonic frequency range were developed for a wide variety of new industrial applications, such as noncontact distance or level measurement, collision avoidance, communication, remote control, intrusion alarms, ultrasonic cleaning, ultrasonic welding, ultrasonic flow detection, and ultrasonic imaging. Different transducers were designed to operate at frequencies as low as 20 kHz, the upper frequency limit of human hearing, to 10 MHz and higher [1–3].

2. FUNDAMENTALS OF ELECTROACOUSTIC TRANSDUCERS

Many factors affect the design of an electroacoustic transducer. For example, a transducer designed to operate in a gaseous medium, such as air, is very different from one designed to operate in a liquid medium, such as water. Likewise, differences in acoustical requirements, such as frequency of operation or radiation pattern, will influence the design. It is, therefore, necessary to first understand some basic acoustical principles in order to properly understand how electroacoustic transducers operate.

2.1. Generation of Sound

Sound is a transfer of energy caused by the vibration of particles in the transmission medium. The particles vibrate back and forth a small distance, which causes a longitudinal wave to travel in the same direction as the vibrating particles.

An electroacoustic transmitting transducer produces sound by vibrating a portion of its surface, which therefore affects the molecules in the transmission medium. When the radiating surface moves forward the molecules are pushed closer together, thus increasing the instantaneous pressure (condensation). When the radiating surface moves back, the molecules expand, thus decreasing the instantaneous pressure (rarefaction). The vibrating molecules near the transducer push against their neighbors, causing them to also vibrate. This process continues, creating a propagating wave in which the instantaneous

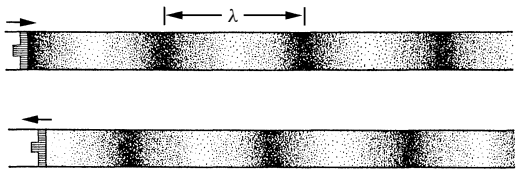


Figure 1. Illustration showing how a piston moving forward (top) compresses the molecules in the transmission medium in a pipe creating condensation, and moving backward (bottom) expands the molecules creating rarefaction.

pressures oscillate between condensation and rarefaction as it progresses outward from the transducer. This is illustrated in Fig. 1, which shows two moments of time of a piston moving back and forth, pushing against the molecules in a transmission medium contained in a pipe. The top picture shows the molecules when the piston is moving forward creating condensation, while the bottom diagram shows the piston moving backward, causing rarefaction.

2.2. Differences in the Characteristics of Sound Propagation in Gaseous and Liquid Media

There are fundamental differences in many of the properties of sound radiating in a liquid as compared to sound propagating in a gas, but there are typically only minor variations in the acoustic properties of sound radiating among various gases or among various liquids. Since the applications of most sound transmissions in gases occur in air, and in liquids occur in water, the characteristics of these two media will be used to illustrate the difference between acoustic radiation in liquids and in gases.

2.2.1. Speed of Sound. The velocity with which the sound waves travel through a transmission medium is called the *speed of sound, c*. The nominal value of c is primarily a function of the composition of the particular medium, but slight changes occur for each medium because of variations in parameters such as temperature or pressure. However, the velocity of sound is much greater in liquids than in gasses. As an example, in air at 20 °C the speed of sound is 343 m/s, and in freshwater at 20 °C it is 1483 m/s [4].

2.2.2. Wavelengths of Sound. The wavelength of sound traveling in a medium is the distance between condensation peaks, as shown in Fig. 1, and is a function of both the frequency and the speed of the sound wave. The wavelength is

$$\lambda = \frac{c}{f} \tag{1}$$

where λ is the wavelength, c is the speed of sound, and f is the frequency.

Figure 2 shows a plot of the wavelength of sound in air and water at room temperature as a function of frequency. As can be seen from the curve, since the speed of sound in water is approximately 4.3 times greater than in air, the wavelength for a given frequency in water is approximately 4.3 times longer than in air.

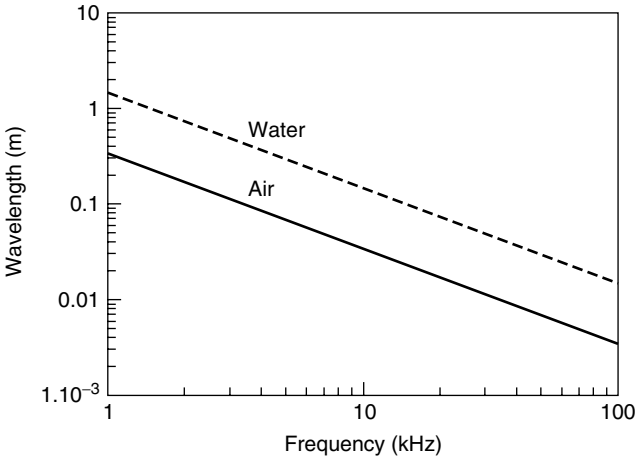


Figure 2. Plot of the wavelength, λ, as a function of frequency for sound in air and water.

2.2.3. Attenuation of Sound. The attenuation of sound traveling through a medium increases as the frequency increases, and the attenuation in a gas at a given frequency is much greater than in a liquid. Figure 3 shows plots of typical attenuations for sound in both air and water as a function of frequency [5,6]. Because the attenuation is much less in water than in air, objects can be detected at much greater ranges using echo location in water than in air. Table 1 compares propagation distances and wavelengths for sound at different frequencies in air and water.

2.2.4. Density. Gasses are much less dense than liquids. The density, ρ_0, of air is only 1.2 kg/m^3. The

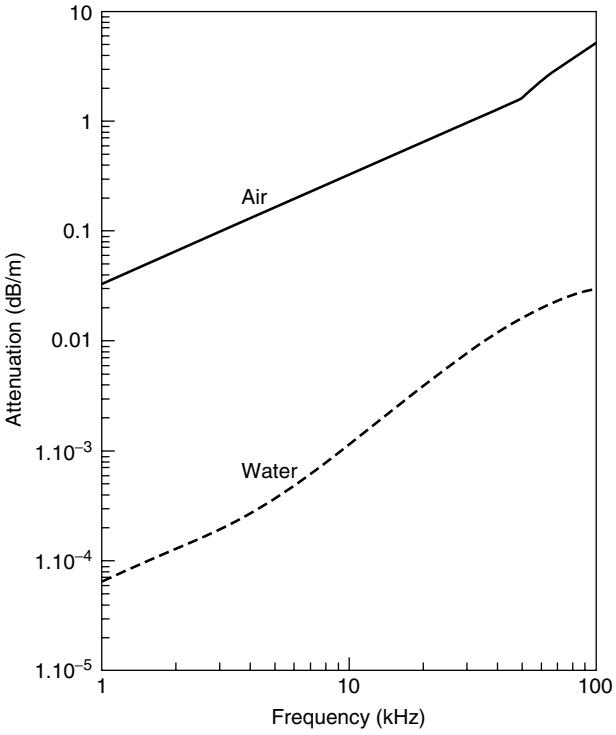


Figure 3. Plot of the attenuation, λ, as a function of frequency for sound in air and water.

Table 1. Sound Propagation Distance and Wavelength Comprison for Air and Water

	Propagation Distance for Planar Sound Wave to Attenuate to Half Its Initial Pressure (No Spreading Loss) (m)		Wavelength (m)	
Frequency (kHz)	Air	Water	Air	Water
0.5	375	330,000	0.68	3
1.0	180	130,000	0.34	1.5
5.0	40	20,000	0.068	0.3
10	18	7,000	0.034	0.15
50	3.8	330	0.0068	0.03
100	1.2	160	0.0034	0.015

density of water is 1000 kg/m^3, which is over 800 times greater [4].

2.2.5. Analogies Between Acoustical Properties of a Transmission Medium and the Electrical Properties of a Circuit. Most people are familiar with Ohm's law in electricity. This law states that for a given voltage E applied across an electrical component that has an impedance Z, the electric current I flowing through the component will be directly proportioned to the value of the electrical impedance Z. Ohm's law can be written as

$$E = IZ \tag{2}$$

where E is the voltage in volts, I is the current in amperes, and Z is the electrical impedance in ohms. Acoustic transmission media have properties analogous to electrical properties in circuits. In acoustics, the sound pressure p of the acoustic wave is equivalent to voltage in an electric circuit, and the particle velocity u is analogous to current. The acoustical impedance of a transmission medium is the product of the density times the speed of sound. It is written as $\rho_0 c$, and the units are acoustic ohms or rayls (after Lord Rayleigh). Ohm's law in acoustics is

$$p = u\rho_0 c \tag{3}$$

where p is the pressure of the sound wave in pascals, u is the particle velocity in meters per second, and $\rho_0 c$ is the acoustic impedance in rayls.

Because gases have much slower sound velocities and lower densities than do liquids, their acoustic impedances are much less. For example, the acoustic impedance of air is nominally 415 rayls, while $\rho_0 c$ for water is approximately 1.48×10^6 rayls. This large disparity causes the fundamental design concepts to be much different from those for use in water. Table 2 summarizes the analogies between acoustical and electrical properties.

2.2.6. Relationship Between Sound Pressure, Particle Displacement, and Partial Velocity. As was discussed previously, sound is transmitted in a wave caused by the particles in the medium vibrating back and forth. In a plane acoustic wave traveling in the x direction, the particle velocity is the real part of

$$\xi = |\xi| e^{j(\omega t - (\omega/c)x)} \tag{4a}$$

which is

$$\xi = |\xi| \cos\left(\omega t - \frac{\omega}{c}x\right) \tag{4b}$$

where ξ is the particle displacement in meters, ω is the frequency of the sound wave in radians per second, and c is the speed of sound in the medium in meters per second.

The particle velocity is the derivative of the displacement, therefore

$$u = -\omega|\xi| \sin\left(\omega t - \frac{\omega}{c}x\right) \tag{5}$$

where u is the particle velocity in meters per second.

Substituting Eq. (5) into Eq. (3), the sound pressure becomes

$$p = -\rho_0 c\omega|\xi| \sin\left(\omega t - \frac{\omega}{c}x\right) \tag{6}$$

where p is the sound pressure of the acoustic wave in pascals and $\rho_0 c$ is the acoustic impedance in rayls. From Equation 6, it can be seen that the magnitude of the sound pressure of an acoustic wave is directly proportional to the acoustic impedance, the frequency, and the magnitude of the particle displacement. To transmit sound at a specific pressure and frequency, an acoustical transducer must vibrate with an amplitude equal to the particle displacement required for the particular medium. Because the acoustic impedance of water is 3600 times greater than that of air, the particle displacement in air must be 3600 times greater than the particle displacement in water to produce the same sound pressure at the same frequency. For either medium, the particle displacement required to produce a constant sound pressure will decrease as the frequency increases.

Because of these relationships, transducer designs are different for operation in fluids than in gases, or for

Table 2. Ohm's Law Analogies Between Electrical and Acoustical Properties

Electrical			Acoustical		
Quantity	Symbol	Units	Quantity	Symbol	Units
Voltage	E	Volts	Pressure	p	μPa
Charge	q	Coulomb	Particle displacement	ξ	m
Current	I	Amperes	Particle velocity	u	m/s
Impedance	Z	Ohms	Acoustic impedance	$\rho \cdot c$	rayls

operation at high frequencies than low frequencies. The radiating surfaces in underwater sonar transducers have to move only small displacements to generate large sound pressures. However, they must generate a relatively large amount of force to compress the dense water. Transducers that operate in air have to vibrate over much larger displacements to generate high sound pressures, but very little force is necessary to compress the gas.

Transducers that radiate at low frequencies in either medium must vibrate for greater distances than those that operate at high frequencies, which can be verified by observing loudspeakers in a typical stereo system. The low-frequency "woofer" can be seen moving large amplitudes when base notes are played, while the motion of the high-frequency "tweeters" appears to be negligible.

2.3. Sound Pressure Levels

Because sound pressures vary more than 10 orders of magnitude, they are expressed by acoustical engineers as logarithmic ratios, which are called sound pressure levels (SPLs). The SPL in decibels for a sound pressure p is calculated as $20\log(p/p_{\text{ref}})$, where p_{ref} is a standard reference sound pressure. Some confusion can occur because several different reference pressures are in use, which results in a given sound pressure being expressed with several different possible sound pressure levels.

Most of the early work in acoustical engineering was associated with the development of audio equipment, so it was natural to use the threshold of human hearing for the reference pressure. In the cgs system, that sound pressure is 0.0002 dyn/cm^2 (0.0002 μbar), so sound pressure levels were expressed in terms of dB/0.0002 μbar [SPL $= 20\log(p/0.0002)$ dB/0.0002 μbar, where p is the pressure in microbars].

During World War II, there were major advances in the development of sonar for detecting submarines. Since the sounds produced by sonar systems are not heard directly by people, sonar engineers began using 1 μbar as a more logical standard reference pressure. Sound pressure levels therefore began being expressed in terms of dB/1 μbar [SPL $= 20\log(p/1)$ dB/1 μbar, where p is pressure in microbars].

In the early 1970s, the SI system of units was adopted in acoustical engineering, so the micropascal (μPa), which is equal to 10^{-6}N/m^2, became the reference pressure. Sound pressure levels therefore began to be expressed in terms of dB/1 μPa [SPL $= 20\log(p/1)$ dB/1 μPa, where p is pressure in μPa]. This is now the most often used reference pressure for acoustic measurements, but it is not unusual to encounter data using any of these three standard reference pressures. To add to the confusion, sometimes the SPL will be improperly stated in terms of decibels only, without indicating the reference pressure used to compute the ratio. It is obviously important to know which reference pressure was used whenever an SPL is expressed, and when comparing transducer responses, all sound pressure levels should be converted to the same reference pressure. It is quite simple to convert sound pressure levels among the three reference pressures by using Table 3.

Table 3. Sound-Pressure-Level Conversion Table

To Convert SPL in	To SPL in	
dB/0.0002 μbar	dB/1 μbar	Subtract 74 dB
dB/0.0002 μbar	dB/1 μPa	Add 26 dB
dB/1 μPa	dB/0.0002 μbar	Subtract 26 dB
dB/1 μPa	dB/1 μbar	Subtract 100 dB
dB/1 μbar	dB/0.0002 μbar	Add 74 dB
dB/1 μbar	dB/1 μPa	Add 100 dB

2.4. Radiation Patterns of Transducers

The acoustic radiation pattern, or beam pattern, is the relative sensitivity of a transducer as a function of spatial angle. This pattern is determined by factors such as the frequency of operation and the size, shape, and acoustic phase characteristics of the vibrating surface. The beam patterns of transducers are reciprocal, which means that the beam will be the same whether the transducer is used as a transmitter or as a receiver.

Transducers can be designed to radiate sound in many different types of patterns, from omnidirectional to very narrow beams. The beam pattern of a transducer is usually calculated and graphed showing the relative reduction in sensitivity as a function of angle, with the maximum sensitivity of the transducer along the main acoustic axis set to equal 0 dB. The beam angle of the transducer is equal to the total arch encompassed by the beam between the angles when the pressure has fallen to a level of −3 dB on either side of the main acoustic axis.

For a transducer with a circular radiating surface vibrating in phase, the narrowness of the beam pattern is a function of the ratio D/λ, the diameter of the radiating surface over the wavelength of sound at the operating frequency. The larger the diameter of the transducer as compared to a wavelength of sound, the narrower the sound beam. For example, if the diameter is twice the dimension of the wavelength, the total beam angle will be approximately 30°, but if either the diameter or frequency is changed so that the ratio becomes 10, the total beam angle will be reduced to approximately 6°. Since the wavelength of sound at a given frequency in water is approximately 4.3 times larger than in air, the diameter of an underwater transducer must be approximately 4.3 times larger than an air transducer to produce the same beam angle at the same frequency.

A transducer large in size compared to a wavelength produces not only a narrow main beam, but also secondary lobes separated by nulls. Figure 4 is a three-dimensional (3D) representation of the beam pattern produced by a transducer with a radiating diameter that is large compared to a wavelength. As can be seen, each secondary lobe is sequentially lower in amplitude than the previous one. The equation for the radiation pattern of a circular rigid piston in an infinite baffle as a function of spatial angle is [7].

$$P(\theta) = \left[\frac{2J_1\left(\pi\dfrac{D}{\lambda}\sin\theta\right)}{\pi\dfrac{D}{\lambda}\sin\theta}\right]^2 \tag{7a}$$

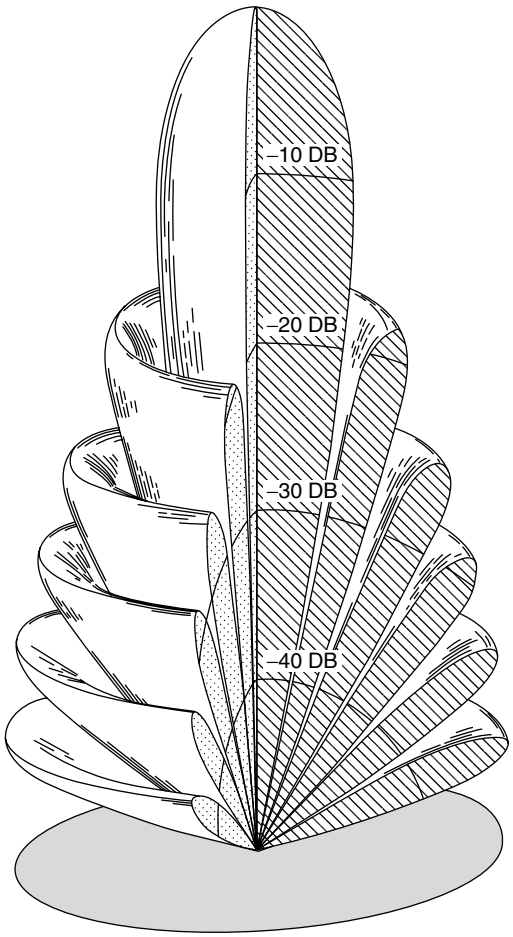


Figure 4. 3D beam pattern produced by a transducer with a circular radiating surface where the diameter is large compared to a wavelength.

where $P(\theta)$ is the relative sound pressure as a function of the angle, θ is the angle of the sound pressure from an axis perpendicular to the center of the piston, D is the diameter of the piston, λ is the wavelength of the sound, and J_1 is the first-order Bessel function.

Beam patterns are usually plotted on a decibel scale where the sound pressure as a function of spatial angle is

$$P_{\mathrm{dB}}(\theta) = 20 \log \left[\frac{2J_1\left(\pi \dfrac{D}{\lambda}\right) \sin\theta}{\pi \dfrac{D}{\lambda} \sin\theta} \right] \tag{7b}$$

where $P_{\mathrm{dB}}(\theta)$ is the relative sound pressure as a function of spatial angle in decibels. The beam angle is usually defined as the measurement of the total angle where the sound pressure level of the main beam has been reduced by 3 dB on both sides from the peak that occurs along the axis perpendicular to the piston. When describing transducer beam patterns, two-dimensional (2D) plots are most commonly used. These show the relative sensitivity of the transducer versus angle in a single plane cut through the axis perpendicular to the center of the piston in the 3D beam pattern. Figure 5 shows 2D plots on rectilinear coordinates of the beam patterns of circular piston radiators for several different values of D/λ.

2.5. Resonance

Many electroacoustic transducers are designed to operate at resonance, which is the natural frequency of vibration of the transducer structure. Transducers will produce a much greater displacement for a given drive voltage when operated at frequencies in the vicinity of resonance. Likewise, when used as receivers they will produce a larger electrical signal for a given sound pressure.

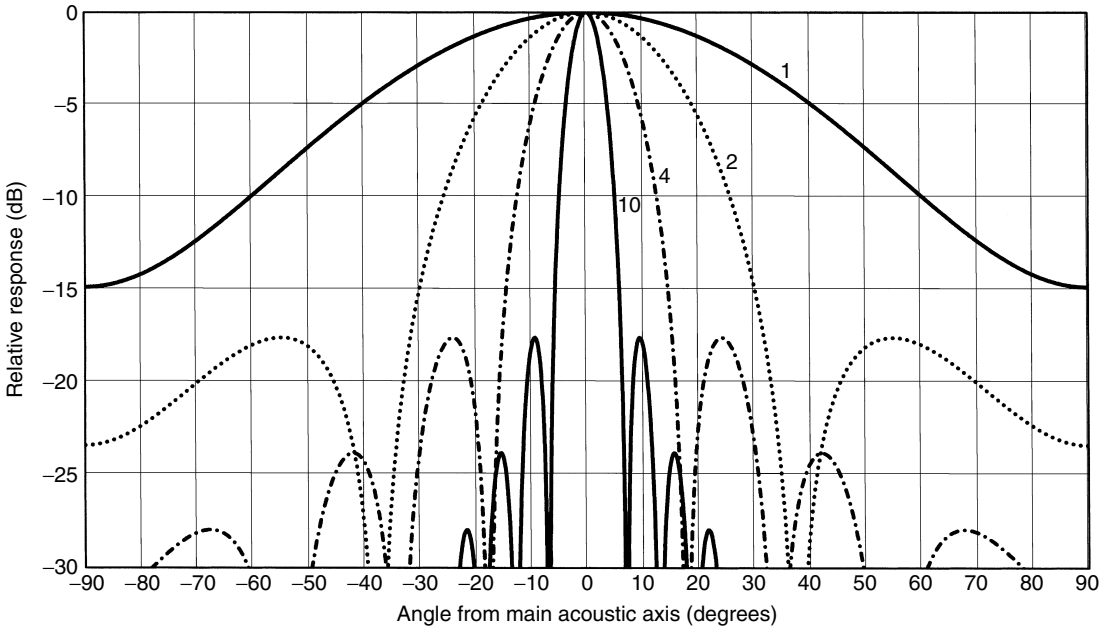


Figure 5. 2D graph showing the beam patterns of four different transducers radiating with circular pistons in an infinite baffle having diameter to wavelength ratios (D/λ) of 1, 2, 4, and 10 [From Eq. (7b)].

The quality factor, Q, of a transducer is a value that indicates the width of the frequency band in the vicinity of resonance over which it can operate with high output. The Q is calculated by dividing the resonant frequency by the bandwidth, which is defined as the frequency band over which the response of the transducer is within 3 dB of the peak response.

The receiving response of transducers is usually constant at frequencies well below resonance, so the output voltage will be constant for a given sound pressure, and proportional to changing sound pressures at all frequencies in this region. Transducers that are used only for receiving (microphones in air, and hydrophones in water) are often operated well below resonance to take advantage of this broadband flat response.

3. TRANSDUCTION

Electroacoustic transducers operate by using a variety of different transduction materials or mechanisms to transform electrical energy into sound and vice versa. For example, in transducers that employ magnetics, an alternating electric current flowing through a coil of wire produces a varying magnetic force that causes the transducer structure to vibrate. In like manner, a sound wave will vibrate the transducer, which moves the coil in a magnetic field, thus generating an electrical signal.

Transducers can also be designed using magnetostrictive materials for transduction. When these materials are placed in a magnetic field, their mechanical dimensions change as a function of the strength of the magnetic field, which in turn can be used to generate sound. Other transducers employ piezoelectric crystals, such as quartz, Rochelle salt, or ammonium dihydrogen phosphate (ADP) for transduction. They develop an electric charge between two surfaces when the crystal is mechanically compressed, and they expand and contract in size in the presence of an applied electric field.

The most commonly used transduction materials for transducers are electrostrictive ceramics. These ceramic materials, such as barium titinate and lead zirconate titanate, are often referred to as *piezoelectric ceramics* and also produce an electric charge when a mechanical stress is applied, and vice versa. However, they must have an internal polarizing electric field established in order for transduction to occur. Their popularity is due to relatively low cost, coupled with the ability to be fabricated into a wide variety of shapes and sizes.

4. A FEW EXAMPLES OF SOME ELECTROACOUSTIC TRANSDUCERS

The following sections contain short descriptions of the construction of a few electroacoustic transducers. Since there are such a wide variety of different types of electroacoustic transducers, it is not possible to provide a description of most of them in this brief overview. Some of the publications in the reading list at the end of this article contain detailed information on many specific types of transducers.

4.1. Moving-Coil Electrodynamic Loud Speaker

The most common loudspeakers used in stereo or public address systems are electrodynamic transducers, which contain a coil of wire suspended in a magnetic field. When an alternating electrical current is passed through the coil, mechanical forces are developed between the coil's electromagnetic field and the field in which it is mounted.

Figure 6 is a cross-sectional sketch illustrating the schematic construction of an electrodynamic speaker [8]. As can be seen, the voice coil (4) is a coil of wire fashioned into a cylindrical tube. It is rigidly connected to a radiating diaphragm (1), which is resiliently mounted to an enclosure (3). This holds the coil within the magnetic field produced by the permanent magnet (4), but allows it to freely vibrate within this field. The magnet is shaped like a disk with a circular groove cut into the surface facing the diaphragm. The tubular voice coil is mounted so that it is held within this groove. A varying electrical current in the coil produces proportional changes in its electromagnetic field, which in turn modulates the magnetic forces between the coil and the permanent magnet. This causes the coil to move back and forth, thus vibrating the diaphragm and generating sound.

4.2. Condenser Microphone

The condenser microphone produces a variation in its electrical capacitance in the presence of an acoustic wave. Figure 7 illustrates the construction of such a transducer. The stretched thin metallic membrane is separated from the rigid backplate by a small airgap. When a sound wave vibrates the membrane, it causes the airgap to change in thickness, producing a variation in the electrical capacitance between it and the backplate. This varying capacitance is converted into an electrical signal that is proportional to the sound pressure wave.

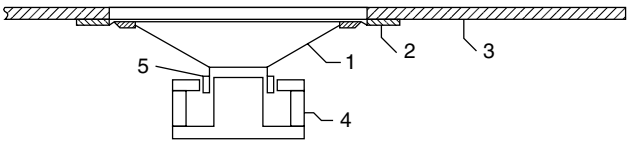


Figure 6. Illustration showing the construction of an electrodynamic speaker (Fig. 1 of U.S. Patent 2,445,276 [8]).

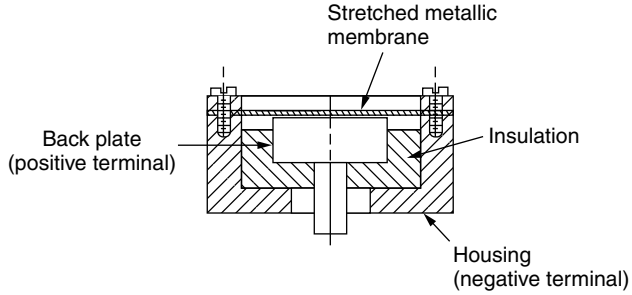


Figure 7. Illustration showing the construction of a condenser microphone.

4.3. Flexural Air Ultrasonic Transducer Using Electrostrictive Ceramics

Flexural ultrasonic transducers use the resonance of a mechanical diaphragm to produce the motion required to generate the required sound pressure. Figure 8 shows a cross-sectional illustration of a typical flexural ultrasonic transducer [9]. The housing is an aluminum cup consisting of an outer cylindrical shell (1) with relatively thick sides and a thin circular diaphragm. This produces a rigid clamped circular disk in which the resonant frequency is controlled primarily by its diameter and stiffness. A thin ceramic disk (5) is cemented to the radiating diaphragm.

As a receiver, the diaphragm is mechanically vibrated by sound pressure, causing it to buckle up and down. Since the ceramic is rigidly attached to the diaphragm, it stretches as the diaphragm buckles, which produces an electrical voltage across it. In like manner, when the ceramic is excited by an electrical voltage it will stretch, causing the diaphragm to vibrate and transmit sound. Because the diaphragm can move large displacements while creating only minor strains in the ceramic, this design allows for generation of large sound pressures without over stressing and cracking the ceramic.

This particular transducer design operates at an overtone of the fundamental resonant frequency of the clamped diaphragm. The frequency of operation can be adjusted by varying the diaphragm thickness.

4.4. Tonpilz Sonar Transducer

A mass-loaded vibratile transducer (Tonpilz transducer) is a common design used in sonar. A typical example is shown in Figure 9 [10]. A ceramic cylinder (12) is cemented between a light aluminum head mass (11), and a heavy steel tail mass (15). The ceramic has electrodes on its two ends. This transducer resonates in much the same way as a large mass attached to a spring. If the mass is reduced, the resonant frequency will lower. If the stiffness of the spring increases, the resonant frequency will be higher.

In the transducer of Fig. 9, the ceramic cylinder is the spring connected to the head mass and the tail mass. If it is made more compliant, for example, by reducing the wall thickness or increasing the length, the resonant frequency will lower. If the head and tail masses are made smaller, the resonant frequency will increase. At resonance, the length of the ceramic will increase and decrease relatively large amounts, causing the head and tail masses to vibrate. Because the head mass is much lighter than the tail mass, it vibrates at much larger amplitude.

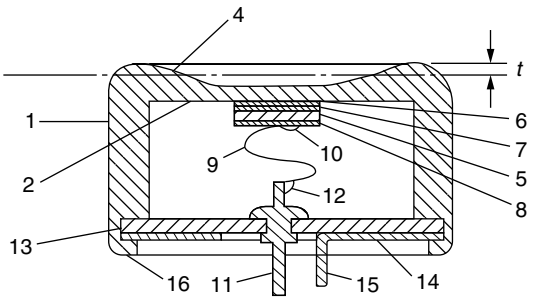


Figure 8. Illustration showing the construction of a flexural ultrasonic transducer designed for operation in air using an electrostrictive ceramic for transduction (Fig. 1 of U.S. Patent 3,943,388 [9]).

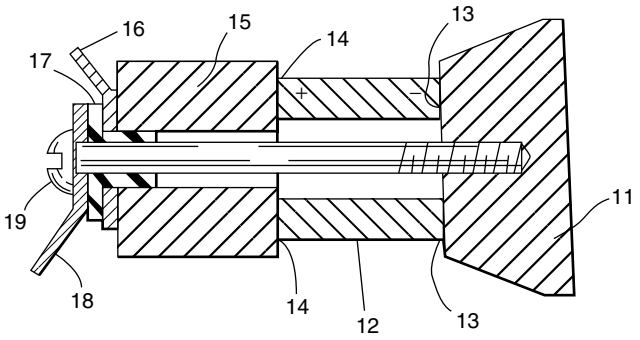


Figure 9. Illustration showing the construction of a tonpilz sonar transducer (Fig. 1 of U.S. Patent 3,739,327 [10]).

In operation the structure is encapsulated in waterproof material, such as rubber, and the radiating head is acoustically coupled to the water. When used as a transmitter, an oscillating electrical voltage is connected across the electrodes of the ceramic, causing it to alternately lengthen and contract. This in turn causes the head mass, which is coupled to the water, to vibrate large amplitudes and produce a sound pressure wave. As a receiver, a sound pressure wave pushes the head mass, causing the transducer structure to vibrate. This in turn causes the length of the ceramic tube to alternately contract and expand, which generates a voltage across the ceramic stack.

BIOGRAPHY

Donald P. Massa received a B.S. and an M.S. degree in electrical engineering from Northeastern University in 1969 and 1972, respectively. His professional career started at Woods Hole Oceanographic Institution as a co-op student, and in the late 1960s he assumed full-time responsibilities as a development engineer for advanced electroacoustic products at Massa Products Corporation. He has been the president and chief technical officer of the company since 1976. He has been responsibilities for the development and production design of more than 100 electroacoustic transducers and systems for both air ultrasonic and underwater sonar applications. Mr. Massa holds 14 patents on electroacoustic devices and systems, and he has published numerous technical articles and presented many invited and contributing papers on electroacoustics to professional societies, the U.S. Navy, and Allied Navies. He was awarded the Outstanding Alumni Award in the Field of Science and Technology by Northeastern University in 1997, and he also serves as a member of Northeastern's Board of Trustees.

BIBLIOGRAPHY

1. F. Massa, Some personal recollections of early experiences on the new frontier of electroacoustics during the late 1920's

and early 1930's, *J. Acoust. Soc. Am.* **77**(4): 1296–1302 (April 1985).

2. F. Massa, Sonar transducers: A history, *Sea Technol.* (Nov. 1989).
3. H. Olson and F. Massa, *Applied Acoustics*, Blakston, Philadelphia, 1934.
4. L. Kinsler and A. Frey, *Fundamentals of Acoustics*, Wiley, New York, 1962.
5. D. Massa, Choosing an ultrasonic sensor for proximity or distance measurement, Parts 1, 2, *Sensors* (Feb., March 1999).
6. R. Urick, *Principles of Underwater Sound for Engineers*, McGraw-Hill, New York, 1967.
7. F. Massa, Radiation of sound, in *American Institute of Physics Handbook*, McGraw-Hill, New York, 1963, Sect. 3, pp. 118–122.
8. U.S. Patent 2,445,276 (July 13, 1948), F. Massa, Electrodynamic loudspeakers.
9. U.S. Patent 3,943,388 (to D. Massa, Trustee Stoneleigh Trust) (March 9, 1976), F. Massa, Electroacoustic transducer of the flexural vibrating diaphragm type.
10. U.S. Patent 3,739,327 (to Massa Div. DCA) (June 12, 1973), D. Massa, Electroacoustic transducers of the mass loaded vibratile piston type.

ACOUSTIC (UNDERWATER) COMMUNICATIONS

MILICA STOJANOVIC
Massachusetts Institute of Technology
Cambridge, Massachusetts

The need for underwater wireless communications exists in applications such as remote control in offshore oil industry, pollution monitoring in environmental systems, collection of scientific data recorded at ocean-bottom stations and unmanned underwater vehicles, speech transmission between divers, and mapping of the ocean floor for detection of objects and discovery of new resources. Wireless underwater communications can be established by transmission of acoustic waves. The underwater acoustic communication channels, however, have limited bandwidth, and often cause signal dispersion in time and frequency [2–7]. Despite these limitations, underwater acoustic communications are a rapidly growing field of research and engineering.

Acoustic waves are not the only means for wireless communication underwater, but they are the best known so far. Radiowaves that will propagate any distance through conductive seawater are the extra-low-frequency ones (30–300 Hz), which require large antennaes and high transmitter powers [1]. Optical waves do not suffer as much from attenuation, but they are affected by scattering. Transmission of optical signals requires high precision in pointing the narrow laser beams, which are still being perfected for practical use. Thus, acoustic waves remain the single best solution for communicating underwater, in applications where tethering is not acceptable.

The idea of sending and receiving information under water is traced back all the way to the time of Leonardo Da Vinci, who is quoted for discovering the possibility of detecting a distant ship by listening on a long tube submerged under the sea. In the modern sense of the word, underwater communications began to develop during World War II, for military purposes. One of the first underwater communication systems was an underwater telephone, developed in 1945 in the United States for communicating with submarines [4]. This device used a single-sideband (SSB) suppressed carrier amplitude modulation in the frequency range of 8–11 kHz, and it was capable of sending acoustic signals over distances of several kilometers. However, it was not until the development of VLSI (very large-scale integration) technology that a new generation of underwater acoustic communication systems began to emerge. With the availability of compact digital signal processors (DSPs) with their moderate power requirements, it became possible for the first time to implement complex signal processing and data compression algorithms at the submerged ends of an underwater communication link.

Since the late 1990s, significant advancements have been made in the development of underwater acoustic communication systems [7], in terms of their operational range and data throughput. Acoustically controlled robots have been used to replace divers in performing maintenance of submerged platforms [16], high-quality video transmission from the bottom of deepest ocean trenches (6500 km) to a surface ship was established [17], and data telemetry over horizontal distances in excess of 200 km was demonstrated [25].

As efficient communication systems are developing, the scope of their applications continues to grow, and so do the requirements on the system performance. Many of the developing applications, both commercial and military, are calling for real-time communication with submarines and autonomous, or unmanned underwater vehicles (AUVs, UUVs). Setting the underwater vehicles free from cables will enable them to move freely and refine their range of operation. The emerging communication scenario in which the modern underwater acoustic systems will operate is that of an underwater data network consisting of both stationary and mobile nodes. This network is envisaged to provide exchange of data, such as control, telemetry, and eventually video signals, between many network nodes. The network nodes, located on underwater moorings, robots, and vehicles, will be equipped with various sensors, sonars, and videocameras. A remote user will be able to access the network via a radio link to a central node based on a surface station.

In attempts to achieve these goals, current research is focusing on the development of efficient communications and signal processing algorithms, design of efficient modulation and coding schemes, and techniques for mobile underwater communications. In addition, multiple-access communication methods are being considered for underwater acoustic networks, as well as the design of network protocols, suited for long propagation delays and strict power requirements encountered in the underwater environment. Finally, data compression algorithms suitable

for low-contrast underwater images, and related image processing methods [18], are expected to enable image transmission through band-limited underwater acoustic channels.

1. SYSTEM REQUIREMENTS

The achievable data throughput and the reliability of an underwater acoustic communication system, as measured by the bit error rate, vary from system to system, but are always subject to bandwidth limitations of the ocean channel. Unlike the situation in the majority of other communication media, the use of underwater acoustic resources has not been regulated yet by standards.

In the existing systems, four kinds of signals usually are transmitted: control, telemetry, speech, and video signals.

1. Control signals include navigation, status information, and various on/off commands for underwater robots, vehicles, and submerged instrumentation such as pipeline valves or deep-ocean moorings. The data rates up to about 1 kilobit per second (kbps) are sufficient for these operations, but very low bit error rates (BERs) may be required.
2. Telemetry data are collected by submerged acoustic instruments such as hydrophones, seismometers, sonars, current meters, and chemical sensors, and it also may include low-rate image data. Data rates on the order of one to several tens of kbps are required for these applications. The reliability requirements are not so stringent as for the command signals, and a probability of bit error of 10^{-3}–10^{-4} is acceptable for many applications.
3. Speech signals are transmitted between divers and a surface station or among divers. While the existing, commercially available diver communication systems use mostly analog communications, based on single-sideband modulation of the 3-kHz audio signal, research is advancing in the area of synthetic speech transmission for divers, as digital transmission is expected to provide better reliability. Transmission of digitized speech by linear predictive coding (LPC) methods requires rates on the order of several kbps to achieve close-to-toll quality. The BER tolerance of $\sim 10^{-2}$ makes it a viable technology for poor-quality band-limited underwater channels [19,20].
4. Video transmission over underwater acoustic channels requires extremely high compression ratios if an acceptable frame transmission rate is to be achieved. Fortunately, underwater images exhibit low contrast and detail, and preserve satisfactory quality if compressed even to 2 bits per pixel. Compression methods, such as the JPEG (Joint Photographic Experts Group) standard discrete cosine transform, have been used to transmit 256×256-pixel still images with 2 bits per pixel, at transmission rates of about one frame per 10 seconds (s^{-1}) [17]. Further reduction of the required transmission rate seems to be possible by using dedicated compression algorithms, such as the discrete wavelet transform [18]. Current achievements report on the development of algorithms capable of attaining compression ratios in excess of 100 : 1. On the other hand, underwater acoustic transmission of television-quality monochrome video would require compression ratios in excess of 1000 : 1. Hence, the required bit rates for video transmission are greater than 10 kbps, and possibly up to several hundreds of kbps. Performance requirements are moderate, as images will have satisfactory quality at bit error rates on the order of 10^{-3}–10^{-4}.

2. CHANNEL CHARACTERISTICS

Sound propagation under water is determined primarily by transmission loss, noise, reverberation, and temporal and spatial variability of the channel. Transmission loss and noise are the principal factors determining the available bandwidth, range, and signal-to-noise ratio. Time-varying multipath influences signal design and processing, which determine the information throughput and communication system performance.

2.1. Range and Bandwidth

Transmission loss is caused by energy spreading and sound absorption. While the energy spreading loss depends only on the propagation distance, the absorption loss increases not only with range but also with frequency, thus limiting the available bandwidth.

In addition to the nominal transmission loss, link condition is influenced largely by the spatial variability of the underwater acoustic channel. Spatial variability is a consequence of the waveguide nature of the channel, which results in such phenomena as formation of shadow zones. Transmission loss at a particular location can be predicted by many of the propagation modeling techniques [2] with various degrees of accuracy. Spatial dependence of transmission loss imposes particularly severe problems for communication with moving sources or receivers.

Noise observed in the ocean consists of human-made noise and ambient noise. In deep ocean, ambient noise dominates, while near shores, and in the presence of shipping activity, human-made noise significantly increases the noise level. Unlike the human-made noise, most of the ambient noise sources can be described as having a continuous spectrum and Gaussian statistics [2]. As a first approximation, the ambient noise power spectral density is commonly assumed to decay at 20 dB/decade in both shallow and deep water, over frequencies that are of interest to communication systems design. The exception are biological sources of noise, such as snapping shrimp, which lives only in certain geographic areas and produces impulsive noise within the range of frequencies used by a typical communication system.

Frequency-dependent transmission loss and noise determine the relationship between the available range, bandwidth, and SNR (signal-to-noise ratio) at the receiver input. This dependence is illustrated in Fig. 1, which shows the frequency dependent portion of SNR for several

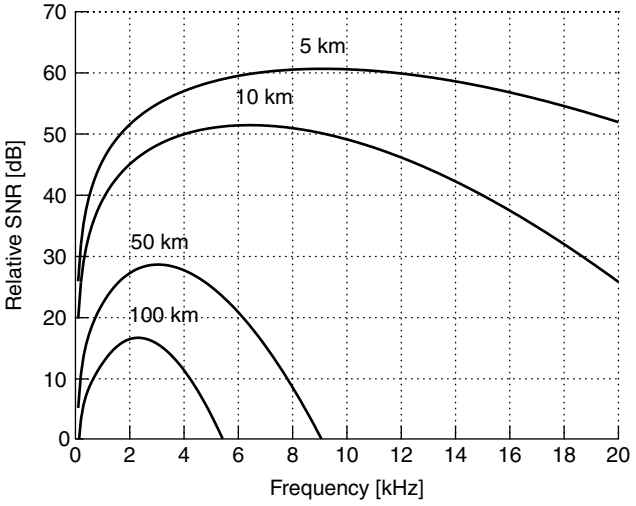


Figure 1. Frequency-dependent portion of SNR.

transmission ranges. (The SNR is evaluated assuming spherical spreading, absorption according to Thorp [2], and a 20-dB/decade decay of the noise power spectral density.) Evidently, this dependence influences the choice of a carrier frequency for the desired transmission range. In addition, it determines the relationship between the available range and frequency band. Underwater acoustic communication links can be classified according to range as very long, long, medium, short, and very short links. For a long-range system, operating over 10–100 km, the bandwidth is limited to few kilohertz (for a very long distance on the order of 1000 km, the available bandwidth falls below 1 kHz). A medium-range system operating over 1–10 km has a bandwidth on the order of 10 kHz, while only at very short ranges below about 100 m, more than 100 kHz of bandwidth may be available.

Within this limited bandwidth, the signal is subject to multipath propagation through a channel whose characteristics vary with time and are highly dependent on transmitter and receiver location. The multipath structure depends on the link configuration, which is designated primarily as vertical or horizontal. While vertical channels exhibit little time dispersion, horizontal channels may have extremely long multipath spreads. Most notable in the long- and medium-range channels, multipath propagation causes severe degradation of the acoustic communication signals. Combating the underwater multipath to achieve a high data throughput is without exception considered to be the most challenging task of an underwater acoustic communication system.

2.2. Multipath

In a digital communication system that uses a single carrier, multipath propagation causes intersymbol interference (ISI), and an important figure of merit is multipath spread in terms of symbol intervals. While typical multipath spreads in the commonly used radio channels are on the order of several symbol intervals, in the horizontal underwater acoustic channels they increase to several tens, or a hundred of symbol intervals for moderate to high data rates. For example, a commonly encountered multipath spread of 10 ms in a medium-range shallow-water channel, causes the ISI to extend over 100 symbols if the system is operating at a rate of 10 kilosymbols per second (ksps).

The mechanisms of multipath formation in the ocean are different in deep and shallow water, and also depend on the frequency and range of transmission. Understanding of these mechanisms is based on the theory and models of sound propagation. Depending on the system location, there are several typical ways of multipath propagation. It is mostly the water depth that determines the type of propagation. The delineation between shallow and deep water is not a strict one, but usually implies the region of continental shelves, with depth less than about 100 m, and the region past the continental shelves, where the water gets deeper. Two fundamental mechanisms of multipath formation are reflection at boundaries (bottom, surface, and any objects in the water), and ray bending (rays of sound always bend towards regions of lower propagation speed). If the water is shallow, propagation will occur in surface–bottom bounces in addition to a possible direct path. If the water is deep, as in the regions past the continental shelves, the sound channel may form by bending of the rays toward the location where the sound speed reaches its minimum, called the *axis of the deep sound channel*. Because there is no loss due to reflections, sound can travel in this way over several thousands of kilometers. Alternatively, the rays bending upward may reach the surface focusing in one point where they are reflected, and the process is repeated periodically. The region between two focusing points on the surface is called a *convergence zone*, and its typical length is 60–100 km.

The geometry of multipath propagation and its spatial dependence are important for communication systems that use array processing to suppress multipath [e.g., 22,23]. The design of such systems is often accompanied by the use of a propagation model for predicting the multipath configuration. Ray theory and the theory of normal modes provide basis for such propagation modeling.

2.3. Time Variation

Associated with each of the deterministic propagation paths (macromultipaths), which can be modeled accurately, are random signal fluctuations (micromultipath), which account for the time variability of the channel response. Some of the random fluctuations can be modeled statistically [2,3]. These fluctuations include surface scattering due to waves, which is the most important contributor to the overall time variability of the shallow-water channel. In deep water, in addition to surface scattering, internal waves contribute to the time variation of the signal propagating along each deterministic path.

Surface scattering is caused by the roughness of the ocean surface. If the ocean were calm, a signal incident on the surface would be reflected almost perfectly, with the only distortion in the form of a phase shift of π. However, wind-driven waves act as the displacement of the reflection point, resulting in signal dispersion. Vertical displacement of the surface can be accurately modeled as a zero-mean Gaussian random variable, whose power spectrum

is completely characterized by the windspeed [2]. Motion of the reflection point results in frequency spreading of the surface-reflected signal, significantly larger than that caused by many other phenomena. Doppler spread of a signal component of frequency f caused by a single surface reflection occurring at an incidence angle θ is $0.0175(f/c)w^{3/2}\cos\theta$, where c is the speed of sound, nominally taken to be 1500 m/s, and w is the windspeed in meters per second [2]. A moderate windspeed is on the order of 10 m/s. Highest Doppler spreads are most likely to be found in short-range links, which use relatively high frequencies. For longer ranges, at which lower frequencies are used, the Doppler spread will be lower; however, multipath spread will increase as there will be more significant propagation paths. The exact values of multipath and Doppler spreads depend on the geometry of multipath on a particular link. Nevertheless, it can be said that the channel spread factor, that is, the product of the Doppler spread and the multipath spread, can in general be expected to decrease with range.

As an example, Figs. 2–4 each show an ensemble of channel impulse responses, observed as functions of delay over an interval of time. These figures describe channel responses obtained at three fundamentally different locations with different mechanisms of multipath formation. Figure 2 shows the impulse responses recorded in deep water of the Pacific Ocean, off the coast of California. In this channel, propagation occurs over three convergence zones, which span 110 nautical miles (nmi). At each fixed time instant, the figure shows a realization of the channel impulse response magnitude as a function of delay. Looking at one channel response reveals that two or more signals arrive at the receiver at any given time. The multipath delay spread in this channel is on the order of 20 ms. The multiple arrivals have comparable energy, thus causing strong ISI. The amplitudes and phases of distinct arrivals may vary independently in time. Along the time axis, variation of the channel response is observed

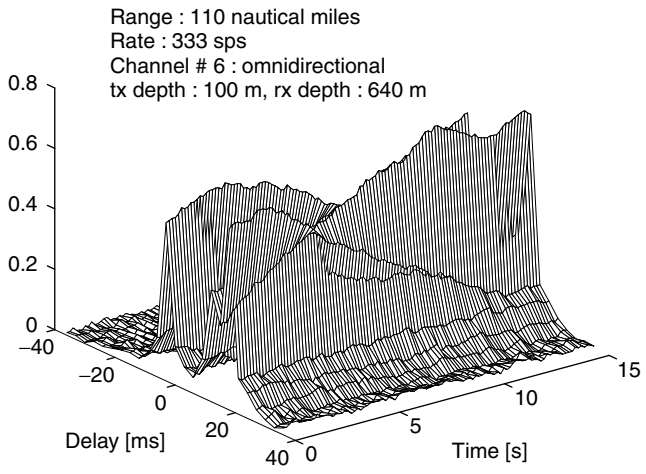


Figure 2. Ensemble of long range channel responses in deep water (~2000 m) off the coast of California, during the month of January. Carrier frequency is 1 kHz. Rate at which quaternary date symbols used for channel estimation were transmitted is given in symbols per second (sps).

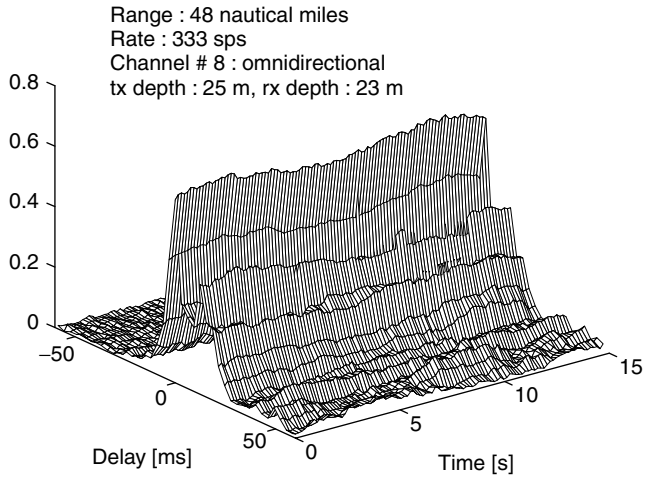


Figure 3. Ensemble of long-range channel responses in shallow water (~50 m) off the coast of New England, during the month of May. Carrier frequency is 1 kHz.

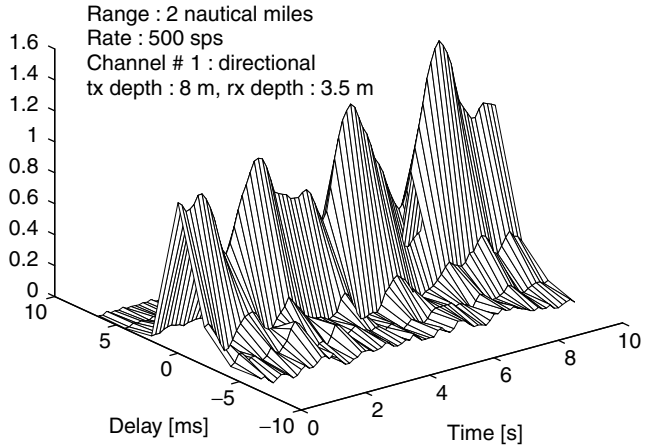


Figure 4. Ensemble of medium-range channel responses in shallow water (~20 m) near the coast of New England, during the month of February. Carrier frequency is 15 kHz.

for each given delay. In this example, significant variation occurs over the shown 15-s interval. This channel does not have a well-defined principal, or strongest, arrival, as evidenced by the fact that the maximum amplitude does not always occur at the same delay. The channel responses shown in Figs. 2–4 are obtained by adaptive channel estimation techniques. In particular, a recursive least-squares algorithm is applied to 4-PSK (phase shift keying) signals transmitted over the channels at rates indicated in the figures. Figure 3 shows the impulse responses obtained in shallow water of the Atlantic Ocean continental shelf, off the coast of New England, over a long distance (48 nmi). This example shows a channel with a well-defined principal arrival, followed by a multipath of lower energy. The extent of multipath is up to 50 ms. It is worth noting that even though the extended multipath may appear to have negligible energy, its contribution to the overall ISI cannot be neglected. This channel shows a slower time-variation than the one observed in Fig. 2. In contrast,

Fig. 4 provides an example of a rapidly time-varying channel. These response were recorded in the shallow water of Buzzards Bay near the coast of New England, over a distance of 2 nmi. Of the three examples shown, this channel demonstrates the fastest time variation, which is typical of a medium-range shallow water environment.

The factor that determines the performance of a digital communication system on a frequency-spread channel is the Doppler spread normalized by the symbol rate. In underwater acoustic channels, the normalized Doppler spread can approach values as high as 10^{-2}. The implications that the time-varying multipath bears on the communication system design are twofold. On one hand, signaling at a high rate causes many adjacent symbols to interfere at the receiver, and requires sophisticated processing to compensate for the ISI. On the other hand, as pulse duration becomes shorter, channel variation over a single symbol interval becomes slower. This allows an adaptive receiver to efficiently track the channel on a symbol-to-symbol basis, providing, of course, a method for dealing with the resulting time dispersion. Hence, time-varying multipath causes a tradeoff in the choice of signaling rate for a given channel. Experimental results obtained on a rapidly varying shallow-water channel [27] demonstrate these observations.

While there exists a vast knowledge of both deterministic and statistical modeling of sound propagation underwater, the use of this knowledge in modeling of communication channels has only recently received more attention [e.g., 8–12]. A time-varying multipath communication channel is commonly modeled as a tapped delay line, with tap spacing equal to the reciprocal of twice the channel bandwidth, and the tap gains modeled as stochastic processes with certain distributions and power spectral densities. While it is known that many radio channels fit well within the model of Rayleigh fading, where the tap gains are derived from complex Gaussian processes, there is no single model accepted to date for any of the underwater acoustic channels. Modeling of the shallow-water medium-range channel has received most attention, as this channel is known to be among the most rapidly varying ones. Most authors consider that this channel is fully saturated, meaning that it exhibits Rayleigh fading [3,5,9]. The deep-water channel has also been modeled as a Rayleigh fading channel; however, the available measurements are scarce, often making channel modeling a controversial issue [10].

The statistical channel measurements available today focus mostly on stationary communication scenarios. In a mobile underwater acoustic channel, vehicle speed will be the primary factor determining the time-coherence properties of the channel, and consequently the system design. Knowledge of a statistical channel model has proved useful in the design and analysis of land-mobile radio systems, and it remains for the future to develop such models for underwater mobile acoustic channels.

3. SYSTEM DESIGN

To overcome the difficulties of time-varying multipath dispersion, the design of commercially available underwater acoustic communication systems has so far relied mostly on the use of noncoherent modulation techniques and signaling methods that provide relatively low data throughput. More recently, phase-coherent modulation techniques, together with array processing for exploitation of spatial multipath diversity, have been shown to provide a feasible means for a more efficient use of the underwater acoustic channel bandwidth. These advancements are expected to result in a new generation of underwater communication systems, with at least an order of magnitude increase in data throughput.

Approaches to system design vary according to the technique used for overcoming the effects of intersymbol interference and signal phase variations. Specifically, these techniques may be classified according to (1) the signal design (i.e., the choice of modulation/detection method) and (2) the transmitter/receiver structure (i.e., the choice of array processing method and the equalization method, if any). In this section, the design of several systems that have been implemented is described. While most of the existing systems operate on the vertical, or the very short-range channels, the systems under development often focus on the severely spread horizontal shallow-water channels. Signal processing methods used in these systems are addressed in the following section.

3.1. Systems Based on Noncoherent Modulation

Noncoherent detection of FSK (frequency shift keying) signals has been used for channels exhibiting rapid phase variation such as the shallow-water long-range and medium-range channels. To overcome the ISI, the existing noncoherent systems employ signal design with guard times, that are inserted between successive pulses to ensure that all the reverberation will vanish before each subsequent pulse is to be received. The insertion of idle periods of time obviously results in a reduction of the available data throughput. In addition, because fading is correlated among frequencies separated by less than the coherence bandwidth (the inverse of the multipath spread), it is desired that only those frequency channels that are separated by more than the coherence bandwidth be used at the same time. This requirement further reduces the system efficiency unless some form of coding is employed so that the adjacent, simultaneously transmitted frequencies belong to different codewords. A representative system [13] for telemetry at a maximum of 5 kbps uses a multiple FSK modulation technique in the 20–30-kHz band. This band is divided into 16 subbands, in each of which a 4-FSK signal is transmitted. Hence, out of a total of 64 channels, 16 are used simultaneously for parallel transmission of 32 information bits (2 information bits per one 4-channel subband). This system has successfully been used for telemetry over a 4-km shallow-water horizontal path, and a 3-km deep-ocean vertical path. It was also used on a <1 km long shallow-water path, where probabilities of bit error on the order of 10^{-2}–10^{-3} were achieved without coding. The system performance may be improved by using error-correction coding (ECC); however, its data throughput will be reduced. This multiple FSK system is commercially available with a maximum data rate of 1200 bps (bits per second). Although bandwidth efficiency of this system does

not exceed 0.5 bps/Hz, noncoherent FSK is a good solution for applications where moderate data rates and robust performance are required. An improved FSK system [14] uses 128 subbands and employs coding. The essence of its coding method is a Hadamard H(20,5) code, in which each 5 input bits are encoded into 20 output bits (the minimum distance of this code is 10). The encoded bits dictate the choice of active subbands for transmission of the given codeword. The 20 subbands that are simultaneously used are chosen (among the 128 available) to be maximally separated, which ensures the least correlated fading, and thus provides diversity on time-varying underwater channels. Because of their robustness and simplicity of implementation, the noncoherent signaling methods are being further developed, and a system has been implemented [15] that uses orthogonal frequency-division multiplexing (OFDM) realized with DFT (discrete-time Fourier transform)-based filter banks. This system was used on a medium-range channel; however, because of the high-frequency separation among the channels (only every fourth channel is used) and relatively long guard times (10-ms guard following a 30-ms pulse), needed to compensate for the multipath fading distortion, the effective data rate is only 250 bps.

3.2. Systems Based on Differentially Coherent and Coherent Modulation

With the goal of increasing the bandwidth efficiency of an underwater acoustic communication system, research focus has shifted toward phase-coherent modulation techniques, such as PSK (phase shift keying) and QAM (quadrature amplitude modulation). Phase-coherent communication methods, previously considered infeasible, were demonstrated to be a viable way of achieving high-speed data transmission over many of the underwater channels, including the severely time-spread horizontal shallow-water channels [24–27]. These methods have the capability to provide raw data throughputs that are an order of magnitude higher than those of the existing noncoherent systems.

Depending on the method for carrier synchronization, phase-coherent systems fall into two categories: differentially coherent and purely phase-coherent. The advantage of using differentially encoded PSK (DPSK) with differentially coherent detection is the simple carrier recovery it allows; however, it has a performance loss as compared to coherent detection. Most of the existing systems employ DPSK methods to overcome the problem of carrier phase extraction and tracking. Real-time systems have been implemented mostly for application in vertical and very short-range channels, where little multipath is observed and the phase stability is good.

In the very short-range channel, where bandwidth in excess of 100 kHz is available, and signal stability is good, a representative system [16] operates over 60 m at a carrier frequency of 1 MHz and a data rate of 500 kbps. This system is used for communication with an undersea robot that performs maintenance of a submerged platform. A 16-QAM modulation is used, and the performance is aided by an adaptive equalizer. A linear equalizer, operating under a least-mean-squares (LMS) algorithm suffices to reduce the bit error rate from 10^{-4} to 10^{-7} on this channel.

A deep-ocean, vertical-path channel is used by an image transmission system [17]. This is 4-DPSK system with carrier frequency 20 kHz, capable of achieving 16 kbps bottom–surface transmission over 6500 m. The field tests of this system indicate the achievable bit error rates on the order of 10^{-4} with linear equalizer operating under an LMS algorithm.

Another example of a successfully implemented system for vertical-path transmission is that of an underwater image and data transmission system [29]. This system uses a binary DPSK modulation at a rate of 19.2 kbps. The carrier frequency of 53 kHz was used for transmission over 2000 m.

More recent advances in digital underwater speech transmission are represented by a prototype system described in Ref. 19. This system uses a code-excited linear prediction (CELP) method to transmit the speech signal at 6 kbps. The modulation method used is 4-DPSK. A decision-feedback equalizer, operating under LMS algorithm is being used in the pool tests. Field tests have not been reported yet. A similar approach has been considered [20].

For applications in shallow-water medium-range channel, a binary DPSK system [21] uses a direct-sequence spread-spectrum (DSSS) method to resolve a strong surface reflection observed in the 1-km-long, 10-m-deep channel. The interfering reflection is only rejected, and not used for multipath recombining. Data throughput of 600 bps within a bandwidth of 10 kHz is achieved. Such high spreading ratios are justified in interference-suppression applications.

Current state of the art in phase-coherent underwater communications is represented by the system described by Johnson et al. [30]. This system is based on purely phase-coherent modulation and detection principles [24] of 4-PSK signals. The signals are transmitted at 5 kbps, using a carrier frequency of 15 kHz. The system's real-time operation in configuration as a six-node network was demonstrated in the under-ice shallow-water environment. To overcome the ISI caused by shallow-water multipath propagation, the system uses a decision feedback equalizer operating under an RLS (recursive least squares) algorithm.

4. SIGNAL PROCESSING METHODS FOR MULTIPATH COMPENSATION

To achieve higher data rates, bandwidth-efficient systems based on phase-coherent signaling methods must allow for considerable ISI in the received signal. These systems employ either some form of array processing, equalization methods, or a combination thereof, to compensate for the distortions. Three main approaches have been taken toward this end. The first two approaches use differentially coherent detection and rely on array processing to eliminate, or reduce, multipath. The third approach is based on purely phase-coherent detection and the use of equalization together with array processing for exploitation of the multipath and spatial diversity.

Array processing for multipath suppression has been used at both the transmitter and receiver ends. Transmitter arrays can be used to excite only a single path of propagation, but very large arrays are required. To overcome the need for a large array, the use of parametric sources has been studied extensively [22]. These highly directive sources rely on the nonlinearity of the medium in the vicinity of a transducer where two or more very high frequencies from the primary projector are mixed. The resulting difference frequency is transmitted by a virtual array formed in the water column in front of the projector. A major limitation of such a source is in its high power requirements. High directivity implies the problem of pointing errors, and careful positioning is required to ensure complete absence of multipath. These systems have been employed in shallow-water channels where equalization is not deemed feasible because of rapid time variation of the signal. Instead, a receiving array is employed to compensate for the possible pointing errors. Binary and quaternary DPSK signals were used achieving data rates of 10 and 20 kbps, respectively, with a carrier frequency of 50 kHz. The estimated bit error rate was on the order 10^{-2}–10^{-3}, depending on the actual channel length. In general, the technique was found to be more effective at shorter ranges.

Multipath rejection using adaptive beamforming at the receiver end only is another possibility. The beamformer [23] uses an LMS algorithm to adaptively steer nulls in the direction of a surface-reflected wave. Similarly as in the case of the transmitter array, it was found that the beamformer encounters difficulties as the range increases relative to depth. To compensate for this effect, the use of an equalizer was considered to complement the performance of the beamformer. The equalizer operates under an LMS algorithm whose low computational complexity permits real-time adaptation at the symbol rate. A separate waveform is transmitted at twice the data rate for purposes of time synchronization. The system was tested in shallow-water at 10 kbps, using a carrier frequency of 50 kHz, and showed the estimated bit error rate of 10^{-2} without, and 10^{-3} with, the equalizer.

A different method, based on purely phase-coherent detection, uses joint synchronization and equalization for combating the effect of phase variations and ISI [24,25]. The equalization method is that of fractionally spaced decision feedback equalization, used with an RLS algorithm. The system incorporates spatial signal processing in the form of multichannel equalization based on diversity combining. The phase-coherent methods have been tested in a variety of underwater channels with severe multipath, showing satisfactory performance regardless of the link geometry. The achieved data rates of up to 2 kbps over long-range channels, and up to 40 kbps over shallow-water medium-range channels, are among the highest reported to date. These methods are discussed in more detail below.

4.1. Design Example: Multichannel Signal Processing for Coherent Detection

In many of the underwater acoustic channels multipath structure may exhibit one or more components that carry the energy similar to that of the principal arrival. As the time progresses, it is not unusual for these components to exceed in energy the principal arrival (e.g., see Fig. 2). The fact that the strongest multipath component may not be well defined makes the extraction of carrier reference a difficult task in such a channel. To establish coherent detection in the presence of strong multipath, a technique based on simultaneous synchronization and multipath compensation may be used [24]. This technique is based on joint estimation of the carrier phase and the parameters of a decision feedback equalizer, where the optimization criterion is minimization of the mean-squared error (MSE) in the data estimation process. In addition, the equalizer/synchronizer structure can be extended to include a number of input array channels [25,26]. Spatial diversity combining has shown superior performance in a number of channels, as well as potentials for dealing with several types of interference. In Fig. 5,

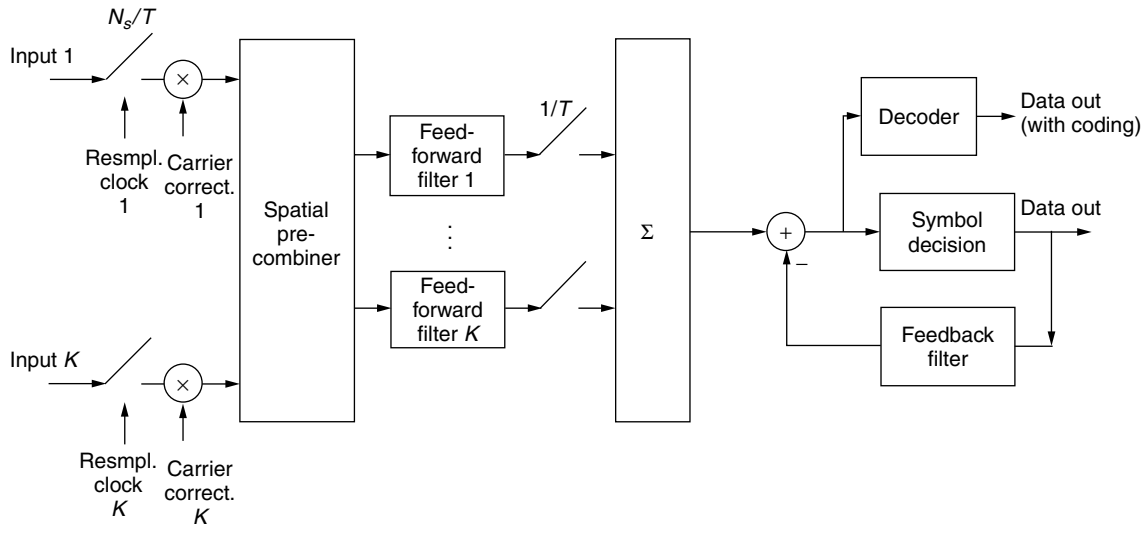


Figure 5. A multichannel receiver for phase-coherent detection.

the multichannel equalizer is shown, preceded by an additional precombiner, which may or may not be used depending on the application and the number of available received channels.

The input signals to the baseband processor are the A/D (analog/digital)-converted array signals, brought to baseband using nominal carrier and lowpass filtering. The signals are frame-synchronized using a known channel probe (usually a short Barker sequence transmitted in phase and quadrature at the data rate). Baseband processing begins with downsampling, which may be carried out to an interval of as few as 2 samples per symbol ($N_s = 2$), since the signals are shaped at the transmitter to have a raised-cosine spectrum that limits their maximal frequency to less than $1/T$. Since there is no feedback to the analog part of the receiver, the method is suitable for an all-digital implementation.

For applications where transmitter and receiver are not moving, but only drifting with water, no explicit adjustment of the sampling clock is needed. This will implicitly be accomplished during the process of adaptive fractionally spaced equalization. The front section of the equalizer will also perform adaptive matched filtering and linear equalization. To correct for the carrier offset, the signals in all channels are phase-shifted by the amount estimated in the process of joint equalization and synchronization. After coherent combining, the ISI resulting from the previously transmitted symbols (postcursors) is canceled in the feedback section of the equalizer. This receiver structure is applicable to any linear modulation format, such as M-PSK, or M-QAM; the only difference is in the way in which symbol decision is performed.

In addition to combining and equalization, signal processing at the receiver includes the operation of decoding if the signal at the transmitter was encoded. For example, in a DSP implementation of the receiver [28] two coding methods are used: concatenated coding of an outer Reed–Solomon code and an inner cyclic block code (Hamming, BCH), and punctured convolutional coding with interleaving. Alternatively, trellis coded modulation, compatible with PSK and QAM signals, provides an effective means of improving performance on a band-limited channel.

The receiver parameters that are adaptively adjusted are the weights of the precombiner, the tap weights of the feedforward filters, the carrier phase estimates, and the tap weights of the feedback filter. A single estimation error is used for the adaptation of all parameters. This error is the difference between the estimated data symbol at the input to the decision device and its true value. During the initial training mode, the true data symbols are known. After the training period, when the receiver parameters have converged, the online symbol decisions are fed back to the equalizer and used to compute the error. The adaptive algorithm used to update the receiver parameters is a combination of the second-order digital phase-locked loop (PLL) for the carrier phase estimates, and the RLS algorithm for the multichannel equalizer tap weights. The complexity of the multichannel equalizer grows with the number of receiver array sensors. For this reason, the spatial precombiner may be used to limit the number of equalizer channels, but still make use of the diversity gain. The precombiner weights can be estimated jointly with the rest of adjustable parameters. The details of the joint adaptation are given in a 1995 paper [26].

The receiver is adaptively adjusted to coherently combine the multiple signal arrivals and thus exploit both spatial and temporal, or multipath, diversity gain. In this manner, it differs from a receiver based on adaptive beamforming that is adjusted to null out the signal replicas arriving from angles different from those of the desired path. The signal isolated by a beamformer usually has to be processed by a separately optimized equalizer to compensate for the residual ISI that arises because the beamformer cannot completely eliminate the multipath interference. Since it is not constrained by angular resolution, the method of multichannel equalization may be used with as few as two input channels, and is applicable to a variety of underwater acoustic channels, regardless of the range : depth ratio. In applications where large arrays are available, the precombiner reduces receiver complexity, while preserving the multichannel diversity gain.

The method of adaptive multichannel combining and equalization was demonstrated to be effective in underwater channels with fundamentally different mechanisms of multipath formation. Experimental results include data rates of 2 kbps over three convergence zones (200 km or 110 nmi) in deep water; 2 kbps over 90 km (50 nmi) in shallow water, and up to 40 kbps over 1–2 km in rapidly varying shallow-water channels [7].

5. ACTIVE RESEARCH TOPICS

At this stage in the development of underwater acoustic communication techniques, with the feasibility of high-rate communications established, a number of research topics are foreseen that will influence the development of future systems. These topics include reduced-complexity receiver structures and algorithms suitable for real-time implementation, techniques for interference suppression, multiuser underwater communications, system self-optimization, development of modulation/coding methods for improved bandwidth efficiency, and mobile underwater acoustic communication systems.

5.1. Reducing the Receiver Complexity

Although the underwater acoustic channels are generally confined to low data rates as compared to many other communication channels, the encountered channel distortions require complex signal processing methods, resulting in high computational load that may exceed the capabilities of the available programmable DSP platforms. Consequently, reducing the receiver complexity to enable efficient real-time implementation has been a focus of many studies.

The problem of reducing the receiver complexity may be addressed on two levels: the design of an efficient receiver structure and the design of an efficient adaptive algorithm. For application to time-varying channels,

the receiver — whether it is based on array processing, equalization, or both methods — must use an adaptive algorithm for adjusting its parameters. Two commonly used types of algorithm are based on the LMS and the RLS estimation principles.

In a majority of the more recent studies, the LMS-based algorithms are considered as the only alternative because of their low computational complexity, which is linear in the number of coefficients N [20,23,33]. However, the LMS algorithm has a convergence time that may become unacceptably long when large adaptive filters are used ($20N$ as opposed to $2N$ of the RLS algorithm). The total number of coefficients N may be very large (more than 100 taps is often needed for spatial and temporal processing in medium and long-range shallow-water channels). In addition, the LMS algorithm is very sensitive to the choice of step size. To overcome this problem, self-optimized LMS algorithms may be used [33], but this results in increased complexity, and increased convergence time.

RLS algorithms, on the other hand, have better convergence properties but higher computational complexity. The quadratic complexity of the standard RLS algorithm is too high when large adaptive filters need to be implemented. In general, it is desirable that the algorithm be of linear complexity, a property shared by the fast RLS algorithms. A numerically stable fast RLS algorithm [31] has been used for the multichannel equalizer [25]. Despite its quadratic complexity, a square-root RLS algorithm [32] has been used for real-time implementation [30]. The advantage of this algorithm is that it allows the receiver parameters to be updated only periodically, rather than every symbol interval, thus reducing the computational load per each detected symbol. In addition, the updating intervals can be determined adaptively, based on monitoring the mean-squared error. Such adaptation methods are especially suitable for use with high transmission rates, where long ISI requires large adaptive filters, but eliminates the need to update the receiver parameters every symbol interval. The square-root RLS algorithm has excellent numerical stability, which makes it a preferable choice for a practical implementation. A different class of adaptive filters, which also have the desired convergence properties and numerical stability, are the lattice filters that use RLS algorithms. These algorithms have been proposed [34], but have not yet been applied to underwater acoustic channel equalization. Choosing an appropriate receiver adaptation method is expected to receive more attention in the future acoustic modem design.

Regardless of the adaptive algorithm used, its computational complexity is proportional to the number of receiver parameters (tap weights). Rather than focusing on low-complexity algorithms only, one may search for a way to reduce the receiver size. Although the use of spatial combining reduces residual ISI and allows shorter-length equalizers to be used, a broadband combiner may still require a large number of taps to be updated, limiting the practical number of receiving channels to only a few. The use of a precombiner [26] is a method for reducing a large number of input channels to a smaller number for subsequent multichannel equalization. By careful design, full diversity gain can be preserved by this technique. More than one channel at the output of the combiner is usually required, but this number is often small (e.g., three). The fact that diversity gain may be preserved is explained by multipath correlation across the receiver array. In addition to the reduced computational complexity, smaller adaptive filters result in less noise enhancement, contributing to improved performance.

A different approach in the design of reduced-complexity receiver structures has been investigated [35], where the focus is on reducing the number of equalizer taps. A conventional equalizer is designed to span all the channel responses. However, if the channel is characterized by several distinct multipath arrivals separated in time by intervals of negligible reverberation, an equalizer may be designed to have fewer taps. By reducing the number of adaptively adjusted parameters, this approach also makes it possible to use simple updating algorithms, such as standard RLS algorithms, which have good numerical stability. Finally, in channels that are naturally sparse, discarding the low-magnitude equalizer taps in fact results in improved performance since no unnecessary noise is processed.

5.2. Interference Cancelation

The sources of interference in underwater acoustic channels include external interference and internal interference, generated within the system. The external sources of interference include noise coming from onboard machinery or other nearby acoustic sources, as well as the propulsion and flow noise associated with the underwater vehicle launch process. The internal noise, which has signal-like characteristics, arises in the form of echo in full-duplex systems, and in the form of multiple-access interference generated by other users operating within the same network.

Methods for cancelation of interference in the form band-limited white noise and multiple sinusoidal interference have been investigated [36]. It was found that the multichannel receiver of Fig. 5 was most effective in canceling the interference while simultaneously detecting the desired signal. Noise cancellation is performed simply by providing a reference of the noise signal to one of the multichannel combiner inputs, while cancellation of the sinusoidal interferer may be performed even without the reference signal. By virtue of having the training sequence, the multichannel combiner is able to adaptively filter the interfering signal out, and extract the desired signal.

5.3. Multiuser Communications and Underwater Networks

A multiple-access communication system represents a special case of structured interference environment. Because of the bandwidth limitation of the underwater acoustic channel, frequency-division multiple access (FDMA) may not be an efficient technique. Time-division multiple access (TDMA) is associated with the problem of efficient time-slot allocation, which arises because of the long propagation delays. A possible solution in such a situation is to allow a number of users to transmit simultaneously in both

time and frequency. The receiver then has to be designed to deal with the resulting multiple-access interference, which may be very strong in an underwater acoustic network. The fact that transmission loss varies significantly with range, and that only very low code-division processing gains are available as a result of bandwidth constraints, both contribute to the enhanced near–far effect in the underwater acoustic channel. The multiuser detection methods suitable for underwater acoustic channels rely on the principles of joint synchronization, channel equalization, and multiple-access interference cancellation [37]. Two categories of multiuser receivers that have been considered are the (1) *centralized receiver*, in which the signals of all the users are detected simultaneously (e.g., uplink reception at a surface buoy, which serves as a central network node), and (2) the *decentralized receiver*, in which only the desired user's signal needs to be detected (e.g., downlink reception by an ocean-bottom node). Similarly as in the case of interference cancellation, the adaptive multichannel receiver of Fig. 5 was experimentally shown to have excellent capabilities in the role of a decentralized multiuser detector, operating without any knowledge of the interfering signal. Array processing plays a crucial role in the detection of multiuser signals, but is associated with the problem of computational complexity.

The advancements in point-to-point communication links have sparked an interest in the development of underwater acoustic communication networks. In addition to the choice of a multiple-access strategy, network design has been addressed on the levels of the data-link layer and the network layer [layers 2 and 3, respectively, of the seven-layer OSI (Open Systems Interconnection) reference model] [8,38]. Typically, packet transmission in a store-and-forward network is considered, and the design of automatic repeat request (ARQ) protocols and routing protocols is influenced by the long propagation times in the underwater channels. Underwater acoustic networks are a young area of research that is only awaiting new developments.

5.4. System Self-Optimization

A receiver algorithm must use a number of parameters that need to be adjusted according to the instantaneous channel conditions before the actual signal detection can begin. These parameters include the number and location of array sensors that provide good signal quality, the sizes of the equalizer filters, and their tracking parameters. The optimal values of receiver parameters depend not only on the general link configuration and location but also on the time of operation. In addition, an increase the background noise level, caused, for example, by a passing ship, may temporarily disable the communication. If the adaptive receiver algorithms are to be used in autonomous systems, external assistance in algorithm initialization, or reinitialization should be minimized. For this reason, the development of self-optimized receiver algorithms is of interest to future research.

The first steps in this direction are evident in the implementation of self-optimized LMS algorithms [23,33], in which the step size is adaptively adjusted, and the periodically updated RLS algorithm [30], self-adjusted to keep a predetermined level of performance by increasing the tracking rate if the channel condition worsens. These strategies provide the receiver with the capability to adjust to the fine channel changes. However, they depend on the availability of a reliable reference of the desired signal. Since a training sequence is inserted only so often in the transmitted signal, a loss of synchronization or convergence during detection of a data packet will cause the entire packet to be lost. An alternative to periodic reinsertion of known data, which increases the overhead, methods for self-optimized, or blind, recovery may be considered.

A blind equalization method based on using the cyclostationary properties of oversampled received signals [39], which requires only the estimation of second-order signal statistics, provides a practical solution for recovering the data sequence in the absence of clock synchronization. Originally developed for linear equalizers, this method has been extended to the case of the decision feedback equalizer, necessary for application in underwater acoustic channels with extreme multipath. These methods have proven successful in preliminary tests with real data [7]. Blind decision feedback equalization for application to underwater acoustic channels has also been investigated [40]. Further work on blind system recovery for underwater acoustic channels will focus on methods for array processing and carrier phase tracking.

5.5. Modulation and Coding

Coding techniques are known to be one of the most powerful tools for improving the performance of digital communication systems on both the additive white Gaussian noise channels and the fading channels. Several well-known techniques have been used for underwater communications with both noncoherent and coherent detection. Turbo codes are also being considered for use in underwater communications. While the performance of various codes is known on Gaussian noise channels and fading channels that can be described by Rayleigh or Rice statistics, it is not known as well on underwater acoustic channels. Future work should provide experimental results necessary for a better understanding of the performance of coded systems on these channels.

Achieving high throughputs over band-limited underwater acoustic channels is conditioned on the use of bandwidth-efficient modulation and coding techniques [41]. Related results documented in contemporary literature are confined to signaling schemes whose bandwidth efficiency is at most 3–4 bps/Hz. Higher-level signal constellations, together with trellis coding, are being considered for use in underwater acoustic communications. While trellis-coded modulation is well suited for vertical channels that have minimal dispersion, their use on the horizontal channels requires further investigation. In the first place, conventional signal mapping into a high-level PSK or QAM constellation may be associated with increased sensitivity of detection on a time-varying channel. Certain results in radio communications show that certain types of high-level constellations are more robust to the channel fading and phase variations than are the conventional rectangular QAM constellations [42]. Another

issue associated with the use of coded modulation on the channels with long ISI is the receiver design that takes full advantage of the available coding gain. Namely, the delay in decoding poses problems for an adaptive equalizer that relies on the feedback of instantaneous decisions. Receiver structures that deal with this problem as it applies to underwater channels are a subject of current studies.

In addition to bandwidth-efficient modulation and coding techniques, the future underwater communication systems will rely on data compression algorithms to achieve high data rates over severely band-limited underwater acoustic channels. This is another active area of research, which, together with sophisticated modulation and coding techniques, is expected to provide solutions for high-rate underwater image transmission.

5.6. Mobile Underwater Communications

The problem of channel variability, already present in applications with a stationary transmitter and receiver, becomes a major limitation for the mobile underwater acoustic communication system. The ratio of the vehicle speed to the speed of sound ($1/10^3$ for a vehicle speed of 30 knots or 54 km/h) often exceeds its counterpart in the mobile radio channels ($1/10^8$ for a mobile moving at 60 mi/h or 100 km/h), making the problem of time synchronization very difficult in the underwater acoustic channel. Apart from the carrier phase and frequency offset, the mobile underwater acoustic systems will have to deal with the motion-induced pulse compression and dilation (time scaling). Successful missions of experimental AUVs that use commercial FSK acoustic modems for vehicle-to-vehicle communication have been reported [43]. In a coherent acoustic modem, a method based on estimating the time-scaling factor from a signal preamble has been implemented and successfully demonstrated in operation with a remotely controlled underwater vehicle [44]. Rather than estimating the motion-induced distortion on a packet-to-packet basis, algorithms for continuous tracking of the time-varying symbol delay in the presence of underwater multipath are under development. One approach is based on a model that relates the instantaneous vehicle speed to the signal phase distortion. Using this relationship and the phase estimate from the PLL, the vehicle speed is calculated, and the corresponding time-scaling factor is used to resample the received signal before equalization. The resampling operation is efficiently implemented using polyphase filters. Other approaches are possible that do not rely on explicit estimation of the vehicle speed to perform adaptive resampling for highly mobile communication scenarios.

While many problems remain to be solved in the design of high-speed acoustic communication systems, more recent advances in this area serve as an encouragement for future work, which should facilitate remote exploration of the underwater world.

BIOGRAPHY

Milica Stojanovic graduated from the University of Belgrade, Belgrade, Yugoslavia, in 1988 and received her M.S. and Ph.D. degrees in electrical engineering from Northeastern University, Boston, Massachusetts, in 1991 and 1993. She is a principal research scientist at the Massachusetts Institute of Technology and a guest investigator at the Woods Hole Oceanographic Institution. Her research interests include digital communications theory and statistical signal processing, and their applications to wireless communication systems.

BIBLIOGRAPHY

1. R. Coates, *Underwater Acoustic Systems*, Wiley, New York, 1989.
2. L. Brekhovskikh and Y. Lysanov, *Fundamentals of Ocean Acoustics*, Springer, New York, 1982.
3. S. Flatte, ed., *Sound Transmission through a Fluctuating Ocean*, Cambridge Univ. Press, Cambridge, UK, 1979.
4. A. Quazi and W. Konrad, Underwater acoustic communications, *IEEE Commun. Mag.* 24–29 (1982).
5. J. Catipovic, Performance limitations in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **15**: 205–216 (1990).
6. A. Baggeroer, Acoustic telemetry — an overview, *IEEE J. Ocean. Eng.* **9**: 229–235 (1984).
7. M. Stojanovic Recent advances in high rate underwater acoustic communications, *IEEE J. Ocean. Eng.* **21**: 125–136 (1996).
8. D. Kilfoyle and A. Baggeroer, The state of the art in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **25**: 4–27 (2000).
9. R. Owen, B. Smith, and R. Coates, An experimental study of rough surface scattering and its effects on communication coherence, *Proc. Oceans'94*, 1994, Vol. III, pp. 483–488.
10. A. Essebbar, G. Loubet, and F. Vial, Underwater acoustic channel simulations for communication, *Proc. Oceans'94*, 1994, Vol. III, pp. 495–500.
11. A. Falahati, B. Woodward, and S. Bateman, Underwater acoustic channel models for 4800 b/s QPSK signals, *IEEE J. Ocean. Eng.* **16**: 12–20 (1991).
12. C. Bjerrum-Niese, L. Bjorno, M. Pinto, and B. Quellec, A simulation tool for high data-rate acoustic communication in a shallow-water, time-varying channel, *IEEE J. Ocean. Eng.* **21**: 143–149 (1996).
13. J. Catipovic, M. Deffenbaugh, L. Freitag, and D. Frye, An acoustic telemetry system for deep ocean mooring data acquisition and control, *Proc. Oceans'89*, 1989, pp. 887–892.
14. K. Scussel, J. Rice, and S. Merriam, A new MFSK acoustic modem for operation in adverse underwater channels, *Proc. Oceans'97*, 1997, Vol. I, pp. 247–254.
15. S. Coatelan and A. Glavieux, Design and test of a multicarrier transmission system on the shallow water acoustic channel, *Proc. Oceans'94*, 1994, Vol. III, pp. 472–477.
16. A. Kaya and S. Yauchi, An acoustic communication system for subsea robot, *Proc. Oceans'89*, 1989, pp. 765–770.
17. M. Suzuki and T. Sasaki, Digital acoustic image transmission system for deep sea research submersible, *Proc. Oceans'92*, 1992, pp. 567–570.
18. D. Hoag, V. Ingle, and R. Gaudette, Low-bit-rate coding of underwater video using wavelet-based compression algorithms, *IEEE J. Ocean. Eng.* **22**: 393–400 (1997).

19. A. Goalic et al., Toward a digital acoustic underwater phone, *Proc. Oceans'94*, 1994, Vol. III, pp. 489–494.

20. B. Woodward and H. Sari, Digital underwater voice communications, *IEEE J. Ocean. Eng.* **21**: 181–192 (April 1996).

21. J. Fischer et al., A high rate, underwater acoustic data communications transceiver, *Proc. Oceans'92*, 1992, pp. 571–576.

22. R. F. W. Coates, M. Zheng, and L. Wang, BASS 300 PARACOM: A model underwater parametric communication system, *IEEE J. Ocean. Eng.* **21**: 225–232 1996.

23. G. S. Howe et al., Sub-sea remote communications utilising an adaptive receiving beamformer for multipath suppression, *Proc. Oceans'94*, 1994, Vol. I, pp. 313–316.

24. M. Stojanovic, J. A. Catipovic, and J. G. Proakis, Phase coherent digital communications for underwater acoustic channels, *IEEE J. Ocean. Eng.* **19**: 100–111 (1994).

25. M. Stojanovic, J. A. Catipovic, and J. G. Proakis, Adaptive multichannel combining and equalization for underwater acoustic communications, *J. Acoust. Soc. Am.* **94**(3)(Pt. 1): 1621–1631 (1993).

26. M. Stojanovic, J. A. Catipovic, and J. G. Proakis, Reduced-complexity multichannel processing of underwater acoustic communication signals, *J. Acoust. Soc. Am.* **98**(2)(Pt. 1): 961–972 (1995).

27. M. Stojanovic, J. G. Proakis, and J. A. Catipovic, Performance of a high rate adaptive equalizer on a shallow water acoustic channel, *J. Acoust. Soc. Am.* **100**(4)(Pt. 1): 2213–2219 (1996).

28. L. Freitag, M. Grund, S. Singh, and M. Johnson, Acoustic communication in very shallow water: Results from the 1999 AUV Fest, *Proc. Oceans'00*, 2000.

29. G. Ayela, M. Nicot, and X. Lurton, New innovative multimodulation acoustic communication system, *Proc. Oceans'94*, 1994, Vol. I, pp. 292–295.

30. M. Johnson, D. Herold, and J. Catipovic, The design and performance of a compact underwater acoustic network node, *Proc. Oceans'94*, 1994, Vol. III, pp. 467–471.

31. D. Slock and T. Kailath, Numerically stable fast transversal filters for recursive least squares adaptive filtering, *IEEE Trans. Signal Process.* **39**: 92–114 (1991).

32. F. Hsu, Square root Kalman filtering for high-speed data received over fading dispersive HF channels, *IEEE Trans. Inform. Theory* **28**: 753–763 (1982).

33. B. Geller et al., Equalizer for video rate transmission in multipath underwater communications, *IEEE J. Ocean. Eng.* **21**: 150–155 (1996).

34. F. Ling and J. G. Proakis, Adaptive lattice decision-feedback equalizers–their performance and application to time-variant multipath channels, *IEEE Trans. Commun.* **33**: 348–356 (1985).

35. M. Stojanovic, L. Freitag, and M. Johnson, Channel-estimation-based adaptive equalization of underwater acoustic signals, *Proc. Oceans'99*, 1999, pp. 590–595.

36. J. Catipovic, M. Johnson, and D. Adams, Noise canceling performance of an adaptive receiver for underwater communications, *Proc. 1994 Symp. AUV Technology*, 1994, pp. 171–178.

37. M. Stojanovic and Z. Zvonar, Multichannel processing of broadband multiuser communication signals in shallow water acoustic channels, *IEEE J. Ocean. Eng.* **21**: 156–166 (1996).

38. E. Sozer, M. Stojanovic, and J. Proakis, Underwater acoustic networks, *IEEE J. Ocean. Eng.* **25**: 72–83 (2000).

39. L. Tong, G. Xu, and T. Kailath, Blind identification and equalization based on second-order statistics, *IEEE Trans. Inform. Theory* **40**: 340–349 (1994).

40. J. Gomes and V. Barroso, Blind decision-feedback equalization of underwater acoustic channels, *Proc. Oceans'98*, 1998, pp. 810–814.

41. J. Proakis, Coded modulation for digital communications over Rayleigh fading channels, *IEEE J. Ocean. Eng.* **16**: 66–74 (1991).

42. W. T. Webb and R. Steele, Variable rate QAM for mobile radio, *IEEE Trans. Commun.* **43**: 2223–2230 (1995).

43. S. Chappell et al., Acoustic communication between two autonomous underwater vehicles, *Proc. 1994 Symp. AUV Technology*, 1994, pp. 462–469.

44. L. Freitag et al., A bidirectional coherent acoustic communication system for underwater vehicles, *Proc. Oceans'98*, 1998, pp. 482–486.

ACTIVE ANTENNAS

ALAN R. MICKELSON
University of Colorado
Boulder, Colorado

The present article is an introduction to the topic of active antennas. The first section is a description of the field suitable for reading by almost any undergraduate science major. The next section is an in-depth reexamination of the subject, including equations and some derivations. Its basic idea is to provide the readers with enough tools to enable them to evaluate whether it is an active antenna that they might need for a specific application. The final section is a discussion of where active antennas are finding and will find application.

We should mention here that, if one really needs to design active antennas, one will need to go further than this article. The set of references to the primary research literature given in this article is by no means complete, nor is it meant to be. A good way to get started on the current literature on this topic would be a reading of the overview monograph of Navarro and Chang [1]. We will not cover active amplifiers in this article. However, this topic is treated in the book edited by York and Popović [2].

1. AN INTRODUCTION TO ACTIVE ANTENNAS

An antenna is a structure that converts electromagnetic energy propagating in free space into voltage and current in an electric circuit and/or vice versa. In a transceiver system, the antenna is used both to receive and to transmit free-space waves. At minimum, a transceiver then must consist of a signal source that serves to drive the antenna as well as a receiver circuit that reads out the signal from the antenna. Previously, practically all antenna systems operating in the microwave frequency regime (operation frequencies greater than 1 billion cycles per second, or 1 GHz) were designed mostly to isolate the antenna from the circuits — that is, to find ways to make system operation independent of the antenna's

electrical characteristics. In contradistinction, an active antenna is one in which the antenna actually serves as a circuit element of either the driver or the readout circuit. To understand why this is different from conventional antenna driving or readout will require us to take a brief historical trip through the last century or so.

Actually, the first antenna was an active one. Heinrich Hertz, back in 1884 [2a], was the first to demonstrate that one could generate radiowaves and that they would propagate from a transmitter to a receiver at the speed of light. The apparatus used is schematically depicted in Fig. 1. The idea of the transmitter is that, by discharging an induction coil (a wire looped about a magnetic core such that the composite device can store significant amounts of magnetic energy) into a spark gap, one can generate a current in the 5-mm-diameter wire. The voltage in the spark gap induces a current in the wires, which in turn induces a voltage in the wires, and this voltage in turn induces current, so that the voltage and current propagate along the two pieces of the wire to either side of the gap as waves, appearing much like a one-dimensional slice through a water wave propagating away from the point where a pebble has struck the water's surface (the spark gap). A wave will propagate rectilinearly until it encounters an obstruction, at which point it can suffer reflection from or transmission into the barrier that the obstruction presents. There will then be reflections off the metal spheres on the ends of the wire. The spark will generate a broad spectrum of frequencies or wavelengths. The reflections off the two ends, though, will tend to cancel each other except at certain special frequencies. The effect at these wrong frequencies is much like the effect of throwing a handful of pebbles into the pond and noting that, in between the points where the pebbles struck, the waves are much more indistinct than they are far from where the handful struck the surface. The special frequencies are ones which just fit into the region between the spheres. The current needs to be zero at the two ends in order to fit, whereas the voltage needs to be maximum

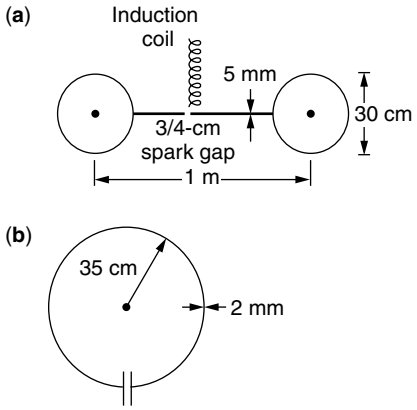


Figure 1. Hertz apparatus for (**a**) transmitting and (**b**) receiving radiowaves, where the transmitting antenna serves to choose a specific frequency of the spark-gap voltage to transmit to the receiving antenna, which also serves to pick out this special frequency from the free-space waveform and turn this electromagnetic disturbance into a voltage across the receiver antenna gap.

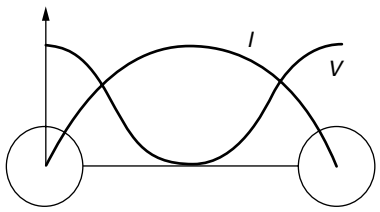


Figure 2. Current and voltage waveforms for the lowest-order (least number of zeros) waveform for the Hertz transmitter of Fig. 1a. The current must go to zero at the points where the wire ends, whereas the potential will be highest there.

at the ends. The current and voltage waves at the right frequencies may appear as depicted in Fig. 2.

The Hertz transmitter is the archetypical active antenna. The source is the spark gap, which is actually placed in the antenna. The antenna then acts as a filter to pick the right frequency out of a large number of frequencies that could be launched from the gap. The receiver is picked to be of a length to also select this primary frequency.

Hertz-style spark-gap transmitters, after further development and popularization by Marconi, were in use for 50 years after Hertz. However, such transmitters exhibit some rather severe drawbacks. The main problem is that the simple resonant dipole antenna (i.e., a straight-wire antenna with a gap or a feeder cable used to feed in current) is a filter with a poor frequency selection. Namely, if one increases the frequency by 50%, there is 75% as much power transmitted at this frequency as at the first resonance, which is called the *fundamental*. There is a second resonance at twice the frequency of the first resonance, and another at each integer multiple of the fundamental. With increasing frequency, the transmitted power decreases a little and then flattens out around the second resonance, decreases a little, flattens out at the third resonance, and so on, as illustrated in Fig. 3. If the spark discharge is really broadband (i.e., if it generates a large number of frequencies where the highest frequency may be many times the lowest), then what is transmitted by the antenna will also be broadband, although with somewhat higher transmission at the fundamental frequency and its harmonics than in between. In the very early days of radio, this was somewhat acceptable, although any information impressed on such a broadband carrier would be rather

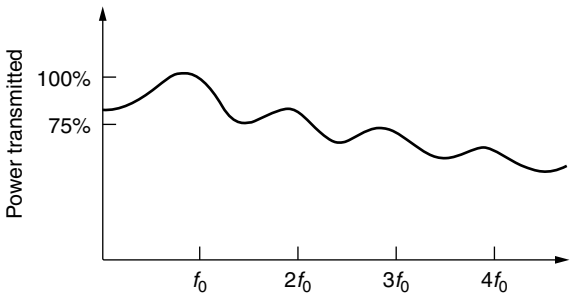


Figure 3. A sketch of what the transmission as a function of frequency might look like for the Hertzian dipole antenna of Figs. 1 and 2.

severely degraded on reception. However, the demise of the spark-gap transmitter was really instigated by the early success of radio, which caused the available frequency bands to begin to fill up rapidly. This band filling led to the formation of the Federal Communications Commission (FCC) in 1934, which was charged with allocation of frequency bands. The allocation by nature led to a ban on spark-gap transmitters, which were needlessly wasting bandwidth.

In a later experiment, Hertz noticed that the waves he was generating would tend to have a component that hugged the ground and could therefore travel over the horizon and, in fact, across the Atlantic Ocean, skimming along the surface of the water. Other researchers noticed that the effect became more pronounced at wavelengths longer than the roughly 2-m wavelength that Hertz originally used. (For the frequencies and wavelengths of some important frequency bands, see Table 1.) In order for wave transmission to be useful, however, the transmitted signal needs to carry information. Impressing information on the wave is called *modulating* the carrier. One can modulate the height (amplitude), the frequency, and so on. The discovery of a technique to *amplitude-modulate* the waves coming off an antenna (in 1906) then led to the inception of AM radio in bands with wavelengths greater than 300 m, which corresponds to roughly 1 MHz. AM radio became commercial in 1920. By the 1930s, other researchers noted that waves with frequencies around 10 MHz, corresponding to a wavelength around 30 m, could be quite efficiently propagated over the horizon by bouncing the wave off the ionosphere. This led to the radio bands known as *shortwave*. In 1939, a researcher realized a technique to modulate the frequency of the wave. This realization led in the 1950s to FM radio, which was allocated the band around 100 MHz with a corresponding wavelength around 3 m. However, the FM technique was used first during World War II as a radar modulation technique. Radars today are at frequencies above roughly 1 GHz or wavelengths below 30 cm.

Table 1. A Listing of the Allocated Microwave and Millimeter-Wave Bands as Defined by the Frequency and Wavelength Range within Each Band

Band Designation	Frequency (GHz)	Wavelength
L	1–2	15–30 cm
S	2–4	7.5–15 cm
C	4–8	3.75–7.5 cm
X	8–12	2.5–3.75 cm
Ku	12–18	1.67–2.5 cm
K	18–26	1.15–1.67 cm
Ka	26–40	0.75–1.15 cm
Q	33–50	6–9 mm
U	40–60	5–7.5 mm
V	50–75	4–6 mm
E	60–80	3.75–5 mm
W	75–110	2.7–4 mm
D	110–170	1.8–2.7 mm
G	140–220	1.4–2.1 mm
Y	220–325	0.9–1.4 mm

There is a fundamental difference between circuits that operate at frequencies whose corresponding wavelengths are less than the maximum circuit dimension and those that are large compared to the carrier wavelength. The effect is closely related to the concept of impedance. As was mentioned above, in the wire antenna, the voltage and current reinforce each other and thereby travel on the antenna as waves. The same effect takes place in a circuit. At any point along the path (line) in a circuit, one defines the ratio of voltage at one frequency to the current at the same frequency as the impedance at that frequency. For a sinusoidal waveform, if the impedance tends to preserve the phase relationship (where the wave peaks lie, relatively), then we say that the impedance is *resistive*. If the impedance tends to drive the voltage peaks forward with respect to the current peaks, we say that the impedance is *capacitive*; in the opposite case we say that the impedance is *inductive*. In a small circuit (small compared to a wavelength), one generally tries to carefully design passive components — resistors, capacitors, and inductors — so that they exhibit large local impedance, that is, large impedance within their physical dimensions. When the circuit is small, one would like to control the phase and amplitude of the wave at discrete points by using lumped elements and thereby minimizing line effects. The lines (wires) between the components have little or no effect on the electromagnetic disturbances passing through the circuit, then, as the impedances in the wires are small and reasonably independent of their lengths. When the circuit is large, the lines themselves effectively become circuit elements, and they themselves must be carefully designed in order to exhibit the proper impedances. To illustrate, consider the parallel-plate capacitor of Fig. 4. The capacitance is maximized by maximizing the permittivity ε (a material parameter equal to the ratio of electrical displacement to applied electric field) and area A while minimizing the plate spacing d. However, the fact that the capacitance depends on the plate spacing d is the important point here. Consider the circuit of Fig. 5 as an example. The only ground in the figure is the one on the battery, but the wires connecting the circuit elements together in essence form at each point a capacitor, with a point on the wire that is carrying charge as the upper plate and the ground as the

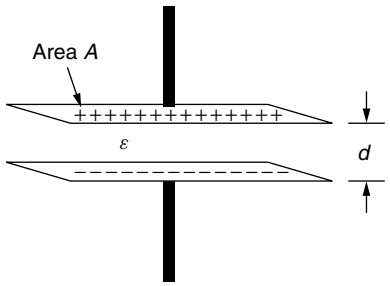


Figure 4. Schematic depiction of a parallel-plate capacitor in which the flow of a current will tend to change the upper plate, causing a voltage difference between upper and lower plates. The capacitance is defined as the ratio of the amount of change of the upper plate to the magnitude of the voltage this change induces between the plates.

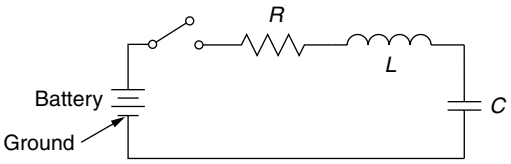


Figure 5. A circuit with lumped elements connected by wire segments.

lower. This capacitance changes as a function of position along the wire. For a small enough circuit (relative to the wavelength of the highest frequency carried by the circuit), the effect is not too important, as the wire–ground pair has small capacitance and the position-varying effect is small. For a large circuit, the effect is disastrous, as we shall consider below. The effect is identical to the effect of Fresnel coefficients in optics.

Consider the circuit of Fig. 6. We will now discuss what happens when impedances are not carefully controlled. This leads to the concept of *impedance matching*. Let us first say that the circuit is short (compared to a wavelength). If the load resistor, R_L, is not matched to (i.e., is not equal to, or, one could say, *not impedance-matched to*) the resistance of the source, R_S, some amount of reflection will occur at R_L, propagate back to R_S, be reflected with a reversal of sign at R_L, propagate back to R_L, and so on. The reflections add up perfectly out of phase (i.e., simply subtract from one another) at the source and load, and the amount of power supplied to the load is less than optimal. In this limit of a small circuit, it is as if the load will not allow the source to supply as much power as it is capable of. Let us now say that the line is "well designed" but long compared to the wavelength used. Then the same argument applies to the reflections, but in this case the source does not know that the load is there until several wave periods have passed (several maxima and minima of the waveform have left the source), so the source supplies all the power it can. The power, though, is not allowed to be fully absorbed by the load, and some of it will rattle around the line until it is radiated or absorbed. As we mentioned above, in a long enough circuit the wire itself becomes a distributed element — that is, one with an impedance of its own. If the distance to the nearest ground is not kept fixed along the line, the inductance and capacitance become dependent on the position. In this case, we have distributed reflections all along the line

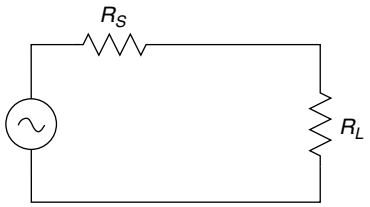


Figure 6. A circuit in which one is trying to supply power from a source with internal resistance R_S to a load with resistance R_L. The power transfer is maximized only when R_S and R_L are equal, in which case half the power supplied by the source is supplied to the load, the other half being dissipated in the source and causing it to heat.

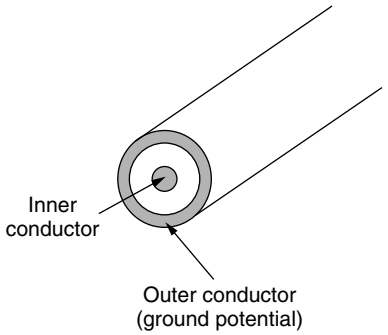


Figure 7. A coaxial cable in which signals are carried on an inner conductor and in which the grounded outer conductor serves to carry the ground plane along with the signal in order to give a constant impedance along the line.

and the circuit will probably not work at all. This spatial variability of the line impedance is remediable, though, as illustrated by the drawing of a coaxial cable in Fig. 7. The idea is that, if the line brings along its own ground plane in the form of a grounded outer conductor, the characteristic impedance of the line can be kept constant with distance. Such a line, which carries its own ground plane, is called a *transmission line*. The problem becomes the connection of the line to the source and load (i.e., impedance matching).

Before going on to discuss the conventional solution versus the new active-antenna solution, perhaps we should summarize a bit. In AM, shortwave, and FM applications, the wavelengths are of order greater than meters. If one considers typical receivers, the whole circuit will generally be small compared to the carrier wavelength. This is also to say that in all of these cases, the antennas will be active in the sense that the antenna presents an impedance to the circuit. (Recall that an active antenna is any antenna in which an active element lies within a wavelength of the antenna and is used as an element to match the antenna impedance to the decoder impedance.) To passively match an antenna to the receiver circuit, one needs pieces of line comparable to a wavelength. However, from here on we shall not be interested in the low-frequency case but rather in the well-above-1-GHz case, as AM, FM, and TV technologies are mature technologies. During World War II, radar was the application that drove the frequencies above 1 GHz (wavelength less than 30 cm). In a radar, one sends out a pulse and, from the returned, scattered wave, tries to infer as much as possible about the target. Target resolution is inversely proportional to wavelength. There has been a constant drive to shorten wavelength. Therefore as is indicated by Table 1, bands have been allocated out to hundreds of gigahertz. Presently, however, there are a plethora of nonmilitary drivers for pushing to higher-frequency communication systems that are compact and have lower power dissipation. However, the conventional solution, which was developed originally for radars, is really not conducive to compactness or to the pressures of cost minimization of the commercial market.

A typical conventional transmitter is schematically depicted in Fig. 8. A main concept here is that the transmission lines and matching networks are being used to isolate the oscillator from the amplifier and the

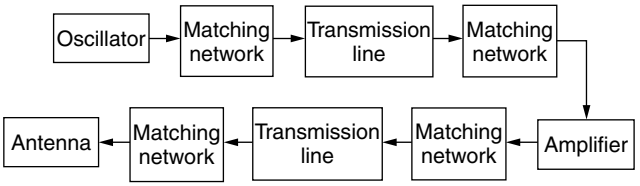


Figure 8. Schematic of a conventional RF microwave transmitter in which each individual element of the transmitter is matched to each other element.

amplifier from the antenna, in contrast to the situation in an active antenna. There were a number of reasons why the conventional solution took on the form it did. Among them was the urgency of World War II. Radar was developed rapidly in both Great Britain and the United States in the 1930s and 1940s. Rapid development required numerous researchers working in parallel. When operating frequencies exceeded 1 GHz (corresponding to 30 cm wavelengths), passive matching networks, whose main requirement is that they must consist of lines of lengths comparable to a wavelength, became convenient to construct (in terms of size) for ground-based radar. In this case, then, the oscillators could be optimized independently of the amplifiers, which in turn could be optimized independently of the antennas and the receiver elements. The impedances of the individual pieces didn't matter, as the matching networks could be used to effectively transform the effective impedances looking into an element into something completely different for purposes of matching pieces of the network to each other. There are costs associated with such a solution, though, such as total system size as well as the tolerances that components must satisfy. However, once the technique was in place, the industry standardized on the conventional solution and perfected it to the point where it was hard to challenge. The reemergence of the active solution owes itself to two independent technologies, the emergence of high-frequency solid-state devices and the development of planar circuit and planar antenna technology.

A single frequency of electromagnetic energy must be generated in a so-called oscillator — that is, a circuit that converts DC electrical power to AC electromagnetic power at the proper frequency. The basic operation of an oscillator can be described with respect to Fig. 9. What is shown here schematically is an amplifier in which a portion $b(<1)$ of the output is fed back to the input with either a plus or a minus sign. When the feedback is off ($b = 0$), then the signal out will be just G times the input. When the feedback is negative, the output will be less than G times the input. However, in the negative-feedback mode, the stability to noise increases, since fluctuations will be damped. That is, if the output fluctuates up, this lowers the effective input, whereas if the output fluctuates down, the output is driven up. The opposite is true in the positive-feedback case. In the positive-feedback case, if there were no fluctuations, any input would cause the output to increase until all of the DC power in as well as all the input signal showed up at the output. (This is all the power that can show up at the output. Such behavior is typical of unstable

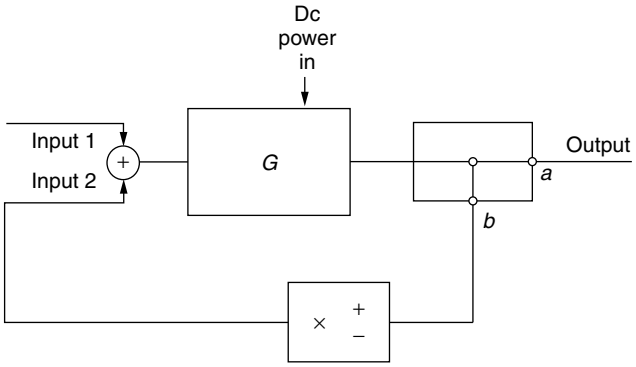


Figure 9. Schematic depiction of a feedback system that can operate as an oscillator when G is greater than 1, the feedback is positive, and there is a delay in feeding back the output to the input.

operation.) This would not be such an interesting case; however, there are always fluctuations of the input, and the positive feedback will cause these to grow. If there is a delay from output to input, then fluctuations with a period corresponding to this delay will be favored, as a rise in the input will show up as a rise in the output one period later, and all the DC power in will be rapidly converted to power out at this magic frequency.

A real circuit operates a bit more interestingly than our ideal one. In a real circuit, as the fluctuations build up, the gain is affected and some elements absorb power, but the oscillations still take place, although perhaps with a different frequency and amplitude from what one would have predicted from nondynamic measurements.

The transistor was first demonstrated in 1947, with publication in 1948 [3], and the diode followed shortly [4]. Although the field-effect transistor (FET) was proposed in 1952 [5], it was not until the mid-1960s that the technology had come far enough that it could be demonstrated [6]. The FET (and variations thereof) is presently the workhorse microwave three-terminal device. Two-terminal transfer electron devices (TEDs) were used before the FET for microwave applications and are still in use, but tend to have a much lower wall plug efficiency (DC/AC conversion), especially as the amplifying device of an oscillator. Radar systems, however, were already in use in the late 1930s. Essentially all of the microwave sources in radars up until the 1970s operated on principles that required that the source have physical dimensions larger than a wavelength, and perhaps many wavelengths. This fact almost required the conventional solution to be used. Transistors, though, can have active areas with dimensions of micrometers; even packaged hybrid devices can have complete packages of dimensions smaller than a millimeter. The transistor can therefore act as an amplifier with dimensions much smaller than a wavelength and does not, therefore, need to be placed in a conventional (passive) solution design.

The last piece of our story of the new active-antenna era involves the development of printed-circuit technology, along with slot and patch antennas. The two most common planar "open waveguide" designs are microstrip line and coplanar waveguide (CPW). Depictions of these waveguide lines are given in Fig. 10. The idea behind the microstrip

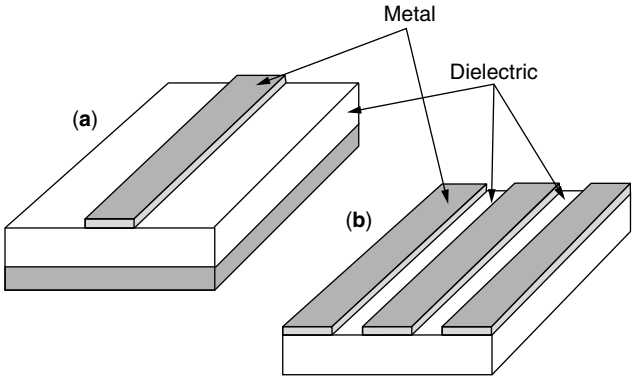


Figure 10. Views of (**a**) a microstrip and (**b**) a coplanar waveguide line. In the microstrip, the ground plane is the lower electrode, whereas in the coplanar waveguide the ground plane is placed on the surface of the dielectric substrate.

line is to propagate electromagnetic energy along the lines by confining the electric field between the upper signal line and a lower ground plane. As the upper line carries current, a magnetic field encircles the upper line. As power flow takes place in a direction perpendicular to the electric and magnetic fields, the power flow is mostly between the signal line and the ground line in the dielectric. On a low-frequency wire (a line whose transverse dimensions are small compared to a wavelength), the voltage and current waveforms reinforce each other. The coupling of the electric and magnetic fields in the microstrip is analogous to the coupling of voltage and current on the Hertz antenna wire, except that the microstrip line can be electrically long in the sense that the distance from the signal line to the ground plane is kept constant so that the impedance can be kept constant, as with the earlier-discussed coaxial cable. Lines that carry along their ground planes are generally referred to as *transmission lines*. Components (i.e., capacitors and inductors) can be built into the line by changing the width, cutting gaps into the upper line, or putting slits in the ground plane. In this sense, we can still describe transmission-line circuits by conventional circuit theory if we use a special circuit model for the line itself. The CPW line is quite similar to the microstrip line except that there the ground planes are on top of the dielectric slab. Either of these line types is reasonably easy to fabricate, as one needs only to buy a metal-coated dielectric plate and then pattern the needed shapes by photographically defining the patterns using a technique known as *photolithography*, a process common to all present-day circuit fabrication. These planar structures are quite compatible with transistor technology, as is indicated by the simple transistor oscillator circuit depicted in Fig. 11. The gap in the line on the drain side is there in order to provide the proper feedback for oscillation. In this case, the total oscillator linear dimension can be less than a wavelength.

In order to have an active antenna, one needs to have a radiating element — that is, a passive antenna element in the active antenna. Certain antenna technologies are compatible with microstrip and CPW technologies, and the resulting antenna types are illustrated in Fig. 12. The idea behind either of these antenna types is that the patch

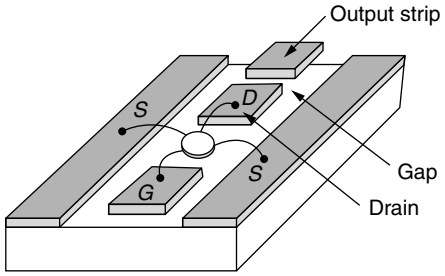


Figure 11. A simple transistor oscillator implemented in CPW technology.

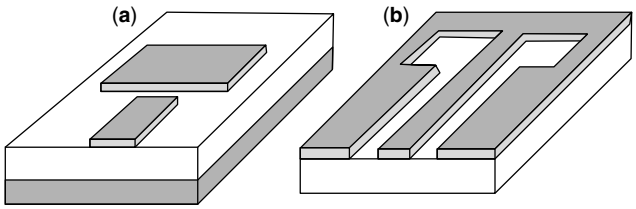


Figure 12. A depiction of (**a**) a patch antenna in a microstrip line and (**b**) a slot antenna in a CPW line.

(slit) is designed to have a transverse length that matches the operating wavelength (as we discussed in conjunction with Hertz dipole antennas). In the case of the patch, the electric field points primarily from the patch to the ground plane, as is illustrated in Fig. 13. The edges of the transverse (to the input line) dimension will then have a field pattern as sketched in Fig. 13a, and the longitudinal edges will have a field pattern as sketched in Fig. 13b, with a composite sketch given in Fig. 13c. The important part of the sketches, however, is really the so-called fringing fields in Fig. 13a — that is, the fields that point neither up nor down but rather across. Beyond the longitudinal edges of the patch are fields, in phase for the two edges, that are

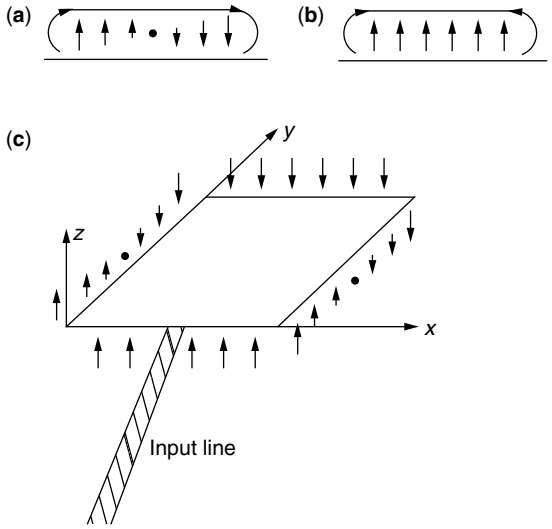


Figure 13. Illustration of the electric field directions along (**a**) the nonradiating edge and (**b**) the radiating edge, and (**c**) a schematic depiction of the edge fields around the patch.

normal to the surface. It is these fields (when combined with transverse magnetic fringe fields in the same strips) that give rise to the upward radiation. Similar arguments describe the operation of the slit antenna if one exchanges the electric and magnetic fields in the argument.

We have now introduced all of the pieces necessary to describe the new resurgence in active antenna research. A possible active-antenna design could appear as in Fig. 14 [7], where the transistor is actually mounted right into the patch antenna element, and therefore the design can be quite compact; that is, the source plus oscillator plus antenna can all be fitted into less than a wavelength. The design of Fig. 14, which comes from R. Compton's group at Cornell [31,32], will be discussed further in the next section.

There are a number of advantages to the use of active antennas. One is that an active antenna can be made compact. Compactness in itself is advantageous, as throughout the history of microelectronics, miniaturization has led to lowered costs. There are two more advantages, though, that relate to compactness. One is that the power-handling capabilities of a device go down with increasing frequency. We would therefore like to find ways to combine the power from several devices. One can try to add together outputs from various oscillators in the circuit before feeding them to the elements, but this goes back to the conventional solution. A more advantageous design is to make an array of antennas, with proper spacing relative to the wavelength and antenna sizes, and add the power of the locked oscillators in the array quasioptically in free space. (In other words, optical radiation tends to radiate into free space, whereas radiofrequency in microwave radiation needs to be kept in guiding waveguides until encroachment on radiating elements. *Quasioptics* uses the principle of the optical interferometer to combine multiple coherent microwave fields in free space.) The locking requires that the oscillators talk to each other so that the phases of all the array elements stay in a given relation. As will be discussed in more detail in the next section, however, an important problem at present in the active-antenna field relates to keeping elements locked yet still being able to modulate the output as well as steer the beam in order to be able to electronically determine on output direction. These issues will be discussed in Section 2 and taken up in more detail in Section 3.

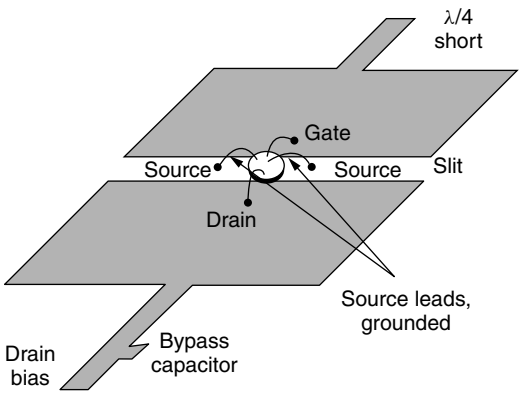


Figure 14. Depiction of the upper surface metallization of a microstrip active patch antenna discussed in Ref. 7. The short circuit on the gate together with the slit between gate and drain provides the proper feedback delay to cause oscillation.

2. SOME QUANTITATIVE DISCUSSION OF ASPECTS OF ACTIVE ANTENNAS

In order to be able to make calculations on active antennas, it is important to know what level of approximation is necessary in order to obtain results. An interesting point is that, although the operating frequency of active antennas is high, the circuit tends to be small in total extent relative to the operating wavelength, and therefore the primary design tool is circuit theory mixed with transmission-line theory. These techniques are approximate, and a most important point in working with high frequencies is to know where a given technique is applicable. Exact treatments of all effects, however, prove to be impossible to carry out analytically. Numerical approaches tend to be hard to interpret unless one has a framework to use. The combined circuit transmission-line framework is the one generally applied. When it begins to break down, one tends to use numerical techniques to bootstrap it back to reality. We will presently try to uncover the basic approximations of transmission-line and circuit theory.

Maxwell's equations are the basic defining equations for all electromagnetic phenomena, and they are expressible in mksA (meter-kilogram-second-ampere) units as [8]

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}$$

$$\nabla \cdot \boldsymbol{D} = \rho$$

$$\nabla \cdot \boldsymbol{B} = 0$$

where $\boldsymbol{E}$ is the electric field vector, $\boldsymbol{B}$ is the magnetic induction vector, $\boldsymbol{H}$ is the magnetic field vector, $\boldsymbol{D}$ is the electric displacement vector, $\boldsymbol{J}$ is the current density vector, and ρ is the volume density of charge. An additional important quantity is $\boldsymbol{S}$, the Poynting vector, defined by

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$$

If one takes the divergence of $\boldsymbol{S}$, one finds

$$\nabla \cdot \boldsymbol{S} = \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H})$$

If one assumes a free-space region

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E}$$

$$\boldsymbol{B} = \mu_0 \boldsymbol{H}$$

which is therefore lossless

$$\boldsymbol{J} = 0$$

and charge-free

$$\rho = 0$$

(where ε_0 is the permittivity of free space and μ_0 is the permeability of free space), one can use vector identities and Maxwell's equations to obtain

$$\nabla \cdot \boldsymbol{S} = -\frac{\varepsilon_0}{2}\frac{\partial}{\partial t}(\boldsymbol{E}\cdot\boldsymbol{E}) - \frac{\mu_0}{2}\frac{\partial}{\partial t}(\boldsymbol{H}\cdot\boldsymbol{H})$$

Integrating this equation throughout a volume V and using Gauss' theorem

$$\int \nabla \cdot \boldsymbol{S}\, dV = \int \boldsymbol{S}\cdot \boldsymbol{dA}$$

where $\boldsymbol{dA}$ is the differential area times the unit normal pointing out of the surface of the volume V, one finds that

$$\int \boldsymbol{S}\cdot\boldsymbol{dA} = -\frac{\partial}{\partial t}W_e - \frac{\partial}{\partial t}W_m$$

where W_e is the electric energy density

$$W_e = \frac{\varepsilon_0}{2}\int \boldsymbol{E}\cdot\boldsymbol{E}\, dV$$

and W_m is the magnetic energy density

$$W_m = \frac{\mu_0}{2}\int \boldsymbol{H}\cdot\boldsymbol{H}\, dV$$

The interpretation of the above is that the amount of $\boldsymbol{S}$ flowing out of V is the amount of change of the energy within. One therefore associates energy flow with $\boldsymbol{S} = \boldsymbol{E}\times\boldsymbol{H}$. This is important in describing energy flow in wires as well as transmission lines and waveguides of all types. As was first described by Heaviside [9], the energy flow in a wire occurs not inside the wire but around it. That is, as the wire is highly conductive, there is essentially no field inside it except at the surface, where the outer layer of oscillating charges have no outer shell to cancel their effect. There is therefore a radial electric field emanating from the surface of the wire, which combines with an azimuthal magnetic field that rings the current flow to yield an $\boldsymbol{E}\times\boldsymbol{H}$ surrounding the wire and pointing down its axis.

It was Pocklington in 1897 [10], who made the formal structure of the fields around a wire a bit more explicit and, in the effort, also formed the basis for the approximation on which most of circuit and transmission-line theory rests, the *quasistatic approximation*. A simplified version of his argument is as follows. Assume an $x-y-z$ Cartesian coordinate system where the axis of the wire is the z axis. One then assumes that all of the field quantities $\boldsymbol{f}(x, y, z, t)$ vary as

$$\boldsymbol{f}(x,y,z,t) = \boldsymbol{f}(x,y)\cos(\beta z - \omega t + \phi)$$

If one assumes that the velocity of propagation of the above-defined wave is $c = (\mu_0\varepsilon_0)^{-1/2}$, the speed of light, then one can write that

$$\beta = \frac{\omega}{c}$$

The assumption here that $\boldsymbol{f}(x,y)$ is independent of z, by substitution of the equation above into Maxwell's equations, can be shown to be equivalent to the assumption that the transverse field components E_x, E_y, B_x, and B_y all satisfy relations of the form

$$\left|\frac{\partial E_x}{\partial z}\right| \ll \beta |E_x|$$

which is the crux of the quasistatic approximation. With the preceding approximation, one finds that

$$\nabla_t \times \boldsymbol{E}_t = \rho$$
$$\nabla_t \times \boldsymbol{H}_t = \boldsymbol{J}$$

where

$$\nabla_t = \hat{\boldsymbol{e}}_x\frac{\partial}{\partial x} + \hat{\boldsymbol{e}}_y\frac{\partial}{\partial y}$$

which is just the transverse, and therefore two-dimensional, gradient operator. These equations are just the electro- and magnetostatic equations for the transverse fields, whereas the propagation equation above shows that these static transverse field configurations are propagated forward as if they corresponded to a plane wave field configuration. If the magnetic field is caused by the current in the wire, it rings the wire, whereas if the electric field is static, it must appear to emanate from charges in the wire and point outward at right angles to the magnetic field. If this is true, then the Poynting vector $\boldsymbol{S}$ will point along the direction of propagation and the theory is self-consistent, if approximate.

If we wish to guide power, then the quasistatic picture must come close to holding, as the Poynting vector is in the right direction for guidance. The more general approximate theory that comes from Pocklington's quasistatic approximation is generally called *transmission-line theory*. To derive this theory, first consider the two-wire transmission line of Fig. 15. If we are to have something that we can actually call a transmission line, then we would hope that we can find equiphase fronts of the electromagnetic disturbance propagating in the gap crossing the gap conductor and that we can find lines along which the current flows on the current-carrying conductor. Otherwise (if the equiphases closed on themselves and/or we had eddies in the current), it would be hard to think of the structure as any form of guiding structure. Let us say

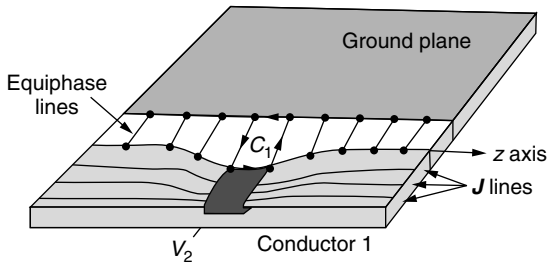


Figure 15. A sketch of a two-conductor transmission line where some equipotentials and some current lines are drawn in, as well as a volume V_1 with outward-pointing normal $d\boldsymbol{A}_1$. There is also an outward-pointing normal $d\boldsymbol{A}_2$ associated with the area bounded by contour C_2.

we form an area in the gap with two walls of the four-sided contour C_1 surrounding this area following equiphases an infinitesimal distance dz from each other. We can then write

$$\int \nabla \times \boldsymbol{E} \cdot \boldsymbol{dA}_1 = -\int \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{dA}_1$$

where $\boldsymbol{dA}_1$ corresponds to an upward-pointing normal from the enclosed area. One generally defines the integral as

$$\int \boldsymbol{B} \cdot \boldsymbol{dA}_1 = \phi$$

where ϕ is the magnetic flux. We often further define the flux as the inductance of the structure times the current:

$$\phi = Li$$

The integral with the curl in it can be rewritten by Stokes' theorem as

$$\int \nabla \times \boldsymbol{E} \cdot \boldsymbol{dA}_1 = \oint_{c_1} \boldsymbol{E} \cdot \boldsymbol{dl}$$

where C_1 is the contour enclosing the area. If we define

$$v = \int \boldsymbol{E} \cdot \boldsymbol{dl}$$

on the two equiphase lines of the contour C_1, where v is an AC voltage (this is the main approximation in the above, as it is only strictly true for truly static fields), then, noting that v does not change along two of the boundaries of the contour (because they are the infinitesimal walls on constant-voltage plates) and making the other two connecting lines infinitesimal, we note that the relation between the curl of $\boldsymbol{E}$ and the magnetic field reduces to

$$v(z + dz) - v(z) = \frac{\partial}{\partial t}(Li)$$

where it has been tacitly assumed that geometric deviations from rectilinearity are small enough that one can approximately use Cartesian coordinates, which can be rewritten in the form

$$\frac{\partial v}{\partial z} = l \frac{\partial i}{\partial t} \tag{1}$$

where l is an inductance per unit length, which may vary with longitudinal coordinate z if the line has longitudinal variation of geometry. A similar manipulation can be done with the second and third of Maxwell's equations. Taking

$$\nabla \cdot (\nabla \times \boldsymbol{H}) = \nabla \cdot \boldsymbol{J} + \frac{\partial}{\partial t} \nabla \cdot \boldsymbol{D}$$

and noting that the divergence of a curl is zero, substituting for $\nabla \cdot \boldsymbol{D}$, we find

$$\nabla \cdot \boldsymbol{J} + \frac{\partial \rho}{\partial t} = 0$$

which is the equation of charge conservation. Integrating this equation over a volume V_2 that encloses the current-carrying conductor whose walls lie perpendicular to the current lines gives

$$\int \nabla \cdot \boldsymbol{J} dV_2 = -\frac{\partial}{\partial t} \int \rho dV_2$$

where the total change Q, given by

$$Q = \int \rho dV_2$$

is also sometimes defined in terms of capacitance C and voltage v by

$$Q = Cv$$

Nothing that

$$\int \nabla \cdot \boldsymbol{J} dV_2 = \int \boldsymbol{J} \cdot \boldsymbol{dA}_2$$

where $\boldsymbol{dA}_2$ is the outward-pointing normal to the boundary of the volume V_2 and where one usually defines

$$i = \int \boldsymbol{J} \cdot \boldsymbol{dA}_2$$

and letting the volume V have infinitesimal thickness, one finds that

$$\int \boldsymbol{J} \cdot \boldsymbol{dA}_2 = i(z + dz) - i(z)$$

Putting this together with the preceding, we find

$$\frac{\partial i}{\partial z} = c \frac{\partial v}{\partial t} \tag{2}$$

where c is the capacitance per length of the structure, and where longitudinal variations in line geometry will lead to a longitudinal variation of c. The system of partial differential equations for the voltage and current have a circuit representation, as is schematically depicted in Fig. 16a. One can verify this by writing Kirchhoff's laws for the nodes with $v(z + dz)$ and $v(z)$ using the relations

$$v = l \frac{\partial i}{\partial t}$$

and

$$i = c \frac{\partial v}{\partial t}$$

Figure 16b illustrates the circuit equivalent for a lossy (and therefore dispersive) transmission line, where r represents the resistance encountered by the current in the metallization and where g represents any conductance of the substrate material that might allow leakage to ground. A major point of the diagram is that the structure need not be uniform in order to have a transmission-line representation, although one may find that irregularities in the structure will lead to longitudinally varying inductances and capacitances.

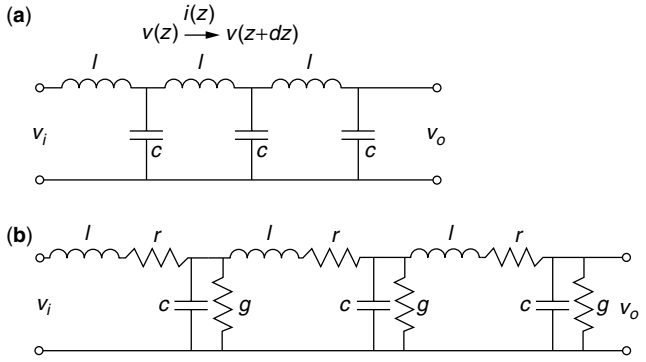


Figure 16. A circuit equivalent for (**a**) a lossless and (**b**) a lossy transmission line. The actual stages should be infinitesimally long, and the l and c values can vary with distance down the line. In reality, one can find closed-form solutions for the waves in nominally constant l and c segments and put them together with boundary conditions.

The solution to the circuit equations will have a wave nature and will exhibit propagation characteristics, which we discussed previously. In a region with constant l and c, one can take a z derivative of Eq. (1) and a t derivative of Eq. (2) and substitute to obtain

$$\frac{\partial^2 v}{\partial z^2} - lc\frac{\partial^2 v}{\partial t^2} = 0$$

which is a wave equation with solutions

$$v(z,t) = v_f \cos(\omega t - \beta z + \phi_f) + v_b \cos(\omega t + \beta z + \phi_b) \qquad (3)$$

where v_f is the amplitude of a forward-going voltage wave, v_b is the amplitude of a backward-going voltage wave, and

$$\frac{\omega}{\beta} = \sqrt{lc}$$

Similarly, taking a t derivative of Eq. (1) and a z derivative of Eq. (2) and substituting gives

$$\frac{\partial^2 i}{\partial z^2} - lc\frac{\partial^2 i}{\partial t^2} = 0$$

which will have a solution analogous to the one in Eq. (3) above, but with

$$v_f = \sqrt{\frac{l}{c}} i_f$$

$$v_b = \sqrt{\frac{l}{c}} i_b$$

which indicates that we can make the identification that the line phase velocity v_p is given by

$$v_p \triangleq \frac{\omega}{\beta} = \sqrt{lc}$$

and the line impedance Z_0 is given by

$$Z_0 = \sqrt{\frac{l}{c}}$$

Oftentimes, we assume that we can write (the sinusoidal steady-state representation)

$$v(z,t) = \mathrm{Re}[v(z)e^{j\omega t}]$$

$$i(z,t) = \mathrm{Re}[i(z)e^{j\omega t}]$$

so that we can write

$$\frac{\partial v}{\partial z} = -j\omega l i$$

$$\frac{\partial i}{\partial z} = -j\omega c v$$

with solutions

$$v(z) = v_f e^{-j\beta z} + v_b e^{j\beta z}$$

$$i(z) = i_f e^{-j\beta z} - i_b e^{j\beta z}$$

Let us say now that we terminate the line with a lumped impedance Z_l at location l. At the coordinate l, then, the relations

$$Z_l i(l) = v_f e^{-j\beta l} + v_b e^{j\beta l}$$

$$Z_0 i(l) = v_f e^{-j\beta l} - v_b e^{j\beta l}$$

hold, and from them we can find

$$v_f = \tfrac{1}{2}(Z_l + Z_0) i(l) e^{j\beta l}$$

$$v_b = \tfrac{1}{2}(Z_l - Z_0) i(l) e^{-j\beta l}$$

which gives

$$v(z) = \frac{i(l)}{2}[(Z_l + Z_0)e^{j\beta(l-z)} + (Z_l - Z_0)e^{-j\beta(l-z)}]$$

$$i(z) = \frac{i(l)}{2Z_0}[(Z_l + Z_0)e^{j\beta(l-z)} - (Z_l - Z_0)e^{-j\beta(l-z)}]$$

allowing us to write that

$$Z(z-\ell) = \frac{v(z-l)}{i(z-l)} = Z_0 \frac{Z_l + jZ_0 \tan\beta(z-l)}{Z_0 + jZ_l \tan\beta(z-l)} \qquad (4)$$

This equation allows us to, in essence, move the load from the plane l to any other plane. This transformation can be used to eliminate line segments and thereby use circuits on them directly. However, note that line lengths at least comparable to a wavelength are necessary in order to significantly alter the impedance. At the plane $z = l$, then, we can further note that the ratio of the reflected voltage coefficient v_b and the forward-going v_f, which is the voltage reflection coefficient, is given by

$$R = \frac{Z_l - Z_0}{Z_l + Z_0}$$

and has the meaning of a Fresnel coefficient [8]. This is the reflection we discussed in the last section, which causes the difference between large and small circuit dimensions.

One could ask what the use was of going at some length into Poynting vectors and transmission lines when

the discussion is about active antennas. The answer is that any antenna system, at whatever frequency or of whatever design, is a system for directing power from one place to another. To direct power from one place to another requires constantly keeping the Poynting vector pointed in the right direction. As we can surmise from the transmission-line derivation, line irregularities may cause the Poynting vector to wobble (with attendant reflections down the line due to attendant variations in the l and c), but the picture must stay close to correct for power to get from one end of the system to another. For this reason, active antennas, even at very high frequencies (hundreds of gigahertz), can still be discussed in terms of transmission lines, impedances, and circuit equivalents, although ever greater care must be used in applying these concepts at increasingly higher frequencies.

The next piece of an active antenna that needs to be discussed is the active element. Without too much loss of generality, we will take our device to be a field-effect transistor (FET). The FET as such was first described by Shockley in 1952 [5], but the MESFET (metal semiconductor FET), which is today's workhorse active device for microwave circuitry, was not realized until 1965 [6], when gallium arsenide (GaAs) fabrication techniques became workable albeit only as a laboratory demonstration. [Although we will discuss the MESFET in this section, it should be pointed out that the silicon MOSFET (metal oxide semiconductor FET) is the workhorse device of digital electronics and therefore the most common of all electronic devices presently in existence by a very large margin.] A top view of an FET might appear as in Fig. 17. As is shown clearly in the figure, an FET is a three-terminal device with gate, drain, and source regions. A cross section of the active region (i.e., where the gate is very narrow) might appear as in Fig. 18. The basic idea is that the saturation-doped n region causes current to flow through the ohmic contacts from drain to source (i.e., electrons flow from source to drain), but the current is controlled in magnitude by the electric field generated by the reverse bias voltage applied to the gate electrode. The situation is described in a bit more detail in Fig. 19, where bias voltages are defined and a typical I–V curve for DC operation is given. Typically the bias is supplied by a circuit such as that of Fig. 20. In what follows, we will simply assume that the biases are properly applied and isolated, and we will consider the AC operation. An AC circuit model is given in Fig. 21. If one uses the proper number of circuit values, these models can be quite accurate, but the values do vary from

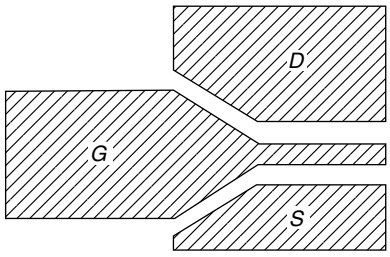


Figure 17. Schematic depiction of a top view of the metallized surface of an FET, where G denotes gate; D, drain; and S, source.

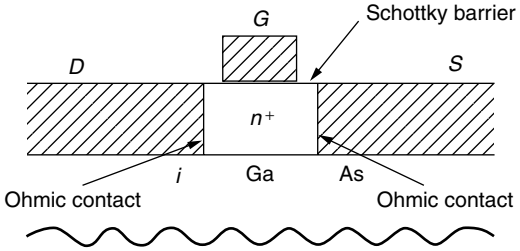


Figure 18. Schematic depiction of the cross section of the active region of a GaAs FET. Specific designs can vary significantly in the field-effect family.

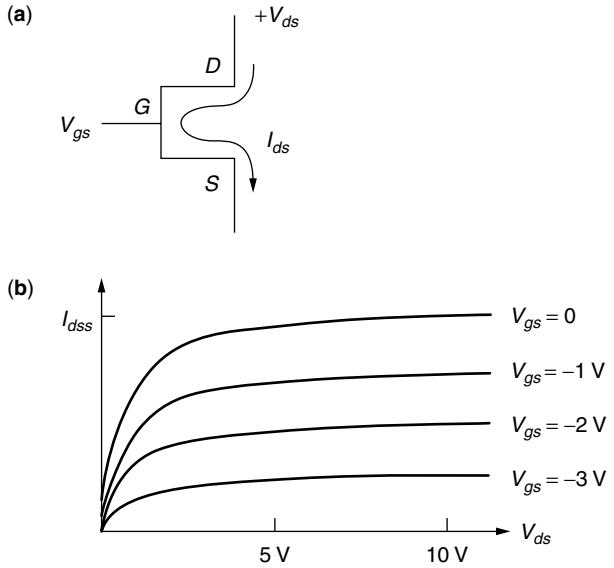


Figure 19. (**a**) Circuit element diagram with voltages and currents labeled for (**b**), where a typical I–V curve is depicted.

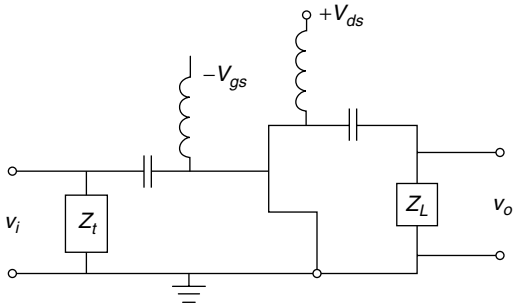


Figure 20. Typical FET circuit including the bias voltages v_{gs} and v_{ds} as well as the AC voltages v_i and v_o, where the conductors represent AC blocks and the capacitors, DC blocks.

device to device, even when the devices were fabricated at the same time and on the same substrate. Usually, the data sheet with a device, instead of specifying the circuit parameters, will specify the parameters of the device S, which are defined as in Fig. 22 and that can be measured in a straightforward manner by a network analyzer. The S parameters are defined by the equation

$$\begin{pmatrix} V_1^- \\ V_2^- \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} V_1^+ \\ V_2^+ \end{pmatrix} \tag{5}$$

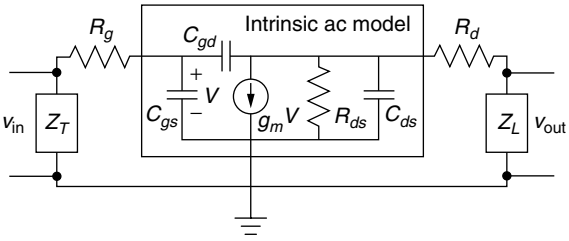


Figure 21. Intrinsic model for a common-source FET with external load and termination impedances and including gate and drain resistive parasitics, where Z_T is the gate termination impedance, R_g is the gate (metallization) resistance, C_{gs} is the gate-to-source capacitance, C_{gd} is the gate-to-drain capacitance, g_m is the channel transconductance, R_{ds} is the channel (drain-to-source) resistance, C_{ds} is the channel capacitance, R_d is the drain (metallization) resistance, and Z_L is the load impedance.

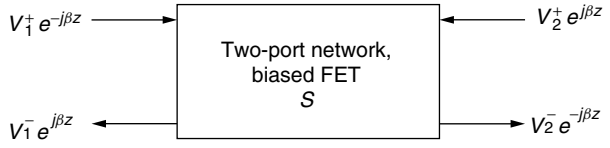


Figure 22. Schematic depiction of an FET as a two-port device that defines the quantities used in the S matrix of Eq. (5).

An important parameter of the circuit design is the transfer function of the transistor circuit, which can be defined as the ratio of v_o to v_i as defined in Fig. 21. To simplify further analysis, we will ignore the package parasitics R_g and R_d in comparison with other circuit parameters, and thereby we will carry out further analysis on the circuit depicted in Fig. 23. The circuit can be solved by writing a simultaneous system of equations for the two nodal voltages v_i and v_o. These sinusoidal steady-state equations become

$$v_i = v$$

$$j\omega C_{gd}(v_o - v_i) + g_m v_i + j\omega C_{ds} v_o + \frac{v_o}{R_{ds}} + \frac{v_o}{Z_L} = 0$$

The system can be rewritten in the form

$$v_o\left(j\omega(C_{gd} + C_{ds}) + \frac{1}{R_{ds}} + \frac{1}{Z_L}\right) = v_i(-g_m + j\omega C_{gd})$$

which gives us our transfer function T in the form

$$T = \frac{v_o}{v_i} = \frac{-g_m + j\omega C_{gd}}{j\omega(C_{gd} + C_{gs}) + \dfrac{1}{R_{ds}} + \dfrac{1}{Z_L}}$$

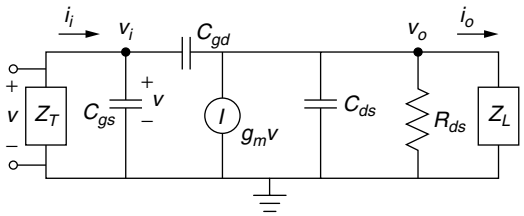


Figure 23. Simplified transistor circuit used for analyzing rather general amplifier and oscillator circuits, where the circuit parameter definitions are as in Fig. 22.

Oftentimes we are interested in open-circuit parameters — for example, the circuit transfer function when Z_L is large compared to other parameters. We often call this parameter G the open-circuit gain. We can write this open-circuit gain in the form

$$G = \left.\frac{v_o}{v_i}\right|_{oc} = \frac{-g_m R_{ds} + j\omega C_{gd} R_{ds}}{j\omega(C_{gd} + C_{gs})R_{ds} + 1}$$

It is useful to look at approximate forms. It is generally true that

$$C_{gd} \ll C_{ds}, C_{gs}$$

and for usual operating frequencies it is also generally true that

$$\frac{1}{\omega C_{ds}} \ll R_{ds}$$

Using both of these in our equations for T and G, we find

$$T = \frac{-g_m R_{ds}}{1 + \dfrac{R}{Z_L}}$$

$$G = -g_m R_{ds}$$

Clearly, one sees that the loaded gain will be lower than the unloaded gain, as we would expect. Making only the first of our two approximations above, we can write the above equations as

$$T = \frac{-g_m R_{ds}}{1 + j\omega\tau_{ds} + \dfrac{R_{ds}}{Z_L}}$$

$$G = \frac{-g_m R_{ds}}{1 + j\omega\tau_{ds}}$$

where τ_{ds} is a time constant given by

$$\tau_{ds} = \frac{1}{C_{ds} R_{ds}}$$

We see that, in this limit, the high-frequency gain is damped. Also, an interesting observation is that, at some frequency ω, an inductive load could be used to cancel the damping and obtain a purely real transfer function at that frequency. This effect is the one that allows us to use the transistor in an oscillator.

Let us now consider an oscillator circuit. The basic idea is illustrated in the one-port diagram of Fig. 24. The transistor's gain, together with feedback to the input loop through the capacitor C_{gd}, can give the transistor an effective negative input impedance, which can lead to oscillation if the real and imaginary parts of the total impedance (i.e., Z_T in parallel with the Z_i of the transistor plus load) cancel. The idea is much like that illustrated in Fig. 25 for a feedback network. One sees that the output of the feedback network can be expressed as

$$v_o = G(j\omega)[v_i - H(j\omega)v_o]$$

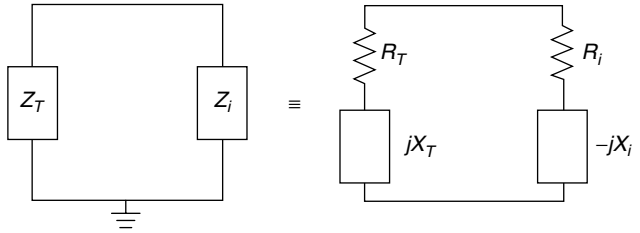


Figure 24. Diagram depicting the transistor and its load as a one-port device that, when matched to its termination so that there is no real or imaginary part to the total circuit impedance, will allow for oscillations.

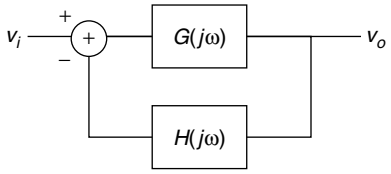


Figure 25. Depiction of a simple feedback network.

or, on rearranging terms

$$\frac{v_o}{v_i} = \frac{G(j\omega)}{1 + G(j\omega)H(j\omega)}$$

which clearly will exhibit oscillation — that is, have an output voltage without an applied input voltage — when

$$H(j\omega) = -\frac{1}{G(j\omega)}$$

What we need to do to see if we can achieve oscillation is to investigate the input impedance of our transistor and load seen as a one-port network. Clearly we can write the input current of Fig. 23 as

$$i_i = j\omega C_{gs} v_i + j\omega C_{gd}(v_i - v_o)$$

and then, using the full expression for T to express v_o as a function of v_i, one finds

$$Z_i = \frac{i_i}{v_i} = j\omega C_{gs} + j\omega C_{gd}\left(1 + \frac{g_m - j\omega C_{gd}}{j\omega(C_{gd} + C_{ds}) + \dfrac{1}{R_{ds}} + \dfrac{1}{Z_L}}\right)$$

which can be somewhat simplified to yield

$$Z_i = j\omega C_{gs} + j\omega C_{gd}\frac{g_m R_{ds} + 1 + j\omega\tau_{ds} + \dfrac{R_{ds}}{Z_L}}{1 + j\omega\tau_{ds} + \dfrac{R_d}{Z_L}}$$

We can again invoke a limit in which $\omega\tau_{ds} \ll 1$ and then write

$$Z_i = j\omega C_{gs} + j\omega C_{gd}\frac{Z_L(1 + g_m R_{ds} + R_{ds})}{R_{ds} + Z_L}$$

Perhaps the most interesting thing about this expression is that if

$$Z_L = j\omega L$$

and

$$g_m R_{ds} \gg 1$$

then clearly

$$R_i < 0$$

Whether X_i can be made to match any termination is another question, which we will take up in the next paragraph.

As was mentioned earlier, generally the data sheet one obtains with an FET has plots of the frequency dependence of the S parameters rather than values for the equivalent-circuit parameters. Oscillator analysis is therefore usually carried out using a model of the circuit such as that depicted in Fig. 26, where the transistor is represented by its measured S matrix. The S matrix is defined as the matrix of reflection and transmission coefficients. That is to say, with reference to the figure, S_{11} would be the complex ratio of the field reflected from the device divided by the field incident on the device. S_{21} would be the field transmitted from the device divided by the field incident on the device. S_{12} would be the field incident from the load side of the device divided by the power incident on the device, and S_{22} would be the power reflected from the load side of the device divided by the power incident on the device. For example, if there is only an input from Z_T, then

$$\Gamma_i = S_{11}$$

If there is only an input from Z_L, then

$$\Gamma_o = S_{22}$$

The condition for oscillation in such a system can be expressed in either of the forms

$$\Gamma_i \Gamma_T = 1$$

or

$$\Gamma_o \Gamma_L = 1$$

where the Γs are defined in the caption of Fig. 26. If both Z_T and Z_L were passive loads — that is, loads consisting of

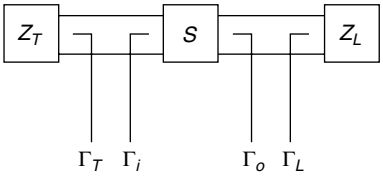


Figure 26. Schematic depiction of an oscillator circuit in which the transistor is represented by its S matrix and calculation is done in terms of reflection coefficients Γ_T looking into the gate termination, Γ_i looking into the gate source port of the transistor, Γ_o looking into its drain source port, and Γ_L looking into the load impedance.

resistance, inductance, and capacitance — then we would have that

$$|\Gamma_T| < 1$$
$$|\Gamma_L| < 1$$

and the conditions for unconditional stability (nonoscillation at any frequency) would be that

$$|\Gamma_i| < 1$$
$$|\Gamma_o| < 1$$

Clearly, we can express Γ_i and Γ_o as series of reflections such that

$$\Gamma_i = S_{11} + S_{12}\Gamma_L S_{21} + S_{12}\Gamma_L S_{22}\Gamma_L S_{21} + S_{12}\Gamma_L S_{22}\Gamma_L S_{22}\Gamma_L S_{21} + \cdots$$
$$\Gamma_o = S_{22} + S_{21}\Gamma_T S_{12} + S_{21}\Gamma_T S_{11}\Gamma_T S_{12} + S_{21}\Gamma_T S_{11}\Gamma_T S_{11}\Gamma_T S_{12} + \cdots$$

Using the fact that

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

we can reexpress the Γs as

$$\Gamma_i = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}$$
$$\Gamma_o = S_{22} + \frac{S_{12}S_{21}\Gamma_T}{1 - S_{22}\Gamma_T}$$

If we denote the determinant of the S matrix by

$$\Delta = S_{11}S_{22} - S_{12}S_{21}$$

and define a transistor parameter κ by

$$\kappa = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|}$$

then some tedious algebra leads to the result that stability requires

$$\kappa > 1$$
$$\Delta < 1$$

At frequencies where the above are not satisfied, oscillation can occur if the load and termination impedances, Z_L and Z_T respectively, are chosen properly. Oscillator design is discussed in various texts [11–14]. Generally, though, oscillator design involves finding instability points and not predicting the dynamics once oscillation is achieved. Here we are discussing only oscillators that are self-damping. External circuits can be used to damp the behavior of an oscillator, but here we are discussing only those that damp themselves independent of an external circuit. The next paragraph will discuss these dynamics.

If a transistor circuit is designed to be unstable, then, as soon as the DC bias is raised to a level where the circuit achieves the set of unstable values, the circuit's output within the range of unstable frequencies rises rapidly and dramatically. The values that we took in the equivalent AC circuit, though, were small-signal parameters. As the circuit output increases, the signal will eventually no longer be small. The major thing that changes in this limit is that the input resistance to the transistor saturates, so that [14]

$$R_i = -R_{i\phi} + mv^2$$

where the plus sign on the nonlinearity is necessary, for if it were negative, the transistor would burn up or else burn up the power supply. Generally, m has to be determined empirically, as nonlinear circuit models have parameters that vary significantly from device to device. For definiteness, let us assume that the Z_T is resistive and the Z_L is purely inductive. At the oscillation frequency, the internal capacitance of the transistor then should cancel the load inductance, but to consider dynamics we need to put in both C and L, as dynamics take place in the time domain. The dynamic circuit to consider is then as depicted in Fig. 27. The loop equation for this circuit in the time domain is

$$L\frac{\partial i}{\partial t} + (R_i + R_T)i + \frac{1}{C}\int i dt = 0$$

Recalling the equivalent circuit of Fig. 23 and recalling that

$$C_{gs} \gg C_{gd}$$

we see that, approximately at any rate, we should have a relation between v_i and i_i of the form

$$i_i = C_{gs}\frac{\partial v_i}{\partial t}$$

Using this $i - v$ relation above, we find that

$$\frac{\partial^2 v}{\partial t^2} - \frac{R_i - R_T}{L}\left(1 - \frac{mv^2}{R_i - R_T}\right)\frac{\partial v}{\partial t} + \frac{v}{LC} = 0$$

which we can rewrite in terms of other parameters as

$$\frac{\partial^2 v}{\partial t^2} - \varepsilon(1 - \gamma^2 v^2)\frac{\partial v}{\partial t} + \omega_0^2 v = 0$$

which is the form of Van der Pol's equation [15,16], which describes the behavior of essentially any oscillator.

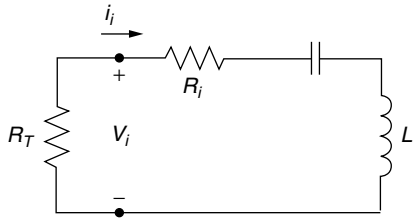


Figure 27. Circuit used to determine the dynamical behavior of a transistor oscillator.

Now that we have discussed planar circuits and dynamical elements that we can put into theory, the time has arrived to discuss planar antenna structures. Perhaps the best way to gain understanding of the operation of a patch antenna is by considering a cavity resonator model of one. A good review of microstrip antennas is given in Carver and Mink [17] and is reprinted in Pozar and Schaubert [18]. Let us consider a patch antenna and coordinate system as is illustrated in Fig. 28. The basic idea behind the cavity model is to consider the region between the patch and ground plane as a resonator. To do this, we need to apply some moderately crude approximate boundary conditions. We will assume that there is only a z-directed electric field underneath the patch and that this field achieves maxima on the edges (open-circuit boundary condition). The magnetic field $\boldsymbol{H}$ will be assumed to have both x and y components, and its tangential components on the edges will be zero. (This boundary condition is the one consistent with the open-circuit condition on the electric field and becomes exact as the thickness of the layer approaches zero, as there can be no component of current normal to the edge at the edge, and it is the normal component of the current that generates the transverse $\boldsymbol{H}$ field.) The electric field satisfying the open-circuit condition can be seen to be given by the modes

$$\boldsymbol{e}_{mn} = \hat{\boldsymbol{e}}_z \frac{\chi mn}{\sqrt{\varepsilon abt}} \cos k_n x \cos k_m y$$

where

$$k_n = \frac{n\pi}{a}$$

$$k_m = \frac{m\pi}{b}$$

$$\chi_{mn} = \begin{cases} 1, & m = 0 \quad \text{and} \quad n = 0 \\ \sqrt{2}, & m = 0 \quad \text{or} \quad n = 0 \\ 2, & m \neq 0 \quad \text{and} \quad n \neq 0 \end{cases}$$

The $\boldsymbol{H}$ field corresponding to the $\boldsymbol{E}$ field then will consist of modes

$$\boldsymbol{h}_{mn} = \frac{1}{j\omega\mu} \frac{\xi_{mn}}{\varepsilon abt} (\hat{e}_x k_m \cos k_n x \sin k_m y - \hat{e}_y k_n \sin k_n x \cos k_m y)$$

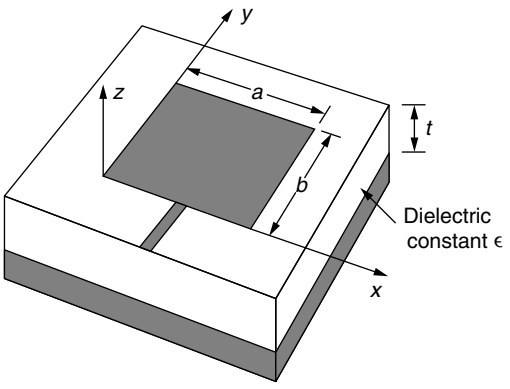


Figure 28. A patch antenna and Cartesian coordinate system.

As can be gathered from Fig. 13, the primary radiation mode is the mode with $m = 1$ and $n = 0$.

The basic operation is described by the fact that the boundary conditions are not quite exact. Recall from the earlier argument that accompanied Fig. 13 that the z-directed field gives rise to a fringe field at the edges $y = 0$ and $y = b$ such that there are strips of y-directed electric field around $y \leq 0$ and $y \geq b$. Because the boundary conditions are not quite correct on $\boldsymbol{H}$, there will also be strips of x-directed magnetic fields in these regions. As the Poynting vector is given by $\boldsymbol{E} \times \boldsymbol{H}$, we note that these strips will give rise to a z-directed Poynting vector. Similar arguments can be applied to the edges at $x = 0$ and $x = a$. However, the x-directed field at $x \leq 0$ has a change of sign at the center of the edge and is pointwise oppositely directed to the x-directed electric field at $x = 0$. These fields, therefore, only give rise to very weak radiation, as there is significant cancellation. Analysis of the slot antenna requires only that we interchange the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields.

The picture of the patch antenna as two radiating strips allows us to represent it with a transmission line as well as a circuit model. The original idea is due to Munson [19]. The transmission-line model is depicted in Fig. 29. The idea is that one feeds onto an edge with an admittance (inverse impedance) $G_1 + jB_1$ and then propagates to a second edge with admittance $G_2 + jB_2$. When the circuit is resonant, then the length of transmission line will simply complex-conjugate the given load [see Eq. (4)], leading to the circuit representation of Fig. 29b. The slot admittance used by Munson [19] was just that derived for radiation from a slit in a waveguide [20] as

$$G_1 + jB_1 = \frac{\pi a}{\lambda_0 Z_0} (1 - j0.636 \ln k_0 t)$$

where Z_0 is the impedance of free space ($\sqrt{\mu_0/\varepsilon_0} = 377\ \Omega$), λ_0 is the free-space wavelength, and k_0 is the free-space propagation vector, and where a and t are defined as in Fig. 28. When the edges are identical (as for a rectangular patch), one can write

$$G_2 + jB_2 = G_1 + jB_1$$

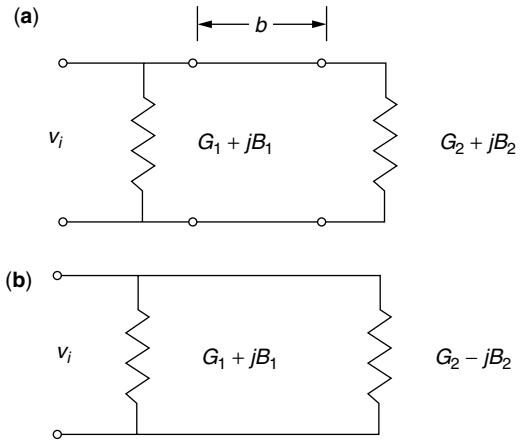


Figure 29. (**a**) A transmission-line model for a patch antenna, and (**b**) its circuital equivalent as resonance.

to obtain the input impedance in the form

$$Z_i = \frac{1}{Y_i} = \frac{1}{2G_1}$$

We have now considered all of the pieces, and therefore it is time to consider a couple of actual active antenna designs. Figure 30 depicts one of the early designs from Kai Chang's group at Texas A&M [21]. Essentially, the patch here is being used precisely as the feedback element of an amplifier circuit (as was described in connection with Fig. 9). A more compact design is that of Fig. 14 [7]. There, the transistor is actually mounted directly into the patch antenna. The slit between the gate and the drain yields a capacitive feedback element such that the effective AC circuit equivalent of this antenna may appear as depicted in Fig. 31. The capacitor–inductor pair attached to the gate lead forms what is often referred to as a *tank circuit*, which (if the load were purely real) defines a natural frequency through the relation

$$\omega = \sqrt{\frac{1}{LC}}$$

As was discussed at some length in Section 1 of this article, a major argument for the use of active antennas is that they are sufficiently compact that they can be arrayed together. Arraying is an important method for free-space power combining, which is necessary because as the frequency increases, the power-handling capability of active devices decreases. However, element size also decreases with increasing frequency so that use of multiple coherently combined elements can allow one to fix the total array size and power more or less independently of frequency, even though the number of active elements to combine increases. In the next paragraph, we shall consider some of the basics of arrays.

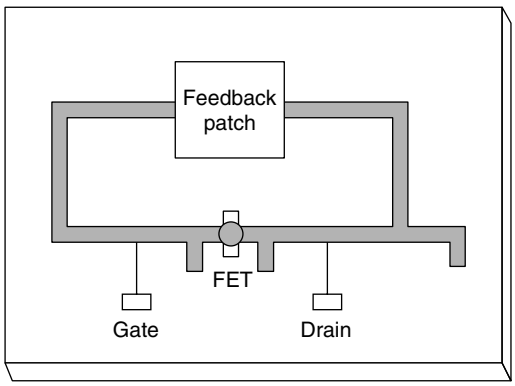


Figure 30. A design of a microstrip active radiating element.

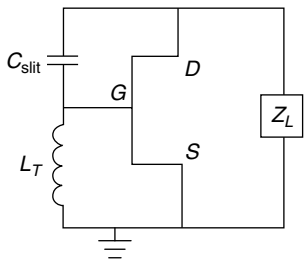


Figure 31. AC circuit equivalent of the active antenna of Fig. 14.

Consider a linear array such as is depicted in Fig. 32. Now let us say that the elements are nominally identical apart from phases that are set by the array operator at each of the elements. The complex electric field far from the nth element due to only the nth element is then given by

$$\boldsymbol{E}_n = \boldsymbol{E}_e e^{i\phi n}$$

where $\boldsymbol{E}_e$ is the electric field of a single element. To find out what is radiated in the direction θ due to the whole array, we need to sum the fields from all of the radiators, giving each radiator the proper phase delay. Each element will get a progressive phase shift $kd \sin\theta$ due to its position (see Fig. 32), where k is the free-space propagation factor, given by

$$k = \frac{2\pi}{\lambda}$$

where λ is the free-space wavelength. With this, we can write for the total field radiated into the direction θ due to all n elements

$$\boldsymbol{E}_t(\theta) = \boldsymbol{E}_e \sum_{n=0}^{N-1} e^{-inkd\sin\theta} e^{i\phi n}$$

The sum is generally referred to as the *array factor*. The intensity, then, in the θ direction is

$$\boldsymbol{I}_t(\theta) = \boldsymbol{I}_e \left| \sum_{n=0}^{N-1} e^{-iknd\sin\theta} e^{i\phi n} \right|^2$$

One notes immediately that, if one sets the phases ϕ_n to

$$\phi_n = nkd\sin\theta$$

then the intensity in the θ direction is N^2 times the intensity due to a single element. This is the effect of coherent addition. One gets a power increase of N plus a directivity increase of N. To illustrate, let us consider the

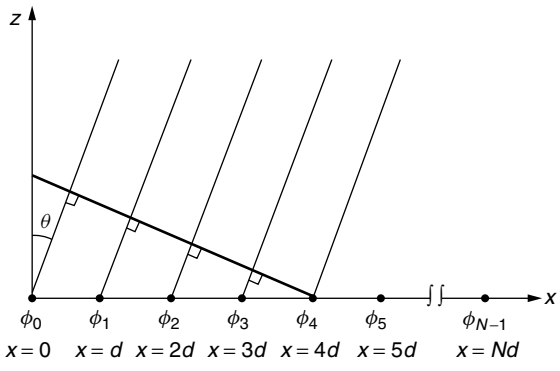


Figure 32. Depiction of a linear array of N identical radiating elements.

broadside case where we take all the ϕ_n to be zero. In this case, we can write the array factor in the form

$$\left|\sum_{n=0}^{N-1} e^{-ind\sin\theta}\right|^2 = \left|\frac{1-e^{-iNkd\sin\theta}}{1-e^{-ikd\sin\theta}}\right|^2$$

which in turn can be written as

$$\mathrm{AF} = \frac{\sin^2\left(N\dfrac{kd}{2}\sin\theta\right)}{\sin^2\left(\dfrac{kd}{2}\sin\theta\right)} \tag{6}$$

which is plotted in Fig. 33. Several interesting things can be noted from the expression and plots. For kd less than π, there is only one central lobe in the pattern. Also, the pattern becomes ever more directed with increasing N. This is called the *directivity effect*. If the array has a power-combining efficiency of 100% (which we have built into our equations by ignoring actual couplings, etc.), then the total power radiated can be only N times that of a single element. However, it is radiated into a lobe that is only $1/N$ times as wide as that of a single element.

If we are to realize array gain, however, we need to be certain that the array elements are identical in frequency and have fixed phase relations in time. This can take place only if the elements are locked together. The idea of locking is probably best understood in relation to the Van der Pol equation [16], with an injected term, such that

$$\frac{\partial^2 v}{\partial t^2} - \frac{R_{i\phi}-R_T}{L}\left(1-\frac{mv^2}{R_{i\phi}-R_T}\right)\frac{\partial v}{\partial t} + \omega_0^2 v = A\cos\omega_i t$$

where $R_{i\phi}$ is the input resistance of the transistor circuit as seen looking into the gate source port and R_T is the external termination resistor placed between the gate and common source. In the absence of the locking term, one can see that oscillation will take place with a primary frequency (and some harmonics) at angular frequency ω_0 with amplitude $\sqrt{R_{i0}-R_T/m}$ such that

$$v(t) \approx \sqrt{\frac{R_{i0}-R_T}{m}}\cos\omega_0 t$$

Without being too quantitative, one can say that, if ω_i is close enough to ω_0 and A is large enough, the oscillation will lock to ω_i in frequency and phase. If ω_i is not quite close enough and A not quite big enough (how big A needs to be is a function of how close ω_i is), then the oscillation frequency ω_0 will be shifted so that

$$v(t) = A_0\cos[(\omega_0+\Delta\omega)t+\phi]$$

where $\Delta\omega$ and ϕ are functions of ω_i and A. These ideas are discussed in a number of places including Refs. 1, 15, 16, 22, 23, and 24. In order for our array to operate in a coherent mode, the elements must be truly locked. This locking can occur through mutual coupling or through the injection of an external signal to each of the elements.

Ideally, we would like to be able to steer the locked beam. A number of techniques for doing this are presently under investigation. Much of the thinking stems from the work Stephan [25–28] and Vaughan and Compton [28a]. One of the ideas brought out in these works was that, if the array were mutually locked and one were to try to inject one of the elements with a given phase, all the elements would lock to that phase. However, if one were to inject two elements at the locked frequency but with different phases, then the other elements would have to adjust themselves to these phases. In particular, if one had a locked linear array and one were to inject the two end elements with phases differing by ϕ, then the other elements would share the phase shift equally so that there would be a linear phase taper of magnitude ϕ uniformly distributed along the array.

A different technique was developed by York [29,30], based on work he began when working with Compton [31,32]. In this technique, instead of injecting the

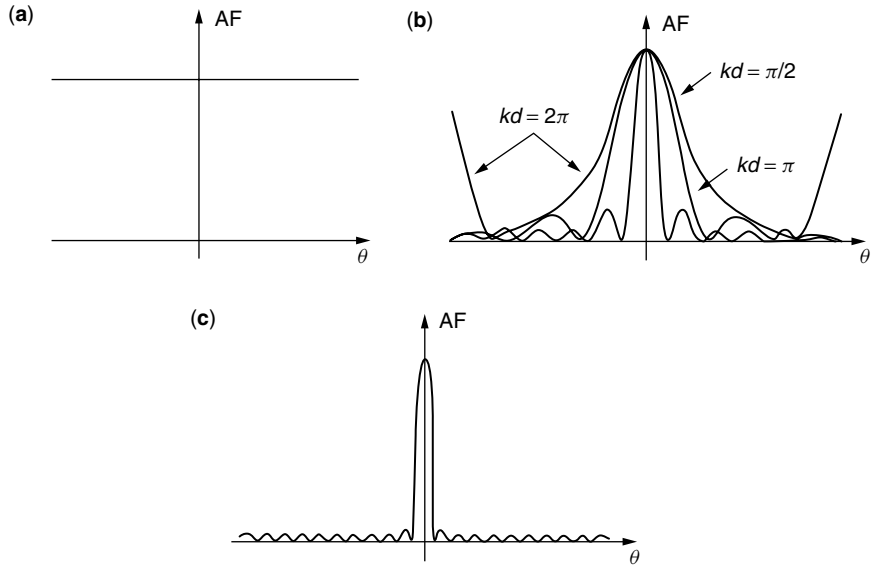


Figure 33. Plots of the array factor of Eq. (6), where (**a**) $N = 1$, (**b**) $N = 5$ and $kd = \pi/2, \pi$, and 2π, and (**c**) $N = 10$ and $kd = \pi$.

end elements with the locked frequency and different phase, one injects with wrong frequencies. If the amplitudes of these injected frequencies are set to values that are not strong enough to lock the elements to this wrong frequency, then the elements will retain their locked frequencies but will undergo phase shifts from the injected signal. If the elements of the array are locked because of mutual feedback, trying to inject either end of the array with wrong frequencies will then tend to give the elements a linear taper — that is, one in which the phase varies linearly with distance down the array — with much the same result as in the technique of Stephan. This will just linearly steer the main lobe of the array off broadside and to a new direction. Such linear scanning is what is needed for many commercial applications such as tracking or transmitting with minimum power to a given location.

Another technique, which again uses locking-type ideas, is that of changing the biases on each of the array's active devices [33–35]. Changing the bias of a transistor will alter the ω_0 at which the active antenna wants to oscillate. For an element locked to another frequency, then, changing the bias will just change the phase. In this way one can individually set the phase on each element. There are still a couple of problems with this approach (as with all the others so far, which is why this area is still one of active research). One is that addressing each bias line represents a great increase in the complexity that we were trying to minimize by using an active antenna. The other is that the maximum phase shift obtainable with this technique is $\pm\pi$ from one end of the array to the other (a limitation that is shared by the phase-shifts-at-the-ends technique). In many phased-array applications, of which electronic warfare is a typical one, one wants to have true time delay, which means that one would like to have as much as a π phase shift between adjacent elements. I do not think that the frequency shifting technique can achieve this either. Work, however, continues in this exciting area.

3. APPLICATIONS OF AND PROSPECTS FOR ACTIVE ANTENNAS

Perhaps the earliest application of the active antenna concept (following that of Hertz) was aimed at solving the small-antenna problem. As we recall, an antenna can be modeled (roughly) by a series RLC network, where the R represents the radiation resistance. The input impedance of such a combination is given by

$$Z_i = \frac{1 - \omega^2/\omega_0^2 + j\omega RC}{j\omega C}$$

and so we see that, when the operation frequency ω is well below the resonant frequency

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

and the reciprocal of the RC time constant

$$\tau = RC$$

then the antenna appears as a capacitor and radiates quite inefficiently. The problem of reception is similar. Apparently already in 1928 Westinghouse had a mobile antenna receiver that used a pentode as an inductive loading element in order to boost the amount of low-frequency radiation that could be converted to circuit current. In 1974, two works discussed transistor-based solutions to the short aerial problem [36,37]. In Ref. 37, the load circuit appeared as in Fig. 34. The idea was to generate an inductive load whose impedance varied with frequency, unlike a regular inductor, but so as to in crease the antenna bandwidth. The circuit's operation is not intuitively obvious. I think that it is possible that most AM, shortwave, and FM receivers employ some short-antenna solution regardless of whether the actual circuit designers were aware that they were employing active antenna techniques.

Another set of applications where active devices are essentially used as loading elements is in the >100-GHz regime. Reviews of progress in this regime are given in Refs. 1 and 38. To date, most work at frequencies greater than 100 GHz has involved radioastronomical receivers. A problem at such frequencies is a lack of components, including circuit elements so basic as waveguides. Microstrip guides already start having extramode problems at Ku band. Coplanar waveguides can go higher, although to date, rectangular metallic waveguides are the preferred guiding structures past about 60 GHz. In W band (normally narrowband, about 94 GHz — see Table 1), there are components, as around 94 GHz there is an atmospheric window of low propagation loss. However, waveguide tolerances, which must be a small percentage of the wavelength, are already severe in W band, where the wavelength is roughly 3 mm. Higher frequencies have to be handled in free space or, as one says, quasioptically. Receivers must therefore by nature be downconverting in this >100-GHz regime. Indeed, these types of solutions are the ones being demonstrated by the group at Michigan [38], where receivers will contain multipliers and downconverting mixers right in the antenna elements in order that CPW can be used to carry the downconverted signals to the processing electronics. Millimeter-wave-terahertz radioastronomy seems to be a prime niche for quasioptical active-antenna solutions.

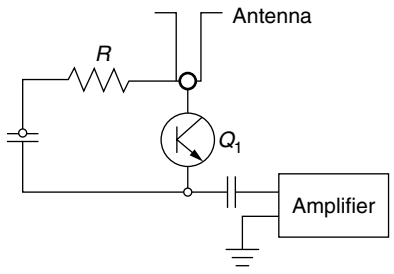


Figure 34. A circuit taken from Ref. 37 in which a transistor circuit is used to load a short antenna. Analysis shows that, in the frequency regime of interest, the loading circuit appears, when looking toward the antenna from the amplifier terminals, to cancel the strongly capacitive load of the short antenna.

The first applications of active antennas where solid-state components were used as gain elements were primarily for power boosting [39–44]. Power combining (see reviews in Refs. 45 and 46) can be hard to achieve. There is a theorem that grew out of the early days of radiometry and radiative transfer (in the 1800s), known variously as the brightness theorem, the Lagrange invariant, or (later) the second law of thermodynamics. (See, e.g., Ref. 8, Chap. 5.) The theorem essentially states that one cannot increase the brightness of a source by passive means. This theorem practically means that, if one tries to combine two nominally identical sources by taking their outputs, launching them into waveguides, and then bringing the two waveguides together in a Y junction into a single waveguide, the power in the output guide, if the output guide is no larger than either of the input guides, can be no greater than that of either of the nominally identical sources. This seems to preclude any form of power combining. There is a bit of a trick here, though. At the time the brightness theorem was first formulated, there were no coherent radiation sources. If one takes the output of a coherent radiation source, splits it in two, and adds it back together in phase, then the brightness, which was halved, can be restored. If two sources are locked, they are essentially one source. (As P. A. M. Dirac said, a photon only interferes with itself. Indeed, the quantum-mechanical meaning of locking is that the locked sources are sharing a wavefunction.) Therefore, locked sources can be coherently added if they are properly phased. We will take this up again in a following paragraph.

An alternative to power combining that obviates the need for locking and precise phase control is amplification of the signal from a single source at each element. By 1960, solid-state technology had come far enough that antennas integrated with diodes and transistors could be demonstrated. The technology was to remain a laboratory curiosity until the 1980s, when further improvements in microwave devices were to render it more practical. More recent research, however, has been more concentrated on the coherent power combining of self-oscillator elements. This is not to say that the element-mounted amplifier may not still be of practical use. The main research issue at present, though, is the limited power available from a single active element at millimeter-wave frequencies.

Another application area is that of proximity detection [47]. The idea is that an oscillator in an antenna element can be very sensitive to its nearby (several wavelengths) environment. As was discussed previously, variation in distances to ground planes changes impedances. The proximity of any metal object will, to some extent, cause the oscillator to be aware of another ground plane in parallel with the one in the circuit. This will change the impedance that the oscillator sees and thereby steer the oscillator frequency. The active antenna of Ref. 47 operated as a self-oscillating mixer. That is, the active element used the antenna as a load, whereas the antenna also used a diode mixer between itself and a low-frequency external circuit. The antenna acted as both a transmitting and a receiving antenna. If there were something moving near the antenna, the signal reflected off the object and rereceived might well be at a different frequency than the shifting oscillator frequency. These two frequencies would then beat in the mixer, be downconverted, and show up as a low-frequency beat note in the external circuit. If such a composite device were to be used in a controlled environment, one could calibrate the output to determine what is occurring. Navarro and Chang [1, p. 130] mention such applications as automatic door openers and burglar alarms. The original paper [47] seemed to have a different application in mind, as the term *Doppler sensor* was in the title. If one were to carefully control the immediate environment of the self-oscillating mixer, then reflections off more distant objects that were received by the antenna would beat with the stable frequency of the oscillator. The resulting beat note of the signals would then be the Doppler shift of the outgoing signal on reflection off the surface of the moving object, and from it one could determine the normal component of the object's velocity. It is my understanding that some low-cost radars operate on such a principle. As with other applications, though, the active-antenna principle, if only due to size constraints, becomes even more appealing at millimeter-wave frequencies, and at such frequencies power constraints favor use of arrays.

An older antenna field that seems to be going through an active renaissance is that of retroreflection. A retroreflector is a device that, when illuminated from any arbitrary direction, will return a signal directly back to the source. Clearly, retroreflectors are useful for return calibration as well as for various tracking purposes. An archetypical passive retroreflector is a corner cube. Another form of passive reflector is a Van Atta array [48]. Such an array uses wires to interconnect the array elements so that the phase progression of the incident signal is conjugated and thereby returned in the direction of the source. As was pointed out by Friis already in the 1930s, though, phase conjugation is carried out in any mixer in which the local oscillator frequency exceeds the signal frequency [49]. (A *phase conjugate* signal is one that takes on negative values at each phase point on the incoming wave.) This principle was already being exploited in 1963 for implementing retroreflection [50]. This work did not catch on, perhaps for technical reasons. A review in 1994 [51] and designs for such arrays were demonstrated and presented at the 1995 International Microwave Symposium [52,53]. Although both demonstrations used transistors and patch-type elements, both also employed circulators for isolation and therefore were not actually active array demonstrations. It would seem that retroreflection should motivate an active self-oscillating mixer solution, which will perhaps appear in the future.

As was mentioned earlier in this article, a quite important application area for active antennas is free-space power combining. As was pointed out then, a number of groups are working on developing compact elements such as those of Fig. 14 [7] and Fig. 30 [21]. As was also mentioned previously, in order to do coherent power combining, the elements must be locked. In designs where the elements are spatially packed tightly enough, proximity can lead to strong enough nearest-neighbor coupling so that the array will lock to a common frequency

and phase. Closeness of elements is also desirable in that arrays with less than $\lambda/2$ spacing will have no sidelobes sapping power from the central array beam. In designs that do not self-lock, one can inject a locking signal either on bias lines or spatially from a horn to try to lock to all elements simultaneously. Of course, the ultimate application would be for a high-bandwidth, steerable, low-cost transceiver.

Another method of carrying out power combining is to use the so-called *grid oscillator* [54,55]. The actual structure of a grid appears in Fig. 35. The operating principle of the grid is quite a bit different from that of the arrays of weakly coupled individual elements. Note that there is no ground plane at all on the back, and there is no ground plane either, per se, on the front side. Direct optical measurements of the potentials on the various lines of the grid [56], however, show that the source bias lines act somewhat like AC grounds. In this sense, either a drain bias line together with the two closest source biases, or a gate bias line together with the two horizontally adjacent bias lines, appears somewhat like CPW. The CPW lines, however, are periodically loaded ones with periodic active elements alternated with structures that appear like slot antennas. The radiating edges of the slots are, for the drain bias lines, the vertical AC connection lines between drain and drain or, for the gate bias CPW, the horizontal AC gate-to-gate connection lines. Indeed, the grid is known to lock strongly between the rows and more weakly between columns. As adjacent row elements are sharing a patch radiator, this behavior should be expected.

In a sense, this strong locking behavior of the grid is both an advantage and a disadvantage. It is advantageous that the grid is compact (element spacing can be $\leq \lambda/6$) and further that it is easy to get the rows to lock to each other. However, the compactness is also a disadvantage in that it is quite hard to get any more functionality on the grid. Much effort has been made in this area to generate functionality by stacking various grid-based active surfaces such as amplifying surfaces, varactor surfaces for frequency shifting and modulation; and doubling surfaces. A problem with stacking is, of course, diffraction as well as alignment. Alignment tolerance adds to complexity. Diffraction tends to ease alignment tolerance, but in an inelegant manner. A 100-transistor array with $\lambda/6$ spacing will have an extent of roughly 1.5λ per side. As the diffraction angle is something like the wavelength divided by the array diameter, the diffraction angle for such an array is a good fraction of a radian. One can say that grids are quasioptical, but in optics one generally doesn't use apertures much smaller than a millimeter (center optical wavelength of micrometers), for which the diffraction angle would be roughly a thousandth of a radian. As far as pure combining efficiency goes, grids are probably the optimal solution. However, more functionality may well be hard to obtain with this solution.

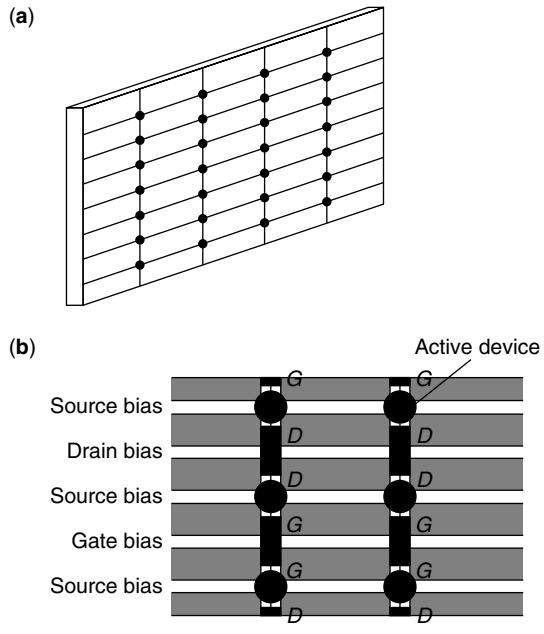


Figure 35. Schematic depiction of (**a**) the active surface of a grid oscillator and (**b**) a breakout of an internal region of the grid showing the active device placement relative to the bias lines.

As we have mentioned, there are a number of techniques for steering being investigated. There seems to be less work on modulation, and I do not know of any simultaneous steering of modulated beams to date. Although the field of active antennas began with the field of radiofrequency, it still seems to be in its infancy. However, as I hope this article has brought across, there is a significant amount of work ongoing, and the field of active antennas will grow in the future.

BIOGRAPHY

Alan Mickelson received his B.S.E.E. degree from the University of Texas, El Paso, in 1973 and his M.S. and Ph.D. degrees in electrical engineering from California Institute of Technology in 1974 and 1978, respectively. He was a National Academy of Sciences, Washington, D.C. visiting scientist at the Byurakan Astrophysical Observatory, Byurakan, Armenian S.S.R. in 1979–1980. Dr. Mickelson was a postdoctoral fellow at the Norwegian Institute of Technology, Norway, in 1980 and 1981 under a grant from the Norwegian National Science and Engineering Foundation and a staff scientist at the Electronics Laboratory of the same institute in 1982 and 1983. In 1984, he joined the faculty of the Electrical and Computer Engineering Department of the University of Colorado, Boulder. Since settling in Colorado, Professor Mickelson has continued to apply his background in electromagnetic theory and technique to a number of technological problems, including novel techniques to control microwave signal transmission and reception.

BIBLIOGRAPHY

1. J. A. Navarro and K. Chang, *Integrated Active Antennas and Spatial Power Combining*, Wiley, New York, 1995.
2. R. A. York and Z. B. Popović, eds., *Active and Quasi-Optical Arrays for Solid-State Power Combining*, Wiley, New York, 1997.

2a. H. Hertz, *Electric Waves*, Macmillan, New York, 1983. (This is a book of reprints of Hertz' work in the 1890s.)

3. J. Bardeen and W. Bratain, The transistor: A semiconductor triode, *Phys. Rev.* **74**: 435 (1948).
4. W. Shockley, The theory of $p-n$ junctions in semiconductors and $p-n$ junction transistors, *Bell Syst. Tech. J.* **28**: 435 (1949).
5. W. Shockley, A unipolar field-effect transistor, *Proc. IEEE* **40**: 1365–1376 (1952).
6. C. A. Mead, Schottky-barrier gate field-effect transistor, *Proc. IEEE* **54**: 307–308 (1966).
7. R. A. York, R. D. Martinez, and R. C. Compton, Active patch antenna element for array applications, *Electron. Lett.* **26**: 494–495 (March 1990).
8. A. R. Mickelson, *Physical Optics*, Van Nostrand Reinholt, New York, 1992, Chap. 2.
9. B. J. Hunt, *The Maxwellians*, Cornell Univ. Press, Ithaca, NY, 1991, Chap. 3.
10. H. C. Pocklington, Electrical oscillations in wires, *Proc. Cambridge Phil. Soc.* 324–333 (1897).
11. D. M. Pozar, *Microwave Engineering*, Addison-Wesley, Reading, MA, 1990.
12. P. E. Gray and C. L. Searle, *Electronic Principles, Physics, Models and Circuits*, Wiley, New York, 1967.
13. R. E. Collin, *Foundations for Microwave Engineering*, 2nd ed., McGraw-Hill, New York, 1992.
14. K. Chang, *Microwave Solid-State Circuits and Applications*, Wiley, New York, 1994.
15. K. Y. Chen et al., Analysis of an experimental technique for determining Van der Pol parameters of a transistor oscillator, *IEEE Trans. Microwave Theory Tech.* **46**: 914–922 (1998).
16. B. Van der Pol, Forced oscillations in a circuit with a nonlinear resistance, *Phil. Mag.* **3**: 65–80 (1927).
17. K. R. Carver and J. W. Mink, Microstrip antenna technology, *IEEE Trans. Antennas Propag.* **AP-29**: 2–24 (1981).
18. D. M. Pozar and D. H. Schauber, eds., *Microstrip Antennas*, IEEE Press, Piscataway, NJ, 1995.
19. R. E. Munson, Conformal microstrip antennas and microstrip phased arrays, *IEEE Trans. Antennas Propag.* **AP-22**: 74–78 (1974).
20. R. F. Harrington, *Time Harmonic Electromagnetic Fields*, McGraw-Hill, New York, 1961, p. 276.
21. K. Chang, K. A. Hammer, and G. K. Gopalakrishnan, Active radiating element using FET source integrated with microstrip patch antenna, *Electron. Lett.* **24**: 1347–1348 (1988).
22. R. Adler, A study of locking phenomena in oscillators, *Proc. IRE* **34**: 351–357 (1946).
23. R. Adler, A study of locking phenomena in oscillators, *Proc. IEEE* **61**: 1380–1385 (1973). (This is a reprint of Ref. 22.)
24. K. Kurokawa, Injection locking of microwave solid-state oscillators, *Proc. IEEE* **61**: 1386–1410 (1973).
25. K. D. Stephan, Inter-injection-locked oscillators for power combining and phased arrays, *IEEE Trans. Microwave Theory Tech.* **34**: 1017–1025 (1986).
26. K. D. Stephan and W. A. Morgan, Analysis of inter-injection-locked oscillators for integrated phased arrays, *IEEE Trans. Antennas Propag.* **35**: 771–781 (1987).
27. K. D. Stephan and S. L. Young, Mode stability of radiation-coupled inter-injection-locked oscillators for integrated phased arrays, *IEEE Trans. Microwave Theory Tech.* **36**: 921–924 (1988).
28. W. A. Morgan and K. D. Stephan, An x-band experimental model of a millimeter-wave inter-injection-locked phase array system, *IEEE Trans. Antennas Propag.* **36**: 1641–1645 (1988).

28a. M. J. Vaughan and R. C. Compton, 28 GHz omnidirectional quasioptical transmitter array, *IEEE Trans. Microwave Theory Tech.* **MTT-43**: 2507–2509 (1995).

29. R. A. York, Nonlinear analysis of phase relationships in quasioptical oscillator arrays, *IEEE Trans. Microwave Theory Tech.* **41**: 1799–1809 (1993).
30. P. Liao and R. A. York, A new phase-shifterless beam scanning technique using arrays of coupled oscillators, *IEEE Trans. Microwave Theory Tech.* **41**: 1810–1815 (1993).
31. R. A. York and R. C. Compton, Quasi-optical power combining using mutual synchronized oscillator arrays, *IEEE Trans. Microwave Theory Tech.* **39**: 1000–1009 (1991).
32. R. A. York and R. C. Compton, Coupled-oscillator arrays for millimeter-wave power-combining and mode-locking, *IEEE MTT-S Int. Microw. Symp. Digest*, 1992, pp. 429–432.
33. P. S. Hall and P. M. Haskins, Microstrip active patch array with beam scanning, *Electron. Lett.* **28**: 2056–2057 (1992).
34. P. S. Hall et al., Phase control in injection locked microstrip active antennas, *IEEE MTT-S Int. Microw. Symp. Digest*, 1994, pp. 1227–1230.
35. A. Zarrang, P. S. Hall, and M. Cryan, Active antenna phase control using subharmonic locking, *Electron. Lett.* **31**: 842–843 (1995).
36. T. S. M. Maclean and P. A. Ransdale, Short active aerials for transmission, *Int. J. Electron.* **36**: 261–269 (1974).
37. P. K. Rangole and S. S. Midha, Short antenna with active inductance, *Electron. Lett.* **10**: 462–463 (1974).
38. G. M. Rebeiz, Millimeter-wave and terahertz integrated circuit antennas, *Proc. IEEE* **80**: 1748–1770 (1996).
39. A. D. Frost, Parametric amplifier antennas, *Proc. IRE* **48**: 1163–1164 (1960).
40. J. R. Copeland and W. J. Robertson, Antenna-verters and antennafiers, *Electronics* 68–71 (1961).
41. M. E. Pedinoff, The negative conductance slot amplifier, *IRE Trans. Microwave Theory Tech.* **9**: 557–566 (1961).
42. W. J. Robertson, J. R. Copeland, and R. G. Verstraete, Antennafier arrays, *IEEE Trans. Antennas Propag.* **2**: 227–233 (1964).
43. K. Fujimoto, Active antennas: Tunnel-diode-loaded dipole, *Proc. IEEE* **53**: 174 (1964).
44. H. H. Meinke, Tunnel diodes integrated with microwave antenna systems, *Radio Electron. Eng.* **31**: 76–80 (1966).
45. K. J. Russell, Microwave power combining techniques, *IEEE Trans. Microwave Theory Tech.* **27**: 472–478 (1979).
46. K. Chang and C. Sun, Millimeter-wave power-combining techniques, *IEEE Trans. Microwave Theory Tech.* **31**: 91–107 (1983).
47. B. M. Armstrong et al., Use of microstrip impedance-measurement technique in the design of BARITT piplex Doppler sensor, *IEEE Trans. Microwave Theory Tech.* **28**: 1437–1442 (1980).

48. E. D. Sharp and M. A. Diab, Van Atta reflector array, *IRE Trans. Antennas Propag.* **8**: 436–438 (1960).

49. H. T. Friis and C. Feldman, A multiple-unit steerable antenna for short-wave reception, *Bell Syst. Tech. J.* **16**: 337–419 (1937).

50. C. Y. Pon, Retrodirective array using the heterodyne technique, *IEEE Trans. Antennas Propag.* **12**: 176–180 (1964).

51. B. S. Hewitt, The evolution of radar technology into commercial systems, *IEEE MTT-S Int. Microw. Symp. Digest*, 1994, pp. 1271–1274.

52. C. W. Poblanz and T. Itoh, A conformal retrodirective array for radar applications using a heterodyne phase scattering element, *IEEE MTT-S Int. Microw. Symp. Digest*, 1995, pp. 905–908.

53. Y. Chang, D. C. Scott, and H. R. Fetterman, Microwave phase conjugation using antenna coupled nonlinear optically pumped surface, *IEEE MTT-S Int. Microw. Symp. Digest*, 1995, pp. 1303–1306.

54. Z. B. Popovic, M. Kim, and D. B. Rutledge, Grid oscillators, *Int. J. Infrared Millimeter Waves* **9**: 647–654 (1988).

55. Z. B. Popovic et al., A 100-MESFET planar grid oscillator, *IEEE Trans. Microwave Theory Tech.* **39**: 193–200 (1991).

56. K. Y. Chen et al., Noninvasive experimental determination of charge and current distributions on an active surface, *IEEE Trans. Microwave Theory Tech.* **44**: 1000–1009 (1996).

ADAPTIVE ANTENNA ARRAYS

KYUNGJUNG KIM
TAPAN K. SARKAR
Syracuse University
Syracuse, New York

MAGDALENA SALAZAR PALMA
Universidad Politecnica de Madrid
Madrid, Spain

1. INTRODUCTION

Adaptive array signal processing has been used in many applications in such fields as radar, sonar, and wireless mobile communication. One principal advantage of an adaptive array is the ability to recover the desired signal while also automatically placing deep pattern nulls along the direction of the interference.

In conventional adaptive algorithms, the statistical approach based on forming an estimate of the covariance matrix of the received antenna voltages (measured voltages at the antenna terminals) without the signal is frequently used. However, these statistical algorithms suffer from two major drawbacks. First, they require independent identically distributed secondary data to estimate the covariance matrix of the interference. The formation of the covariance matrix is quite time-consuming, and so is the evaluation of its inverse. Unfortunately, the statistics of the interference may fluctuate rapidly over a short distance, limiting the availability of homogeneous secondary data. The resulting errors in the covariance matrix reduce the ability to suppress interference. The second drawback is that the estimation of the covariance matrix requires the storage and processing of the secondary data. This simply cannot be accomplished in real time for most applications.

Recently, a direct data domain algorithm has been proposed to overcome these drawbacks of a statistical technique [1–7]. In that approach one adaptively minimizes the interference power while maintaining the gain of the antenna array along the direction of the signal. Not having to estimate a covariance matrix leads to an enormous savings in memory and computer processor time and makes it possible to carry out an adaptive process in real time. The novelty of the proposed approach is that we analyze the antenna systems as spatial filters instead of treating them as temporal channels.

The use of real antenna elements and not omnidirectional point sources in an actual antenna array will also require an investigation into the capabilities of the direct data domain algorithms to perform adaptivity in nonideal situations such as in the presence of mutual coupling between the elements of the array, near-field scatterers, and obstacles located close to the array. This could also involve the various platform effects on which the antenna array is mounted.

Most adaptive algorithms assume that the elements of the receiving array are independent isotropic omnidirectional point sensors that do not reradiate the incident electromagnetic energy. It is further assumed that the array is isolated from its surroundings. However, in a practical case, array elements have a finite physical size and reradiate the incident fields. The reradiated fields interact with the other elements, causing the antennas to be mutually coupled. Adve and Sarkar [7] observed the degradation in the capabilities of direct data domain algorithms and suggested ways to improve it under some circumstances.

Gupta and Ksienski [8] and Pasala and Friel [9] compensate for the effects of mutual coupling by relating the open-circuit voltages (voltages at the ports of the array as if all were open-circuited) with the voltages measured at the ports in an adaptive antenna array used for direction of arrival (DoA) estimation. Adve and Sarkar [7] used the method of moments (MOM) to analyze the antenna array in which the entries of the MOM impedance matrix measure the interaction between the basis functions; that is, they quantize the mutual coupling. In these works the compensation matrix is in general considered to be independent of the angle of arrival of the signals. However, in a more practical environment the presence of near field scatterers (i.e., buildings the structure on which the array is mounted) will have effects on the array elements. The effects of these near field elements are similar to the effects of mutual coupling between the elements of the array. These environmental scatterers necessitate the development of a compensation matrix, which depends on the direction of arrival of signals including the undesired ones. In this article, we shall use the measured steering vector in an interpolation technique, which is contaminated by the presence of near field scatters as well as by the mutual coupling between

the elements of the real array, to obtain the compensation matrix for a more accurate numerical analysis.

This presentation is divided into two distinct parts. In the first part we use the electromagnetic analysis along with an interpolation algorithm to transform the voltages that are measured or computed in a real array containing realistic antenna elements receiving signals in the presence of near field scatterers to a uniform linear virtual array (ULVA) consisting of isotropic omnidirectional point radiators. In this way we take into account not only the effects of mutual coupling and the near-field scattering effects of the antenna array produced by the signal of interest but also those due to the strong coherent interferers whose directions of arrival are unknown. The first stage preprocesses the voltages induced in the real elements and transforms them to a set of voltages that would be induced in an ULVA of isotropic omnidirectional point radiators radiating in free space.

During the second phase of the processing, these transformed voltages induced in the ULVA are processed by a direct data domain least-squares method. In this phase, the goal is to estimate the complex signal amplitude given the direction of arrival in a least-squares fashion when the signal of interest (SoI) is contaminated by strong interferers that may come through the mainlobe of the array, clutter and thermal noise. The advantage of this methodology is that no statistical description of the environment is necessary, and since we are processing the data on a snapshot-by-snapshot basis, this new technique can be applied to a highly dynamic environment.

The article is organized as follows. In Section 2 we formulate the problem. In Section 3 we present the transformation technique incorporating mutual coupling effects between the array elements and near field scatterers. Section 4 describes the direct data domain least-squares approach. In Section 5 we present simulation results illustrating the performance of the proposed method in a real environment. Finally, in Section 6 we present the conclusion.

2. PROBLEM FORMULATION

Consider an array composed of N sensors separated by a distance d as shown in Fig. 1. We assume that narrowband signals consisting of the desired signal plus possibly coherent multipaths and jammers with center frequency f_0 are impinging on the array from various angles θ, with the constraint $0 \leq \theta \leq 180^\circ$. For sake of simplicity we assume that the incident fields are coplanar and that they are located in the far field of the array. However, this methodology can easily be extended to the noncoplanar case without any problem including the added polarization diversity.

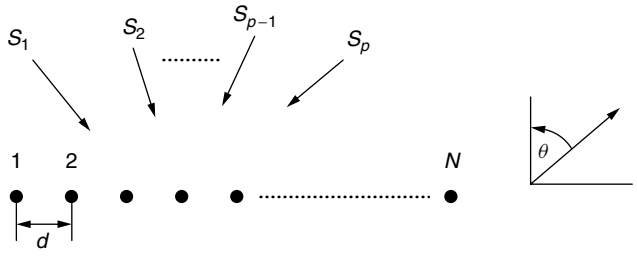


Figure 1. A linear uniform array.

Using the complex envelope representation, the $N \times 1$ complex vectors of phasor voltages $[\mathbf{x}(t)]$ received by the antenna elements at a single time instance t can be expressed by

$$[\mathbf{x}(t)] = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_N(t) \end{bmatrix} = \sum_{k=1}^{P} [\mathbf{a}(\theta_k)] s_k(t) + [n(t)] \tag{1}$$

where $s_k(t)$ denotes the incident signal from the kth source directed towards the array at the instance t, and P stands for the number of sources, $[\mathbf{a}(\theta)]$ denotes the steering vector of the array toward direction θ, and $[n(t)]$ denotes the noise vector at each of the antenna elements. It is important to note that the array elements can be dissimilar and they can be noncoplanar and may even be nonuniformly spaced. Here, the angle θ is measured from the broadside direction as shown in Fig. 1. We now analyze the data using a single snapshot of voltages measured at the antenna terminals.

Using a matrix notation, (1) becomes

$$[\mathbf{x}(t)] = [\mathbf{A}(\theta)][s(t)] + [n(t)] \tag{2}$$

where $[\mathbf{A}(\theta)]$ is the $N \times p$ matrix of the steering vectors, referred to as the array manifold

$$[\mathbf{A}(\theta)] = [a(\theta_1), a(\theta_2), \ldots, a(\theta_p)] \tag{3}$$

In a typical calibration methodology, a far-field source $s_k(t)$ is placed along the directions θ_k and then $[\mathbf{x}(t)]$ is the voltage measured at the feed point of the antenna elements in the array. Here $[s(t)]$ is a $p \times 1$ vector representing the various signals incident on the array at time instance t. In practice, this array manifold for a real array is contaminated by both effects of nonuniformity in the individual elements, and the interelement spacing may also be nonuniform to achieve greater aperture efficiency. Furthermore, there are mutual couplings between the antenna elements in the array, which undermine the performance of any conventional adaptive signal processing algorithm.

Hence, our problem can be stated as follows. Given the sampled data vector snapshot $[\mathbf{x}(t)]$ at a specific instance of time, how do we recover the desired signal arriving from a given look direction while simultaneously rejecting all other interferences that may be coherent? Most signal processing techniques are based on the fact that a far-field source presents a linear phase front at the elements of the antenna array. However, we shall demonstrate that the nonuniformity of a real array and the presence of mutual coupling between the elements of the real array and scatterers located close to the array undermines the ability of any adaptive algorithm to maintain the gain of the array along the direction of the signal while

simultaneously rejecting the interferences. To compensate for the lack of nonuniformity of the real array and mutual coupling effects, we propose an interpolation technique based on the method of least squares that incorporates all the electromagnetic coupling effects as outlined in Section 3. With appropriate preprocessing using Maxwell's equations, any adaptive technique can be applied to real antenna arrays located in any arbitrary environment. However, the use of a direct data domain least-squares procedure makes it possible to implement the algorithm in hardware, and the solution can be obtained in almost real time.

3. AN ARRAY TRANSFORMATION TECHNIQUE USING LEAST SQUARES THAT ACCOUNTS FOR ALL THE ELECTROMAGNETIC EFFECTS SUCH AS MUTUAL COUPLING AND PRESENCE OF NEAR FIELD SCATTERERS

For the first step of this adaptive method we transform the voltages that are induced in the actual antenna elements operating in any environment to a set of equivalent voltages that would be induced in a ULVA consisting of omnidirectional point radiators located in free space. The presence of mutual coupling between the antenna elements and existence of near field scatterers also disturb the capability of any algorithm to maintain the gain of the array along the direction of the signal while simultaneously rejecting the strong time varying coherent interferences. Hence, we need to preprocess the data to account for these undesired electromagnetic effects.

The preprocessing is to compensate for the lack of nonuniformity in a real array contaminated by the mutual coupling effects between the various elements. The methodology is similar to the one described by Friedlander [10–12]. The procedure is based on transforming the nonuniformly spaced array into a uniform linear virtual array (ULVA) consisting of isotropic omnidirectional point radiators operating in vacuum through the use of a transformation matrix. Our basic assumption is that electrical characteristics of the array corresponding to the ULVA can be obtained through an interpolation of the real array, which is disturbed by various undesired electromagnetic couplings. The goal is to select the best-fit transformation $[T]$ between the real array manifold $[A(\theta)]$ and the array manifold corresponding to a uniform linear virtual array (ULVA) consisting of isotropic omnidirectional point radiators $[\overline{A}(\theta)]$ such that $[T][A(\theta)] = [\overline{A}(\theta)]$ for all possible angles θ within a predefined sector. In this way we not only compensate for the various electromagnetic effects associated with the SoI but also correct for the interactions associated with coherent strong interferers whose direction of arrival we do not know. Since such a transformation matrix is defined within a predefined sector, the various undesired electromagnetic effects such as nonuniformity in spacing and mutual coupling between the elements and presence of near field obstacles for an array is made independent of the angular dependence.

The following is a step-by-step description of what needs to be done to obtain the transformation matrix $[T]$ that will transform the real array manifold that is disturbed by various undesired electromagnetic effects such as mutual coupling and various near-field effects to that of a ULVA:

1. The first step in designing the ULVA is to divide the field of view of the array into Q sectors. If the field of view is 180°, it can be divided into 6 sectors of 30° each. Then, each of the Q sectors is defined by the interval $[\theta_q, \theta_{q+1}]$, for $q = 1, 2, \ldots, Q$. Or equivalently, only one sector of 180° extent can also be used. In that case $Q = 1$.

2. Next we define a set of uniformly defined angles to cover each sector:

$$\mathbf{\Theta}_q = [\theta_q, \theta_q + \Delta, \theta_q + 2\Delta, \ldots, \theta_{q+1}] \tag{4}$$

where Δ is the angular step size.

3. We measure/compute the steering vectors associated with the set $\mathbf{\Theta}_q$ for the real array. This is done by placing a signal in the far field for each angle of arrival θ_q, $\theta_q + \Delta, \theta_q + 2\Delta, \ldots, \theta_{q+1}$. The measured/computed vector is different from the ideal steering vector, which is devoid of any undesired electromagnetic effects such as the presence of the mutual coupling between the nonuniformly spaced elements and other near-field coupling effects. Then, we obtain either through measurement or by using an electromagnetic analysis tool such as WIPL-D [13], to obtain the measured voltages at the antenna elements from

$$[\mathbf{A}_q(\mathbf{\Theta}_q)] = [a(\theta_q), a(\theta_q + \Delta) \cdots a(\theta_{q+1})] \tag{5}$$

This can be either actually measured or simulated and includes all the undesired electromagnetic coupling effects. Hence, each column of $[\mathbf{A}_q(\mathbf{\Theta}_q)]$ represents the relative signal strength received at each of the antenna elements for an incident signal arriving the angular direction θ_q. The elements of the matrix are a function of only the incident angle of an incoming plane wave within that predefined sector.

4. Next we fix the virtual elements of the interpolated array. We always assume that the ULVA consists of omnidirectional isotropic sources radiating in free space. We denote by the section of the array manifold of the virtual array obtained for the set of angles $\mathbf{\Theta}_q$:

$$[\overline{\mathbf{A}}_q(\mathbf{\Theta}_q)] = [\overline{a}(\theta_q), \overline{a}(\theta_q + \Delta), \ldots, \overline{a}(\theta_{q+1})] \tag{6}$$

where $[\overline{\mathbf{a}}(\theta)]$ is a set of theoretical steering vectors corresponding to the uniformly spaced linear array.

5. Now we evaluate the transformation matrix $[\mathbf{T}_q]$ for the sector q such that $[\mathbf{T}_q][\mathbf{A}(\Theta_q)] = [\overline{\mathbf{A}}(\Theta_q)]$ using the least-squares method. This is achieved by minimizing the functional

$$\min_{T_q} \|[\overline{\mathbf{A}}_q] - [\mathbf{T}_q][\mathbf{A}_q]\| \tag{7}$$

In order to have a unique solution for (7), the number of direction vectors in a given sector must be greater than or equal to the number of the elements of array. The least square solution to (7) is given by [14]

$$[\mathbf{T}] = [\overline{\mathbf{A}}(\Theta_q)][\mathbf{A}(\Theta_q)]^H\{[(\mathbf{A}(\Theta_q)][\mathbf{A}(\Theta_q)]^H\}^{-1} \tag{8}$$

where the superscript H represents the conjugate transpose of a complex matrix. Computationally it is more efficient and accurate to carry out the solution of (7)

through the use of the total least squares implemented through the singular value decomposition [15]. The transformation matrix needs to be computed only once *a priori* for each sector and the computation can be done offline. Hence, once [**T**] is known, we can compensate for the various undesired electromagnetic effects such as mutual coupling between the antenna elements, including the effects of near-field scatterers, as well as nonuniformity in the spacing of the elements in the real array simultaneously. The transformation matrix [**T**] is thus characterized within the predefined angle. However, if there is only one sector, specifically, $Q = 1$, then there will be only one transformation matrix [**T**].

6. Finally, using (8), one can obtain the corrected input voltages in which all the undesired electromagnetic effects are accounted for and the measured snapshot of the voltages are transformed to that which will be obtained for a ULVA. Let that set be denoted by $[\mathbf{x}_c(t)]$. Its value can be obtained through

$$[\mathbf{x}_c(t)] = [\mathbf{T}][\mathbf{x}(t)] \tag{9}$$

Once (9) is obtained, we can apply the direct data domain algorithms to the preprocessed corrected voltages $[\mathbf{x}_c(t)]$ without any significant loss of accuracy.

Next a direct data domain least-squares algorithm is applied to the processed voltage sets $[\mathbf{x}_c(t)]$ to obtain the complex amplitude corresponding to the signal of interest in a least-squares fashion.

4. THE DIRECT DATA DOMAIN LEAST-SQUARES PROCEDURE

Let us assume that the signal of interest (SoI) is coming from the angular direction θ_d and that our objective is to estimate its complex amplitude while simultaneously rejecting all other interferences. The signal arrives at each antenna at different times dependent on the direction of arrival of the SoI and the geometry of the array. We make the narrowband assumption for all the signals including the interferers. At each of the N antenna elements, the received signal given by (9) is a sum of the SoI, interference, and thermal noise. The interference may consist of coherent multipaths of SoI along with clutter and thermal noise. Here we model clutter as a bunch of electromagnetic waves coming through an angular sector. Hence, this model of clutter does not require any statistical characterization of the environment [1–7]. Therefore, we can reformulate (9) as

$$[\mathbf{x}(t)] = [\mathbf{s}_d] + \sum_{p=1}^{P-1} s_p\, a(\theta_p) + [\mathbf{n}(t)] \tag{10}$$

where s_p and θ_p are the amplitude and direction of arrival of the pth interference, respectively, and s_d is the SoI. If θ_d is the assumed DoA of the SoI, then we can represent the received voltage solely due to the desired signal at the kth sensor element as

$$s_d = s_d(t)e^{j\psi(\theta_d)} \tag{11}$$

The strength of the SoI, $s_d(t)$, is the desired unknown parameter that will be estimated for the given snapshot at the time instance t. $\psi(\theta_d)$ does not provide a linear phase regression along the elements of the real array, when the elements deviate from isotropic omnidirectional point sensors. This deviation from phase linearity undermines the capabilities of the various signal processing algorithms. For a conventional adaptive array system, we can now estimate the SoI by a weighted sum given by

$$y(t) = \sum_{k=1}^{K} w_k x_k(t) \tag{12}$$

or in a compact matrix form as

$$[\mathbf{y}(t)] = [\mathbf{W}]^T[\mathbf{X}] = [\mathbf{X}]^T[\mathbf{W}] \tag{13}$$

where the superscript T denotes the transpose of a matrix. The two vectors [**W**] and [**X**] are given by

$$[\mathbf{W}]^T = [w_1, w_2, \ldots, w_K] \tag{14}$$

$$[\mathbf{X}]^T = [x_1, x_2, \ldots, x_K] \tag{15}$$

Let [**V**] be a matrix whose elements consist of the complex voltages measured at a single time instance t at all the N elements of the array simultaneously. The received signals may also be contaminated by thermal noise. Let us define another matrix [**S**] whose elements comprise of the complex voltages received at the antenna elements of the ULVA due to a signal of unity amplitude coming from the desired direction θ_d. However, the actual complex amplitude of the signal is α, which is to be determined. Then if we form the matrix pencil using these two matrices, we have

$$[\mathbf{V}] - \alpha[\mathbf{S}] \tag{16}$$

where

$$[\mathbf{V}] = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & & & \\ x_K & x_{K+1} & \cdots & x_N \end{bmatrix}_{K\times K} \tag{17}$$

$$[\mathbf{S}] = \begin{bmatrix} s_{d1} & s_{d2} & \cdots & s_{dK} \\ s_{d2} & s_{d3} & \cdots & s_{dK+1} \\ \vdots & & & \\ s_{dK} & s_{dK+1} & \cdots & s_{dN} \end{bmatrix}_{K\times K} \tag{18}$$

represent only the undesired signal components. The difference between each element of $\{[\mathbf{V}]-\alpha[\mathbf{S}]\}$ represents the contribution of all the undesired signals due to coherent mutipaths, interferences, clutter, and thermal noise (i.e., all undesired components except the signal). It is assumed that there are K equivalent interferers, and so the number of degrees of freedom is $K = (N+1)/2$. One could form the undesired noise power from (16) and estimate a value of α by using a set of weights [**W**], which minimizes the noise power. This results in [1–7]

$$([\mathbf{V}] - \alpha[\mathbf{S}])[\mathbf{W}] = [\mathbf{0}] \tag{19}$$

Alternately, one can view the left-hand side of (19) as the total noise signal at the output of the adaptive processor due to interferences and thermal noise:

$$[\mathbf{N}_{\text{out}}] = [\mathbf{R}][\mathbf{W}] = \{[\mathbf{V}] - \alpha[\mathbf{S}]\}[\mathbf{W}] \tag{20}$$

Hence, the total undesired power would be given by

$$[\mathbf{P}_{\text{undesired}}] = [\mathbf{W}]^H\{[\mathbf{V}] - \alpha[\mathbf{S}]\}^H\{[\mathbf{V}] - \alpha[\mathbf{S}]\}[\mathbf{W}] \tag{21}$$

where the superscript H denotes the conjugate transpose of a matrix. Our objective is to set the undesired power to a minimum by selecting [**W**] for a fixed signal strength α. This yields the generalized eigenvalue equation given by (19). Therefore

$$[\mathbf{V}][\mathbf{W}] = \alpha[\mathbf{S}][\mathbf{W}] \tag{22}$$

where α, the strength of the signal, is given by the generalized eigenvalue and the weights [**W**] are given by the generalized eigenvector. Even though (22) represents a $K \times K$ matrix, the matrix [**S**] is only of rank 1. Hence, (22) has only one eigenvalue, and that generalized eigenvalue is the solution for the SoI.

For real-time applications, it may be computationally difficult to solve the reduced-rank generalized eigenvalue problem in an efficient way, particularly if the dimension K, i.e., the number of weights is large. For this reason we convert the solution of a nonlinear eigenvalue problem in (22) to the solution of a linear matrix equation.

We observe that the first and the second elements of the matrix [**R**] in (20) is given by

$$R(1) = x_1 - \alpha s_{d1} \tag{23}$$

$$R(2) = x_2 - \alpha s_{d2} \tag{24}$$

where x_1 and x_2 are the voltages received at the antenna elements 1 and 2 due to the signal, interferences and thermal noise, whereas s_{d1} and s_{d2} are the values of the signals only at the same elements due to an assumed incident signal of unit strength.

Define

$$Z = \exp\left[j2\pi\frac{d}{\lambda}\sin\theta_d\right] \tag{25}$$

where θ_d is the angle of arrival corresponding to the desired signals. Then $R(1) - Z^{-1}R(2)$ contains no components of the signal as

$$s_{d1} = \exp\left[j2\pi\frac{id}{\lambda}\sin\theta_d\right] \quad \text{with } i = 1 \tag{26}$$

$$s_{d2} = \exp\left[j2\pi\frac{id}{\lambda}\sin\theta_d\right] \quad \text{with } i = 2 \tag{27}$$

Therefore one can form a reduced-rank matrix $[\mathbf{U}]_{(K-1)\times K}$ generated from [**R**] such that

$$[\mathbf{U}] = \begin{bmatrix} X_1 - Z^{-1}X_2 & X_2 - Z^{-1}X_3 & \cdots & X_K - Z^{-1}X_{K+1} \\ \vdots & \vdots & & \\ X_{K-1} - Z^{-1}X_K & X_K - Z^{-1}X_{K+1} & \cdots & X_{N-1} - Z^{-1}X_N \end{bmatrix}_{(K-1)\times K} = [0] \tag{28}$$

In order to make the matrix full rank, we fix the gain of the subarray by forming a weighted sum of the voltages $\sum_{i=1}^{K} W_iX_i$ along the direction of arrival of the SoI. Let us say that the gain of the subarray is C in the direction of θ_d. This provides an additional equation resulting in

$$\begin{bmatrix} 1 & \cdots & Z^{K-1} \\ X_1 - Z^{-1}X_2 & \cdots & X_K - Z^{-1}X_{K+1} \\ \vdots & \vdots & \vdots \\ X_{K-1} - Z^{-1}X_K & \cdots & X_{N-1} - Z^{-1}X_N \end{bmatrix}_{K\times K} \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_K \end{bmatrix}_{K\times 1} = \begin{bmatrix} C \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{K\times 1} \tag{29}$$

or, equivalently,

$$[\mathbf{F}][\mathbf{W}] = [\mathbf{Y}] \tag{30}$$

Once the weights are solved by using (30), the signal component α may be estimated from

$$\alpha = \frac{1}{C}\sum_{i=1}^{K} W_iX_i \tag{31}$$

The proof of (29–31) is available in Ref. 1. As noted in that article [1], (29) can be solved very efficiently by applying the FFT and the conjugate gradient method, which may be implemented to operate in real time utilizing, for example, a DSP32C signal processing chip [16].

So for the solution of $[\mathbf{F}][\mathbf{W}] = [\mathbf{Y}]$ in (30), the conjugate gradient method starts with an initial guess $[W]_0$ for the solution and lets [16]

$$[\mathbf{P}]_0 = -b_{-1}[\mathbf{F}]^H[\mathbf{R}]_0 = -b_{-1}[\mathbf{F}]^H\{[\mathbf{F}][\mathbf{W}]_0 - [\mathbf{Y}]\}, \tag{32}$$

At the nth iteration the conjugate gradient method develops

$$t_n = \frac{1}{\|[\mathbf{F}][\mathbf{P}]_n\|^2} \tag{33}$$

$$[\mathbf{W}]_{n+1} = [\mathbf{W}]_n + t_n[\mathbf{P}]_n \tag{34}$$

$$[\mathbf{R}]_{n+1} = [\mathbf{R}]_n + t_n[\mathbf{Z}][\mathbf{P}]_n \tag{35}$$

$$b_n = \frac{1}{\|[\mathbf{F}]^H[\mathbf{R}]_{n+1}\|^2} \tag{36}$$

$$[P]_{n+1} = [\mathbf{P}]_n - b_n[\mathbf{F}]^H[\mathbf{R}]_{n+1} \tag{37}$$

The norm is defined by

$$\|[\mathbf{F}][\mathbf{P}]_n\|^2 = [\mathbf{P}]_n^H[\mathbf{F}]^H[\mathbf{F}][\mathbf{P}]_n \tag{38}$$

These iterative procedures continue until the error criterion is satisfied. In our case, the error criterion is defined by

$$\frac{\|[\mathbf{F}][\mathbf{W}]_n - [\mathbf{Y}]\|}{\|\mathbf{Y}\|} \le \sigma \tag{39}$$

where σ denotes the number of effective bits of data associated with the measured voltages. Hence, the iteration is stopped when the normalized residuals are of the same order as the error in the data. The computational bottleneck in the application of the conjugate gradient method is to carry out the various matrix vector products. That is where the FFT comes in as the equations involved have a Hankel structure and therefore use of the FFT reduces the computational complexity by an order of magnitude without sacrificing accuracy [17].

The advantage of using the conjugate gradient method is that the iterative solution procedure will converge even if the matrix [**F**] is exactly singular. Hence, it can be used for real time implementations. Also, the number of iterations taken by the conjugate gradient method to converge to the solution is dictated by the number of independent eigenvalues of the matrix [**F**]. This often translates into the number of dominant signal components in the data. So, the conjugate gradient method has the advantage of a direct method as it is guaranteed to converge after a finite number of steps and that of an iterative method as the roundoff and the truncation errors in the computations are limited only to the last stage of iteration.

Next we illustrate how to increase the degrees of freedom associated with (30). It is well known in the parametric spectral estimation literature that a sampled sequence can be estimated by observing it in either the forward or reverse direction [1–7]. This we term as the backward model as opposed to the forward model just outlined. If we now conjugate the data and form the reverse sequence, then we get an equation similar to (29) for the solution of weights W_m:

$$\begin{bmatrix} 1 & Z & \cdots & Z^{K-1} \\ X_N^* - Z^{-1}X_{N-1}^* & X_{N-1}^* - Z^{-1}X_{N-2}^* & \cdots & X_K^* - Z^{-1}X_{K+1}^* \\ \vdots & \vdots & & \\ X_{K+1}^* - X_K^* & X_K^* - X_{K-1}^* & \cdots & X_2^* - Z^{-1}X_1^* \end{bmatrix}_{K\times K} \times \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_K \end{bmatrix}_{K\times 1} = \begin{bmatrix} C \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{K\times 1} \tag{40}$$

where $*$ denotes the complex conjugate, or equivalently

$$[\mathbf{B}][\mathbf{W}] = [\mathbf{Y}] \tag{41}$$

The signal strength α can again be determined by (31), once (40) is solved for the weights. C is the gain of the antenna array along the direction of the arrival of the signal. Note that in both cases of equations (30) and (41) [**F**] and [**B**], $K = (N+1)/2$, where N is the total number of antenna elements. So we now increase the number of weights significantly by combining the forward–backward model. In this way we double the amount of data by not only considering the data in the forward direction but also conjugating it and reversing the increment direction of the independent variable. This type of processing can be done as long as the series to be approximated can be fit by exponential functions of purely imaginary argument. This is always true for the adaptive array case. There is an additional benefit. For both the forward and the backward methods, the maximum number of weights we can consider is given by $(N-1)/2$, where N is the number of antenna elements. Hence, even though all the antenna elements are being utilized in the processing, the number of degrees of freedom available is essentially half that of the number of antenna elements. For the forward–backward method, the number of degrees of freedom can be significantly increased without increasing the number of antenna elements. This is accomplished by considering the forward–backward version of the array data. For this case, the number of degrees of freedom M can reach $(N+0.5)/1.5$. This is approximately equal to 50% more weights or numbers of degrees of freedom than the two previous cases. The equation that needs to be solved for the weights in the combined forward–backward model is obtained by combining (29) and (40) into

$$\begin{bmatrix} 1 & Z & \cdots & Z^{M-1} \\ X_1 - Z^{-1}X_2 & X_2 - Z^{-1}X_3 & \cdots & X_M - Z^{-1}X_{M+1} \\ \vdots & \vdots & \vdots & \\ X_{M-1} - Z^{-1}X_M & X_M - Z^{-1}X_{M+1} & \cdots & X_{M-1} - Z^{-1}X_N \\ X_N^* - Z^{-1}X_{N-1}^* & X_{N-1}^* - Z^{-1}X_{N-2}^* & \cdots & X_M^* - Z^{-1}X_{M+1}^* \\ \vdots & \vdots & & \\ X_{M+1}^* - X_M^* & X_M^* - X_{M-1}^* & \cdots & X_2^* - Z^{-1}X_1^* \end{bmatrix}_{M\times M} \times \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix}_{M\times 1} = \begin{bmatrix} C \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{M\times 1} \tag{42}$$

or, equivalently,

$$[\mathbf{U}][\mathbf{W}] = [\mathbf{Y}] \tag{43}$$

Once the increased degrees of freedom are used to compute the weights the complex amplitude for the signal of interest is determined from Eq. (31), where in the summation N is replaced by the new degrees of freedom M. Also as before the matrix [**U**] now has a block Hankel structure.

This illustrates how the direct data domain least-squares approach can be implemented in real time by using single snapshots of data.

5. NUMERICAL EXAMPLES

In this section we illustrate the principles described above through some numerical simulations.

5.1. Application in Absence of Mutual Coupling

As a first example consider a signal of unit amplitude arriving from $\theta = 0^\circ$. We consider a 17-element array of element spacing of $\lambda/2$ as shown in Fig. 1. The magnitude of the incident signal is varied from 1 to 10.0 V/m in steps of 0.1 V/m while maintaining the jammer intensities constant, which are arriving from -50°, -30°, 20°. The signal-to-thermal noise ratio at each antenna element is set at 20 dB. All signal intensities and directions of arrival are summarized in Table 1.

Table 1. Parameters of the Incident Signals

	Magnitude(V/m)	Phase	DoA
Signal	1.0–10.0	0.0	0°
Jammer	1.0	0.0	−50°
Jammer	1.0	0.0	−30°
Jammer	1.0	0.0	20°

Here, we assume that we know the direction of arrival of the signal but need to estimate its complex amplitude.

If the jammers have been nulled correctly and the signal recovered properly, it is expected that the recovered signal will have a linear relationship with respect to the intensity of the incident signal. Figure 2 plots the results of using the direct data domain approach presented in Section 3. The magnitude and the phase are shown. As can be seen, the magnitude displays the expected linear relationship, and the phase varies within very small values. The beam pattern associated with this example is shown in Fig. 3. Here, we set the magnitude of the desired signal to be 1 V/m and the other parameters are as given in Table 1. The nulls are deep and occur along the correct directions.

For the second example, the intensity of the jammer signal is varied from 1 to 1000 V/m in 10-V/m increments while the intensity of the desired signal is fixed at 1 V/m. All signal intensities and directions of arrival are summarized in Table 2.

The signal-to-thermal noise ratio at each antenna element is set at 20 dB.

Figure 4 shows the results of using the direct data domain approach. The fluctuations of the magnitude and phase of the estimated signal are very small even when there is a very strong interference.

The beam pattern associated with this adaptive system when the field strength of the strong jammer is 500 V/m is shown in Fig. 5. This demonstrates that the strong jammer

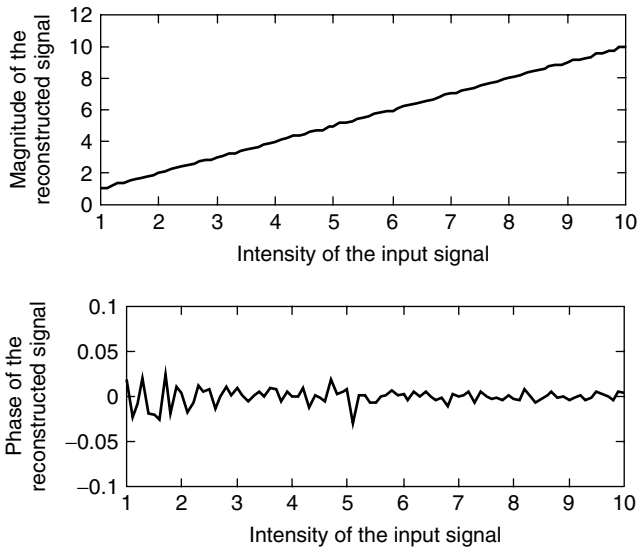


Figure 2. Estimation of the signal of interest (SoI) in the presence of jammers and thermal noise.

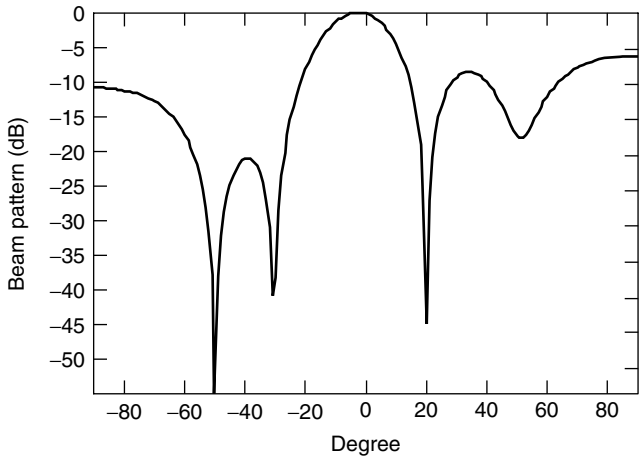


Figure 3. Adapted beam pattern in the presence of the jammers and thermal noise.

Table 2. Parameters for the SoI and Interference

	Magnitude(V/m)	Phase	DoA
Signal	1.0	0.0	0°
Jammer	1	0.0	−50°
Jammer	1	0.0	−30°
Jammer	1–1000	0.0	20°

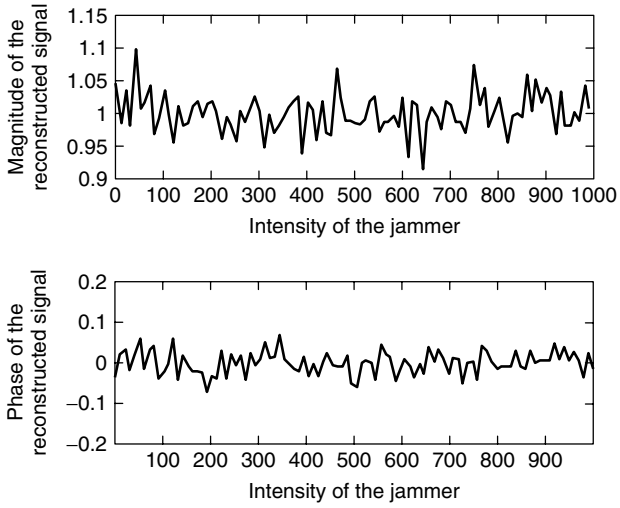


Figure 4. Estimation of the reconstructed signal in the presence of the strong jammer.

has been suppressed enough so as to recover the proper amplitude of the signal.

5.2. Application in Presence of Mutual Coupling

Next we consider a semi circular array consisting of half-wave dipoles as shown in Fig. 6. It consists of 24 half-wave thin-wire dipole antenna elements. Each element is identically loaded at the center by 50 Ω. The radius of the semicircular array is 3.82 wavelength. The dipoles are z-directed, of length $L = \lambda/2$ and radius $r = \lambda/200$, where λ is the wavelength of operation. The details of the semicircular are presented in Table 3.

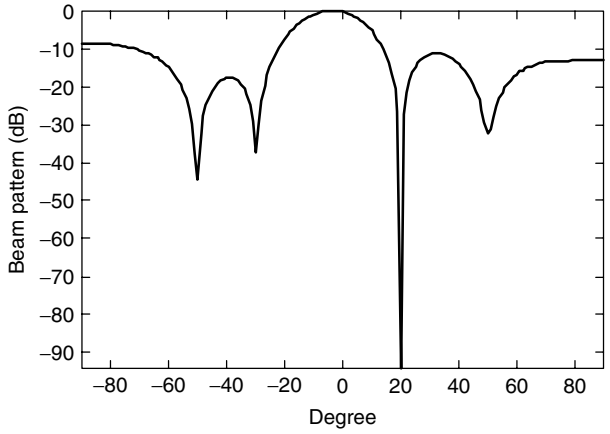


Figure 5. Adapted Beam pattern in the presence of the jammers.

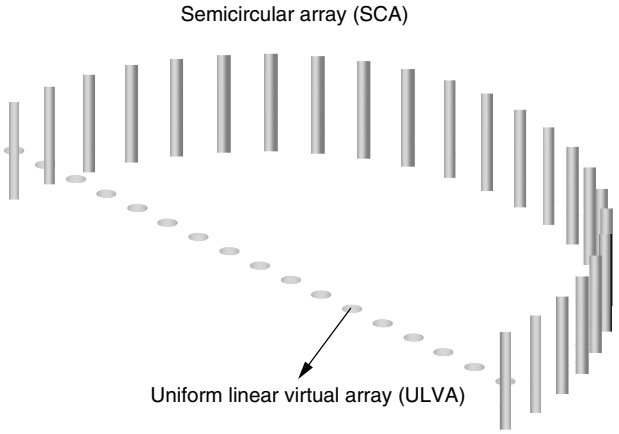


Figure 6. A semicircular array.

Table 3. Physical Sizes of the Elements for the Semicircular Array

Number of elements in semicircular array	24
Length of z-directed wires	$\lambda/2$
Radius of wires	$\lambda/200$
Loading at the center	50Ω

Then, as described in Section 4, the real array is interpolated into a ULVA consisting of $M = 17$ isotropic omnidirectional point sources separated by a distance d/λ. Typically d is chosen to be close to $\lambda/2$. By choosing the reference point of the ULVA at the center of the real array, the steering vectors associated with the virtual array are given by

$$\overline{a}(\theta) = \begin{bmatrix} \exp\left(\frac{j2\pi kd}{\lambda}\cos\theta\right), \ldots, \exp\left(\frac{j2\pi 2d}{\lambda}\cos\theta\right), \\ \times \exp\left(\frac{j2\pi d}{\lambda}\cos\theta\right), \quad 1, \\ \exp\left(\frac{j2\pi d}{\lambda}\cos\theta\right), \exp\left(\frac{j2\pi 2d}{\lambda}\cos\theta\right) \ldots, \\ \times \exp\frac{j2\pi kd}{\lambda}\cos\theta \end{bmatrix}^T_{N\times 1}, \tag{44}$$

The distance d between the elements of the ULVA was chosen to be 0.4775λ. The incremental size Δ in the interpolation region, $\Theta = [-\theta_q, \theta_{q+1}] = [-60, 60^\circ]$, is chosen to be 1°. In this case $Q = 1$. The sector chosen here, for example, is of 120° symmetrically located. Then, a set of real steering vectors are computed for the sources located at each of the angles $\theta_q, \theta_q + \Delta, \theta_q + 2\Delta, \ldots, \theta_{q+1}$. The computed vector $[\overline{a}(\theta)]$ is then distorted from the ideal steering vector due to the presence of mutual coupling between the elements of the real array. The actual steering vectors having all the undesired electromagnetic effects are computed using WIPL-D [13]. Then, using (8) we obtain the transformation matrix to compensate for the effects of nonuniformity in spacing and the presence of mutual coupling between the elements of the real array. Finally, using (69), we can obtain the corrected input voltage in which the nonuniformity in spacing and mutual coupling effects are eliminated from the actual voltage.

Next, the magnitude of the incident SoI is varied from 1 to 10.0 V/m in increments of 0.01 V/m while maintaining the jammer intensities constant, which are arriving from -20°, 40°, 50°. The direction of arrival of the SoI is 10°. The signal-to-thermal noise ratio at each antenna element is set at 20 dB. All signals intensities and directions of arrival are given by the data in Table 4.

If the jammers have been nulled correctly and the SoI recovered properly, it is expected that the recovered signal will have a linear relationship with respect to the intensity of the incident signal, implying that the various electromagnetic effects have been properly accounted for. The estimate for the SoI in Fig. 7a shows that the mutual coupling between the elements of the real array undermines the performance of the direct data domain approach if the various electromagnetic effects are not accounted for. Fig. 7b illustrates the superiority of the results when the direct data domain least-squares approach is used after preprocessing the data to take into account the various mutual coupling effects. The estimated amplitude and the phase for the SoI are shown in Fig. 7b. Here, the amplitude displays the expected linear relationship and the phase changes very little from zero degrees.

The beam patterns associated with this example are shown in Fig. 8a,b. Here, we set the amplitude of the SoI to be 1 V/m, and the other parameters are as given in Table 4. Figure 8b illustrates that the nulls are deep and are located along the correct directions. This indicates that the two-step procedure illustrated in this article have properly modeled the real environment nulling the relevant interferers. However, we see that in Fig. 8a the nulls in the beam pattern are shallow and are displaced from their desired locations, as the undesired

Table 4. Parameters of the Signals

	Magnitude(V/m)	Phase	DoA
Signal	1.0 −10.0	0.0	10°
Jammer	1.0	0.0	-20°
Jammer	1.0	0.0	40°
Jammer	1.0	0.0	50°

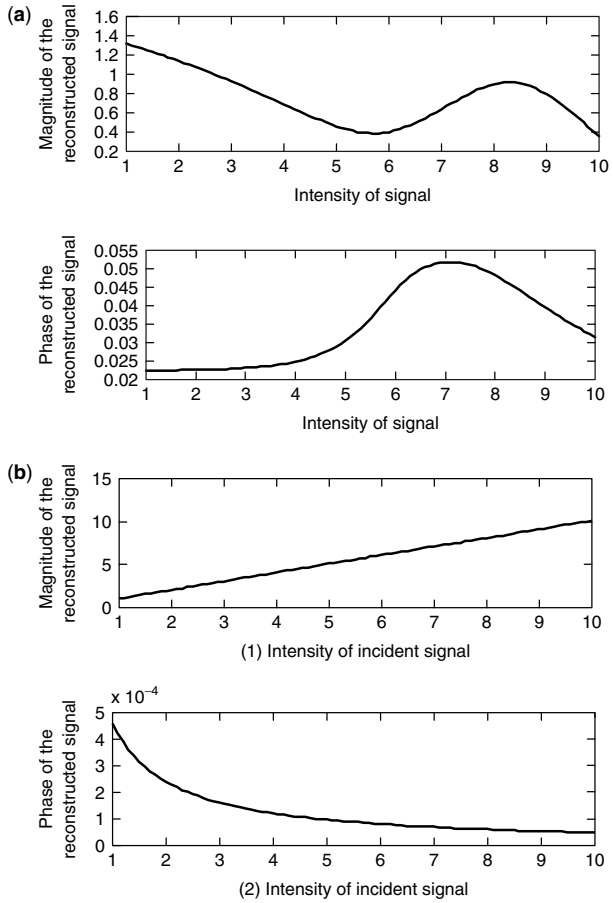


Figure 7. Estimation of the SoI (**a**) without and (**b**) after compensating for mutual coupling.

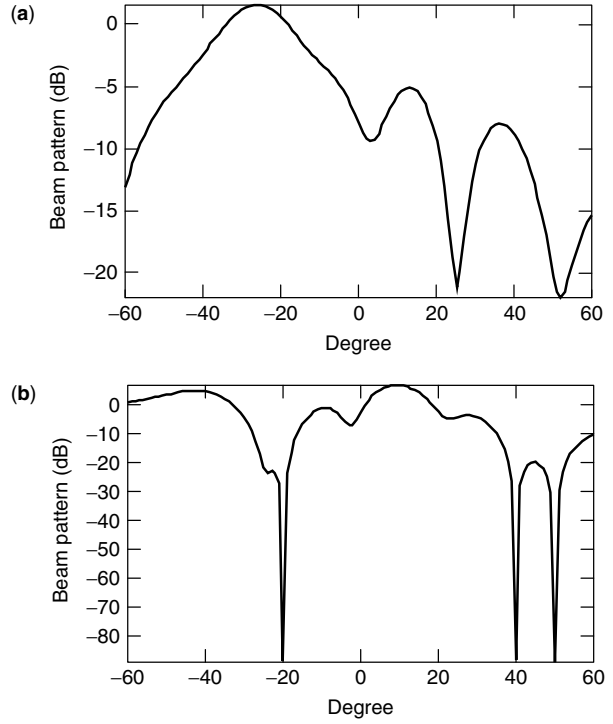


Figure 8. Adapted beam pattern (**a**) without and (**b**) after compensating for the mutual coupling.

electromagnetic effects have not been properly taken into account. Hence, in that case the SoI cannot be recovered.

For the final example we consider the effects of large near-field scatterers located close to the semicircular array. As shown in Fig. 9, there is a large structure located within a distance, that is 5 times the radius of the semicircular array and is oriented along the direction of -20°. The width of the structure is 7.64 wavelengths, and the height is 15.28 wavelengths. Hence, the semicircular array and the scatterer have strong electromagnetic coupling in addition to the presence of mutual coupling between the elements. We again consider the case of four incoming signals from -20°, 10°, 40°, 50°. The parameters for the desired signal and the jammers are summarized in Table 4.

After we compensate for the various electromagnetic couplings and project the data to that due to a ULVA, we solve for the weights [**W**] using (43). Then, we estimate the amplitude of the desired signal through (31). Fig. 10a,b plots the amplitude and the phase for the SoI in both presence and absence of mutual coupling between the

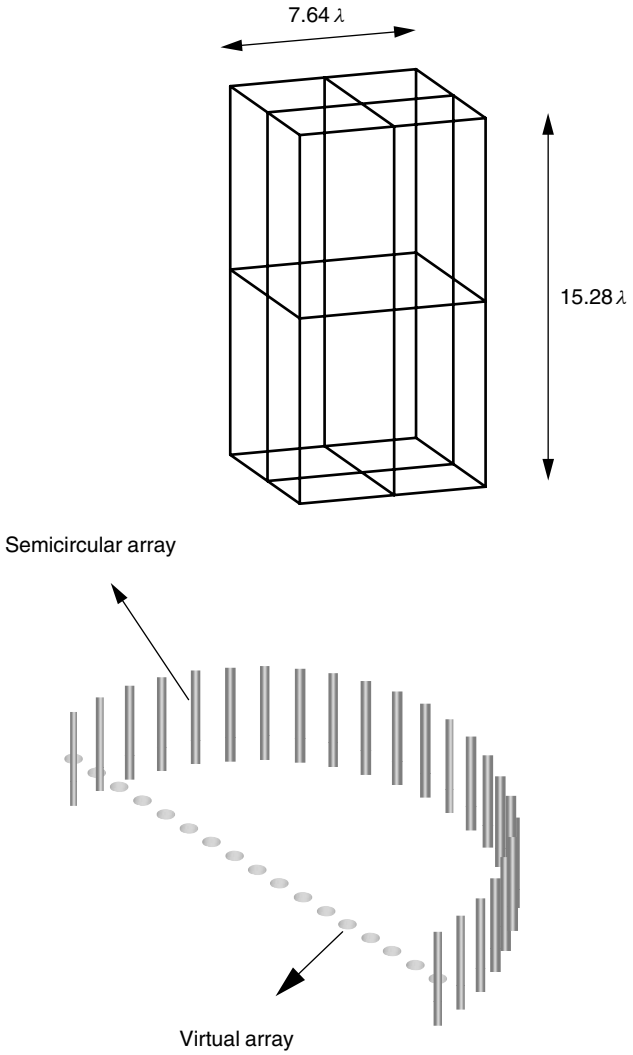


Figure 9. A semicircular array operating in the presence of a large obstacle.

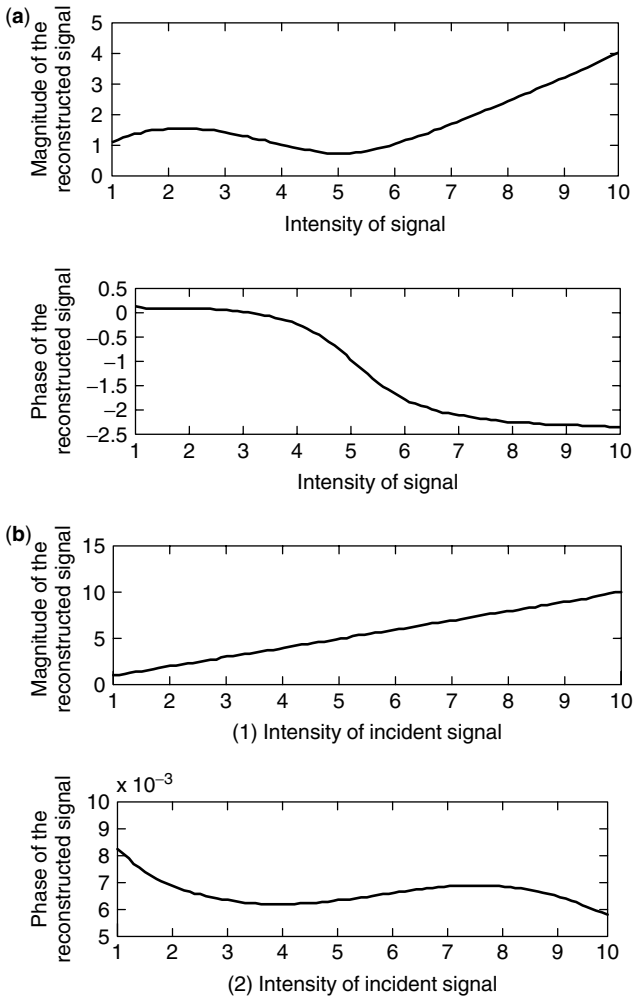


Figure 10. Estimation of SoI (**a**) without and (**b**) after compensating for the mutual coupling and the near-field scatterer.

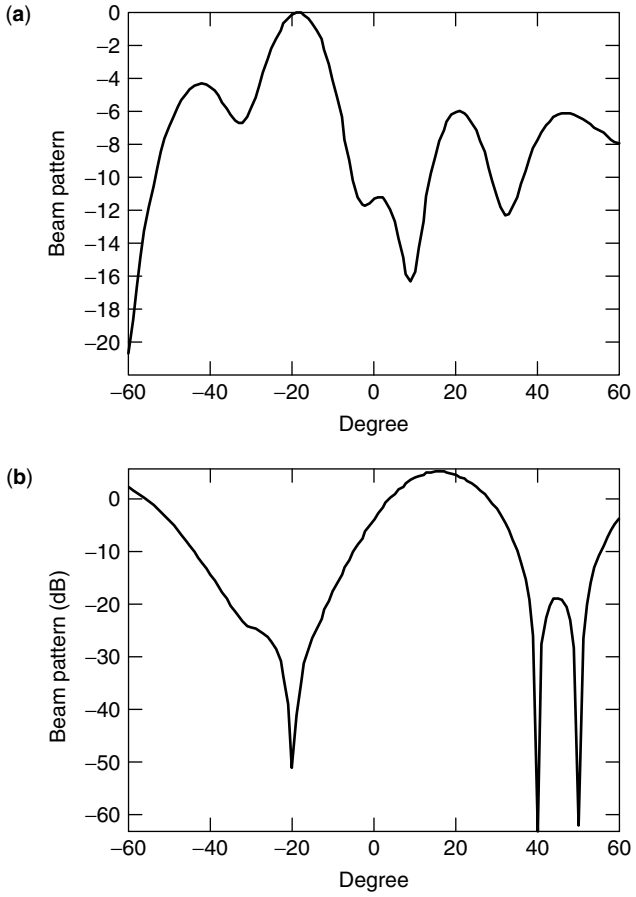


Figure 11. Adapted beam pattern (**a**) without and (**b**) after compensating for the mutual coupling and the near-field scatterer.

elements of the array and the scatterer located close to the array. As can be seen, after compensation of the undesired electromagnetic effects, the expected linear relationship is clearly seen, implying that the jammers have been nulled and the SoI estimated with a good accuracy.

The adapted beam patterns associated with this example are shown in Fig. 11a,b for the two cases considered above. When the mutual coupling is neglected, the beam pattern in Fig. 11a clearly indicates that the interferers have not been nulled in a correct fashion. However, when the electromagnetic effects have been appropriately accounted for, the beam points correctly along the direction of SoI while simultaneously placing deep nulls along the direction of the interferers. By comparing the adapted beam patterns in Figs. 8b and 11b it is seen that the presence of a large near-field scatterer has indeed produced a wide null along that direction due to the diffraction effects of the interferer.

6. CONCLUSION

The objective of this article has been to present a two-step process for using adaptive antenna arrays operating in a real environment. In the first step a transformation matrix is determined that transforms the voltages induced at the feed points of the antenna elements operating in the presence of near-field scatterers and the presence of mutual coupling between the antenna elements to that of the voltages induced in a uniform linear virtual array consisting of isotropic omnidirectional point radiators operating in free space. Such a transformation takes into account all the electromagnetic interactions between the antenna elements and its environment. The next step in the solution procedure involves applying a direct data domain least-squares approach that estimates the complex amplitude of the signal of interest given its direction of arrival. The signal of interest may be accompanied by coherent multipaths and interferers, which may be located in an angle quite close to the direction of arrival of the signal. In addition, there may be clutter and thermal noise at each of the antenna elements. In this approach, since no statistical methodology is employed, there is no need to compute a covariance matrix. Therefore, this procedure can be implemented on a general-purpose digital signal processor for real-time implementations.

BIOGRAPHIES

Kyungjung Kim was born in Seoul, Korea. He received the B.S. degree from Inha University, Korea, and the

M.S. degree from Syracuse University, Syracuse, New York. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering at Syracuse University.

He was a research assistant at Syracuse University from 1998 to 2001 and a graduate fellow from 2001 to 2002. His current research interests include adaptive array signal processing and wavelet transform.

Tapan Kumar Sarkar received the B. Tech. degree from the Indian Institute of Technology, Kharagpur, India, in 1969; the M.Sc.E. degree from the University of New Brunswick, Fredericton, Canada, in 1971; and the M.s. and Ph.D. degrees from Syracuse University; Syracuse, New York, in 1975. He is now a professor in the Department of Electrical and Computer Engineering at Syracuse University. His current research interests deal with numerical solutions of operator equations arising in electromagnetics and signal processing with application to system design. He obtained one of the "best solution" awards in May 1977 at the Rome Air Development Center (RADC) Spectral Estimation Workshop. He has authored or coauthored more than 250 journal articles and numerous conference papers and has written chapters in 28 books and 10 books including the latest ones on *Iterative and Self Adaptive Finite-Elements in Electromagnetic Modeling* and *Wavelet Applications in Engineering Electromagnetics* by Artech House.

Dr. Sarkar is a registerd professional engineer in the State of New York. He received the Best Paper Award of the IEEE Transactions on Electromagnetic Compatibility in 1979 and in the 1997 National Radar Conference. He received the College of Engineering Research Award in 1996 and the chancellor's citation for excellence in research in 1998 at Syracuse University. He received the title Docteur Honoris Causa from Universite Blaise Pascal, Clermont Ferrand, France, in 1998, and he was awarded the medal of Friends of Clermont Ferrand by the mayor of the city in 2000.

Magdalena Salazar-Palma receive dhte degree of *Ingeniero de Telecomunicación* and the Ph.D. degree from the *Universidad Politècnica de Madrid* (Madrid, Spain), where she is a *Profesor Titular* of the *Departamento de Señales, Sistemas y Radiocomunicaciones* (Signals, Systems and Radiocommunications Department) at the *Escuela Técnica Superior de Ingenieros de Telecomunicación*. Her research is in the areas of electromagnetic field theory, computational and numerical methods for microwave passive components and filter design, antenna analyis and design. A number of times she has been a visiting professor at the Electrical Engineering and Computer Science, Syracuse University (Syracuse, New York).

She has authored one book and a total of 15 contributions for chapters and articles in books, 25 papers in international journals, and 113 papers in international conferences, symposiums, and workshops. She is a member of the editorial board of three scientific journals. She is a registered engineer in Spain. She is a senior member of the Institute of Electrical and Electronics Engineers (IEEE). She has served as vice chairman and chairman of IEEE MTT-S/AP-S (Microwave Theory and Techniques Society/Antennas and Propagation Society) Spain joint chapter and chairman of IEEE Spain Section. She is a member of IEEE Region 8 Nominations and Appointments Committee, IEEE Ethics and Member Conduct Committee, and IEEE Women in Engineering Committee (WIEC). She is acting as liaison between IEEE Regional Activities Board and IEEE WIEC.

BIBLIOGRAPHY

1. T. K. Sarkar et al., A pragmatic approach to adaptive antennas, *IEEE Antennas Propag. Mag.* **42**(2): 39–55 (April 2000).
2. T. K. Sarkar, S. Park, J. Koh, and R. A. Schneible, A deterministic least squares approach to adaptive antennas, *Digital Signal Process. Rev. J.* **6**: 185–194 (1996).
3. S. Park and T. K. Sarkar, Prevention of signal cancellation in an adaptive nulling problem, *Digital Signal Process. Rev. J.* **8**: 95–102 (1998).
4. T. K. Sarkar, S. Nagaraja, and M. C. Wicks, A deterministic direct data domain approach to signal estimation utilizing nonuniform and uniform 2-D arrays, *Digital Signal Process. Rev. J.* **8**: 114–125 (1998).
5. T. K. Sarkar et al., A deterministic least squares approach to space time adaptive processing (STAP), *IEEE Trans. Antennas Propag.* **49**: 91–103 (Jan. 2001).
6. T. K. Sarkar and R. Adve, Space time adaptive processing using circular arrays, *IEEE Antennas Propag. Mag.* **43**: 138–143 (Feb. 2001).
7. R. S. Adve and T. K. Sarkar, Compensation for the effects of mutual coupling on direct data domain adaptive algorithms, *IEEE Trans. Antennas Propag.* **48**(1): (Jan. 2000).
8. I. J. Gupta and A. A. Ksienski, Effects of mutual coupling on the performance of adaptive arrays, *IEEE Trans. Antennas Propag.* **31**: 785–791 (Sept. 1983).
9. K. M. Pasala and E. M. Friel, Mutual coupling effects and their reduction in wideband direction of arrival estimation, *IEEE Trans. Aerospace Electron. Syst.* **30**: 1116–1122 (April 1994).
10. B. Friedlander, The root-MUSIC algorithm for direction finding with interpolated arrays, *Signal Process.* **30**: 15–29 (1993).
11. T.-S. Lee and T.-T. Lin, Adaptive beamforming with interpolation arrays for multiple coherent interferes, *Signal Process.* **57**: 177–194 (1997).
12. M. Wax and J. Sheinvald, Direction finding of coherent signals via spatial smoothing for uniform circular arrays, *IEEE Trans. Antennas Propag.* **42**(5): (May 1994).
13. B. M. Kolundzija, J. S. Ognjanovic, and T. K. Sarkar, *WIPL-D: Electromagnetic Modeling of Composite Metallic and Dielectric Structures*, Artech House, Norwood, MA, 2000.
14. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Press, Baltimore, 1989.
15. S. Van Huffel, *Analysis of the Total Least Squares Problem and Its Use in Parameter Estimations*, PhD thesis, Dept. Electrical Engg, Katholicke Univ. Leuven, 1990.
16. R. Brown and T. K. Sarkar, Real time deconvolution utilizing the fast Fourier transform and the Conjugate Gradient

method, 5th *ASSP Workshop on Spectral Estimation and Modeling*, Rochester, NY, 1980.

17. T. K. Sarkar, Application of the conjugate gradient method to electromagnetics and signal analysis, *Progress in Electromagnetics Research*, Vol. 5, Elsevier, 1991.

ADAPTIVE EQUALIZERS

KRZYSZTOF WESOŁOWSKI
Poznań University of Technology
Poznań, Poland

1. INTRODUCTION

Physical channels used in transmission of digital signals can be rarely represented by a nondistorting channel model with additive noise as the only impairment. However, the vast majority of channels are characterized by a limited bandwidth in which particular frequency components of transmitted signals are nonequally attenuated (causing *amplitude distortion*) and nonequally delayed (creating *delay distortion*). These effects are the results of the physical properties of the transmission medium and of the imperfect design of transmit and receive filters applied in the transmission system. A good example of the first is the radio channel, in which the transmitted signal reaches the receiver along many different paths through reflections, diffractions, and dispersion on the terrain obstacles. As a result, particular signal path components arriving with various attenuations and delays are combined at the receiver. The delayed components can be considered as echoes that cause time dispersion of the transmitted signal. If time dispersion is greater than a substantial fraction of the signaling period, the channel responses to the subsequent data signals overlap. This effect is known as *intersymbol interference*. Thus, the signal observed at the receiver input contains information on a certain number of data signals simultaneously. In many cases the channel impulse response spans even tens of signaling periods and intersymbol interference appears to be a major impairment introduced by the channel.

The destructive influence of intersymbol interference on a digital communication system performance has to be counteracted by special receiver and/or transmitter design. Thus, a fundamental part of the receiver is the channel *equalizer*. Very often transmission channel characteristics are either not known at the beginning of a data transmission session or they are time-variant. Therefore, it is advantageous to make the equalizer *adaptive*. The adaptive equalizer is able to adaptively compensate the distorting channel characteristics and simultaneously track the changes of channel characteristics in time. The latter property is a key feature of equalizers used in digital transmission over nonstationary radio channels.

Since the invention of an equalizer in the early 1960s, hundreds of papers have been devoted to this subject. Adaptive equalization is usually a topic of a separate chapter in leading books on digital communication systems [1–3], and separate books tackle this subject as well [5,6]. Adaptive equalization is also a well-documented application example in books devoted to adaptive filters [4,7]. In this tutorial we are able to present a general survey of adaptive equalizers only and give reference to the key papers and books. Interested readers are advised to study the wide literature, quoted, for example, in such papers as those by Qureshi [8] and Taylor et al. [9].

In this tutorial we will concentrate on basic structures and algorithms of adaptive equalizers. We will start with the model of a transmission system operating in the intersymbol interference channel. Subsequently, we will divide the equalizers into several classes. We will continue our considerations with the basic analysis of adaptation criteria and algorithms for linear and decision feedback equalizers. Then we will concentrate on adaptive algorithms and equalizer structures applying the MAP (*maximum a posteriori*) symbol-by-symbol detector and the MLSE (*maximum-likelihood sequence estimation*) detector. Finally, we will describe basic structures and algorithms of adaptive equalization without a training sequence (*blind equalization*).

2. SYSTEM MODEL

Generally, we can consider two types of data transmission. In the first one the channel transmits signal spectral components close to zero frequency and no modulation of the sinusoidal carrier is necessary. A good example is data transmission over a PSTN (public switched telephone network) subscriber loop, realizing the basic or primary ISDN access. Figure 1 shows a model of the data transmission system operating in the baseband channel. The binary sequence is transformed into the data symbol sequence in a symbol mapper. Data symbols are fed to the transmit filter $p(t)$ with the signaling period of T seconds. This filter shapes the transmitted signal spectrum. The data pulses are subsequently transmitted through the channel with the impulse response $g(t)$. The receive filter is usually a filter matched to the transmit filter, so its impulse response is $p(-t)$ (assuming that the additive white Gaussian noise $n(t)$ is added at the output of the channel). Replacing the cascade of the transmit filter, the transmission channel and the receive filter by a single equivalent

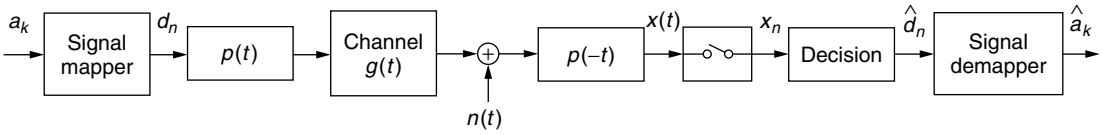


Figure 1. Model of the baseband transmission system.

filter, we receive the following equation describing the operation of the transmission system

$$x(t) = \sum_{i=-\infty}^{+\infty} d_i h(t - iT) + \nu(t) \tag{1}$$

where $h(t)$ is a convolution of the filter and channel impulse responses $[h(t) = p(t) * g(t) * p(-t)]$ and $\nu(t)$ is the noise $n(t)$ filtered by the receive filter $(\nu(t) = n(t) * p(-t))$.

In the second type of data transmission system a bandpass channel is used, so data signals modulate a sinusoidal carrier. Figure 2 presents a system model in which sinusoidal and cosinusoidal carriers of the frequency f_c are modulated by a pair of in-phase and quadrature data symbols d_n^I and d_n^Q, respectively. These symbols are received from the signal mapper, which associates the binary data blocks with a pair of data symbols resulting from the applied modulation format. As in the baseband transmission system, the spectrum of the transmitted signal is shaped in the baseband by the transmit filters $p(t)$. One can easily prove that the system model contained between lines A and B can be represented by the same equation as (1); however, in this case the variables and functions in Eq. (1) are complex: $d_i = d_i^I + jd_i^Q$, and $h(t) = h_{\text{re}}(t) + jh_{\text{im}}(t)$. In the passband transmission system model the channel impulse response $h(t)$ is a convolution of the transmit and receive filter responses $p(t)$ and $p(-t)$, respectively, and the so called *baseband equivalent channel* impulse response $g_B(t)$. The baseband equivalent channel impulse response is associated with the impulse response $g_P(t)$ of the bandpass channel (see Fig. 2) by the equation

$$g_P(t) = g_B(t) \exp[j2\pi f_c t] + g_B^*(t) \exp[-j2\pi f_c t] \tag{2}$$

Concluding, with respect to intersymbol interference, both baseband and passband transmission systems can be modeled by Eq. (1), in which variables and functions of time are real- or complex-valued depending on whether the system is implemented in the baseband or whether it uses the bandpass channel.

3. INTERSYMBOL INTERFERENCE

The task of the digital system receiver is to find the most probable data symbols d_n^I and d_n^Q on the basis of the observed signal $x(t)$ at its input. If the channel were nondistorting, then, assuming appropriate shaping of the transmit filter $p(t)$ and the filter $p(-t)$ matched to it, it would be possible to find such periodic sampling moments at the outputs of the receive filters that the samples of signals $x^I(t)$ and $x^Q(t)$ would contain information on single data symbols only. However, the distortion introduced by the channel renders the finding of such sampling moments impossible. Thus, it is necessary to apply a special system block denoted in Fig. 3 as equalizer, which is able to detect data symbols on the basis of $x(t)$ [or equivalently $x^I(t)$ and $x^Q(t)$] or its samples. An equalizer is in fact a kind of receiver that either minimizes the influence of intersymbol interference or uses it constructively in the decisions concerning the transmitted data.

Although the signals $x^I(t)$ and $x^Q(t)$ are represented in Fig. 3 as continuous time functions, typically, due to digital implementation, the equalizer accepts their samples only. Let us temporarily assume that the equalizer input samples are given with the symbol period of T seconds with the time offset τ with respect to the zero moment. Then the equalizer input signal is expressed by the equation

$$\begin{aligned} x_n = x(nT + \tau) &= \sum_{i=-\infty}^{\infty} d_i h(nT + \tau - iT) + \nu(nT + \tau) \\ &= \sum_{i=-\infty}^{\infty} d_i h_{n-i} + \nu_n = \sum_{i=-\infty}^{\infty} h_i d_{n-i} + \nu_n \end{aligned} \tag{3}$$

$$x_n = h_0 d_n + \sum_{i=-\infty, i\neq 0}^{\infty} h_i d_{n-i} + \nu_n \tag{4}$$

where $h_{n-i} = h(nT + \tau - iT)$. The first term in Eq. (4) is proportional to the data symbol to be decided on. The second term is a linear combination of previous and future data symbols and expresses intersymbol interference. It should be eliminated or constructively used by the

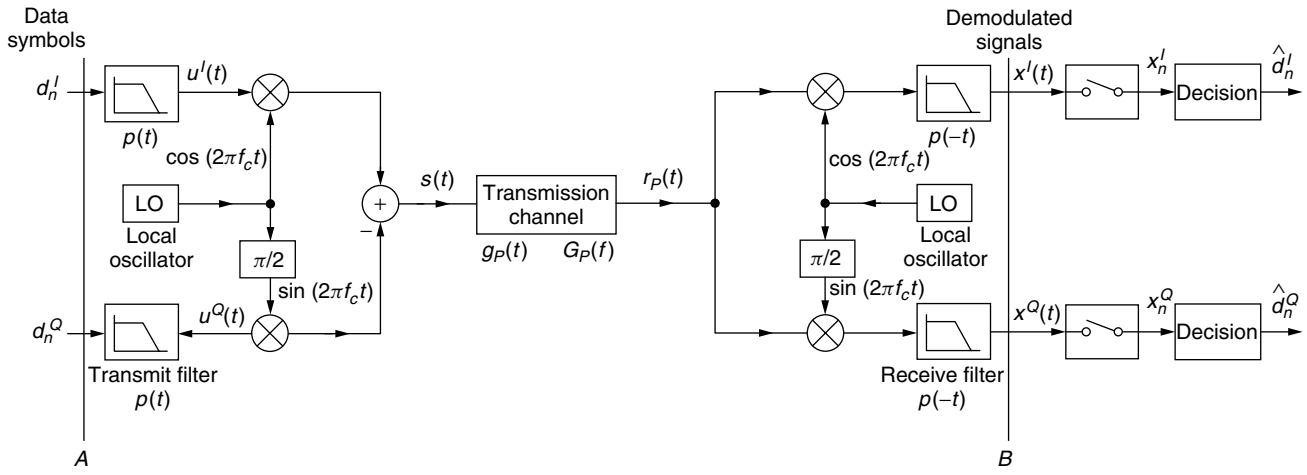


Figure 2. Model of the passband transmission system.

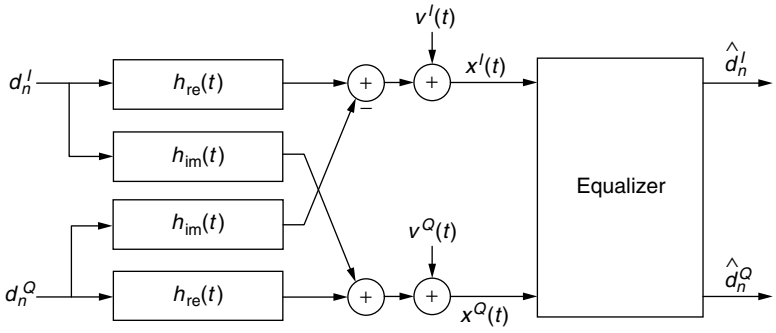


Figure 3. Equivalent transmission system model with the equalizer.

equalizer. The third term is the additive noise and cannot be eliminated.

4. CLASSIFICATION OF EQUALIZER STRUCTURES AND ALGORITHMS

Channel equalization can be performed by linear or nonlinear methods. Decisions regarding the data symbols can be made by the equalizer on the symbol-by-symbol basis or can be performed on the whole sequence. Figure 4 presents the classification of equalization structures.

Within the class of linear receivers the equalizer based on an FIR (Finite impulse response) *transversal filter* is of great importance. It is implemented using symbol-spaced or fractionally spaced taps. A lot of attention has also been paid in the literature to the linear equalizer applying a *lattice filter* [10]. The latter, although more complicated than the transversal filter, assures faster convergence of the adaptation algorithm.

In case of channels characterized by the occurrence of deep notches, nonlinear receivers are used. The simplest version of a nonlinear receiver is the *decision-feedback equalizer* (DFE) [11]. The MLSE equalizer, which is more computationally intensive is applied for example in GSM receivers and high-speed telephone modems. It detects a whole sequence of data symbols, usually using the *Viterbi algorithm* [12]. If the intersymbol interference is caused by a long-channel impulse response or if the data symbol alphabet is large, the MLSE equalizer becomes unfeasible due to excessive computational complexity. Several suboptimal structures and procedures can be applied instead, such as *reduced state sequence estimation* (RSSE) [13], *delayed decision feedback sequence estimation* (DDFSE) [14], or the *M* algorithm [15]. Another approach is a nonlinear symbol-by-symbol detection using the *maximum a posteriori* (MAP) criterion. The algorithm of Abend and Fritchman [22] is one example of such an approach. The MAP algorithms are usually computationally complex.

The key feature of all equalization structures is their ability of initial adaptation to the channel characteristics (*startup equalization*) and tracking it in time. In order to acquire the initial adaptation, an optimization criterion has to be defined. Historically, the first criterion was minimization of the maximum value of intersymbol interference (*minimax criterion*) resulting in the *zero-forcing* (ZF) equalizer. The most popular adaptation criterion is minimization of the *mean-squared error*, resulting in the *MSE* equalizer. In this criterion the expectation of the squared error signal at the equalizer output is minimized. Finally, the criterion used in the fastest adaptation algorithms relies on minimization of the *least squares* (LS)

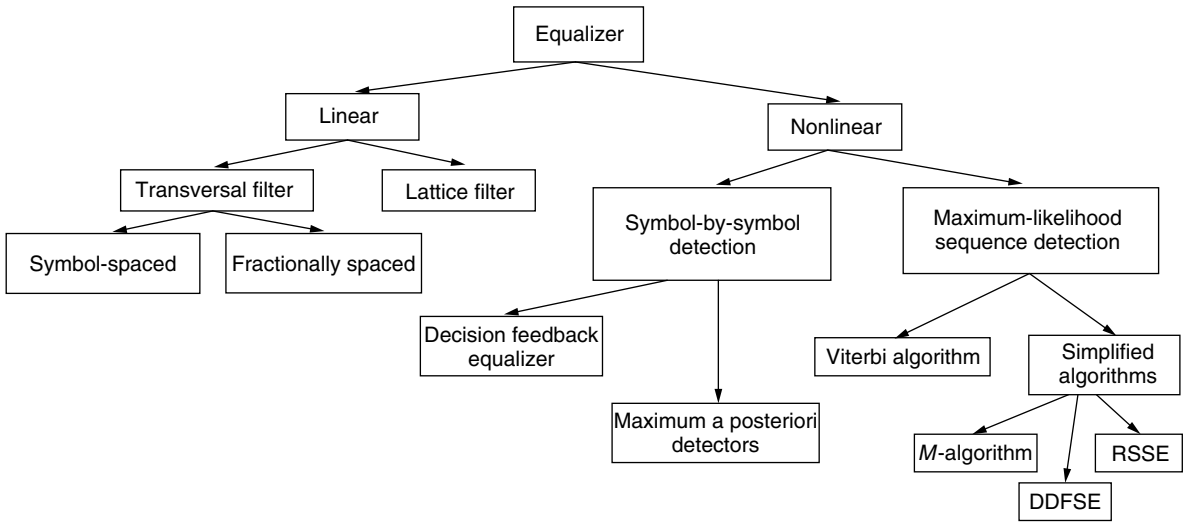


Figure 4. Classification of the equalization structures.

of errors. The equalizer using the algorithm based on this criterion is called an *LS* equalizer. The equalizer parameters are selected to minimize the squared sum of the equalizer output signal errors that would be achieved if these parameters were used starting from the initial moment of adaptation. Some other cost functions can be selected if the equalizer coefficients are derived without the knowledge of the transmitted data symbols.

The equalizer parameters are derived in accordance with a chosen adaptation criterion by an *adaptation algorithm*. Most of the algorithms are *recursive* — the adaptation at a given moment is performed iteratively, taking advantage of the results achieved in the previous adaptation step. In special cases *fast startup equalization* algorithms are applied, resulting in extremely fast calculation of coarse equalizer parameters that are good enough to start regular data transmission and that are later refined. Some of these algorithms are known as *noniterative*; some others are *recursive least-squares* (RLS) algorithms.

The adaptation of a typical equalizer can be divided into two phases. In the first phase, the training data sequence known to the receiver is transmitted. The adaptation algorithm uses this sequence as a reference for the adjustment of the equalizer coefficients; thus the equalizer is in the *training mode*. After achieving the equalizer parameters that result in a sufficiently low probability of errors made by the equalizer decision device, the second phase of adaptation begins in which the equalizer starts to use the derived decisions in its adaptation algorithm. We say that the equalizer is then in the *decision — directed* mode.

In some cases, in particular in point-to-multipoint transmission, sending the training sequence to initiate a particular receiver is not feasible. Thus, the equalizer must be able to adapt without a training sequence. Its algorithm is based exclusively on the general knowledge about the data signal statistics and on the signal reaching the receiver. Such an equalizer is called *blind*. Blind adaptation algorithms are generally either much slower or much more complex than data-trained algorithms.

5. LINEAR ADAPTIVE EQUALIZERS

The linear equalizer is the simplest structure, most frequently used in practice. Let us consider the receiver applying a transversal filter. The scheme of such an equalizer is shown in Fig. 5. The output signal y_n at the nth moment depends on the input signal samples x_{n-i} and the equalizer coefficients $c_{i,n}(i = -N, \ldots, N)$ according to the equation

$$y_n = \sum_{i=-N}^{N} c_{i,n} x_{n-i} \tag{5}$$

The equalizer output signal is a linear combination of $2N + 1$ subsequent samples of the input signal. Indexing the equalizer coefficients from $-N$ to N reflects the fact that the reference tap is located in the middle of the tapped delay line of the equalizer and that, typically, not only previous data symbols with respect to the reference one but also some future symbols influence the current input signal sample.

5.1. ZF Equalizers

Historically, the earliest equalizers used the *minimax* adaptation criterion. It resulted in the simplest algorithm that is still used in the equalizers applied in line-of-sight microwave radio receivers. Let us neglect the additive noise for a while. Taking into account Eqs. (4) and (5), we obtain

$$y_n = \sum_{i=-N}^{N} c_{i,n} \sum_{k=-\infty}^{\infty} h_k d_{n-i-k} \tag{6}$$

Substituting $j = i + k$, we get

$$y_n = \sum_{i=-N}^{N} c_{i,n} \sum_{j=-\infty}^{\infty} h_{j-i} d_{n-j} \tag{7}$$

or equivalently

$$y_n = \sum_{j=-\infty}^{\infty} g_{j,n} d_{n-j} \qquad \text{where} \qquad g_{j,n} = \sum_{i=-N}^{N} c_{i,n} h_{j-i} \tag{8}$$

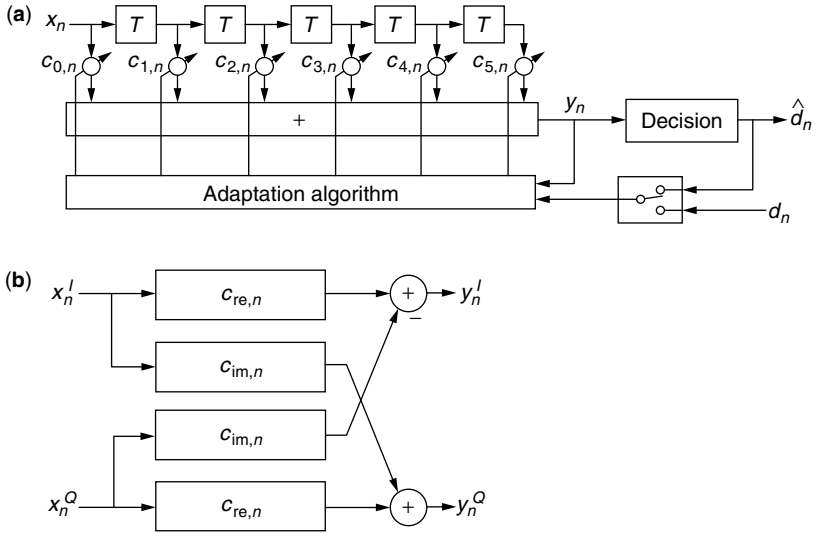


Figure 5. Linear adaptive equalizer: (**a**) basic structure; (**b**) structure equivalent to the complex filter applying real filters.

and $g_{j,n}$ are the samples of cascade connection of the discrete channel and the equalizer. In the *minimax* criterion the equalizer coefficients $c_{i,n} (i = -N, \ldots, N)$ are adjusted to minimize the expression

$$I = \frac{1}{g_{0,n}} \sum_{j=-\infty,\ j\neq 0}^{\infty} |g_{j,n}| \tag{9}$$

Let us note that, because of the finite number of the adjustable equalizer coefficients, it is possible to set to zero only part of the intersymbol interference samples observed at the output of the equalizer filter. One can show that in order to set the intersymbol interference samples to zero, at the assumption that the data symbols are uncorrelated and equiprobable, it suffices to set the equalizer coefficients to force to fulfil the following equality

$$E[e_n d_{n-i}] = 0 \qquad \text{for} \qquad i = -N, \ldots, N \tag{10}$$

where the error e_n in the training mode is given by the expression

$$e_n = y_n - d_n \tag{11}$$

or $e_n = y_n - \mathrm{dec}(y_n)$ in the decision-directed mode. In fact, substituting in (10) the expression for e_n and y_n, we obtain from (8)

$$E[e_n d_{n-i}] = E\left[\left(\sum_{j=-\infty}^{\infty} g_{j,n} d_{n-j} - d_n\right) d_{n-i}\right]$$
$$= \begin{cases} 0 & \text{for} \quad i = 0, \quad \text{if} \quad g_{0,n} = 1 \\ 0 & \text{for} \quad i \neq 0, i \in \langle -N, N\rangle, \quad \text{if} \quad g_{i,n} = 0 \end{cases} \tag{12}$$

Forcing condition (12), $2N$ intersymbol interference samples can be set to zero. Therefore, such an equalizer is called *zero-forcing equalizer*. If the equalizer was infinitely long it would be able to completely eliminate the ISI at its output. The cascade connection of the channel and equalizer would have the discrete impulse response in form of a unit pulse. Therefore, the equalizer would ideally inverse the channel frequency characteristics. Such an equalizer could be adjusted iteratively according to the equation

$$c_{i,n+1} = c_{i,n} - \alpha E[e_n d_{n-i}] \qquad \text{for} \qquad i = -N, \ldots, N \tag{13}$$

However, replacing the ensemble average with its stochastic estimate, we receive the following equation for the coefficients' adjustment, which is easily implementable even at a very high symbol rate

$$c_{i,n+1} = c_{i,n} - \alpha e_n d_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{14}$$

for real equalizers, and

$$c_{i,n+1} = c_{i,n} - \alpha e_n d^*_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{15}$$

for complex ones. More details on the ZF equalizer can be found in Ref. 16. The ZF equalizer attempting to inverse the channel characteristics amplifies the noise in these frequency regions in which the channel particularly attenuates the signal.

5.2. MSE Equalizers

As we have already mentioned, the most frequent adaptation criterion is minimization of the mean square error:

$$\min_{\{c_{i,n}, i=-N,\ldots,N\}} E[|e_n|^2] \tag{16}$$

where the error is given by equation (11). Direct calculations of the mean square error (MSE) $\mathcal{E}_n^{\mathrm{MSE}} = E[|e_n|^2]$ with respect to the equalizer coefficients $\mathbf{c}_n = [c_{-N,n}, \ldots, c_{0,n}, \ldots, c_{N,n}]^T$ lead to the following dependence of the MSE on the coefficients for the real equalizer

$$\mathcal{E}_n^{\mathrm{MSE}} = \mathbf{c}_n^T A \mathbf{c}_n - 2\mathbf{b}^T \mathbf{c}_n + E[|d_n|^2] \tag{17}$$

where $A = E[\mathbf{x}_n \mathbf{x}_n^T]$ ($\mathbf{x}_n = [x_{n+N}, \ldots, x_n, \ldots, x_{n-N}]^T$) is the input signal autocorrelation matrix and $\mathbf{b} = E[d_n \mathbf{x}_n]$ is the vector of cross-correlation between the current data symbol and the equalizer input samples. The autocorrelation matrix A is positive definite (its all eigenvalues are positive). It is well known from algebra that for such a matrix expression (17) has a single and global minimum. The minimum can be found if we set the condition

$$\frac{\partial \mathcal{E}_n^{\mathrm{MSE}}}{\partial c_{i,n}} = 0 \qquad \text{for} \qquad i = -N, \ldots, N \tag{18}$$

The result is the well-known Wiener–Hopf equation for the optimum equalizer coefficients

$$A\mathbf{c}_{\mathrm{opt}} = \mathbf{b} \tag{19}$$

An efficient method of achieving the optimum coefficients and the minimum MSE is to update the equalizer coefficients iteratively with adjustments proportional to the negative value of the gradient of $\mathcal{E}_n^{\mathrm{MSE}}$ calculated for the current values of the coefficients:

$$c_{i,n+1} = c_{i,n} - \alpha_n \frac{\partial \mathcal{E}_n^{\mathrm{MSE}}}{\partial c_{i,n}} \qquad \text{for} \qquad i = -N, \ldots, N \tag{20}$$

where α_n is a small positive value called the adjustment step size. Generally, it can be time variant, which is expressed by the time index n. The calculation of the gradient $\partial \mathcal{E}_n^{\mathrm{MSE}} / \partial c_{i,n}$ leads to the result

$$\frac{\partial \mathcal{E}_n^{\mathrm{MSE}}}{\partial c_{i,n}} = \frac{\partial E\left[|e_n|^2\right]}{\partial c_{i,n}} = 2E\left[e_n \frac{\partial e_n}{\partial c_{i,n}}\right] = 2E[e_n x_{n-i}]$$
$$\text{for} \qquad i = -N, \ldots, N \tag{21}$$

Replacing the gradient calculated in Eq. (21) by its stochastic estimate $e_n x_{n-i}$ $(i = -N, \ldots, N)$ for the real equalizer we receive the stochastic gradient [*least mean-square*(LMS)] algorithm

$$c_{i,n+1} = c_{i,n} - \gamma_n e_n x_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{22}$$

where $\gamma_n = 2\alpha_n$. One can show that the analogous equation for the complex equalizer is

$$c_{i,n+1} = c_{i,n} - \gamma_n e_n x^*_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{23}$$

Figure 6 presents the scheme of the linear transversal equalizer with the tap coefficients adjusted according to algorithm (23). The switch changes its position from 1 to 2 after a sufficiently long training mode.

The convergence rate of the LMS algorithm depends on the value of the step size γ_n. This problem has been thoroughly researched. Generally, the value of the step size depends on the eigenvalue distribution of the input signal autocorrelation matrix A [1]. G. Ungerboeck [17] derived a simple "engineering" formula for the step size, which results in fast and stable convergence of the LMS adaptive equalizer. The initial step size is described by the formula

$$\gamma_0 = \frac{1}{(2N+1)E[|x_n|^2]} \tag{24}$$

where $E[|x_n|^2]$ is the mean input signal power and is equal to the elements of the main diagonal of the autocorrelation matrix A. When the equalizer taps are close to their optimum values, the step size should be decreased in order to prevent an excessively high level of the residual mean square error (e.g., $\gamma_\infty = 0.2\gamma_0$).

5.3. LS Equalizers

Particularly fast initial equalizer convergence is achieved if the *least-squares* adaptation criterion is applied. The coefficients of a linear equalizer are set in order to minimize the following cost function with respect to the filter coefficient vector $\mathbf{c}_n$:

$$\mathcal{E}_n^{LS} = \sum_{i=0}^{n} \lambda^{n-i} |\mathbf{c}_n^T \mathbf{x}_i - d_i|^2 \tag{25}$$

For each moment n, that weighted summed squared error starting from the initial moment up to the current moment n is minimized, which would be achieved if the current coefficient vector calculated on the basis of the whole signal knowledge up to the nth moment were applied in the equalizer from the initial moment. The window coefficient $\lambda^{n-i} (\lambda \leq 1)$ causes gradual forgetting of the past errors and is applied for nonstationary channels to follow the changes in the channel characteristics. The calculation of (25) leads to equations similar to (17) and (19):

$$\varepsilon_n^{LS} = \mathbf{c}_n^T R_n \mathbf{c}_n - 2\mathbf{c}_n^T \mathbf{q}_n + \sum_{i=0}^{n} \lambda^{n-i} |d_i|^2 \tag{26}$$

$$R_n \mathbf{c}_{n,\text{opt}} = \mathbf{q}_n \tag{27}$$

where

$$R_n = \sum_{i=0}^{n} \lambda^{n-i} \mathbf{x}_i^T \mathbf{x}_i = \lambda R_{n-1} + \mathbf{x}_n^T \mathbf{x}_n \qquad \text{and}$$

$$\mathbf{q}_n = \sum_{i=0}^{n} \lambda^{n-i} d_i \mathbf{x}_i \tag{28}$$

Instead of solving the set of linear equations (27) at each subsequent moment, the optimum coefficients can be found iteratively using the results derived at the previous time instant. Below we list only the equations of the standard RLS (Kalman) algorithm proposed by Godard [18] for fast adaptive equalization. The algorithm is quoted after Proakis [1].

For convenience, let us define $P_n = R_n^{-1}$. Let us also assume that before adaptation at the nth moment we have the filter coefficients $\mathbf{c}_{n-1}$ and the inverse matrix P_{n-1} at our disposal. The algorithm steps are as follows:

- Initialization: $\mathbf{c}_0 = [0, \ldots, 0]^T$, $\mathbf{x}_0 = [0, \ldots, 0]^T$, $R_0 = \delta I$.

Do the following for $n \geq 1$

- Shift the contents of the filter tapped delay line by one position and accept the new input signal x_n,
- Compute the filter output signal:

$$y_n = \mathbf{c}_{n-1}^T \mathbf{x}_n \tag{29}$$

- Compute the error at the filter output:

$$e_n = d_n - y_n \tag{30}$$

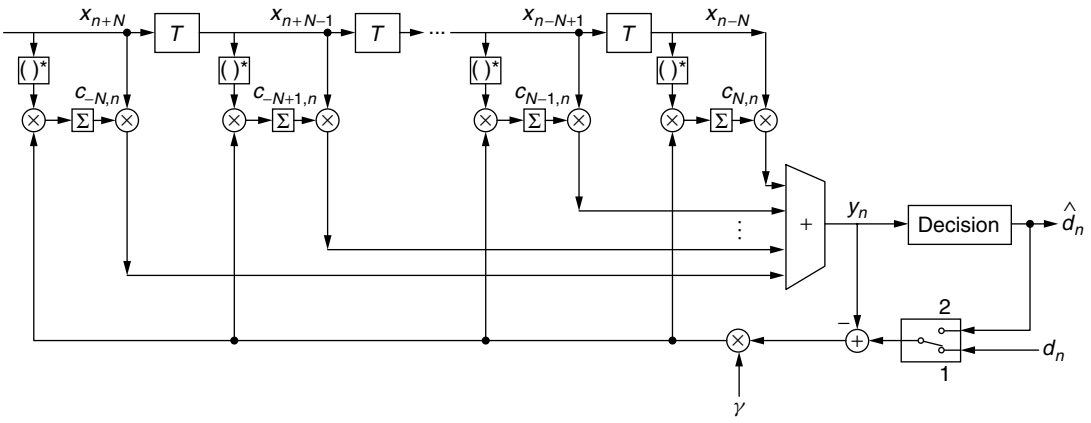


Figure 6. Adaptive MSE gradient equalizer.

- Compute the so called Kalman gain vector $\mathbf{k}_n = P_n\mathbf{x}_n$:

$$\mathbf{k}_n = \frac{P_{n-1}\mathbf{x}_n}{\lambda + \mathbf{x}_n^T P_n \mathbf{x}_n} \tag{31}$$

- Update the inverse of the autocorrelation matrix:

$$P_n = \frac{1}{\lambda}\left[P_{n-1} - \mathbf{k}_n \mathbf{x}_n^T P_{n-1}\right] \tag{32}$$

- Update the filter coefficients:

$$\mathbf{c}_n = \mathbf{c}_{n-1} + \mathbf{k}_n e_n \tag{33}$$

Formulas (29)–(33) summarize the RLS Kalman algorithm for the real equalizer. The complex version of this algorithm can be found in Proakis' handbook [1]. Knowing that $\mathbf{k}_n = P_n\mathbf{x}_n$, we find that the coefficients update is equivalent to the formula

$$\mathbf{c}_n = \mathbf{c}_{n-1} + R_n^{-1}\mathbf{x}_n e_n \tag{34}$$

Comparing the equalizer update using the LMS algorithm (23) and the RLS algorithm (34) we see that the Kalman algorithm speeds up its convergence because of the inverse matrix $P_n = R_n^{-1}$ used in each iteration. In the LMS algorithm this matrix is replaced by a single scalar γ_n. Figure 7 presents the convergence rate for both the LMS and RLS algorithms used in the linear transversal equalizer. The step size of the LMS algorithm was constant and selected to ensure the same residual mean-square error as that achieved by the RLS algorithm. The difference in the convergence rate is evident. However, we have to admit that for the channel model used in the simulations shown in Fig. 7, the application of the step size according to formula (24) and switching it to a small fraction of the initial value after the appropriate number of iterations improves the convergence of the LMS equalizer considerably. On the other hand, tracking abilities of the Kalman algorithm are much better than these of the LMS algorithm. However, the RLS Kalman algorithm is much more demanding computationally. Moreover, because of the roundoff noise, it becomes numerically unstable in the long run, particularly if the forgetting factor λ is lower than one. Solving the problem of excessive computational complexity and ensuring the numerical stability have been the subject of intensive research. Cioffi and Kailath's paper [19] is only one representative example of numerous publications in this area.

Besides the transversal filter, a lattice filter can also be applied in the adaptive equalizer using both the LMS [10] and RLS [20] adaptation algorithms.

5.4. Choice of the Reference Signal

The selection of the reference signal plays an important role in the equalizer adaptation process. In fact, the reference signal tests the unknown channel. Its spectral properties should be selected in such a way that the channel characteristics is fully reflected in the spectrum of the signal at the input of the equalizer. Thus far in our analysis we have assumed that the data symbols are uncorrelated and equiprobable:

$$E[d_n d_{n-k}^*] = \begin{cases} \sigma_d^2 & \text{for} \quad k = 0 \\ 0 & \text{for} \quad k \neq 0 \end{cases} \tag{35}$$

This means that the power spectrum of the test signal is flat and the channel characteristic is "sampled" by a constant power spectrum of the input signal. In practice this theoretical assumption is only approximately fulfilled. Typically, the data sequence is produced on the basis of the *maximum-length* sequence generator. The test generator is usually implemented by a scrambler contained in the transmitter that is based on a *linear feedback shift register* (LFSR). As a result, a *pseudonoise* (PN) binary sequence is generated. Typically, subsets of very long PN sequences are used as a training sequence.

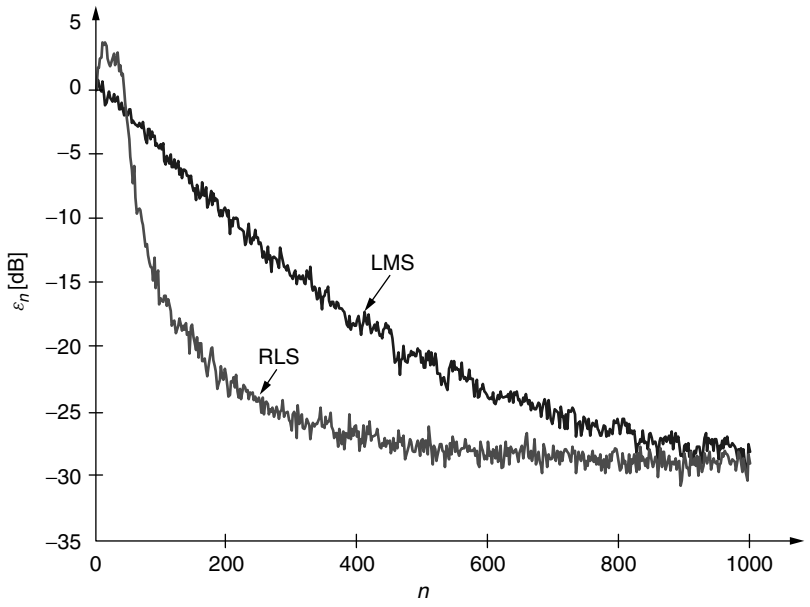


Figure 7. Convergence of the constant-step-size LMS and Kalman RLS equalizer of the length $2N = 30$.

Special attention has focused on very short test sequences that allow for fast, coarse setting of the equalizer coefficients. These sequences are periodic and are constructed in such a way that their deterministic autocorrelation function is zero apart from its origin.

5.5. Fast Linear Equalization Using Periodic Test Signals

In certain applications extremely fast initial equalization is of major importance. A good example is the master modem in a computer network, which receives data blocks from many tributary modems communicating through different channels. Such communication can be effective if the block header is a small part of the whole transmitted data block. A part of the header is a training sequence necessary to acquire the equalizer settings. Let us neglect the influence of the noise for a while. In many cases the SNR in the channel is around 30 dB and the ISI plays a dominant role in the signal distortion. Let the reference signal be periodic. In fact, no more than two periods of the test signal should be transmitted in order to acquire coarse equalizer settings. The period M of the test signal is at least as long as the highest expected length of the channel impulse response L. With periodic excitation the channel output (neglecting the influence of the additive noise) is also a periodic signal. This fact is reflected by the formula

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \end{bmatrix} = \begin{bmatrix} d_0 & d_1 & d_2 & \cdots & d_{M-1} \\ d_{M-1} & d_0 & d_1 & \cdots & d_{M-2} \\ \vdots & & \ddots & & \vdots \\ d_1 & d_2 & \cdots & d_{M-1} & d_0 \end{bmatrix} \cdot \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{M-1} \end{bmatrix} \tag{36}$$

If the length of the channel impulse response is shorter than the length of the test signal, we can assume that some of the last elements in the vector $\mathbf{h}^T = [h_0, h_1, \ldots, h_{M-1}]$ are equal to zero. Due to the periodic nature of the signal transmitted through the channel, a cyclic convolution of the sequence $\mathbf{h}$ and the data sequence $\mathbf{d} = [d_0, d_1, \ldots, d_{M-1}]$ is realized. In the frequency domain this operation is equivalent to the multiplication of two *discrete Fourier transform* (DFT) spectra:

$$X(k\Delta f) = D(k\Delta f) \cdot H(k\Delta f), k = 0, 1, \ldots, M-1 \tag{37}$$

where $\Delta f\ T = 1/M$ and

$$X(k\Delta f) = \frac{1}{M} \sum_{i=0}^{M-1} x(iT) \exp\left[-j2\pi k \Delta f i T\right] \tag{38}$$

Dependencies similar to (38) are held for the data and channel impulse response sequences. Knowing the spectrum of the data sequence, one can easily calculate the spectrum of the channel and, after reversing it, the characteristics of the ZF equalizer can be achieved, i.e.

$$C(k\Delta f) = \frac{1}{H(k\Delta f)} = \frac{D(k\Delta f)}{X(k\Delta f)} \quad k = 0, 1, \ldots, M-1 \tag{39}$$

On the basis of the equalizer characteristics $\mathbf{C}^T = [C(0), C(\Delta f), \ldots, C((M-1)\Delta f)]$, the equalizer coefficients $\mathbf{c}^T = [c_0, c_1, \ldots, c_{M-1}]$ can be calculated using the inverse DFT. If the length of the training sequence and of the equalizer is a power of 2 then all the DFT and IDFT calculations can be effectively performed by the FFT/IFFT algorithms. More detailed considerations on fast startup equalization using the periodic training sequence can be found [21].

5.6. Symbol-Spaced Versus Fractionally Spaced Equalizers

Thus far we have considered equalizers which accepted one sample per symbol period at their input. In fact the spectrum of the transmitted signal, although usually carefully shaped, exceeds half of the signaling frequency by 10–50%. Thus, the Nyquist theorem is not fulfilled, and as a result of sampling at the symbol rate, the input signal spectra overlap. In consequence, the symbol spaced equalizer is able to correct the overlapped spectrum only. In some disadvantageous cases the overlapping spectra can result in deep nulls in the sampled channel characteristic, which is the subject of equalization. In these spectral intervals the noise will be substantially amplified by the equalizer, which results in deterioration of the system performance.

Derivation of the optimum MSE receiver in the class of linear receivers results in the receiver structure consisting of a filter matched to the impulse observed at the receiver input and an infinite T-spaced transversal filter (see Ref. 3 for details). This derivation also shows that the characteristics $W_0(f)$ of the optimum MSE linear receiver are given by the formula

$$W_0(f) = \frac{\sigma_d^2}{\sigma_v^2} H^*(f) \left(\sum_{i=-\infty}^{\infty} c_i \exp\left[-j2\pi f i T\right] \right) \exp[-j2\pi f t_0] \tag{40}$$

where σ_d^2 is the data symbol mean power and σ_v^2 is the noise power. Lack of a matched filter preceding the transversal filter results in the suboptimality of the receiver and in performance deterioration. In practice, a sufficiently long but finite transversal filter is applied.

The question of whether an optimum receiver can be implemented more efficiently was answered by Macchi and Guidoux [23] as well as by Qureshi and Forney [24].

As we have mentioned, typically the spectrum of the received input signal is limited to the frequency $f_{\max} = (1/2T)(1+\alpha)$, where $\alpha \leq 1$. Let us assume that the noise is also limited to the same bandwidth because of the band-limiting filter applied in the receiver front end. Thus, the bandwidth of the optimal receiver is also limited to the same frequency $f_{\max}$. Because the input signal is spectrally limited to $f_{\max}$, the optimum linear receiver can be implemented by the transversal filter working at the input sampling frequency equal at least to $2f_{\max}$. Let the sampling period $T' = (KT/M)$ be selected to fulfill this condition: $(1/2T') \geq f_{\max}$. K and M are integers of possibly small values. As a result, the following equation holds:

$$H(f) \cdot W_0(f) = H(f) \cdot C_{\text{opt}}(f) \tag{41}$$

where

$$C_{\text{opt}}(f) = \sum_i W_0\left(f - i\frac{1}{T'}\right) \tag{42}$$

We must stress that although the input sampling frequency is $1/T'$, the data symbols are detected each T seconds, so the output of the equalizer is processed at the rate of $1/T$. It is important to note that the channel characteristics is first equalized by the T'-spaced filter and then its output spectrum is overlapped due to sampling the output at the symbol rate. Figure 8 illustrates these processes for $K = 1$ and $M = 2$, specifically, the equalizer is $T/2$-spaced. One can also show that the performance of the fractionally spaced equalizer is independent of the sampling phase [25].

Because the input signal spectrum is practically limited to $|f_{\max}|$, the equalizer can synthesize any characteristics in the frequency range

$$\left(-\frac{1}{2T'}, -f_{\max}\right) \cup \left(f_{\max}, \frac{1}{2T'}\right)$$

without any consequences for the system performance. Thus the optimum fractionally spaced equalizer can have many sets of the optimum coefficients. This phenomenon is disadvantageous from the implementation point of view because the values of the coefficients can slowly drift to unacceptable values. To stabilize the operation of the gradient algorithm, a *tap leakage algorithm* was introduced [26].

6. DECISION-FEEDBACK EQUALIZER

The decision-feedback equalizer (DFE) is the simplest nonlinear equalizer with a symbol-by-symbol detector. It was first described by Austin in 1967 [27]. The in-depth treatment of decision feedback equalization can be found in Ref. 28. It was found that intersymbol interference arising from past symbols can be canceled by synthesizing it using already detected data symbols and subtracting the received value from the sample entering the decision device. Figure 9 presents the basic scheme of the decision-feedback equalizer.

The equalizer input samples are fed to the linear (usually fractionally spaced) adaptive filter which performs matched filtering and shapes the ISI on its output in such a way that the symbol-spaced samples given to the decision device contain the ISI arising from the past symbols only. The ISI resulting from the joint channel and linear filter impulse response is synthesized in the transversal decision-feedback filter. The structure of the DFE is very

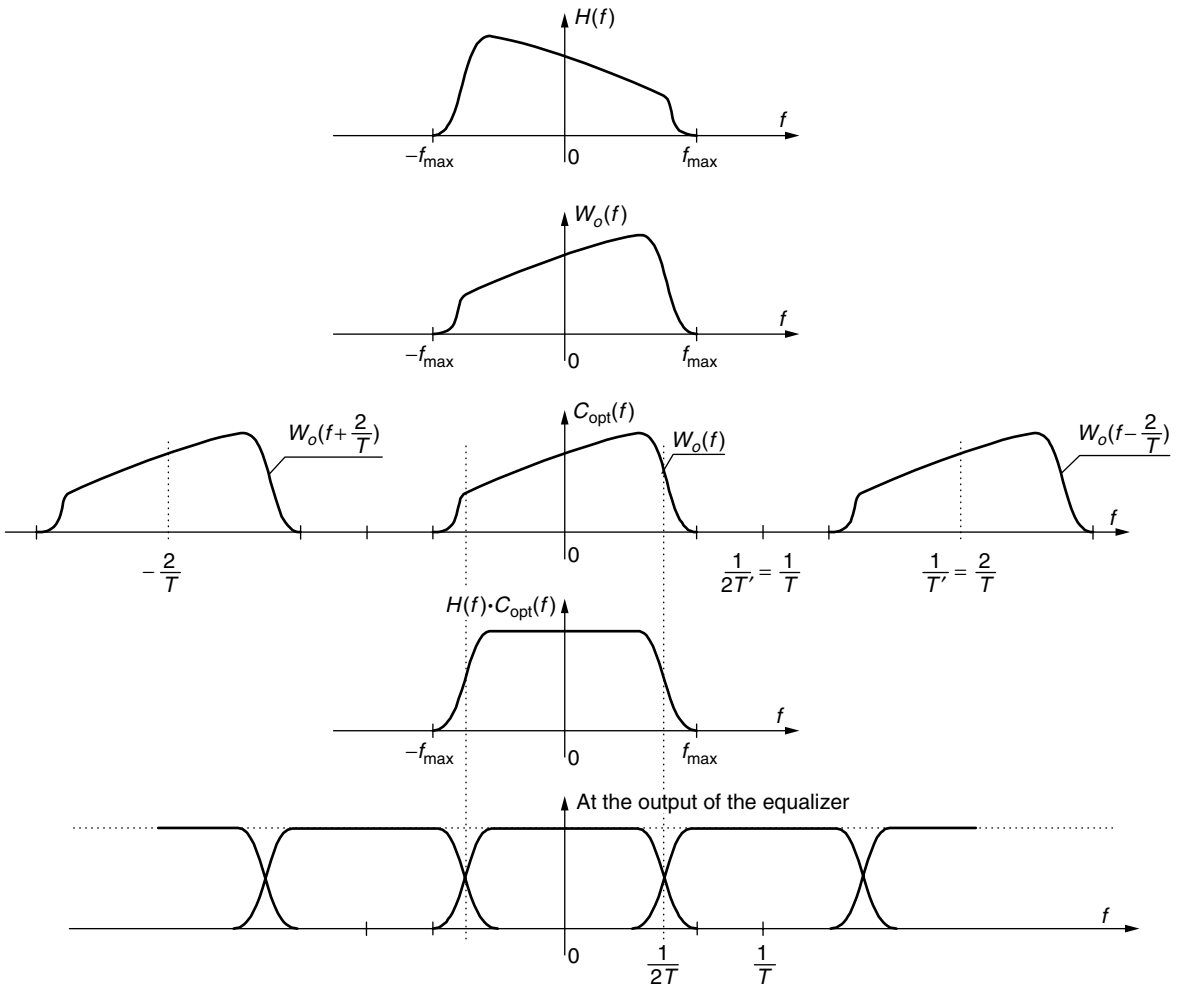


Figure 8. Equalization of the channel spectrum using $T/2$-spaced equalizer.

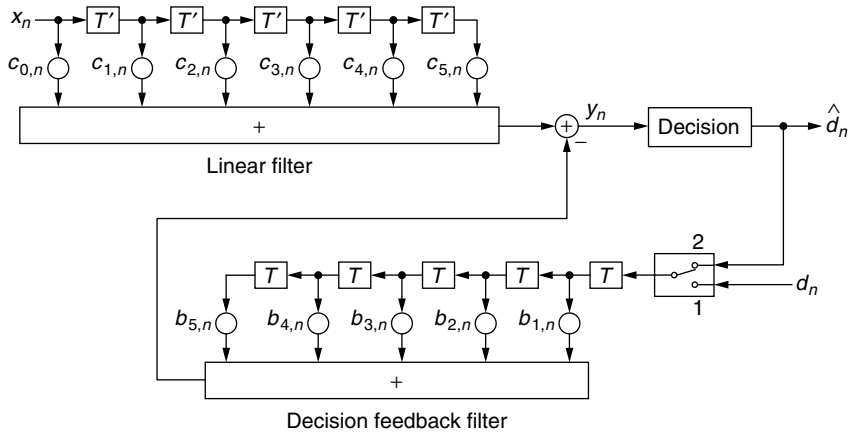


Figure 9. Structure of the decision-feedback equalizer.

similar to the infinite impulse response filter; however, the decision device is placed inside the filter loop, causing the whole structure to be nonlinear. Generally, the operation of the decision feedback equalizer is described by the equation

$$y_n = \sum_{k=-N_1}^{N_2} c_{k,n} x(nT - kT') - \sum_{j=1}^{N_3} b_{j,n} \hat{d}_{n-j} \tag{43}$$

where $c_{k,n}$ are the tap coefficients of the linear filter, $b_{j,n}$ are the tap coefficients of the decision-feedback filter and $\hat{d}_n$ is a data symbol estimate produced by the decision device. In the training mode the data estimates are replaced by the training data symbols.

The decision-feedback equalizer is applied in digital systems operating on channels with deep nulls [29]. Such channels cannot be effectively equalized by the linear equalizers attempting to synthesize the reverse channel characteristics. Instead, the DFE cancels a part of the ISI without inverting the channel and, as a result, the noise in the frequency regions in which nulls in channel characteristics occur is not amplified. Although the DFE structure is very simple and improves the system performance in comparison to that achieved for the linear equalizer, it has some drawbacks as well: (1) part of the signal energy is not used in the decision process because of its cancellation by the decision feedback filter; and (2) because of the decision feedback, errors made in the decision device take part in the synthesis of the ISI as they propagate along the decision feedback filter delay line. Thus, the errors contained in the tapped delay line increase the probability of occurrence of next errors. The phenomenon of error propagation effect can be observed if the signal to noise ratio is not sufficiently high. This effect is discussed, for example, by Lee and Messerschmitt [2].

The DFE tap coefficients can be adjusted according to the ZF or MSE criterion. As for the linear equalizer, the LMS and RLS adaptation algorithms can be used in the DFE. The DFE can be based on transversal or lattice filter structures [30]. Let us concentrate on the LMS algorithm only. We can combine the contents of the tapped delay lines of the linear and decision feedback filters as well as the filter coefficients into single vectors:

$$\mathbf{z}_n = \begin{bmatrix} \mathbf{x}_n \\ \cdots \\ \mathbf{d}_n \end{bmatrix} \quad \mathbf{w}_n = \begin{bmatrix} \mathbf{c}_n \\ \cdots \\ -\mathbf{b}_n \end{bmatrix} \tag{44}$$

where $\mathbf{x}_n = [x_{n+N_1}, \ldots, x_{n-N_2}]^T$, $\mathbf{d}_n = [d_{n-1}, \ldots, d_{n-N_3}]^T$, $\mathbf{c}_n = [c_{-N_1,n}, \ldots, c_{N_2,n}]^T$ and $\mathbf{b}_n = [b_{1,n}, \ldots, b_{N_3,n}]^T$. Then equation (43) can be rewritten in the form

$$y_n = \mathbf{z}_n^T \mathbf{w}_n \tag{45}$$

and the LMS gradient algorithm can be described by the recursive expression

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \beta_n e_n \mathbf{z}_n^* \tag{46}$$

where $e_n = y_n - d_n$. Knowing (44), we can break Eq. (46) into two separate LMS adjustment formulas for the feedforward and feedback filters.

Besides the regular DFE structure shown in Figure 9 there exists the so called predictive DFE [1,28], which, although featuring slightly lower performance, has some advantages in certain applications. Figure 10 presents the block diagram of this structure. The feedforward filter works as a regular linear equalizer according to the ZF or MSE criterion. Its adaptation algorithm is driven by the error signal between the filter output and the data decision (or training data symbol). As we remember, the linear equalizer more or less inverts the channel characteristics, which results in noise amplification. The noise contained in the feedforward filter output samples is correlated due to the filter characteristics. Therefore, its influence can be further minimized applying the linear predictor. Assuming that the decision device makes correct decisions, the noise samples contained in the feedforward filter output are the error samples used in the adaptation of this filter. The linear combination of the previous noise samples allows to predict the new sample, which is subsequently subtracted from the feedforward filter output. This way the effective SNR is increased. The result of subtraction constitutes the basis for decisionmaking.

Let us note (see Fig. 10) that the feedforward filter and the predictor are adjusted separately, so the performance

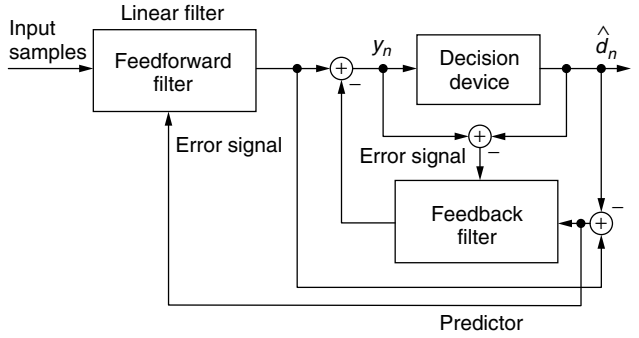


Figure 10. Predictive DFE.

of the predictive DFE is worse than the performance of the conventional DFE for which the taps adjustments are realized on the basis of the final output error. It has been shown that the predictive DFE is useful in realization of the joint trellis code decoder and channel equalizer [31].

7. EQUALIZERS USING MAP SYMBOL-BY-SYMBOL DETECTION

The decision-feedback equalizer is a particularly simple version of a nonlinear receiver in which the decision device is some kind of an M-level quantizer, where M is the number of data symbols. Much more sophisticated detectors have been developed which minimize the symbol error probability. This goal is achieved if the *maximum a posteriori probability* (MAP) criterion is applied. Let us consider the receiver structure shown in Fig. 11. The linear filter preceding the detection algorithm is a *whitened matched filter* (WMF). Its function is very similar to that of the linear filter applied in the decision feedback equalizer. It shapes the joint channel and linear filter impulse response to receive ISI arising from the past data symbols only. At the same time the noise samples at the output of the WMF are white. We say that the signal at the output of the whitened matched filter constitutes a *sufficient statistic* for detection. This roughly means that that part of the received signal that has been removed by the WMF is irrelevant for detection. Assuming that the number of interfering symbols is finite, we can write the following equation describing the sample y_n at the detector input:

$$y_n = \sum_{i=0}^{N} b_i d_{n-i} + \nu_n \tag{47}$$

where ν_n is a white Gaussian noise sample. Let us note that the information on the data symbol d_n is "hidden" in the samples $y_n, y_{n+1}, \ldots, y_{n+N}$. Generally, according to the MAP criterion the detector finds that $\hat{d}_n$ among

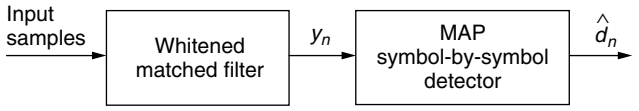


Figure 11. Basic scheme of the MAP symbol-by-symbol equalizer.

all possible M data symbols for which the following a posteriori probability is maximum:

$$\Pr\{d_n|\mathbf{y}_{n+N}\} \tag{48}$$

where $\mathbf{y}_{n+N} = (y_{n+N}, y_{n+N-1}, \ldots, y_1)$ is the vector of the observed input samples. From the Bayes theorem we know that for expression (48) the following equality holds:

$$\Pr\{d_n = m|\mathbf{y}_{n+N}\} = \frac{p(\mathbf{y}_{n+N}|d_n = m)\Pr\{d_n = m\}}{p(\mathbf{y}_{n+N})} \tag{49}$$

Because $p(\mathbf{y}_{n+N})$ is common for all possible probabilities (49), it has no meaning in the search for the data symbol featuring the MAP probability. Thus the task of the MAP detector can be formulated in the following manner

$$\hat{d}_n = \arg\left\{\max_{d_n} p(\mathbf{y}_{n+N}|d_n)\Pr\{d_n\}\right\} \tag{50}$$

Finding the data estimate (50) is usually computationally complex. Several algorithms have been proposed to realize (50). Abend and Fritchman [22] as well as Chang and Hancock [32] algorithms (the last being analogous to the well known BCJR algorithm [33] applied in convolutional code decoding) are good examples of these methods. We have to stress that all of them require the knowledge of the impulse response $\{b_i\}, (i = 1, \ldots, N)$ to calculate values of the appropriate conditional probability density functions. This problem will also appear in the MLSE receiver discussed in the next section.

8. MAXIMUM-LIKELIHOOD EQUALIZERS

Instead of minimizing the data symbol error, we could select minimization of error of the whole sequence as the optimization goal of the receiver. Thus, the MAP criterion yields the form

$$\max_{\mathbf{d}_n} P(\mathbf{d}_n|\mathbf{y}_n) \tag{51}$$

If the data sequences are equiprobable, our criterion is equivalent to the selection of such a data sequence that maximizes the conditional probability density function $p(\mathbf{x}_n|\mathbf{d}_n)$. Namely, we have

$$\begin{aligned}\hat{\mathbf{d}}_n &= \arg\left\{\max_{\mathbf{d}_n} P(\mathbf{d}_n|\mathbf{y}_n)\right\} = \arg\left\{\max_{\mathbf{d}_n} \frac{p(\mathbf{y}_n|\mathbf{d}_n)P(\mathbf{d}_n)}{p(\mathbf{y}_n)}\right\} \\ &= \arg\left\{\max_{\mathbf{d}_n} p(\mathbf{y}_n|\mathbf{d}_n)\right\}\end{aligned} \tag{52}$$

where, as before, $\mathbf{y}_n = (y_1, \ldots, y_n)^T$, $\mathbf{d}_n = (d_1, \ldots, d_n)^T$. Because noise at the WMF output is white and Gaussian, the conditional probability density function can be expressed by the formula

$$\begin{aligned}p(\mathbf{y}_n|\mathbf{d}_n) &= \prod_{i=1}^{n} p(y_i|\mathbf{d}_i) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \prod_{i=1}^{n} \exp\left[-\frac{\left|y_i - \sum_{k=0}^{N} b_k d_{i-k}\right|^2}{2\sigma^2}\right]\end{aligned} \tag{53}$$

Calculating the natural logarithm of both sides of (53), we obtain

$$\hat{\mathbf{d}}_n = \arg\left\{\max_{\mathbf{d}_n} \ln p(\mathbf{y}_n|\mathbf{d}_n)\right\} = \arg\left\{\min_{\mathbf{d}_n} \sum_{i=1}^{n} \left| y_i - \sum_{k=0}^{N} b_k d_{i-k} \right|^2 \right\} \tag{54}$$

Concluding, from all possible equiprobable data sequences $\mathbf{d}_n$ this sequence $\hat{\mathbf{d}}_n$ is selected for which the sum

$$S_n = \sum_{i=1}^{n} \left| y_i - \sum_{k=0}^{N} b_k d_{i-k} \right|^2 \tag{55}$$

is minimum. It was found by Forney [12] that the effective method of searching for such a sequence is the *Viterbi algorithm* See VITERBI ALGORITHM. Let us note that in order to select the data sequence the samples of the impulse response $\{b_k\}$, $k = 0, \ldots, N$ have to be estimated. They are usually derived on the basis of the channel impulse response $\{h_k\}$, $k = -N_1, \ldots, N_2$. The scheme of such a receiver is shown in Figure 12. The heart of the receiver is the Viterbi detector fed with the impulse response samples $\{b_k\}$ calculated on the basis of the channel impulse response samples $\{h_k\}$ estimated in the *channel estimator*. The channel estimator is usually an adaptive filter using the LMS or RLS algorithm for deriving the impulse response samples. From the system theory point of view it performs system identification. The channel estimator input signal is the data reference signal or the final or preliminary decision produced by the Viterbi detector. The channel output signal acts as a reference signal for the channel estimator. Usually, the reference signal has to be appropriately delayed in order to accommodate the decision delay introduced by the Viterbi detector.

For example, let us consider the channel estimator using the LMS algorithm and driven by ideal data symbols. Let us neglect the delay with which the data symbols are fed to the estimator. Assume that the data symbols are uncorrelated. Then, applying the mean-square error as the criterion for the estimator, we have

$$\mathcal{E}_n = E\left[|e_n|^2\right] = E\left[\left| x_n - \sum_{j=-N}^{N} \hat{h}_{j,n} d_{n-j} \right|^2\right] \tag{56}$$

where x_n is the channel output sample [see Eq. (4)] and $\hat{h}_{j,n}, j = -N, \ldots, N$ are the estimates of the channel impulse response at the nth moment. The calculation of the gradient of error $\mathcal{E}_n$ with respect to the channel impulse response estimate $\hat{h}_j$ gives

$$\frac{\partial \mathcal{E}_n}{\partial \hat{h}_{j,n}} = -2E[e_n d^*_{n-j}] \tag{57}$$

Therefore the stochastic gradient algorithm for the adjustment of channel impulse response estimates is

$$\hat{h}_{j,n+1} = \hat{h}_{j,n} - \alpha_n e_n d^*_{n-j} \quad j = -N, \ldots, N \tag{58}$$

where α_n is an appropriately selected step size. It can be shown that the initial step size should be $\alpha_0 = 1/(2N+1)E[|d_n|^2]$.

Another solution for deriving the channel impulse response is to use a zero-autocorrelation periodic training sequence. A fast channel estimator using such sequence is applied, for example, in the GSM receiver. Part of the known sequence placed in the middle of the data burst, called *midamble*, is a zero-autocorrelation periodic training sequence. In this case the channel impulse response samples are estimated on the basis of the following formula:

$$\hat{h}_i = \sum_{j=-N}^{N} x_j d^*_{i-j} \tag{59}$$

Thus, the received signal, which is the response of the channel to the periodic training signal, is cross-correlated

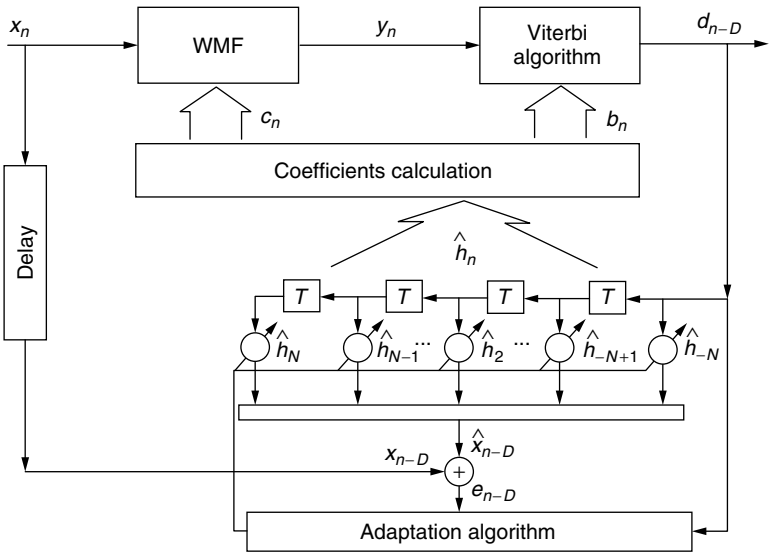


Figure 12. Basic scheme of the MLSE receiver with the whitened matched filter and the Viterbi algorithm.

with the complex conjugate of the training sequence. On the basis of the estimated impulse response samples $\hat{h}_i$ the receiver calculates the WMF coefficients and the weights $\{b_k\}$ used by the Viterbi detector.

An alternative equivalent structure of the MLSE equalizer was proposed by Ungerboeck [34]. Its derivation following Ungerboeck's considerations can be also found in Proakis' handbook [1]. Instead of the whitened matched filter, the filter matched to the channel impulse response is applied and the Viterbi detector maximizes the following cost function C_n with respect to the data sequence

$$C_n = \sum_{i=1}^{n} \mathrm{Re}\left[d_i^* \left(2y_i - g_0 d_i - 2\sum_{k=1}^{N} g_k d_{i-k} \right) \right] \qquad (60)$$

where y_i is the sample at the matched filter output at the ith moment and g_k $(k = 0, \ldots, N)$ are the samples of the channel impulse response autocorrelation function. See Ref. 34, 1, or 3 for details and for derivation of the algorithm.

Closer investigation of formula (54) allows us to conclude that in order to minimize the cost function and find the optimum data sequence, M^N operations (multiply and add, compare etc.) have to be performed for each timing instant. M is the size of the data alphabet. If modulation is binary $(M = 2)$ and the length of ISI is moderate, the detection algorithm is manageable. This is the case of the GSM receiver. However, if M is larger and/or the ISI corrupts a larger number of modulation periods, the number of calculations becomes excessive and suboptimal solutions have to be applied. Three papers [13,14,35] present examples of suboptimum MLSE receivers.

9. EQUALIZERS FOR TRELLIS-CODED MODULATIONS

In 1982 Ungerboeck published a paper [36] in which he proposed a joint approach to modulation and coding applied on band-limited channels. At the cost of expansion of the data signal constellation, application of a convolutional code and an appropriate binary data block-to-data symbol mapping, interdependence of subsequent data signals is obtained. Therefore, in order to select the maximum likelihood sequence among all possible sequences, whole data sequences have to be compared in the decision process. The distance between two closest data sequences is larger than between two uncoded data symbols and, in consequence, the system using *trellis-coded modulation* (TCM) is more robust against errors than the uncoded system transmitting data at the same data rate. The detection of the trellis-coded data stream requires sequential algorithm such as the Viterbi algorithm.

Using TCM signals on the ISI channels requires adaptive equalization and TCM decoding. The TCM detection process of the whole symbol sequences creates some problems in the selection of the equalizer structure and in the adjustment of the equalizer coefficients. The standard solution is to apply a linear equalizer minimizing the ISI followed by the TCM Viterbi decoder. The equalizer coefficient updates can be done using unreliable tentative decisions or the reliable but delayed decisions from the

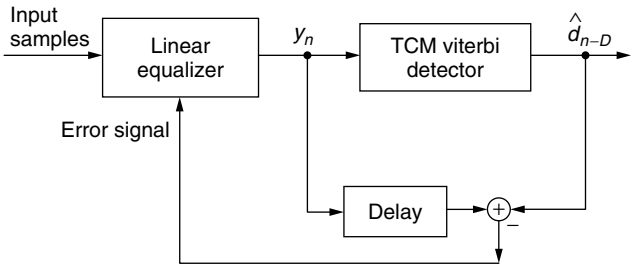


Figure 13. Linear equalizer with trellis-coded modulation decoder.

TCM Viterbi decoder [37]. In case of the LMS algorithm applied in the equalizer, the consequence of the delayed error signal (see Fig. 13) is the necessity of decreasing the step size [37].

On some channels, in particular those featuring a long tail in the channel impulse response or possessing deep nulls in their characteristics, applying a decision-feedback equalizer is more advantageous. Using joint DFE and trellis coding requires some special solutions because the DFE uses symbol-by-symbol decisions in its feedback filter with a single delay. One solution to this problem is to apply an interleaver between the TCM encoder and the modulator at the transmitter and the predictive DFE with the deinterleaver between the linear part of the equalizer and the decision-feedback part incorporating the TCM Viterbi decoder and predictor [1]. Another solution which is applicable in systems with a feedback channel and operating on transmission channels that are stationary or slowly change in time, is to share the DFE equalization between transmitter and receiver. In this case the concept of Tomlinson precoding applied jointly with the TCM coding is very useful [38].

The optimum receiver for TCM signals received in the presence of the ISI has been shown [31]. Its structure is basically the same as that shown in Fig. 12; however, now the Viterbi detector operates on the supertrellis resulting from concatenation of the ISI and TCM code trellises. See VITERBI ALGORITHM. Because the number of supertrellis states and the computational complexity associated with them are very high, suboptimum solutions have to be applied. The most efficient one is to incorporate intersymbol interference into the decision feedback for each supertrellis state, using the data sequences that constitute the "oldest" part of the maximum-likelihood data sequence associated with each state (so called *survivor*). In fact, this idea is already known from the delayed decision-feedback sequence estimation [14] used for uncoded data.

10. BLIND ADAPTIVE EQUALIZATION

As we have already mentioned, in some cases sending a known data sequence to train the equalizer can result in wasting of a considerable part of transmission time. One of the cases where adaptive equalization without a training sequence is applied is the transmission of *digital video broadcasting* (DVB) datastream in a DVB cable distribution system. A DVB cable receiver, after being

switched on, has to compensate intersymbol interference on the basis of the received signal and the general knowledge of its properties.

Blind equalization algorithms can be divided into three groups:

- The Bussgang [39] algorithms, which apply the gradient-type procedure with nonlinear processing of the filter output signal in order to obtain a reference signal conforming to the selected criterion,
- Second- and higher-order spectra algorithms, which apply higher-order statistics of the input signals in order to recover the channel impulse response and subsequently calculate the equalizer coefficients,
- Probabilistic algorithms, which realize the ML or MAP sequence estimation or suboptimum methods.

The algorithms belonging to the first category are easiest to implement. They will be described below.

The theory of blind equalization presented by Benveniste et al. [40] shows that in order to adjust the linear equalizer properly, one should drive its coefficients in such a way that the instantaneous probability distribution of the equalizer output y_n converges to the data input signal probability distribution $p_D(y)$. However, one important condition has to be fulfilled — the probability density function of the input signal d_n must be different from the Gaussian one. It has been found that the ISI introduced by the channel distorts the shape of the input probability density function unless it is Gaussian.

The main difficulty in designing the equalizer's adaptation algorithm is finding a criterion which, when minimized with respect to equalizer's coefficients, results in (almost) perfect channel equalization. One approach is to calculate the error

$$e_n = y_n - g(y_n) \tag{61}$$

which is to be minimized in the MSE sense, where $g(y_n)$ is an "artificially" generated "reference signal" and $g(.)$ is the memoryless nonlinearity. Thus, the general criterion that is the subject of minimization with respect to the coefficient vector $\mathbf{c}_n$ is

$$\mathcal{E}_n = E[|e_n|^2] = E[|y_n - g(y_n)|^2] \tag{62}$$

A typical approach to finding the minimum of $\mathcal{E}_n$ is to change the equalizer's coefficients in the direction opposite that indicated by the current gradient of ε_n, calculated with respect to $\mathbf{c}_n$. If we assume that all the signals and filters are complex, we get the following "reference" and error signals:

$$\tilde{y}_n = g(\mathrm{Re}(y_n)) + jg(\mathrm{Im}(y_n)) \quad \tilde{e}_n = y_n - \tilde{y}_n \tag{63}$$

Calculation of the gradient of $\mathcal{E}_n$ leads to the result

$$\operatorname*{grad}_{\mathbf{c}_n} \mathcal{E}_n = 2E\left\{\left[\mathrm{Re}(\tilde{e}_n)(1 - g'(\mathrm{Re}(y_n))) + j\,\mathrm{Im}(\tilde{e}_n)(1 - g'(\mathrm{Im}(y_n)))\right]\mathbf{x}_n^*\right\} \tag{64}$$

In practice the derivative $g'(.)$ is equal to zero except for a few discrete values of its argument. Thus, the stochastic version of the gradient algorithm achieves the well-known form

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha\tilde{e}_n\mathbf{x}_n^* \tag{65}$$

where this time the error signal $\tilde{e}_n$ is described by Eq. (63). Unfortunately, the optimum nonlinear function $g(.)$ is difficult to calculate. Bellini [39] investigated this function with several simplifying assumptions. Generally, function $g(.)$ should vary during the equalization process. Most of the gradient-based adaptation algorithms are in fact examples of the Bussgang technique, although they were found independently of it. Below we list the most important versions of the gradient algorithms, quoting the error signals that are characteristic for them.

- *Sato algorithm*: $\tilde{e}_n = e_n^S = y_n - A_S\mathrm{csgn}(y_n)$, where $\mathrm{csgn}(y_n) = \mathrm{sgn}(\mathrm{Re}(y_n)) + j\mathrm{sgn}(\mathrm{Im}(y_n))$, A_S is the weighting center of the in-phase and quadrature data signal components,
- *Benveniste–Goursat algorithm*: $\tilde{e}_n = e_n^B = k_1 e_n + k_2 |e_n| e_n^S$, $e_n = y_n - \mathrm{dec}(y_n)$, k_1, k_2 are properly selected weighting coefficients,
- *Stop-and-go algorithm*: $\tilde{e}_n = e_n^{SG} = f_n^R\mathrm{Re}(e_n) + jf_n^I\mathrm{Im}(e_n)$, $e_n = y_n - \mathrm{dec}(y_n)$, the weighting factors f_n^R, f_n^I turn on and off the in-phase and quadrature components of the decision error depending on the probability of the event that these components indicate the appropriate direction of the coefficients' adjustment,
- *Constant-modulus (CM) algorithm*: $\tilde{e}_n = e_n^G = (|y_n|^2 - R_2)y_n$, where R_2 is a properly selected data constellation radius.

The CM algorithm, although the most popular among the four described above, loses information about the phase of the received signal. Therefore, it has to be supported by the phase-locked loop in order to compensate for the phase ambiguity.

The second group of blind algorithms applied in channel estimation or equalization are the algorithms using the methods of the higher-order statistics of the analyzed signal.

A survey of the higher-order statistics applied in adaptive filtering can be found in Haykin's book [4]. Let us concentrate on the second-order statistics methods, showing an example of such an algorithm [41]. If the signal on the channel output

$$x(t) = \sum_{k=0}^{L-1} d_k h(t - kT) + n(t) \tag{66}$$

is sampled once per data symbol period, it is not possible to identify the samples of the channel impulse response based on the autocorrelation function of these samples. However, it is possible to do this if the signal is oversampled or received from the antenna arrays, that is, if more samples per data symbol are processed by the algorithm. If the

signal is sampled m times in each data period T, then it can be expressed in the vector form as

$$\mathbf{x}_n = \begin{bmatrix} x_{1,n} \\ \cdots \\ x_{m,n} \end{bmatrix} = \sum_{k=0}^{L-1} \begin{bmatrix} h_{1,k} \\ \cdots \\ h_{m,k} \end{bmatrix} d_{n-k} + \begin{bmatrix} \nu_{1,n} \\ \cdots \\ \nu_{m,k} \end{bmatrix} = \sum_{k=0}^{L-1} \mathbf{h}_i d_{n-i} + \boldsymbol{\nu}_n \tag{67}$$

On the basis of the signal vectors $\mathbf{x}_n$, we can estimate the matrices

$$C_x(i) = \frac{1}{N-i} \sum_{k=i+1}^{N} \mathbf{x}_k \mathbf{x}_{k-i}^* \; i = -L+1, \ldots, L-1 \tag{68}$$

and calculate the power density spectrum estimate in the matrix form

$$Q(e^{j\omega}) = \sum_{i=-L+1}^{L-1} C_x(i) e^{j\omega i} \tag{69}$$

By eigendecomposition of Q we can obtain the principal eigenvector, which is also described by the equation

$$\mathbf{c}(\omega) = e^{j\alpha(\omega)} \sum_{k=0}^{L-1} \mathbf{h}_k e^{j\omega k} \tag{70}$$

In order to calculate the channel impulse response samples, we take $N > 2L+1$ uniform samples of $\mathbf{c}(\omega_i)$ and form an appropriate system of linear equations. The solution of the system is the set of samples of the channel impulse response and weights $a_k = \exp(j\alpha(\omega_i))$. The estimated channel impulse response can be subsequently applied in the sequential algorithm or serve as the basis for the calculation of the equalizer coefficients. The simulation results reported by Xu et al. [41] show that satisfactory results can be achieved already with $N = 50$ sample blocks. A low number of input signal samples necessary for reliable channel estimation is the main advantage of the methods using second-order statistics as compared with the Bussgang techniques using the gradient algorithm. The price paid for fast channel estimation and equalization is high computational complexity of the algorithms.

The algorithms applying higher than second-order statistics generally use the cumulants of the input signal samples and their Fourier transforms called polyspectra. Polyspectra provide the basis for the nonminimum phase channel identification, thanks to their ability to preserve phase information of the channel output signal. The main drawback of the higher-order statistical methods is an extensive number of signal samples necessary to estimate the cumulants with sufficient accuracy and high computational complexity of the algorithms.

The third group of blind equalization algorithms relies on the joint channel estimation and data detection. Usually the maximum-likelihood criterion is applied, which in the blind case is expressed in the form

$$\arg p(\mathbf{x}_n | \mathbf{h}, \mathbf{d}_n) = \arg_{\mathbf{h},\mathbf{d}_n} \frac{1}{(2\pi\sigma^2)^n} \times \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left| x_i - \sum_{k=0}^{L-1} h_k d_{i-k} \right|^2 \right] \tag{71}$$

The channel identification can be performed after reception of the signal sequence $\mathbf{x}_n = (x_1, \ldots, x_n)$ by averaging the probability density function over all possible data sequences of length n. Subsequently, the Viterbi algorithm can be performed that finds the best data estimate in the ML sense. There are several other versions of joint channel data sequence estimation (see Ref. 42 as a representative example).

11. CONCLUSIONS

In this tutorial we have concentrated on the problem of equalization for point-to-point transmission. Limited space did not allow us to describe many other important issues such as *multiple-input/multiple-output* (MIMO) equalizers [43], the principle of per survivor processing [44], or equalization and MLSE detection performed jointly with diversity reception in mobile radio channels [e.g., 45]. Other interesting subjects are adaptive equalization in the frequency domain and adaptive channel equalization in the OFDM (orthogonal frequency-division multiplexing) systems.

BIOGRAPHY

Krzysztof Wesołowski was born in 1952. He received a M.Sc. degree in electrical engineering from Poznań University of Technology, Poznań, Poland, in 1976, a M.A. in Mathematics (*cum laude*) from Adam Mickiewicz University, Poznań, Poland, in 1978, a Ph.D. in 1982, and a Dr *Habilitus* degrees in 1989 in telecommunications. Currently, he holds a position of the professor of electrical engineering at the same university and leads the research group of wireless communications. He has published over 90 papers in Polish, English, and German. He is the author of the book *Mobile Communication Systems* published in Polish (1998, 1999). Its updated translation has been published in 2002 by John Wiley & Sons, Ltd. He spent his sabbatical leaves at Northeastern University, Boston, (Postdoctoral Fulbright Scholarship) and at the University of Kaiserslautern, Germany, (Alexander von Humboldt Scholarship). At the latter university he also served as a visiting professor teaching courses on adaptive equalization, information theory, and coding.

His main interests concentrate on the physical layer of digital communication systems, in particular on adaptive equalization, signal detection, error control coding, and other transmit and receive techniques applied to wireless communications.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1996.
2. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, 2nd ed., Kluwer, Boston, 1995.
3. R. D. Gitlin, J. F. Hayes, and S. B. Weinstein, *Principles of Data Communications*, Plenum Press, New York, 1992.
4. S. Haykin, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, Englewood-Cliffs, NJ, 1991.

5. A. P. Clark, *Equalizers for Digital Modems*, Pentech Press, London, 1985.
6. Zh. Ding and Ye Li, *Blind Equalization and Identification*, Marcel Dekker, New York, 2001.
7. O. Macchi, *Adaptive Processing*, Wiley, Chichester, UK, 1995.
8. S. U. H. Qureshi, Adaptive equalization, *Proc. IEEE* **53**: 1349–1387 (1985).
9. D. P. Taylor, G. M. Vitetta, B. D. Hart, and A. Mämmelä, Wireless channel equalisation, *Eur. Trans. Telecommun.* **9**: 117–143 (1998).
10. E. H. Satorius and S. T. Alexander, Channel equalization using adaptive lattice algorithms, *IEEE Trans. Commun.* **COM-27**: 899–905 (1979).
11. P. Monsen, Feedback equalization for fading dispersive channels, *IEEE Trans. Inform. Theory* **IR-17**: 56–64 (1971).
12. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363–378 (1972).
13. M. V. Eyuboglu and S. U. H. Qureshi, Reduced-state sequence estimation with set partitioning and decision feedback, *IEEE Trans. Commun.* **36**: 13–20 (1988).
14. A. Duel-Hallen and C. Heegard, Delayed decision-feedback sequence estimation, *IEEE Trans. Commun.* **37**: 428–436 (1989).
15. J. B. Anderson and S. Mohan, Sequential decoding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**: 169–176 (1984).
16. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*, McGraw-Hill, New York, 1968.
17. G. Ungerboeck, Theory on the speed of convergence in adaptive equalizers for digital communication, *IBM J. Res. Devel.* **16**: 546–555 (1972).
18. D. N. Godard, Channel equalization using a Kalman filter for fast data transmission, *IBM J. Res. Devel.* **18**: 267–273 (1974).
19. J. M. Cioffi and T. Kailath, Fast recursive least-squares transversal filter for adaptive filtering, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-32**: 304–337 (1984).
20. E. H. Satorius and J. D. Pack, Application of least squares lattice algorithms to adaptive equalization, *IEEE Trans. Commun.* **COM-29**: 136–142 (1981).
21. P. R. Chevillat, D. Maiwald, and G. Ungerboeck, Rapid training of a voiceband data-modem receiver employing an equalizer with fractional-T spaced coefficients, *IEEE Trans. Commun.* **COM-35**: 869–876 (1987).
22. K. Abend and B. D. Fritchman, Statistical detection for communication channels with intersymbol interference, *Proc. IEEE* 779–785 (1970).
23. O. Macchi and L. Guidoux, A new equalizer and double sampling equalizer, *Ann. Telecommun.* **30**: 331–338 (1975).
24. S. U. H. Qureshi and G. D. Forney, Jr., Performance and properties of a T/2 equalizer, *Conf. Record, National Telecommunication Conf.*, 1977.
25. G. Ungerboeck, Fractional tap-spacing equalizer and consequence for clock recovery in data modems, *IEEE Trans. Commun.* **COM-24**: 856–864 (1976).
26. R. D. Gitlin, H. C. Meadors, and S. B. Weinstein, The tap leakage algorithm: An algorithm for the stable operation of a digitally implemented, fractionally spaced adaptive equalizer, *Bell Syst. Tech. J.* **61**: 1817–1839 (1982).
27. M. E. Austin, *Equalization of Dispersive Channels Using Decision Feedback*, Research Laboratory of Electronics, MIT, Cambridge, MA., QPR 84, 1967, pp. 227–243.
28. C. A. Belfiore and J. H. Park, Jr., Decision feedback equalization, *Proc. IEEE* **67**: 1143–1156 (1979).
29. P. Monsen, Feedback equalization for fading dispersive channels, *IEEE Trans. Inform. Theory* **IT-17**: 56–64 (1971).
30. F. Ling and J. G. Proakis, Adaptive lattice decision-feedback equalizers — their performance and application to time-variant multipath channels, *IEEE Trans. Commun.* **COM-33**: 348–356 (1985).
31. P. R. Chevillat and E. Eleftheriou, Decoding of trellis-encoded signals in the presence of intersymbol interference and noise, *IEEE Trans. Commun.* **37**: 669–676 (1989).
32. R. W. Chang and J. C. Hancock, On receiver structures for channel having memory, *IEEE Trans. Inform. Theory* **IT-12**: 463–468 (1966).
33. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (1974).
34. G. Ungerboeck, Adaptive maximum-likelihood receiver for carrier-modulated data transmission systems, *IEEE Trans. Commun.* **COM-22**: 624–636 (1974).
35. K. Wesolowski, An efficient DFE & ML suboptimum receiver for data transmission over dispersive channels using two-dimensional signal constellations, *IEEE Trans. Commun.* **COM-35**: 336–339 (1987).
36. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**: 55–67 (1982).
37. G. Long, F. Ling, and J. G. Proakis, The LMS algorithm with delayed coefficient adaptation, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-37**: (1989).
38. A. K. Aman, R. L. Cupo, and N. A. Zervos, Combined trellis coding and DFE through Tomlinson precoding, *IEEE J. Select. Areas Commun.* **9**: 876–883 (1991).
39. S. Bellini, Bussgang techniques for blind equalization, *Proc. GLOBECOM'88*, 1988, pp. 1634–1640.
40. A. Benveniste, M. Goursat, and G. Ruget, Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications, *IEEE Trans. Autom. Control* **AC-25**: 385–398 (1980).
41. G. Xu, L. Tong, and H. Liu, A new algorithm for fast blind equalization of wireless communication channels, *Proc. GLOBECOM'94*, 1994, pp. 544–548.
42. N. Seshadri, Joint data and channel estimation using blind trellis search techniques, *IEEE Trans. Commun.* **42**: 1000–1011 (1994).
43. A. Duel-Hallen, Equalizers for multiple input/multiple output channels and PAM Systems with cyclostationary input sequences, *IEEE J. Select. Areas Commun.* **10**: 630–639 (1992).
44. R. Raheli, A. Polydoros, and Ch.-K. Tzou, The principle of per-survivor processing: A general approach to approximate and adaptive MLSE, *Proc. GLOBECOM'91*, 1991, pp. 1170–1175.
45. R. Krenz and K. Wesołowski, Comparative study of space-diversity techniques for MLSE receivers in mobile radio, *IEEE Trans. Vehic. Technol.* **46**: 653–663 (1997).

ADAPTIVE RECEIVERS FOR SPREAD-SPECTRUM SYSTEMS

URBASHI MITRA
Communication Sciences Institute
Los Angeles, California

1. INTRODUCTION

The explosive growth of wireless communications has motivated the "re"consideration of spread-spectrum techniques for multiuser communications. As the phrase suggests, "multiuser" communication systems offer communication services to multiple users simultaneously. Our focus is on a system as depicted by Fig. 1. In such a system, multiple users share a communications channel. The term "channel" is both abstract and physical; it describes the link between the transmitter(s) and the receiver(s). Thus, it could refer to free space or even a body of water. For free space, the channel is typically defined by a band of frequencies and characterized by the physical topology between the transmitter(s) and the receiver(s). This multiuser system can also be termed a *multipoint-to-point* communications system in contrast with a *point-to-multipoint* or *broadcast* system as employed for broadcast radio transmission and broadcast television. In broadcast channels, a single information stream is transmitted from a centralized transmitter to be received by multiple receivers. The objective of the receiver in our multipoint-to-point or *multiple-access* system is to ultimately demodulate the information stream of one, some, or all of the active users in the system. The receiver is thus the recipient of the different information signals of multiple users. Examples of multiuser communication systems include cellular communications, local-area networks, computer communications networks (such as the Internet), telephone networks, and packet-radio networks. Note that while Fig. 1 distinguishes the interference, or additive noise, that can be contributed by the channel, from the contributions from the individual users, in a sense, each active user in the system can represent a noise source for every other user in the system. The challenge of designing a receiver that operates well in a multiuser environment is to mitigate the effects of both the interfering users as well as the effects inherent to the wireless channel due to propagation and ambient channel noise.

Spread-spectrum signaling while somewhat bandwidth inefficient relative to more traditional narrowband signaling schemes offers certain advantages for radio communication systems. The wideband nature of the signal facilitates channel estimation and enables the resolution of multipath (described in more detail in Section 5). Multipath occurs to the presence of obstructions such as buildings, trees, and vehicles in the path between transmitter and receiver. Because of these obstructions, the transmitted signal is reflected, absorbed, and diffused; the received signal is in fact a sum of delayed and attenuated replicas of the originally transmitted signal. The access schemes associated with spread-spectrum technology tend to be more flexible. Statistical multiplexing can be exploited since all active users have bursty communication. Thus, there is potential for a capacity increase relative to narrowband signaling systems.

We shall focus on adaptive schemes for data detection; however, adaptive algorithms can also be developed for adaptive estimation of key communication parameters such as the channel, the number of active users, timing information, etc.

1.1. Access Methods

The emphasis of this entry is on signaling and detection methods appropriate for multiuser radio communications; however, it is observed that multiuser demodulation methods have found application in a variety of other fields including radar signal processing and medical imaging. There are many ways in which the radio channel resource can be shared. Two more classical methods are frequency-division multiple access (FDMA) and time-division multiple access (TDMA) (see Fig. 2). For illustrative purposes, one can view the communications resource as having two dimensions: frequency and time. In reality, other dimensions are available such as space [48,60].

The first wireless mobile communications system, the advanced mobile phone service (AMPS) employed FDMA as the multiuser access technology in 1977. In FDMA, each user is assigned a frequency band; the user can communicate continuously employing this frequency "slot." Commercial radio and broadcast television employ FDMA as a method of transmitting many different station signals to an individual receiver. In TDMA, each active user is assigned a nonoverlapping time slot. During its

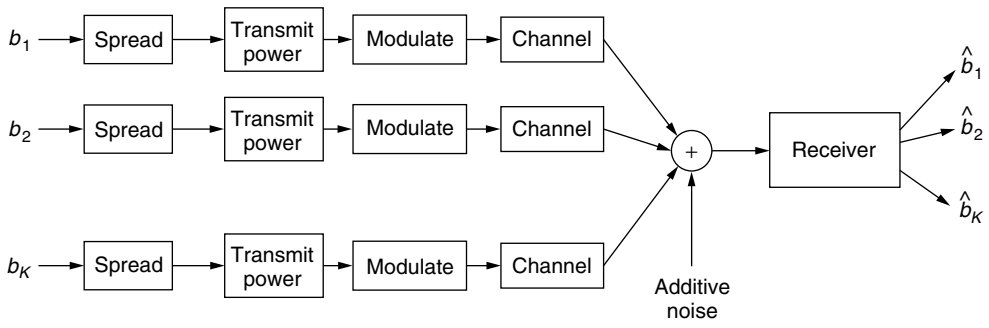


Figure 1. Multiuser system.

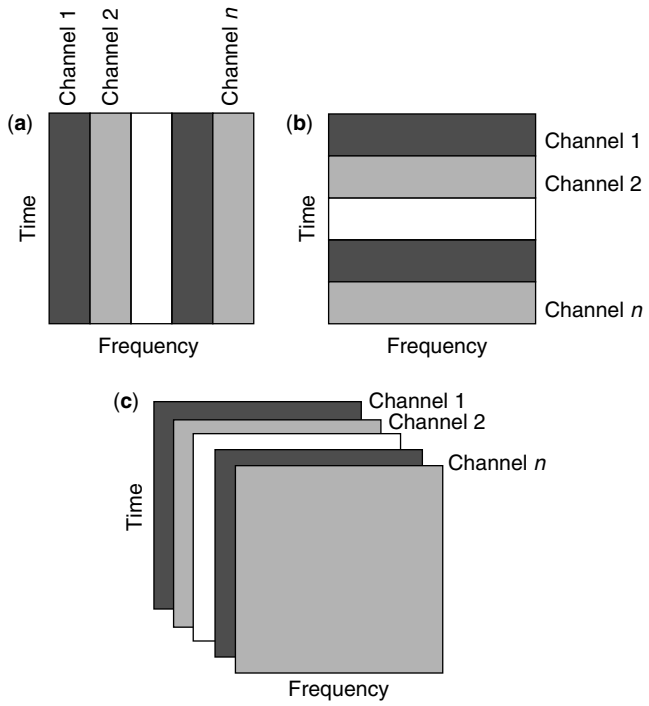


Figure 2. Multiple-access methods: (**a**) frequency division; (**b**) time division; (**c**) code division.

assigned time slot, the active user transmits over the entire frequency band allocated to the TDMA service. The TDMA access scheme can be considered to be the dual of FDMA: users are assigned non-overlapping time slots and utilize the entire frequency band during this time slot. Users share the resource by communicating one at a time, in round-robin fashion. Currently, there are three variations of TDMA in use for commercial wireless communications. The Global System for Mobile Communications (GSM) is widely deployed in Europe, North America, and parts of Asia in the 900-, 1800-, and 1900-MHz frequency bands. In place in Japan, is the Pacific Digital Cellular system, which also employs TDMA. And finally, North American Digital Cellular exists in North America at the 800- and 1900-MHz bands.

Both of these methods, FDMA and TDMA, assign orthogonal channels to each active user and have a predetermined maximum number of users that can be serviced simultaneously. We thus consider TDMA and FDMA to be fixed resource allocation schemes. We note that if there are fewer users than the maximum number of "slots," resources can be wasted (not used) for these fixed resource allocation schemes.

The multiple access strategy of interest for this work, is code-division multiple access (CDMA), also illustrated in Fig. 1. In CDMA, each active user is assigned a waveform that exploits the total time and frequency bands allocated for the service. Both TDMA and FDMA can be considered as special cases of CDMA. By allowing for user waveforms with more general characteristics, CDMA signals can be designed to be more immune to the effects of the wireless channel and to allow more users to share the radio channel. This additional user capacity, however, will come at the expense of degradation in performance or at the expense of a more sophisticated and thus generally more complex receiver structure.

The particular type of CDMA considered here is direct-sequence CDMA (DS-CDMA). In the present implementations of standardized DS-CDMA, (long code) the spreading sequence is time-varying and has a period that is equal to that of many symbol intervals. In short code DS-CDMA, the waveform assigned to each user is generally a sequence, $\mathbf{s}$, drawn from a finite alphabet (e.g., $\{\pm 1/\sqrt{N}\}$, where N is the length of the sequence) and modulated onto a pulseshape $p(t)$ (e.g., a rectangular pulse shape or a raised cosine pulseshape). To provide a consistent definition of signal-to-noise ratio per bit, the spreading waveforms are normalized $\|\mathbf{s}\|^2 = 1$. The parameter N is the length of the spreading sequence and is also known as the *processing gain*. It is a measure of the bandwidth expansion offered by the spreading operation. In distinguishing "short code" DS-CDMA, we focus on systems where the same spreading sequence is used for each bit transmitted. An example is

$$\mathbf{s} = \frac{1}{\sqrt{N}}[-1, -1, +1, -1, +1, -1, +1, +1]$$

$$p(t) = \begin{cases} 1 & t \in [0, T_c) \\ 0 & \text{else} \end{cases}$$

The parameter T_c is called the *chip duration* and the symbol duration is thus $T = NT_c$. The spreading sequences are chosen to have desirable autocorrelation and cross-correlation properties [50].

While versions of FDMA and TDMA have been standardized for some years, standards for DS-CDMA are relatively recent. In 1993, IS95 was the first interim standard for the CDMA protocol. Since then several revisions have occurred. The current, second generation, CDMA personal communications system (PCS) is in the 1.8- and 2.0-GHz bands.

The focus on DS-CDMA is motivated by the imminent adoption of the DS-CDMA-type signaling in a variety of third generation wireless standards [1,11,38]. It is observed that for both the frequency-division duplex (FDD) and the time-division duplex (TDD) modes of UMTS, DS-CDMA multiple access is laid over the FDD and TDD duplexing schemes [16].

As a concluding note to the discussion of multiple-access schemes, we observe that TDMA and FDMA are special cases of CDMA, where typically the "spreading waveforms" are mutually orthogonal. Thus, with proper signal description, the methods described herein have utility for TDMA and FDMA systems where there is adjacent or co-channel interference (CCI) caused by dispersion or insufficient frequency reuse.

1.2. The Need for Adaptive Systems in Wireless Communications

The objective of this chapter is to introduce methods for the adaptive demodulation of data in DS-CDMA systems. Adaptive algorithms are instrumental for providing consistent performance in unknown or time-varying environments. Adaptive methods can implicitly reveal unknown parameters of a system or can be

used to track these parameters as they change over time. These characteristics of adaptive methods make them particularly suitable for wireless communications systems. In contrast to communication environments with relatively fixed characteristics, as is found in the wired environment of classical telephony, the wireless communication channel is time-varying [60]. This time-variation stems from a variety of sources: mobility of the user, changes in the active user population, and the potential mobility of interference sources. As a mobile user moves, the orientation of the user and the structures that absorb, reflect, and diffuse the user's transmitted radio signal change, thus changing the number of replica signals impinging on the receiver from the mobile user and further changing the amount of attenuation experienced by each signal. Further variability is induced by the fact that users rarely maintain a truly constant speed, especially in an urban environment. In a wireless communications system, users enter and exit communication in a bursty fashion. As noted earlier, each active user can be considered a form of interference for other users in a DS-CDMA system. Thus, with a time-varying user population, the interference experienced by a single user also varies with the population. Furthermore, as vehicles and other sources of scattering move themselves, this movement will also affect the channel experienced by the user. Another possible scenario is one in which the receiver has only partial knowledge of the communication or interference environment. Thus, in a cellular system, a base station may have accurately estimated parameters for the active users within that cell, but may not have information about out-of-cell interferers. Thus, it is clear that the wireless channel is diverse and changing. Because of these inherent characteristics of the wireless communications environment, we shall see that adaptive algorithms can be used to mitigate the effects of this channel.

2. SIGNAL MODEL

For illustrative purposes, we initially focus on a simplified communication scenario: bit-synchronized multiple users communicating over an additive white Gaussian noise (AWGN) channel. We shall assume that the front-end filter of the receiver is synchronized to the users and is coherent. The chip-matched filtered signal can be represented as a sequence of vectors:

$$\begin{aligned}\mathbf{r}(i) &= \sum_{k=1}^{K} A_k b_k(i)\mathbf{s}_k + \mathbf{n}(i)\\ &= \mathbf{SAb}(i) + \mathbf{n}(i)\end{aligned}$$

$$\text{where } \mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K]$$

$$\mathbf{A} = \text{diag}\,[A_1, A_2, \ldots, A_K]$$

$$\mathbf{b}(i) = [b_1(i), b_2(i), \ldots, b_K(i)]^T$$

where K is the number of active users and $\mathbf{n}(i)$ is the additive noise process, modeled as a white, Gaussian random vector process with zero mean and covariance $\sigma^2\mathbf{I}_N$. The matrix $\mathbf{I}_L$ is an $L \times L$ identity matrix. User k's received amplitude, data bit at time i and spreading sequence are denoted by A_k, $b_k(i)$, and $\mathbf{s}_k$, respectively. Herein, we shall assume binary phase shift keying (BPSK) data; that is $b_k(i) = \pm 1$. If required, the statistical model for the data is that $b_k(i)$ takes on its binary values with equal probability ($\frac{1}{2}$).

To facilitate the description of some classical nonadaptive multiuser receivers, we introduce the following notions. It can be shown that a set of sufficient statistics for detecting the data of a single user or all users in a multiuser DS-CDMA system is the output of a bank of filters matched to the spreading code of each active user. These matched filter outputs are defined as

$$\mathbf{y}(i) = \mathbf{S}^H\mathbf{r}(i) = \mathbf{RAb}(i) + \tilde{\mathbf{n}}(i)$$

where $\mathbf{R} = \mathbf{S}^H\mathbf{S}$. The noise present in the system is now colored, $\tilde{\mathbf{n}}(i) \sim \mathcal{N}(\mathbf{0}_K, \sigma^2\mathbf{R})$, where $\mathbf{0}_L$ is a $L \times 1$ vector of zeros. The $K \times K$ cross-correlation matrix $\mathbf{R}$ can be interpreted as a catalog of how far apart pairs of spreading codes are in the multiuser system. The Euclidean distance between two spreading codes is

$$\begin{aligned}\|\mathbf{s}_j - \mathbf{s}_k\|^2 &= \|\mathbf{s}_j\|^2 + \|\mathbf{s}_k\|^2 - 2\mathbf{s}_j^H\mathbf{s}_k\\ &= 2 - 2\mathbf{R}[j,k]\end{aligned}$$

We shall see that the performance of any multiuser or single-user receiver is very dependent on the values of the components of $\mathbf{R}$. Note that a set of K mutually orthogonal spreading codes leads to $\mathbf{R} = \mathbf{I}_K$.

Another key matrix for multiuser detection is the data correlation matrix:

$$\mathbf{C} = \mathbf{E}\{\mathbf{rr}^H\} = \mathbf{SA}^2\mathbf{S}^H + \sigma^2\mathbf{I}_N. \tag{1}$$

The operator $\mathbf{E}\{\cdot\}$ denotes expectation over all random quantities in the argument, unless otherwise specified. We next briefly review five important static multiuser receivers: the conventional receiver, the decorrelating detector, the minimum-mean-squared-error receiver, the jointly optimal detector, and the individually optimal detector. These receivers vary in their offered performance and attendant complexity. In the subsequent sections, we study various adaptive versions of a subset of these receivers.

2.1. Conventional Receiver

This is the conventional receiver that was considered optimal prior to a deeper understanding of multiple-access interference (MAI). If the central-limit theorem [e.g., 42,59] holds, then, it was originally argued, the MAI could be modeled as AWGN Gaussian noise [e.g., 43,56]. The optimal single-user receiver for transmission in AWGN is the matched-filter receiver. The *conventional* or *matched filter* receiver output is given by

$$\hat{\mathbf{b}}(i) = \text{sgn}\,(\mathbf{y}(i)) \tag{2}$$

where the operator $\text{sgn}(\cdot)$ outputs the sign of each component of its argument. The conventional receiver

is MAI-limited and incurs a high probability of error if the power of the received signal of the desired user is significantly less than that of the interfering users. This undesirable property of the conventional receiver is termed the *near-far problem*. We note that if a receiver is insensitive to the near far problem it is deemed *near-far-resistant*. The conventional receiver also suffers if the system is highly loaded. We note that there exists one scenario in which the conventional receiver is actually optimal in terms of probability of bit error — this occurs when the spreading codes of the active users are mutually orthogonal. In a realistic wireless system, this property is difficult to maintain. It is noted, however, that the conventional receiver is very straightforward to implement and requires knowledge only of the desired user's spreading waveform and timing.

2.2. Decorrelating Detector

The decorrelating detector [29,30,55] is the receiver that *zero-forces* the MAI — that is, it completely nulls out the MAI at the possible expense of removing some of the signal energy of the user of interest. It can also be viewed as the maximum-likelihood estimator of the vector $\mathbf{Ab}(i)$. This receiver is formed by

$$\hat{\mathbf{b}}(i) = \text{sgn}\,(\mathbf{R}^{-1}\mathbf{y}(i)) \tag{3}$$

The direct construction of the decorrelator requires the knowledge of the spreading waveforms of all the active users and the associated timing information. Despite the simplicity of this receiver, the decorrelator shares many properties with that of the optimal detector to be discussed in the sequel.

2.3. Minimum Mean-Squared Error Receiver

To introduce the linear minimum-mean-squared error (MMSE) receiver [32,67], we first discuss a generic linear receiver. Let $\mathbf{z}(i)$ be the soft output of a general linear receiver $\mathbf{M}$; then

$$\mathbf{z}(i) = \mathbf{M}\mathbf{r}(i) \tag{4}$$

$$\hat{\mathbf{b}}(i) = \text{sgn}\,(\mathbf{z}(i)) \tag{5}$$

Then, the MMSE receiver is determined by

$$\mathbf{M} = \arg\min_{\mathbf{M}} \mathbf{E}\left\{\|\mathbf{b}(i) - \mathbf{z}(i)\|^2\right\} \tag{6}$$

Two equivalent solutions, ignoring positive scalings, are given by

$$\mathbf{M} = \mathbf{S}^T(\mathbf{S}\mathbf{A}^2\mathbf{S}^T + \sigma^2\mathbf{I}_N)^{-1} \tag{7}$$

$$= (\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}\mathbf{S}^T \tag{8}$$

The two forms of the MMSE receiver can be shown to be equivalent through the use of the matrix inversion lemma [15,25]. Thus, the MMSE estimate of the data is given by

$$\hat{\mathbf{b}}(i) = \text{sgn}\,\left((\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}\mathbf{y}(i)\right) \tag{9}$$

In addition to the information required by the decorrelating detector, the MMSE receiver is also a function of the received amplitudes of the active users and the noise variance. In general, the MMSE receiver outperforms the decorrelating detector. As will be observed below, decentralized implementations of the MMSE receiver (and decorrelator) exist. In considering these decentralized receivers, a few observations can be made in regard to asymptotic behavior. As the noise variance grows ($\sigma^2 \to \infty$) or as the interfering amplitudes diminish ($A_2, \ldots, A_K \to 0$), the MMSE receiver approaches the conventional receiver. Alternatively, as the noise variance decreases ($\sigma^2 \to 0$), the MMSE receiver converges to the decorrelating detector. The MMSE receiver performs the optimal tradeoff in combating multiple access interference versus suppressing ambient channel noise in a linear receiver.

2.4. A Few Points on Linear Receivers

We note that the prior receiver algorithms are all linear in nature. To summarize, the bit decision for a particular user — say, user 1 — can be written as

$$\hat{b}_1(i) = \text{sgn}\,\left(\mathbf{c}_1^H\mathbf{r}(i)\right) \tag{10}$$

Note that for data demodulation, a positive scaling of the receiver does not affect the decision. Thus for $\alpha > 0$, both $\mathbf{c}_1$ and $\alpha\mathbf{c}_1$ yield the same decision. For the sequel, we shall note the soft-decision statistic for user k as

$$z_k(i) = \mathbf{c}_k^H\mathbf{r}(i) \tag{11}$$

Thus, for the joint receivers discussed above, we can define the following single user (decentralized) linear receiver vectors:

1. Conventional receiver $\mathbf{c}_k = \mathbf{s}_k$
2. Decorrelating detector $\mathbf{c}_k = \mathbf{S}\rho_k$ where ρ_k the kth column of $\mathbf{R}^{-1}$
3. MMSE receiver $\mathbf{c}_k = \mathbf{S}\mathbf{m}_k$, where $\mathbf{m}_k$ the kth column of $(\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}$

In addition, we also define the error sequence for a time-varying linear receiver $\mathbf{c}_k(i)$:

$$e_k(i) = b_k(i) - \mathbf{c}_k(i)^H\mathbf{r}(i) \tag{12}$$

Many of the adaptive algorithms to be discussed herein update the ith instantiation of the parameter vector by a function of the value of the error.

We next consider nonlinear receivers. These receivers are sometimes more complex to implement than the linear receivers discussed above; the benefit of this added complexity is improved performance.

2.5. Jointly Optimal Detector

The jointly optimal multiuser receiver is formed by determining the maximum likelihood estimate of

$\mathbf{b}$ [63,64]. Thus

$$\hat{\mathbf{b}}(i) = \arg \max_{\mathbf{b}} p(\mathbf{y}(i)|\mathbf{b}) \tag{13}$$

$$= \arg \max_{\mathbf{b}(i)} 2\mathbf{b}(i)^H \mathbf{A}\mathbf{y}(i) - \mathbf{b}(i)^H \mathbf{ARAb}(i) \tag{14}$$

$$= \arg \max_{\mathbf{b}(i)} \Omega(\mathbf{b}(i)) \tag{15}$$

where

$$\Omega(\mathbf{b}) = 2\mathbf{b}^T \mathbf{A}\mathbf{y} - \mathbf{b}^T \mathbf{ARAb} \tag{16}$$

The spreading waveforms and amplitudes of all active users are necessary to construct the jointly optimal receiver. Note that the decorrelating detector requires only the spreading waveforms of the active users and not the amplitudes.

2.6. Individually Optimal Detector

The individually optimum receiver achieves the minimum probability of error for the user of interest [63,64]. Without loss of generality, we shall assume that user 1 is the intended user. Then, the individually optimum receiver is obtained by

$$\hat{\mathbf{b}}_1(i) = \text{sgn}\left[\sum_{\mathbf{b},\mathbf{b}_1=1} \exp\left(\frac{\Omega(\mathbf{b})}{2\sigma^2}\right) - \sum_{\mathbf{b},\mathbf{b}_1=-1} \exp\left(\frac{\Omega(\mathbf{b})}{2\sigma^2}\right)\right] \tag{17}$$

The individually optimal detector requires the same set of parameter knowledge as the MMSE detector: user spreading waveforms, amplitudes, the noise variance, and timing. We observe that the individually optimal detector converges to the jointly optimum receiver as $\sigma \to 0$.

Other nonlinear receiver structures exist such as serial or parallel interference cancellation schemes. In such algorithms, estimates of certain bits are used to reconstruct the contribution due to a subset of users, which is in turn subtracted from the received signal, thus diminishing the multiple access interference. Adaptive versions of such receivers have been considered [e.g., 26,40,68], although we do not focus on them here.

3. ADAPTIVE SYSTEMS

We begin by discussing a generic adaptive receiver algorithm and desirable properties of such an adaptive system. As in any equalizer design, adaptive multiuser receivers can be *direct* or *indirect*. In the direct adaptive detectors, the adaptive algorithm demodulates the data directly. In the indirect implementations, the adaptive subsystems adaptively estimate parameters that are then utilized in a receiver of fixed form. An illustration of these two methods is provided in Fig. 3.

Let the *objective function*[1] of interest be defined as $J(\mathbf{c})$. Our goal is to determine the parameter vector $\mathbf{c}$ such that the objective function is minimized. Typical cost

[1] The objective function can also be termed the *cost function*.

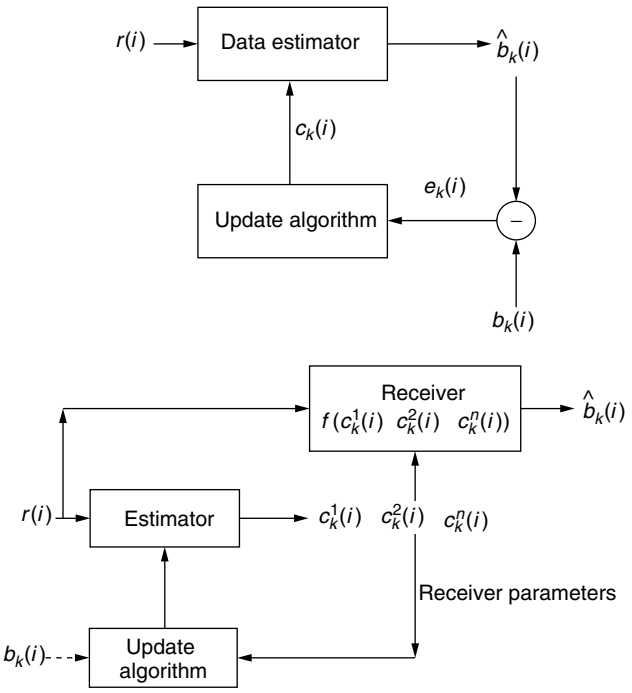


Figure 3. Direct and indirect implementations of adaptive receiver structures.

functions include the mean-squared error and the mean output energy. These cost functions will be discussed in more detail in the sequel.

Because of the stochastic nature of communication signals, the cost functions of interest will typically have the following form:

$$J(\mathbf{c}) = \mathbf{E}\{j(\mathbf{c}, \mathbf{r})\} \tag{18}$$

The expectation is taken with respect to all random quantities present in the received signal $\mathbf{r}$. If the cost function is convex in the unknown parameter vector, we can consider the method of steepest descent to determine the desired stationary point [17]. The method of steepest descent updates the parameter vector estimate in the direction opposite to the gradient of the cost function.

Ideally, update of the multidimensional parameter vector $\mathbf{c}(i)$ is conducted by

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu \nabla J(\mathbf{c}(i-1)) \tag{19}$$

The scalar μ, called the *step size*, is selected to ensure convergence of the adaptive algorithm while offering a reasonable convergence rate. These two objectives are conflicting.

The true gradient is often impossible to determine as it may require the statistics of the received signal (which are often presumed unknown as they could be employed to derive a nonadaptive receiver, if such quantities were known). Thus an approximation to the gradient is used:

$$\nabla J(\mathbf{c}(i-1) \approx \nabla j(\mathbf{c}(i-1), \mathbf{r}(i)) \tag{20}$$

The justification for such an approximation is provided by the fact that under sufficient regularity of the cost

functions and the probability distribution functions of the embedded random processes

$$\nabla J(\mathbf{c}(i-1)) = \mathbf{E}\{\nabla j(\mathbf{c}(i-1), \mathbf{r}(i))\} \tag{21}$$

4. ADAPTIVE DETECTION ALGORITHMS

Two types of adaptive algorithm are possible. In the first case, a *training signal* is used to update the adaptive algorithm. This data signal is known at both the transmitter and the receiver. This training signal is essentially a known data sequence for one or some of the active users [e.g., $b_k(i)$]. The adaptive algorithm typically employs this training signal to form an error signal that is used to drive the update. In contrast, in *blind* adaptive algorithms, there is no training signal and cost functions based on the statistics of the signal are formed and updated. In training-based systems, no meaningful data are transmitted during training; thus it is sometimes argued that training-based systems are not bandwidth-efficient. However, many blind algorithms require a considerable amount of observations (with embedded unknown data) before reliable estimates are achieved. Our focus is on training-based algorithms.

4.1. Adaptive Least Mean Squares

We shall assume a linear receiver for user 1. Thus the data are estimated via

$$\hat{b}_1(i) = \text{sgn}\,(\mathbf{c}_{\text{LMS}}(i)^H \mathbf{r}(i)) \tag{22}$$

For the adaptive least-mean-squares (LMS) algorithm, the cost function is the mean-squared error:

$$J(\mathbf{c}) = \mathbf{E}\left\{|b_1(i) - \mathbf{c}^H \mathbf{r}(i)|^2\right\} \tag{23}$$

The parameter vector is the desired linear receiver. The estimate of the parameter vector at time i ($\mathbf{c}(i)$) is obtained by employing the method of steepest descent. Thus

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu \nabla J(\mathbf{c}(i-1)) \tag{24}$$

$$= \mathbf{c}(i-1) - \mu \mathbf{E}\{\mathbf{r}(i)(b_1(i-1) - \mathbf{c}(i-1)^H \mathbf{r}(i))\} \tag{25}$$

$$\approx \mathbf{c}(i-1) - \mu \mathbf{r}(i)(b_1(i-1) - \mathbf{c}(i-1)^H \mathbf{r}(i)) \tag{26}$$

The cost function of interest here, the mean-squared error, is a special case of the least-mean p norm:

$$J(\mathbf{c}) = \mathbf{E}\left\{|b_1(i) - \mathbf{c}^H \mathbf{r}(i)|^p\right\} \tag{27}$$

It is noted that the cost function above is convex for $1 \le p < \infty$. The nonstochastic form of the cost function above was investigated for determining adaptive multiuser receivers for systems with non-Gaussian additive noise modeled as a symmetric alpha-stable process [27]. The resultant adaptive algorithm is given by

$$\mathbf{c}(i) = \mathbf{c}(i-1) + \mu p |b_1(i) - \mathbf{c}(i-1)^H \mathbf{r}(i)|^{p-1} \times \text{sgn}\,(b_1(i) - \mathbf{c}(i-1)^H \mathbf{r}(i)) \mathbf{r}(i)$$

Clearly when $p = 2$, the algorithm above reduces to the adaptive LMS algorithm noted above.

4.1.1. Convergence Analysis of LMS. The sequence $\{\mathbf{c}(i)\}_{i=1}^{\infty}$ is a sequence of random vectors. It is of interest to investigate the limiting behavior of this random sequence to determine the efficacy of the update algorithm. The typical convergence analysis provided for adaptive LMS algorithms is the study of the asymptotic bias of the update algorithm. Thus, we seek to determine whether the following is in fact true:

$$\lim_{i \to \infty} \mathbf{E}\{\mathbf{c}(i)\} \stackrel{?}{=} \mathbf{c}_{\text{MMSE}} \tag{28}$$

Recall that the optimal MMSE receiver is given by

$$\mathbf{c}_{\text{MMSE}} = (\mathbf{S}\mathbf{A}^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{s}_1 \tag{29}$$

To study the convergence behavior of LMS to this desired parameter vector, we define the parameter error vector, $\mathbf{v}(i) = \hat{\mathbf{c}}(i) - \mathbf{c}_{\text{MMSE}}$. The evolution of the mean error vector can thus be described as

$$\mathbf{E}\{\mathbf{v}(i)\} = \mathbf{E}\{\mathbf{v}(i-1)\} - \mu \mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H \mathbf{v}(i-1)\} + \mu \mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H\} \mathbf{c}_{\text{MMSE}} \tag{30}$$

$$= [\mathbf{I}_N - \mu(\mathbf{S}A^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N)] \mathbf{E}\{\mathbf{v}(i-1)\} \tag{31}$$

where

$$\mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H\} = \mathbf{S}\mathbf{A}^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N \tag{32}$$

The vector $\hat{\mathbf{c}}(i-1)$ is a function of all prior received signal vectors, $\mathbf{r}(0), \mathbf{r}(1), \ldots, \mathbf{r}(i-1)$, but is independent of the current observation $\mathbf{r}(i)$. We define the following matrix and its associated eigenvalue decomposition:

$$\mathbf{C}(\mu) = \mathbf{I}_N - \mu(\mathbf{S}\mathbf{A}^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N) \tag{33}$$

$$= \mathbf{V} \Lambda(\mu) \mathbf{V}^H \tag{34}$$

where $\Lambda(\mu)$ is a diagonal matrix of the eigenvalues ($\lambda_i(\mu)$) of $\mathbf{C}(\mu)$ and $\mathbf{V}$ is a marix whose columns correspond to the eigenvectors of $\mathbf{C}(\mu)$. Thus

$$\mathbf{E}\{\mathbf{v}(i)\} = \mathbf{C}(\mu) \mathbf{E}\{\mathbf{v}(i-1)\} \tag{35}$$

Because of the orthonormal property of the eigenvectors, we obtain

$$\mathbf{V}^H \mathbf{E}\{\mathbf{v}(i)\} = \Lambda(\mu) \mathbf{V}^H \mathbf{E}\{\mathbf{v}(i-1)\} \tag{36}$$

$$\mathbf{E}\{\tilde{\mathbf{v}}(i)\} = \Lambda(\mu) \mathbf{E}\{\tilde{\mathbf{v}}(i-1)\} \tag{37}$$

where

$$\mathbf{V}^H \mathbf{E}\{\mathbf{v}(i)\} = \mathbf{E}\{\tilde{\mathbf{v}}(i)\} \tag{38}$$

We can now rewrite the linear transformation of the error vector at time i as a function of the initial error:

$$\mathbf{E}\{\tilde{\mathbf{v}}(i)\} = \Lambda(\mu)^i \mathbf{E}\{\tilde{\mathbf{v}}(0)\} \tag{39}$$

$$= \text{diag}\,[\lambda_1^i(\mu), \lambda_2^i(\mu), \ldots, \lambda_N^i(\mu)] \mathbf{E}\{\tilde{\mathbf{v}}(0)\} \tag{40}$$

The system is stable, that is, the expected error vector converges to zero if $0 < |\lambda_i(\mu)| < 1$, this implies that

$$0 < \mu < \frac{2}{\lambda_i(\mu)} \forall i \tag{41}$$

If we denote $\lambda_{\max}$ as the maximum eigenvalue of the matrix $\mathbf{SA}^2\mathbf{S}^H$, then a fixed step size that achieves the desired convergence must fall within the following range

$$0 < \mu < \frac{2}{\sigma^2 + \lambda_{\max}} \tag{42}$$

A key concern about the implementation of an adaptive algorithm is that the rate of adaptation of the algorithm be matched to the underlying rate of change of the time-varying system. For certain fast-fading channels (see Sections 4.5 and 5.1), certain families of adaptive algorithms cannot be used because they cannot track the channel variations.

4.2. Recursive Least-Squares Methods

The recursive least-squares algorithm differs from the adaptive LMS algorithm just derived in that the cost function is a deterministic one. For an observation record corresponding to M symbols, it is desired to minimize the metric

$$J(\mathbf{c}) = \sum_{i=1}^{M} \lambda^{M-i} |e(i)|^2, \tag{43}$$

where λ is deemed the *forgetting factor* and is assumed to be $0 < \lambda < 1$. With this form of weighting, more recent observations are given more weight than are previous observations. As before, the desired parameter vector is a linear receiver, thus

$$\hat{b}_1(i) = \text{sgn}\,(\mathbf{c}_{\text{RLS}}(i)^H \mathbf{r}(i)) \tag{44}$$

The resultant algorithm, which exploits the matrix inversion lemma (to avoid a computationally expensive direct matrix inverse) is given as follows:

$$\mathbf{k}(i) = \frac{\lambda^{-1}\mathbf{P}(i-1)\mathbf{r}(i)}{1 + \lambda^{-1}\mathbf{r}(i)^H\mathbf{P}(i-1)\mathbf{r}(i)} \tag{45}$$

$$\zeta(i) = b_1(i) - \mathbf{c}(i-1)^H\mathbf{r}(i) \tag{46}$$

$$\mathbf{c}(i) = \mathbf{c}(i-1) + \mathbf{k}(i)\zeta(i) \tag{47}$$

$$\mathbf{P}(i) = \lambda^{-1}\mathbf{P}(i-1) - \lambda^{-1}\mathbf{k}(i)\mathbf{r}(i)^H\mathbf{P}(i-1) \tag{48}$$

The typical initialization of the algorithm is with $\mathbf{P}(0) = \delta^{-1}\mathbf{I}_N$ and $\mathbf{c}(0) = \mathbf{0}_N$, where δ is a small positive constant. In the multiuser receiver context, we can also initialize the weight vector to be $\mathbf{c}(0) = \mathbf{s}_1$; thus, the receiver is initialized as the conventional matched-filter receiver. The vector $\mathbf{k}(i)$ is called the *gain vector*. We also note that $\mathbf{P}(i)$ is the current estimate of the inverse of the weighted (each observation is weighted by λ) data correlation matrix.

There is a subtle distinction between the *tentative* error sequence $\zeta(i)$ and the *current* error sequence $e(i)$, in that $\zeta(i)$ employs the past value of receiver vector $\mathbf{c}(i-1)$ while $e(i)$ is formed with $\mathbf{c}(i)$. This error is often termed the *prediction error* as it uses the past estimate of the receiver to predict the current data.

For the general case (linear estimation of a scalar process), one can show that the RLS algorithm has a rate of convergence that is far superior to that of the LMS algorithm [17]; however, this comes as the expense of greater computational complexity. Theoretically, the RLS algorithm produces zero excess mean-squared error. This result is dependent on the presence of a statistically stationary environment, which is not typically experienced in a wireless communications channel. Finally, the convergence of the RLS algorithm is not dependent on the eigenvalue spread of the data correlation matrix. In contrast, the LMS algorithm's convergence speed is dependent on this eigenvalue spread. The implication for DS-CDMA systems is the fact that the eigenvalue spread can be quite large in a system where there is a great disparity in received powers amongst the users. Thus, if tight power control is not in place, the LMS algorithm will experience slow convergence. However, for many scenarios, this improved convergence speed for RLS is in fact not observed [65].

We conclude this subsection by noting a work [5] that considers the steady-state behavior of LMS and RLS operating in a multiuser environment by employing the results on steady-state excess mean-squared error provided elsewhere [12,17].

4.3. Adaptive Linear Minimum Probability of Error Receivers

Through performance comparisons and analysis, it can be shown that the MMSE receiver (and thus the convergent adaptive LMS receiver) offers strong performance and can combat multiple-access interference (MAI). However, in a digital communications system, the true performance metric of interest is the probability of bit detection error rather than the mean-squared error. Thus, we next consider a set of adaptive linear receivers that endeavor to minimize the probability of bit detection error.

Let the transmitted, noiseless, multiuser signal be represented as

$$\mathbf{m}(i) = \sum_{k=1}^{K} A_k b_k(i)\mathbf{s}_k = \mathbf{SAb}(i) \tag{49}$$

Then the received signal is simply $\mathbf{r}(i) = \mathbf{m}(i) + \mathbf{n}(i)$. Conditioned on knowing all of the parameters of the interfering users, that is, if we know the matrices $\mathbf{S}$, $\mathbf{A}$, $\mathbf{b}$ completely, then the probability of error for the bit interval i for a linear receiver $\mathbf{c}(i)$ is given by,

$$P_e(\mathbf{b}, \mathbf{c}) = Q\left(\frac{b_1(i)\mathbf{c}(i)^H\mathbf{m}(i)}{\sigma\|\mathbf{c}(i)\|}\right) \tag{50}$$

where $Q(x)$ is the complementary cumulative distribution function[2] of a zero-mean, unit-variance Gaussian random variable.

[2] $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} dv.$

The goal of the adaptive minimum probability of error receiver is to determine the optimal receiver vector $\mathbf{c}^*$ that satisfies [34]

$$\mathbf{c}^* = \arg \min_{\mathbf{c}} J(\mathbf{c}) \tag{51}$$

where

$$J(\mathbf{c}) = \mathbf{E}\{P_e(\mathbf{b}, \mathbf{c})\} \tag{52}$$

The method for updating is also based on steepest descent; thus we invoke Eq. (19). We also approximate $\nabla \mathbf{E}\{P_e(\mathbf{b}, \mathbf{c})\} \approx \nabla P_e(\mathbf{c}(i), \mathbf{b}(i))$. It has been shown [34] that this approximation is an unbiased estimator of the true gradient. The desired update procedure is

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu \nabla P_e(\mathbf{c}(i), \mathbf{b}(i)) \tag{53}$$

$$= \mathbf{c}(i-1) - \mu - \frac{1}{2\sqrt{2\pi}} \times \left\{ \exp\left(-\frac{1}{2} \left(\frac{b_1(i)\mathbf{c}(i-1)^H \mathbf{m}(i)}{\sigma \|\mathbf{c}(i-1)\|} \right)^2 \right) \right. \tag{54}$$

$$\left. \times \left[\frac{b_1(i)\mathbf{m}(i)}{\sigma \|\mathbf{c}(i-1)\|} + \frac{b_1(i)\mathbf{c}(i-1)^H \mathbf{m}(i)}{\sigma^2 \|\mathbf{c}(i-1)\|^3} \mathbf{c}(i-1) \right] \right\} \tag{55}$$

This detector is deemed the *clairvoyant* adaptive receiver [34] as it requires the knowledge of the transmitted signal $\mathbf{m}(i)$ and not just the training sequence $b_1(i)$. This unrealistic assumption is removed by performing a maximum-likelihood estimation (MLE) operation to determine an estimate for $\mathbf{m}(i)$. Thus $\mathbf{m}(i)$ above is replaced with $\hat{\mathbf{m}}(i)$, where

$$\hat{\mathbf{m}}(i) = \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^H \mathbf{r}(i) \tag{56}$$

Therefore, the estimated transmitted signal vector is simply the projection of the received signal onto the subspace spanned by the spreading codes of the active users.

The function $J(\mathbf{c})$ given above is not strictly a convex function. However, in [34], a series of conditions are established that ensure convexity. In brief, at each iteration of the update algorithm, the receiver must be near-far resistant.

An alternative approach to minimizing the probability of error through an adaptive linear receiver has been investigated in [49], where the cost function of interest is the expected value of the *single-letter distortion measure*, $\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c})$. For simplicity, we assume that the prior probabilities of the BPSK transmitted data for the desired user are $\pi_0 = \pi_1 = \frac{1}{2}$. Then

$$\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c}) = \tfrac{1}{4}\{[1 + \operatorname{sgn}(\mathbf{c}^H \mathbf{r}_0)] + [1 - \operatorname{sgn}(\mathbf{c}^H \mathbf{r}_1)]\} \tag{57}$$

The vectors $\mathbf{r}_0, \mathbf{r}_1$ represent training signals given that $b_1 = -1$ and $b_1 = 1$ respectively:

$$\mathbf{r}_0 = \mathbf{r}(i) | b_1 = -1 \tag{58}$$

$$\mathbf{r}_1 = \mathbf{r}(i) | b_1 = +1 \tag{59}$$

Thus, if the receiver vector, $\mathbf{c}$, achieves correct decisions, $\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c}) = 0$. The cost function is strictly positive if errors occur. Note that

$$P_e = \mathbf{E}\{\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c})\} = J(\mathbf{c}) \tag{60}$$

This cost function is independent of the underlying noise distribution. Thus the methods described below can be applied in scenarios where non-Gaussian (impulsive) noise is present. Steepest descent methods can be considered where the noisy estimate of the gradient of $J(\mathbf{c})$ is used. In this case

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu_n \nabla \zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c}) \tag{61}$$

However, in contrast to the adaptive minimum probability of error receiver discussed earlier, determining the true gradient poses difficulty. Thus the adaptive receiver update algorithm is constructed by determining one or two-sided difference approximations to the gradient of $J(\mathbf{c})$. For example, construct the following vector function at time i

$$\mathbf{x}(\mathbf{c}(i)) = [\mathbf{x}(\mathbf{c}(i))_1, \mathbf{x}(\mathbf{c}(i))_2, \ldots, \mathbf{x}(\mathbf{c}(i))_N,] \tag{62}$$

$$\mathbf{x}(\mathbf{c}(i))_j = \frac{1}{2\alpha(i)}\{[\zeta(\mathbf{r}_0(i), \mathbf{r}_1(i); \mathbf{c}(i) + \alpha(i)\mathbf{e}_j)] - [\zeta(\mathbf{r}_0(i), \mathbf{r}_1(i); \mathbf{c}(i) - \alpha(i)\mathbf{e}_j)]\} \tag{63}$$

where the vector $\mathbf{e}_j$ is the jth coordinate unit vector; $\alpha(i)$ is selected such that $\alpha(i) = \beta i^{-(1/4)}$ for some $\beta > 0$. The receiver update algorithm with the approximate gradient in place is given by

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu_i \mathbf{x}(\mathbf{c}(i-1)) \tag{64}$$

With key conditions on the step size sequence (μ_i) and the perturbation rate $\alpha(i)$, it can be shown that the recursion in Eq. (64) converges with probability 1 to the minimizing value of $\mathbf{c}^*$. Under the assumption of AWGN, necessary conditions can be established to ensure the convexity of the cost function of interest.

Numerical results show that for the scenario of a strong desired user, the adaptive algorithm in Eq. (64) achieves a significant performance improvement over the adaptive LMS algorithm, the true MMSE solution and the decorrelating detector. However, as the near-far ratio increases, these three algorithms achieve comparable probability of error.

Finally, it is noted that the work of Yeh and Barry [69] can be viewed as the generalization of these two approaches for multiuser detection to the equalization of single-user intersymbol interference channels. Thus a steepest-descent approach for the probability of error in Gaussian channels as well as a steepest-descent method based on a single-letter distortionlike measure are considered. Further, low-complexity approximations and convergence rate increasing alternatives are also investigated.

4.4. Adaptive Optimal Receivers

Because of the assumption of additive Gaussian noise, the form of the individually optimal receiver described by Eq. (17) is the same as a radial basis function (RBF) network. Thus, methods previously considered to adaptively update the parameters of such a network can

be employed to form an adaptive multiuser detector [36]. The general form of a RBF network output is given by

$$z(i) = \sum_{j=1}^{N} w_j \Phi\left(\frac{\|\mathbf{r}(i) - \mathbf{m}_j\|}{\sigma_j}\right). \tag{65}$$

Here $\Phi(\cdot)$ is a continuous, nonlinear function from $\Re^+ \rightarrow \Re^+$ (other conditions on Φ, which stem from regularization and approximation theory can be found in [45]). The input vector is $\mathbf{r}(i)$, $\mathbf{m}_j$ is called the *center* of the RBF neuron, σ_j is the *spread* of the neuron, and the w_j are the *weights* that optimize some performance criterion. The methods by which the $\Phi(\cdot)$, $\mathbf{m}_j$, σ_j, and w_j are selected, constitute much of the research on RBF networks. Traditionally, the centers were randomly chosen from the given data set; the spreads were then calculated by determining the minimum distance between centers using the appropriate distance metric and the weights could be solved for given a simple error criterion (e.g., MMSE) [28]. Methods for determining these network parameters are often unique to the application for which the RBF network is used.

The application of RBF networks as multiuser receivers is inspired by the work of Chen et al. [7] and Chen and Mulgrew [8], who used these networks and modifications thereof to perform equalization of intersymbol interference (ISI) channels. While intersymbol interference is analogous to MAI, the distinctions between these two noise sources imply that modifications of the previous RBF techniques are necessary to ensure good performance from the RBF network as an adaptive multiuser detector.

By inspecting Eq. (17), the decision rule for the individually optimum receiver, we see that we can rewrite the decision rule as a function of the received signal (versus the matched filter outputs) as follows:

$$\hat{b}_1(i) = \operatorname{sgn}\left[\sum_{j=1}^{2^{K-1}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{r}(i) - \underline{\Theta}_1 - \underline{\mu}_j\|^2\right\} - \sum_{j=1}^{2^{K-1}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{r}(i) - \underline{\Theta}_0 - \underline{\mu}_j\|^2\right\}\right]$$

This decision rule can then be mapped to the RBF network with the following definitions of key functions and parameters:

$$\Phi(x) = \exp\{-x^2\} \tag{66}$$

$$\sigma_j = \sqrt{2}\sigma \tag{67}$$

(where σ^2 is the Gaussian noise variance)

$$\underline{\Theta}_d = (-1)^{d+1} A_1 \mathbf{s}_1 \tag{68}$$

$$\underline{\mu}_j = \sum_{k=2}^{K} A_k b_k \mathbf{s}_k \tag{69}$$

(for some permutation of the b_k values)

$$\mathbf{m}_j \in \{\underline{\mu}_j + \Theta_0, \underline{\mu}_j + \Theta_1 | j = 1, \ldots, 2^{K-1}\} \tag{70}$$

$$w_j = \begin{cases} 1 & \text{if } \mathbf{m}_j = \underline{\mu}_l + \Theta_1 \\ -1 & \text{if } \mathbf{m}_j = \underline{\mu}_l + \Theta_0 \end{cases} \tag{71}$$

With this mapping in hand, we turn to methods used to train RBF networks to determine the centers and the weights. We first present a *supervised* learning algorithm whereby the data of *all active users* must be known; this is akin to the clairvoyant receiver of Ref. [34]. With known bits, it is known to which center a received signal corresponds. Given i observations of the received signal, we update the centers as follows:

$$\mathbf{b}(i) \leftrightarrow \text{index}, j \tag{72}$$

$$\hat{\mathbf{m}}_j(i) = \frac{i-1}{i}\hat{\mathbf{m}}_j(i-1) + \frac{1}{i}\mathbf{r}(i) \tag{73}$$

This supervised algorithm is somewhat impractical as it requires coordination of all active users to send training data simultaneously. Next the *k-means clustering algorithm* [31] is described:

$$\text{index}, j^* = \arg\min_l \|\hat{\mathbf{m}}_l(i-1) - \mathbf{r}(i)\|^2 \tag{74}$$

$$\hat{\mathbf{m}}_{j^*}(i) = \frac{i-1}{i}\hat{\mathbf{m}}_{j^*} l(i-1) + \frac{1}{i}\mathbf{r}(i) \tag{75}$$

The convergence of the supervised algorithm to the true centers trivially follows from the law of large numbers [42,59]. The convergence of the k-means algorithm has been investigated [31]. To speed up convergence, the k-means algorithm can be initialized with estimates of the centers using matched-filter outputs to perform coarse amplitude estimation. Then all possible permutations of the noiseless received signal are constructed.

To adaptively determine the weights, a LMS update is considered:

$$\mathbf{w}(i+1) = \mathbf{w}(i) + \mu(z_1(i) - b_1(i))\underline{\Phi}(\mathbf{r}(i))$$

where μ is the adaptation gain, $z_1 = \mathbf{w}(i)^H \underline{\Phi}(\mathbf{r}(i))$ is the output of the RBF network at time i, and $\underline{\Phi}(\cdot)$ is the vector RBF nonlinear functions applied to the input. It has been shown [36] that even with estimated centers, the mean weight vector is a positively scaled version of the desired weights.

4.5. Reduced-Rank Adaptive MMSE Filtering

We return to linear adaptive receivers to consider reduced-rank adaptive MMSE algorithms. To introduce the reduced-rank methods, we recall the full rank-fixed MMSE receiver for the desired user 1. This linear MMSE receiver is an $N \times 1$ vector $\mathbf{c}$ that minimizes the mean-squared error (MSE):

$$\mathbf{c}_{\text{MMSE}} = \arg\min_{\mathbf{c}} \text{MSE} = \arg\min_{\mathbf{c}} \mathbf{E}\{|b_1(i) - \mathbf{c}^H \mathbf{r}(i)|^2\} = \mathbf{C}^{-1}\mathbf{p} \tag{76}$$

Recall that $\mathbf{C} = \mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H\}$ is the data cross-correlation matrix and $\mathbf{p} = \mathbf{E}\{b_1(i)\mathbf{r}(i)\}$ is termed the *steering vector* [35].

The adaptive algorithms [35,70] can be used to estimate $\mathbf{c}_{\text{MMSE}}$, even in a time-varying channel. When N is large, convergence is slow. Reduced rank techniques reduce the

number of taps to be adaptively tracked by projecting the received signal vector onto a lower-dimensional subspace. Let D be the resultant lower dimension, where $D < N$, the projection is

$$\tilde{\mathbf{r}}(i) = \mathbf{P}_D^H \mathbf{r}(i) \tag{77}$$

where $\mathbf{P}_D$ is the $N \times D$ projection matrix and the D dimensional signal is denoted by a tilde. The vector $\tilde{\mathbf{r}}(i)$ is then the input to a length D tap delay linear filter. When the MMSE criterion is applied, the optimum coefficients for the D dimensional space are given by

$$\tilde{\mathbf{c}}_{\text{MMSE}} = \tilde{\mathbf{C}}^{-1}\tilde{\mathbf{p}}, \quad \text{where} \quad \tilde{\mathbf{C}} = \mathbf{P}_D^H \mathbf{C} \mathbf{P}_D, \tilde{\mathbf{p}} = \mathbf{P}_D^H \mathbf{p}. \tag{78}$$

4.5.1. Projection Matrix Selection. A number of methods for selecting the projection matrix $\mathbf{P}_D$ are considered here.

The multistage Wiener filtering (MWF) algorithm for DS-CDMA has been presented [20]. The resultant algorithm is a specialization of the work by Goldstein et al. [14]. The MWF projection matrix is given by

$$\begin{aligned} \mathbf{P}_D^{MW} &= [\mathbf{g}_{MW,1}\, \mathbf{g}_{MW,2} \cdots \mathbf{g}_{MW,D}] \\ &= \left[\mathbf{h}_1 \; \mathbf{B}_1^H \mathbf{h}_2 \; \cdots \prod_{j=1}^{D-1} \mathbf{B}_j^H \mathbf{h}_D \right] \end{aligned} \tag{79}$$

where $\mathbf{g}_{MW,j}$, $j = 1, \ldots, D$ are implicitly defined. The matrix $\mathbf{B}_j$ is an $(N-j) \times (N-j+1)$ blocking matrix, namely, $\mathbf{B}_j \mathbf{h}_j = 0$. The vector $\mathbf{h}_j$ is the normalized correlation vector $E\{\tilde{d}_{j-1}(i)\mathbf{r}_{j-1}(i)\}$, where $\mathbf{r}_j(i) = \mathbf{B}_j \mathbf{r}_{j-1}(i)$ with $\mathbf{r}_0(i) = \mathbf{r}(i)$, and $\tilde{d}_j(i) = \mathbf{h}_i^H \mathbf{r}_{j-1}(i)$ with $\tilde{d}_0(i) = b_1(i)$ [20].

The projection matrix for the auxiliary vector filtering (AVF) algorithm is given by [41]

$$\mathbf{P}_D^{AV} = [\mathbf{g}_{AV,1}\, \mathbf{g}_{AV,2} \cdots \mathbf{g}_{AV,D}] \tag{80}$$

where $\mathbf{g}_{AV,1}$ is also the normalized correlation vector $E\{b_1(i)\mathbf{r}(i)\} = \mathbf{h}_1$, and $\mathbf{g}_{AV,j}$, $j = 2, \ldots, D$ are the auxiliary vectors, given by [41]

$$\begin{aligned} &\mathbf{g}_{AV,j+1} \\ &= \frac{\mathbf{C}\mathbf{g}_{AV,j}^{Eq} - (\mathbf{g}_{AV,1}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,1} - \sum_{l=2}^{j} (\mathbf{g}_{AV,l}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,l}}{\| \mathbf{C}\mathbf{g}_{AV,j}^{Eq} - (\mathbf{g}_{AV,1}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,l} - \sum_{l=2}^{j} (\mathbf{g}_{AV,l}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,l} \|} \end{aligned} \tag{81}$$

where $\mathbf{g}_{AV,j}^{Eq} = \mathbf{g}_{AV,1} - \sum_{l=2}^{j} w_l \mathbf{g}_{AV,j}, w_l, l = 2, \ldots, j$ are the optimized constants [41] and $\| \cdot \|$ is the vector norm. By construction, the $\mathbf{g}_{AV,j}$, $j = 1, \ldots, D$ are normalized orthogonal vectors.

Using the Cayley–Hamilton (CH) theorem, Moshavi et al. proposed the following projection matrix [37]:

$$\mathbf{P}_D^{CH} = [\mathbf{g}_{CH,1}\; \mathbf{g}_{CH,2} \; \cdots \; \mathbf{g}_{CH,D}] = [\mathbf{h}_1 \; \mathbf{C}\mathbf{h}_1 \cdots \mathbf{C}^{D-1}\mathbf{h}_1] \tag{82}$$

The vector $\mathbf{h}_1$ is the one defined previously.

The projection matrix can also be selected by performing an eigendecomposition on $\mathbf{C}$:

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^H \tag{83}$$

The matrix $\Lambda = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_N]$ contains the eigenvalues associated with the eigenvectors which are the columns of $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$. If the eigenvalues are ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, the desired projection matrix is then $\mathbf{P}_D = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D]$. This method is titled the method of *principal components*. A related method, termed the *cross-spectral reduced-rank method*, selects the D eigenvectors that result in a minimum mean-squared error estimate of the bit. This method requires knowledge of the desired user's spreading waveform, but offers improved performance over the principal-components methods [13]. Eigendecomposition methods, in general, are not attractive due to their high attendant computational complexity.

Finally, a very simple method of rank reduction is to employ partial despreading as proposed in [57]. The desired user's spreading code is decomposed into multiple subvectors: $\mathbf{s}_1 = [\mathbf{s}_1^{(1)}, \mathbf{s}_1^{(2)}, \ldots, \mathbf{s}_1^{(D)}]$. The projection matrix is then constructed as

$$\mathbf{P}_D = \begin{bmatrix} \mathbf{s}_1^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{s}_1^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{s}_1^{(D)} \end{bmatrix} \tag{84}$$

Through a key simplification of the AVF algorithm, it can be shown that the AVF method is equivalent to the MWF algorithm [10]. The work by Chen et al. [10] provides a simplified proof of the equivalence of the static MWF receiver and that based on the Cayley Hamilton expansion of the correlation matrix inverse [37]. This observation was made previously [20]. Because of its simplicity of presentation, we focus on adaptive implementations of the MWF method as presented [20].

4.5.2. Adaptive Reduced-Rank Detection. In general, the following two cost functions can be set up for determining adaptive reduced rank receivers:

$$J(\tilde{\mathbf{c}})_{\text{LS}} = \sum_{i=1}^{M} \| b_1(i) - \tilde{\mathbf{c}}^H \tilde{\mathbf{r}}(i) \|^2 \tag{85}$$

$$J(\tilde{\mathbf{c}})_{\text{MMSE}} = \mathbf{E}\{ \| b_1(i) - \tilde{\mathbf{c}}^H \tilde{\mathbf{r}}(i) \|^2 \} \tag{86}$$

where

$$\tilde{\mathbf{r}}(i) = \hat{\mathbf{P}}^H(i)\mathbf{r}(i) \tag{87}$$

Note that $\hat{\mathbf{P}}(i)$ is an estimate at time i of the projection matrix $\mathbf{P}$. The methods discussed in the previous sections can be employed to derive adaptive algorithms.

A host of adaptive algorithms for the reduced rank multistage Wiener filter for DS-CDMA signaling has been provided [20]. Batch and recursive algorithms based on least-squares as well stochastic gradient descent algorithms are considered. In addition, both blind and training-based methods are provided. The training-based stochastic gradient method, which offers strong performance, is described here. The stochastic gradient algorithm of Ref. 20 is given by two sets of "recursions". We note that the subscript n indicates the stage number

of the multistage system, while the index i corresponds to the symbol interval timing.

Initialization:

$$d_0(i) = b_1(i)\,\mathbf{r}_0(i) = \mathbf{r}(i) \tag{88}$$

Forward recursion: at each i, for $n = 1, \ldots D$

$$\hat{\mathbf{p}}_n(i) = (1-\mu)\hat{\mathbf{p}}_n(i-1) + \mu d^*_{n-1}(i)\mathbf{r}_{n-1}(i) \tag{89}$$

$$\hat{\mathbf{c}}_n(i) = \frac{\hat{\mathbf{p}}_n(i)}{\|\hat{\mathbf{p}}_n(i)\|} \tag{90}$$

$$\mathbf{B}_n(i) = \text{null}\,[\hat{\mathbf{c}}^H_n(i)] \tag{91}$$

$$d_n(i) = \hat{\mathbf{c}}_n(i)^H \mathbf{r}_{n-1}(i) \tag{92}$$

$$\mathbf{r}_n(i) = \mathbf{B}^H_n(i)\mathbf{r}_{n-1}(i) \tag{93}$$

Backward recursion: decrementing $n = D, \ldots, 1$

$$\zeta_n(i) = (1-\mu)\zeta_n(i-1) + \mu|\varepsilon_n|^2 \tag{94}$$

$$w_n(i) = \frac{\|\hat{\mathbf{p}}_n(i)\|}{\zeta_n(i)} \tag{95}$$

$$\varepsilon_{n-1}(i) = d_{n-1}(i) - w^*_n(i)\varepsilon_n(i) \tag{96}$$

$$\text{where } \varepsilon_D(i) = d_D(i) \tag{97}$$

The estimated data is given by

$$\hat{b}_1(i) = \text{sgn}(w^*_1(i)\varepsilon_1(i)) \tag{98}$$

For the MWF, the matrices $\mathbf{B}_n$ are *blocking matrices*, which are selected to be orthogonal to the nested filers $\hat{\mathbf{c}}_n$. The scalar sequence $d_n(i)$ is the output of the filter $\hat{\mathbf{c}}_n$ and the filter in the next stage would be selected as

$$\hat{\mathbf{c}}_{n+1} = \frac{\mathbf{E}\,\{d^*_n\mathbf{r}_n\}}{\|\mathbf{E}\,\{d^*_n\mathbf{r}_n\}\|} \tag{99}$$

where $\mathbf{r}_n$ is the output of the blocking matrix $\mathbf{B}_n$. Thus, the stochastic gradient algorithm above provides adaptive estimators for these key quantities. The choice of blocking matrices is not unique. For example, one can select $\mathbf{B}_n = \mathbf{I} - \mathbf{c}_n\mathbf{c}^H_n$. However, it is shown in [10], that the projection matrix for the D-stage MWF algorithm is in fact independent of the choice of the blocking matrices within the class of row orthonormal blocking matrices.

4.6. Decision-Directed and Decision Feedback Methods

In the most simplistic of views, one could convert the training based adaptive algorithms previously described into blind adaptive algorithms by considering the "hard" decision of the receiver to be the true data and comparing this to the associated "soft" decision. This notion is depicted in Fig. 4. However, it should be noted that if this adaptive receiver is not initialized to a receiver vector that is close in some sense to the desired receiver vector, the algorithm could converge to the receiver of a more powerful user in a near-far environment. Algorithms that employ this idea are termed *decision-directed*.

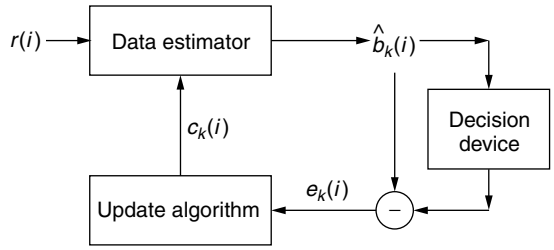


Figure 4. Decision-directed adaptive receiver.

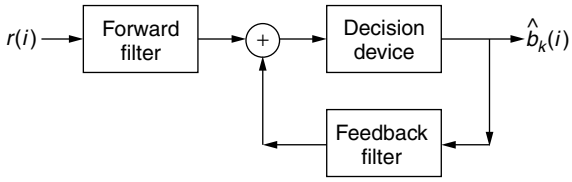


Figure 5. Decision feedback adaptive receiver.

Another set of algorithms that also feed back decisions, but are far more robust than the simple decision-directed scheme exhibited in Fig. 4, are decision feedback algorithms, depicted in Fig. 5. Adaptive schemes are used to determine the feedforward and feedback weight vectors. Examples of the design of such structures specifically for multiuser spread-spectrum systems can be found in the literature [44,51,52,58].

5. MULTIPATH FADING ENVIRONMENTS

In practice, the AWGN channel is too idealized of a model. In this section, the multipath channel model is introduced and the resultant signal model is provided. In addition to the assumption of a more realistic channel, we also consider possible asynchronism amongst the active users.

The received baseband signal corresponding to a single-symbol interval, which is coherently demodulated, chip-matched filtered and sampled at the chip rate is now described as

$$\mathbf{r}(i) = \sum_{k=1}^{K}\sum_{l=1}^{L_k} A_k[h_{kl}(i)b_k(i)\mathbf{s}^+_{kl} + h_{kl}(i-1)b_k(i-1)\mathbf{s}^-_{kl}] + \mathbf{n}(i) \tag{100}$$

We note that this received vector is of dimension $N \times 1$. In general for asynchronous, multipath systems, such short observations lead to degraded performance. We note, however, that the signal descriptions provided are easily generalized to the consideration of received signal vectors that correspond to multiple symbol intervals. The complex coefficients $h_{kl}(i)$ correspond to the multipath coefficients. The number of multipaths for user k is denoted by L_k. The partial spreading vectors $\mathbf{s}^+_{kl}$ and $\mathbf{s}^-_{kl}$ correspond to truncated and shifted versions of the spreading codes. For example, if we consider a two user system, with $L_1 = 2$

and $L_2 = 1$, employing spreading codes of length $N = 7$, a possible realization of the partial spreading codes would be

$$\begin{array}{lccccccc}
\mathbf{s}_{11}^{+} = [& 0 & 0 & \mathbf{s}_1(1) & \mathbf{s}_1(2) & \mathbf{s}_1(3) & \mathbf{s}_1(4) & \mathbf{s}_1(5) &] \\
\mathbf{s}_{11}^{-} = [& \mathbf{s}_1(6) & \mathbf{s}_1(7) & 0 & 0 & 0 & 0 & 0 &] \\
\mathbf{s}_{21}^{+} = [& 0 & 0 & 0 & 0 & \mathbf{s}_2(1) & \mathbf{s}_2(2) & \mathbf{s}_2(3) &] \\
\mathbf{s}_{22}^{+} = [& 0 & 0 & 0 & 0 & 0 & \mathbf{s}_2(1) & \mathbf{s}_2(2) &] \\
\mathbf{s}_{21}^{-} = [& \mathbf{s}_2(4) & \mathbf{s}_2(5) & \mathbf{s}_2(6) & \mathbf{s}_2(7) & 0 & 0 & 0 &] \\
\mathbf{s}_{22}^{-} = [& \mathbf{s}_2(3) & \mathbf{s}_2(4) & \mathbf{s}_2(5) & \mathbf{s}_2(6) & \mathbf{s}_2(7) & 0 & 0 &]
\end{array} \tag{101}$$

Here we note that user 1 has a delay of $\tau_1 = 2$ chips with respect to the receiver clock, while user 2 has a delay of $\tau_2 = 4$ chips. We can denote the discrete baseband equivalent representation of the multipath channel for user k as the complex vector $\mathbf{h}_k = [h_{k1}, h_{k2}, \ldots h_{kL_k}]^T$. This model assumes that the multipath channel can be represented as a finite-impulse response filter whose taps are spaced one chip (T_c) apart [48,60]. The *effective spreading code* is the convolution of the transmitted spreading sequence with this channel vector and is denoted by $\overline{\mathbf{s}}_k = \mathbf{h}_k \star \mathbf{s}_k$, where $\star$ denotes convolution. Note that $\overline{\mathbf{s}}_k$ is of length $N + L_k - 1$. If we let

$$\overline{\mathbf{s}}_k^{+} = \sum_{l=1}^{L_k} h_{kl}\mathbf{s}_{kl}^{+} \tag{102}$$

$$\overline{\mathbf{s}}_k^{-} = \sum_{l=1}^{L_k} h_{kl}\mathbf{s}_{kl}^{-} \tag{103}$$

$$\overline{\mathbf{S}}^{+} = \left[\overline{\mathbf{s}}_1^{+}, \overline{\mathbf{s}}_2^{+}, \cdots \overline{\mathbf{s}}_K^{+}\right] \tag{104}$$

$$\overline{\mathbf{S}}^{-} = \left[\overline{\mathbf{s}}_1^{-}, \overline{\mathbf{s}}_2^{-}, \cdots \overline{\mathbf{s}}_K^{-}\right] \tag{105}$$

then we can rewrite the received signal vector as

$$\mathbf{r}(i) = \overline{\mathbf{S}}^{+}\mathbf{A}\mathbf{b}(i) + \overline{\mathbf{S}}^{-}\mathbf{A}\mathbf{b}(i-1) + \mathbf{n}(i) \tag{106}$$

From this description it is clear to see that the K user asynchronous multiuser system can be considered as a $2K$ user synchronous system if observations corresponding to N chips (one symbol interval) are employed. We also observe the following relationship between $\overline{\mathbf{s}}_k$ and the vectors $\overline{\mathbf{s}}_k^{+}, \overline{\mathbf{s}}_k^{-}$. Let τ_k be the delay of user k in integer multiples of a chip; then

$$[\overline{\mathbf{s}}_k^{+}, \overline{\mathbf{s}}_k^{-}] = \left[\underbrace{0, 0, \ldots 0}_{1\times\tau_k}, \overline{\mathbf{s}}_k, \underbrace{0, 0, \ldots, 0}_{(N-\tau_k-L_k+1)\times 1}\right] \tag{107}$$

We note that if the channel of the desired user is completely known, that is, if we have knowledge of $\mathbf{h}_k$ and τ_k, then the previous algorithms can be applied directly where observations of length $N + L_k - 1$ are collected and the desired user's spreading code, $\mathbf{s}_k$, is replaced by the effective spreading code, $\overline{\mathbf{s}}_k$. If information about the interfering users is necessary, the interfering users are modeled as $2(K-1)$ synchronous users with spreading waveforms $\overline{\mathbf{s}}_k^{+}$ and $\overline{\mathbf{s}}_k^{-}$. The information required to implement the various adaptive receivers considered herein in the asynchronous multipath environment is noted in Table 1. This table is constructed with the view that each user is viewed as a *single* user with spreading code $\overline{\mathbf{s}}_k$. We note that an alternative view is possible, which can obviate the need for the channel vector $\mathbf{h}_k$, but necessitates knowledge of the channel length L_k and the relative delay τ_k. In this approach, each user is considered to be L_k users with a spreading code that is a shifted version of the other spreading codes. Then, receivers can be designed for each path. The final decision is made by combining the soft decisions from each L_k adaptive receiver. The choice of combining coefficients remains an open question. A typical approach is to consider *equal-gain combining* as in Barbosa and Miller [2]. This approach can be applied to the bulk of the receivers considered.

5.1. MMSE Receivers for Multipath Fading Channels

Because of its simplicity of implementation, strong performance, and amenability to adaptive implementation, there has been significant interest in applying the linear MMSE receiver to multipath fading channels. The construction of the true linear MMSE receiver requires knowledge of all the active users' spreading codes, timing, and channel state information [33,66]. However, an adaptive implementation of the this receiver can be achieved with prior information comparable to that of the conventional matched filter (MF) receiver, namely, information of the user of interest only and not that of the interfering users. Table 3 summarizes the applicability of the algorithms in section 4 to the multipath channel case.

Barbosa and Miller [2] proposed a modified MMSE receiver for flat fading channels in which the channel phase of the desired user is estimated and then compensated for in the MMSE receiver input. However, in frequency-selective fading channels, determining accurate estimates of the channel phases for all resolvable paths is at best challenging, and often impossible. As a result, noncoherent MMSE receivers become a more favorable choice for rapidly fading multipath environments. Several works have considered training-signal-independent, or blind approaches to developing MMSE-based receivers for multipath channels. Such receivers are robust to deep

Table 1. Information Required to Construct Nonadaptive Multiuser Receivers

Receiver	Signature of User 1, $\mathbf{s}_1$	Signature of AllUsers, $\mathbf{S}$	Relative Amplitudes, $\mathbf{A}$	Noise Variance, σ^2	Timing Information, τ_i
Matched filter	Yes	No	No	No	Yes
MMSE receiver	Yes	Yes	Yes	Yes	Yes
Decorrelator	Yes	Yes	No	No	Yes
Jointly optimal	Yes	Yes	Yes	No	Yes
Individually optimal	Yes	Yes	Yes	Yes	Yes

Table 2. Taxonomy of Adaptive Receivers in Terms of Direct or Indirect Implementation

Direct	Indirect
LMS	RBF [36]
RLS	—
Linear/letter distortion MPER[a]	—
MWF/AVF	—

[a]Minimum probility of error rate.

Table 3. Knowledge Required for Implementation of Adaptive Multiuser Receivers in an Asynchronous Multipath Environment

Need τ_k	Need $\bar{\mathbf{s}}_k, \tau_k$	Need $\bar{\mathbf{s}}_k, \tau_k \forall k$
LMS/RLS	RBF[a]	RBF
Letter distortion MPER	MWF/AVF	Linear MPER
MWF[a]	—	—

[a]An implementation is possible, but degraded performance will be experienced.

fades as they do not attempt to track the amplitude and phase fluctuations of the user of interest [18,21,46]. However, this robustness comes at the cost of higher excess MSE for adaptive implementations of such blind receivers. This feature is in contrast to training sequence based adaptive MMSE algorithms [46]. The training sequence based differential least-squares (DLS) algorithm proposed in [21] suffers from robustness to deep fades. Thus, to achieve acceptable performance and robustness simultaneously, the DLS algorithm is switched to a blind adaptive implementation when the receiver is in deep fade [21,46].

We next discuss two advances in the development of adaptive DS-CDMA receivers with a view to moderately fast fading environments. With the challenge of accurately estimating the channel phase in a fast fading environment, both methods consider differentially encoded phase shift keying (DPSK) to avoid estimating the channel phase. The first method [70] is initially developed for flat fading channels and then extended to multipath channels using the multipath combining technique noted above. The second method makes a key observation based on the technique of Zhu et al. [70] to develop an improved method for estimating the data correlation matrix, $\mathbf{C}$ [9].

5.1.1. Differential MMSE. As noted previously, the development of the differential MMSE criterion presumes a flat fading channel, thus the contribution in the received signal due to user k can be modeled as, $\alpha_k(i)b_k(i)\mathbf{s}_k$. The random process $\alpha_k(i)$ is complex-valued and represents the channel fading. The proposed modified MMSE-based cost function is

$$J(\mathbf{c}) = \mathbf{E}\left[|b_1(i-1)\mathbf{c}^H\mathbf{r}(i) - b_1(i)\mathbf{c}^H\mathbf{r}(i-1)|^2\right] \tag{108}$$

subject to

$$\mathbf{c}^H\mathbf{C}\mathbf{c} = 1 \tag{109}$$

where $\mathbf{C}$ remains the data correlation matrix; however, expectation is now taken over the channel coefficients as well as the data and the noise. The objective of this cost function is to suppress the multiple access interference while endeavoring to recover a scaled version of the datastream of the desired user. Thus the resultant soft decision will include an unknown complex scalar. By assuming that $\alpha_1(i) \approx \alpha_1(i-1)$, we can employ heuristic arguments to show the suppression of the multiple-access interference. By invoking some simple assumptions, it can be shown [70] that the solution to the cost function above is a scaled version of the true MMSE receiver in Eq. (76). The required assumptions are

1. $\mathbf{E}\left[\text{Re}\ (\alpha_1(i)\alpha_1^*(i-1))\right] > 0$
2. $\mathbf{E}[\text{Re}\ (b_1(i)b_1^*(i-1)\alpha_k(i)d_k(i)\alpha_j^*(i-1)d_j^*(i-1))]$
   $= \gamma\delta(k-1)\delta(j-1)$

where γ is a positive constant and $\delta(\cdot)$ is the Krönecker delta function.

The general solution to the minimization of Eq. (109) is the determination of the following generalized eigenvalue problem:

$$\mathbf{Q}\mathbf{c} = \lambda\mathbf{C}\mathbf{c} \tag{110}$$

where

$$\mathbf{Q} = \text{Re}\ \{\mathbf{E}\left[b_1(i)b_1^*(i-1)\mathbf{r}(i)\mathbf{r}(i-1)^H + b_1(i-1)b_1^*(i)\mathbf{r}(i-1)\mathbf{r}(i)^H\right]\} \tag{111}$$

An efficient algorithm, denoted the *power algorithm* [15], can be utilized to determine the desired generalized eigenvector $\mathbf{c}$. The power algorithm requires a key matrix $\mathbf{M} = \mathbf{C}^{-1}\mathbf{Q}$. Either block adaptive or recursive methods can be employed to estimate $\hat{\mathbf{C}}$ and $\hat{\mathbf{Q}}$ from the data. In addition, a gradient-based, recursive least-squares-type algorithm can be employed:

$$\mathbf{k}(i) = \frac{\mathbf{P}(i-1)\mathbf{r}(i)b_1^*(i-1)}{\lambda + |b_1(i-1)|^2\mathbf{r}(i)^H\mathbf{P}(i-1)\mathbf{r}(i)} \tag{112}$$

$$\zeta(i) = b_1(i)\mathbf{c}(i-1)^H\mathbf{r}(i-1) - b_1(i-1)\mathbf{c}(i-1)^H\mathbf{r}(i) \tag{113}$$

$$\mathbf{P}(i) = \lambda^{-1}\mathbf{P}(i-1) - \lambda^{-1}b_1(i-1)\mathbf{k}(i)\mathbf{r}(i)^H\mathbf{P}(i-1) \tag{114}$$

$$\mathbf{c}(i) = \frac{\mathbf{c}(i-1) + \beta_a\mathbf{k}(i)\zeta(i)^*}{|\mathbf{c}^H(i-1)\mathbf{r}(i-1)|} \tag{115}$$

The differential MMSE adaptive methods for flat fading channels can be extended to the multipath environment by constructing a correlator for each path and then combining the soft outputs.

5.1.2. Improved Correlation Matrix Estimation. While the modified differential MMSE criterion proposed in Ref. 70 offers solid performance in flat fading channels, and also enables various adaptive implementations, the receiver experiences significant degradation over the true MMSE receiver in frequency-selective fading channels. A further challenge to consider is that in the presence of unknown multipath (or imperfectly estimated multipath), there is performance degradation for MMSE based

receivers [21,35]. This is because in the fast multipath fading environment, an interfering user appears as multiple virtual users for the adaptive MMSE receiver, a phenomenon known as *interferer multiplication* [70]; it has been observed that the performance of the MMSE receiver degrades with the number of effective users in the system [2]. We note that this problem is not an issue in the flat fading environment where both training-sequence-based and blind adaptive MMSE detectors can achieve performance close to that of the true MMSE receiver, although the detector may not track the channel parameter of each interfering user perfectly [21,35].

The approach to be discussed herein makes the following key observation about the cost function of [70]. The new approach is based on observations for flat fading channels, however, the method offers strong performance improvements even in multipath fading channels [9]. In flat fading channels, where $L = 1$ [see Eq. (100)], the true $\mathbf{R}$ is given by

$$\mathbf{R} = \mathbf{R}_u + \mathbf{R}_I = P_1\mathbf{s}_1\mathbf{s}_1^T + \left\{\sum_{k=2}^{K} P_k[\mathbf{s}_k^+(\mathbf{s}_k^+)^T + \mathbf{s}_k^-(\mathbf{s}_k^-)^T] + \sigma^2\mathbf{I}_N\right\} \quad (116)$$

where $\mathbf{R}_u = P_1\mathbf{s}_1\mathbf{s}_1^T$ is the correlation matrix for user 1, $P_k = A_k^2$ and $\mathbf{R}_I$ is the interference correlation matrix which is defined implicitly in this equation. For this scenario, it can be shown through use of the matrix inversion lemma

$$\frac{\mathbf{R}^{-1}\mathbf{s}_1}{\mathbf{s}_1^T\mathbf{R}^{-1}\mathbf{s}_1} = \frac{\mathbf{R_I}^{-1}\mathbf{s}_1}{\mathbf{s}_1^T\mathbf{R_I}^{-1}\mathbf{s}_1} \quad (117)$$

In other words, $\mathbf{c}_{\text{MMSE}}$ can be expressed as a function of $\mathbf{R}_I$ only and not as a function of $\mathbf{R}_u$ [22]. This important property will form the basis of the proposed correlation matrix estimation scheme.

We next recall the objective function of Zhu et al. [70] in Eq. (109). One can show that

$$E\{\hat{b}_1(m)\hat{b}_1(m-1)\mathbf{y}(m)\mathbf{y}(m-1)^H\} \approx (\check{\mathbf{s}}_{1,1:L}^+)^H\check{s}_{1,1:L}^+ = \mathbf{R}_u \quad (118)$$

where the assumptions $\hat{b}_1(m) \approx b_1(m)$, $\gamma_{1l}(m) \approx \gamma_{1l}(m-1)$, $l = 1, \ldots, L$ and $E\{|\gamma_{1l}(m)|^2\} = 1$ have been used, and $\check{\mathbf{s}}_{1,1:L}^+ = \mathbf{s}_{1,1:L}^+\mathbf{A}_1$, $\mathbf{A}_1 = \text{diag}([A_{11}, A_{12}, \ldots, A_{1L_1}])$ and diag(.) denotes to diagonalize.

Now, the new correlation matrix estimation scheme is given by

$$\hat{\mathbf{R}}_I(m) = \hat{\mathbf{R}}(m) - \hat{\mathbf{R}}_u(m) = \lambda\hat{\mathbf{R}}_I(m-1) + \mathbf{r}(m)\mathbf{z}(m)^H \quad (119)$$

where

$$\mathbf{z}(m) = \mathbf{r}(m) - \hat{b}_1(m)\hat{b}_1(m-1)\mathbf{r}(m-1) \quad (120)$$

for the blind adaptive MMSE detector. The new correlation matrix estimate given in Eq. (120) results in significantly improved performance for the blind adaptive MMSE receiver. One might think that for asynchronous systems, the performance improvement is due to an equivalent observation window enlargement since $\hat{\mathbf{R}}_u(m)$ uses $\mathbf{r}(m-1)$ as well as $\mathbf{r}(m)$ for estimation. However, it turns out that even for synchronous flat fading environments, the performance gain is still evident.

We note that data decisions of $b_1(m)$ are needed for the adaptation of the receivers, as shown in Eq. (120). Similar to the decision-directed receivers, the proposed MMSE receivers also start with the conventional RAKE receiver [47] if the initialization of the estimate of $\mathbf{R}$ is given by a small identity matrix [17]. However, in contrast to the decision-directed adaptive MMSE receivers, which might lose track of the user of interest and lock on the user with the strongest signal instead, the blind adaptive MMSE receivers will always lock on the intended user since the steering vector is assumed to be known and we are always in the right direction.

6. DIFFERENT ENVIRONMENTS AND FURTHER EXTENSIONS

In this section, we discuss modifications of the previously discussed linear adaptive receiver algorithms based on minimizing the mean-squared error. These modifications are inspired by characteristics of the particular communications environment, signaling and/or reception scheme. In particular, we shall consider receivers tailored to: multiple data rates, binary phase shift keying (BPSK), and multiple sensors.

6.1. Multiple Data Rates

With the discussion of future standards [1,11,39], there has been heightened interest in wireless systems that offer multiple data rates in an integrated fashion. In such a system, users can consider transmitting at one of a class or possible data rates. The methods by which one can modify a data rate are varied. Herein, we shall focus on systems where the symbol rates among users are different. Digital communication signals are, in general, cyclostationary; that is, the received signal $r(t)$, is wide-sense cyclostationary with period T if

$$\mathbf{E}\{r(t)\} = \mathbf{E}\{r(t+T)\}$$

$$\mathbf{E}\{r(t_1)r^*(t_2)\} = \mathbf{E}\{r(t_1+T)r^*(t_2+T)\}$$

For a digital communications signal, this period T is the symbol duration. In a multiuser environment where all users transmit at the same data rate, the period of cyclostationarity is also T; if multiple users have different symbol rates, then the received sum signal retains its cyclostationary nature. However, the period of cyclostationarity is now the least common multiple of the individual symbol rates, that is, $T^* = \text{LCM}(T_1, T_2, \ldots, T_K)$. It can be shown that the optimal MMSE detector is time-varying, but periodic with period T^* [4,53,54]. For the

design of MMSE receivers, the desired time-varying filter will be the solution to

$$\mathbf{c}_k(i)_{\text{MMSE}} = \arg\ \min \mathbf{E}\{|b_k(i) - \mathbf{c}(i)^H \mathbf{r}(i)|^2\}$$

The optimal receiver $\mathbf{c}_k(i)_{\text{MMSE}}$, due to the fact that it is periodic, will have the following Fourier series representation:

$$\mathbf{c}_k(i)_{\text{MMSE}} = \sum_{q=1}^{R} \mathbf{c}_k^{(q)} \exp\left(\frac{j2\pi qi}{R}\right)$$

Let R be such that $T = R \min_i T_i$. The subfilters $\mathbf{c}_k^{(q)}$ are not time-varying. Thus, we can rewrite our desired criterion in (121) as

$$\mathbf{c}_k(i)_{\text{MMSE}} = \arg\ \min \mathbf{E}\{|b_k(i) - \tilde{\mathbf{c}}^H \tilde{\mathbf{r}}(i)|^2\}$$

where

$$\tilde{\mathbf{c}} = \left[\mathbf{c}_k^{(0)^H}, \mathbf{c}_k^{(1)^H}, \ldots, \mathbf{c}_k^{(R-1)^H}\right]^H$$

$$\tilde{\mathbf{r}}(i) = \mathbf{r}(i) \odot \theta(i)$$

$$\theta(i) = \left[1, e^{j2\pi i}, \ldots, e^{j2\pi(R-1)i}\right]^T$$

The Schur product operator is denoted by $\odot$; in the sequel, the Krönecker product operator will also be required, $\otimes$. The desired static receiver is given by

$$\tilde{\mathbf{c}} = \mathbf{E}\left[\tilde{\mathbf{r}}(i)\tilde{\mathbf{r}}(i)^H\right]^{-1} \mathbf{E}\left[b_k(i)\tilde{\mathbf{r}}(i)\right]$$

where

$$\mathbf{E}\left[b_K(i)\tilde{\mathbf{r}}(i)\right] = \mathbf{s}_1 \otimes \theta(i)$$

The receiver can be implemented as depicted in Fig. 6. These receiver structures have been investigated for both multimedia applications and narrowband interference suppression [4,53,54]. The extension to the case of of multipath channels is straightforward if the channel is known; the strategies for unknown channels as discussed above can be applied. The adaptive implementation of this multirate receiver is discussed by Buzzi et al. [4]. A host

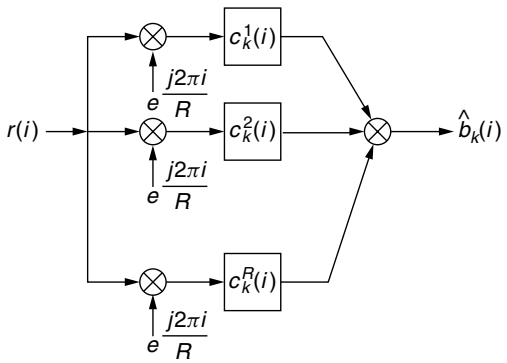


Figure 6. Implementation of MMSE receiver for multirate signaling.

of adaptive implementations are possible, as described in the prequel, with the modification that the observation is no longer $\mathbf{r}(i)$, but rather $\tilde{\mathbf{r}}(i)$.

6.2. Binary Signaling

In the presence of flat fading or multipath, the received signal is complex-valued; thus the use of the decision statistic for a linear receiver, under the assumption of BPSK signaling, is

$$\hat{b}_1(i) = \text{sgn}\{\text{Re}(\mathbf{c}^H \mathbf{r})\} \tag{121}$$

Thus, in contrast to the MMSE cost function in Eq. (76), it is of interest to consider optimizing the following cost function:

$$J(\mathbf{c}) = \mathbf{E}\left\{[b_1(i) - \text{Re}(\mathbf{c}^H \mathbf{r}(i))]^2\right\} \tag{122}$$

The desired receiver is found by forming the alternative, but equivalent optimization problem:

$$\mathbf{c} = \arg \min_{\mathbf{c}_a} \mathbf{E}\left\{|b_1(i) - \mathbf{c}_a^H \mathbf{r}_a(i)|^2\right\} \tag{123}$$

where

$$\mathbf{c}_a = [\mathbf{c}^H, \mathbf{c}^T]^H \tag{124}$$

$$\mathbf{r}_a(i) = [\mathbf{r}(i)^H, \mathbf{r}(i)^T]^H \tag{125}$$

The resulting solution is

$$\mathbf{c}_a = \mathbf{C}_a^{-1} \mathbf{p}_a \tag{126}$$

where

$$\mathbf{p}_a = \mathbf{E}\{b_1(i)\mathbf{r}(i)\} \tag{127}$$

$$\mathbf{C}_a = \begin{bmatrix} \mathbf{C} & \mathbf{C}' \\ \mathbf{C}'^* & \mathbf{C}^* \end{bmatrix} \tag{128}$$

$$\mathbf{C}' = \mathbf{E}[\mathbf{r}(i)\mathbf{r}(i)^T] \tag{129}$$

As in the case of multirate data signals as described in Section 6.1, the previous adaptive implementations of an MMSE-based algorithm can be constructed by replacing the observation $\mathbf{r}(i)$ with $\mathbf{r}_a(i)$. Given that the algorithm exploits further structure in the received signal, the binary signaling-based MMSE receiver outperforms the conventional MMSE receiver [3].

7. CONCLUSIONS

In this article, methods for adaptive reception of DS-CDMA multiuser signals has been provided. Key static receivers were reviewed and their adaptive counterparts provided. As there is much structure embedded in multiuser DS-CDMA receivers, a variety of specialized adaptive receivers are possible and have been pursued. For further reading, several tutorials on

adaptive multiuser detection have been written. Perhaps the most extensive is the paper by Woodward and Vucetic [65]. The Honig–Tsatsanis article [19] focuses on blind adaptive algorithms based on second order statistics as well as the reduced-rank methods described in Section 4.5. An extensive bibliography, coupled with the derivation and description of the MOE algorithm [23], has been provided [62]. In addition, two texts on adaptive receivers are suggested: Refs. 24 and 17. These texts consider the derivation of a host of adaptive algorithms based on a variety of cost functions and also describe the properties of the derived adaptive receivers.

BIBLIOGRAPHY

1. F. Adachi, M. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communication systems, *IEEE Commun. Mag.* 56–69 (Sept. 1998).
2. A. N. Barbosa and S. L. Miller, Adaptive detection of DS-CDMA signals in fading channels, *IEEE Trans. Commun.* **46**(5): 115–124 (Jan. 1998).
3. S. Buzzi, M. Lops, and A. Tulino, A new family of MMSE multiuser receivers for interference suppression in DS-CDMA systems employing BPSK modulation, *IEEE Trans. Commun.* **49**(1): 154–167 (Jan. 2001).
4. S. Buzzi, M. Lops, and A. Tulino, Partially blind adaptive MMSE interference rejection in asynchronous DS-CDMA networks over frequency selective fading channels, *IEEE Trans. Commun.* **49**(1): 94–108 (Jan. 2001).
5. G. Caire, Adaptive linear receivers for DS-CDMA, Part 1: Steady-state performance analysis, *IEEE Trans. Commun.* **48**(10): 1712–1724 (Oct. 2000).
6. D. S. Chen and S. Roy, An adaptive multi-user receiver for CDMA systems, *IEEE J. Select. Areas Commun.* **12**(5): 808–816 (June 1994) (issue on code-division multiple-access networks II).
7. S. Chen, G. J. Gibson, C. F. N. Cowan, and P. M. Grant, Reconstruction of binary signals using an adaptive radial-basis-function equalizer, *Signal Process.* **22**: 77–93 (1991).
8. S. Chen and B. Mulgrew, Overcoming co-channel interference using an adaptive radial basis function equalizer, *Signal Process.* **28**: 91–107 (1992).
9. W. Chen and U. Mitra, An improved blind adaptive MMSE receiver for fast fading DS-CDMA channels, *IEEE J. Select. Areas Commun.* **2**: 758–762 (2001).
10. W. Chen, U. Mitra, and P. Schniter, Reduced rank detection schemes for DS-CDMA communication systems, *IEEE Trans. Inform. Theory* (in press).
11. E. Dahlman, B. Gudmundson, M. Nilsson, and J. Sköld, UMTS/IMT2000 based on wideband CDMA, *IEEE Commun. Mag.* 70–80 (Sept. 1998).
12. E. Eleftherious and D. D. Falconer, Tracking properties and steady-state performance of RLS adaptive filter algorithms, *IEEE Trans. Acoust. Speech Signal Process.* **34**(5): 1097–1110 (Oct. 1986).
13. J. S. Goldstein and I. S. Reed, Reduced rank adaptive filtering, *IEEE Trans. Signal Process.* **45**(2): 492–496 (Feb. 1997).
14. J. S. Goldstein, I. S. Reed, and L. L. Scharf, A multistage representation of the Wiener filter based on orthogonal projections, *IEEE Trans. Inform. Theory* **44**(7): 2943–2959 (Nov. 1998).
15. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins Univ. Press, Baltimore, 1989.
16. M. Haardt et al., The TD-CDMA based UTRA TDD mode, *IEEE J. Select. Areas Commun.* **18**(8): 1375–1386 (Aug. 2000).
17. S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
18. M. Honig, U. Madhow, and S. Verdu, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **41**(4): 944–960 (July 1995).
19. M. Honig and M. Tsatsanis, Adaptive techniques for multiuser CDMA receivers, *IEEE Signal Process. Mag.* **17**(3): 49–61 (May 2000).
20. M. L. Honig and J. S. Goldstein, Adaptive reduced-rank interference suppression based on the multi-stage Weiner filter, 2000.
21. M. L. Honig, S. L. Miller, M. J. Shensa, and L. B. Milstein, Performance of adaptive linear interference suppression in the presence of dynamic fading, **49**(4): 635–645 (April 2001).
22. M. L. Honig and W. Xiao, Performance of reduced-rank linear interference suppression for DS-CDMA, **47**(5): 1928–1946 (July 2001).
23. Michael Honig, Upamanyu Madhow, and Sergio Verdú, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **41**(4): 944–960 (July 1995).
24. M. L. Honig and D. G. Messerschmitt, *Adaptive Filters*, Kluwer, Boston, 1984.
25. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1985.
26. W.-S. Hou and B.-S. Chen, Adaptive detection in asynchronous code-division multiple-access system in multipath fading channels, *IEEE Trans. Commun.* **48**(5): 863–874 (May 2000).
27. S. Lee and J. Dickerson, Adaptive minimum dispersion interference suppression for DS-CDMA systems in non-gaussian impulsive channels, *Proc. of Milcom'97*, IEEE, Nov. 1997, Vol. 2, pp. 857–861.
28. D. Lowe, Adaptive radial basis function nonlinearities, and the problem of generalization, *Proc. 1st IEE Int. Conf. Artificial Neural Networks*, IEE, Oct. 1989, pp. 171–175.
29. R. Lupas and S. Verdú, Linear multiuser detectors for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **35**(1): 123–136 (Jan. 1989).
30. R. Lupas and S. Verdú, Near-far resistance of multi-user detectors in asynchronous channels, *IEEE Trans. Commun.* **38**: 496–508 (April 1990).
31. J. B. MacQueen, Some methods of classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.
32. U. Madhow and M. L. Honig, MMSE interference suppression for direct-sequence spread spectrum CDMA, *IEEE Trans. Commun.* **42**(12): 3178–3188 (Dec. 1994).

33. U. Madhow and M. L. Honig, MMSE interference suppression for DS/SS CDMA, *IEEE Trans. Commun.* **42**: 3178–3188 (Dec. 1994).

34. N. B. Mandayam and B. Aazhang, Gradient estimation for sensitivity analysis and adaptive multiuser interference rejection in code-division multiple-access systems, *IEEE Trans. Commun.* **45**(7): 848–858 (July 1997).

35. M. L. Miller, S. L. Honig, and L. B. Milstein, Performance analysis of MMSE receivers for DS-CDMA in frequency-selective fading channels, **48**(11): 1919–1929 (Nov. 2000).

36. U. Mitra and H. V. Poor, Neural network techniques for adaptive multi-user demodulation, *IEEE J. Select. Areas Commun.* **12**(9): 1460–1470 (Dec. 1994) (issue on intelligent communications systems).

37. S. Moshavi, E. G. Kanterakis, and D. L. Schilling, Multistage linear receivers for DS-CDMA systems, *Int. J. Wireless Inform. Networks* **3**(1): 1–17 (1996).

38. T. Ojanperä and R. Prasad, An overview of air interface multiple access for IMT-2000/UMTS, *IEEE Commun. Mag.* 82–95 (Sept. 1998).

39. Y. Okamura and F. Adachi, Variable-rate data transmission with blind rate detection for coherent DS-CDMA mobile radio, *IEICE Trans. Commun.* **E81-B**: 71365–71373 (July 1998).

40. T.-B. Oon, R. Steele, and Y. Li, Performance of an adaptive successive serial-parallel CDMA cancellation scheme in flat rayleigh fading channels, *Vehic. Technol. Conf.* **49**(1): 130–147 (Jan. 2000).

41. D. A. Pados, F. J. Lombardo, and S. N. Batalama, Auxiliary-vector filters and adaptive steering for DS-CDMA single-user detection, *IEEE Trans. Vehic. Technol.* **48**(6): 1831–1839 (Nov. 1999).

42. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed., McGraw-Hill, New York, 1984.

43. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread spectrum communications — a tutorial. *IEEE Trans. Commun.* **30**(5): 855–884 (May 1982).

44. H. Ping, T. Tjhung, and L. Rasmussen, Decision-feedback blind adaptive multiuser detector for synchronous CDMA system, *Vehic. Technol.* **49**(1): 159–166 (Jan. 2000).

45. T. Poggio and F. Girosi, Networks for approximation and learning, *Proc. IEEE* **78**(9): 1481–1497 (1990).

46. H. V. Poor and X. Wang, Code-aided interference suppression for DS-CDMA communications — Part II: Parallel blind adaptive implementations, *IEEE Trans. Commun.* **45**(9): 1112–1122 (Sept. 1997).

47. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

48. J. G. Proakis, *Digital Communications*, 2nd ed., McGraw Hill Series in Communications and Signal Processing, McGraw-Hill, New York, 1989.

49. I. N. Psaromiligkos, S. N. Batalama, and D. A. Pados, On adaptive minimum probability of error linear filter receivers for DS-CDMA channels, *IEEE Trans. Commun.* **47**(7): 1092–1102 (1999).

50. M. B. Pursley and D. V. Sarwate, Performance evaluation of phase-coded spread-spectrum multiple-access communication — Part II: Code sequence analysis, *IEEE Trans. Commun.* **25**(8): 800–803 (Aug. 1977).

51. P. Rapajic, M. Honig, and G. Woodward, Multiuser decision-feedback detection: performance bounds and adaptive algorithms, *Proc. ISIT*, IEEE, Nov. 1998, p. 34.

52. P. B. Rapajic and B. S. Vucetic, Adaptive receiver structures for asynchronous CDMA systems, *IEEE J. Select. Areas Commun.* **12**(4): 685–697 (May 1994).

53. A. Sabharwal, U. Mitra, and R. Moses, Low complexity MMSE receivers for multirate DS-CDMA systems, *Proc. 2000 Conf. Information Sciences and Systems*, Princeton, NJ, March 2000, Vol. 1, pp. TA3–TA18.

54. A. Sabharwal, U. Mitra, and R. Moses, MMSE receivers for multirate DS-CDMA systems, *IEEE Trans. Commun.* **49**(12): 2184–2197 (Dec. 2001).

55. K. S. Schneider, Optimum detection of code division multiplexed signals, *IEEE Trans. Aerospace Electron. Syst.* **16**: 181–185 (Jan. 1979).

56. K. Simon, J. Omura, R. Scholtz, and B. Levitt, *Spread Spectrum Communications*, 2nd ed., Vol. III, Computer Science Press, New York, 1985.

57. R. Singh and L. B. Milstein, Interference suppression for DS-CDMA, *IEEE Trans. Commun.* **47**(3): 446–453 (March 1999).

58. J. E. Smee and S. C. Schwartz, Adaptive feedforward/feedback architectures for multiuser detection in high data rate wireless CDMA networks, *IEEE Trans. Commun.* **48**(6): 996–1011 (June 2000).

59. H. Stark and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1994.

60. G. L. Stüber, *Principles of Mobile Communication*, Kluwer, Boston, 1996.

61. S. Ulukus and R. D. Yates, A blind adaptive decorrelating detector for CDMA systems, *IEEE J. Select. Areas Commun.* **16**(8): 1530–1541 (Oct. 1998).

62. S. Verdú, Adaptive multiuser detection, *Proc. IEEE ISSSTA*, IEEE, July 1994, Vol. 1, pp. 43–50.

63. S. Verdú, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **32**(1): 85–96 (Jan. 1986).

64. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, Cambridge, UK, 1998.

65. G. Woodward and B. S. Vucetic, Adaptive detection for DS-CDMA, *Proc. IEEE* **86**(7): 1413–1434 (July 1998).

66. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multi-user communications, *IEEE J. Select. Areas Commun.* **8**: 683–690 (May 1990).

67. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multiuser communications, *IEEE J. Select. Areas Commun.* **8**(4): 683–690 (May 1990).

68. G. Xue, J. Weng, T. Le-Ngoc, and S. Tahar, Adaptive multistage parallel interference cancellation for CDMA, *IEEE J. Select. Areas Commun.* **17**(10): 1815–1827 (Oct. 1999).

69. C.-C. Yeh and J. R. Barry, Adaptive minimum bit-error rate equalization for binary signaling, *IEEE Trans. Commun.* **48**(7): 1226–1235 (July 2000).

70. L. J. Zhu, U. Madhow, and L. Galup, Differential MMSE: Two applications to interference suppression for direct-sequence CDMA, manuscript in preparation, 2000.

ADMISSION CONTROL IN WIRED NETWORKS

SYMEON PAPAVASSILIOU
JIE YANG
New Jersey Institute of Technology
University Heights
Newark, New Jersey

1. INTRODUCTION

Within the wired communications infrastructure, where switches and/or routers are deployed to support voice and data transmission, there are basically two key switching techniques: circuit switching and packet switching.[1] Traditionally, circuit switching is mostly used in telephone networks and packet switching is used in data networks including today's Internet. With the development of Broadband Integrated Services Digital Network (B-ISDN), where voice and various data services are provided in a common network infrastructure, packet switching has been widely deployed. Independent of the applied switching technique, to provide end-to-end connection and communication through the network, some network resources, such as link capacity, switching bandwidth, and buffers, are utilized. Since the amount of such resources are limited compared to the fast-increasing demand on voice and data communication, if at some point the requests on the resources exceed the available network resources, the network is considered as "congested." When the network is congested, either the connection can not go through or the quality of service (QoS) degrades. The technique to avoid congestion in a network is referred to as "congestion control." In general, there are two approaches to perform congestion control: (1) the *preventive approach*, where each connection reserves resources in advance, from which the idea of admission control stems; and (2) the *reactive approach*, where when congestion occurs flow control is performed via end-to-end closed-loop control mechanism or open-loop control mechanism at the intermediate network nodes.

Admission control is a more effective way to perform congestion control in high-speed networks because when congestion occurs, even though the network could react promptly, a large amount of data may be affected. Moreover, combined with resource allocation, admission control can reserve sufficient resources in advance for a connection so that its QoS can be guaranteed. Admission control can also be utilized as a mechanism to check and enforce policies before providing services to the users.

In traditional telephone networks, where circuit-switching is applied, congestion control is usually achieved via the first approach: admission control. Before users can talk to each other, a path from the sender to the receiver has to be set up and the required resources have to be reserved first. If many users want to use the telephone network at the same time and there are insufficient resources to be allocated, the admission control mechanism will turn down some new requests so that the admitted calls can be supported with satisfactory QoS.

In data networks including B-ISDN networks and the Internet, two scenarios may occur. If the network provides connection-oriented service, such as the ATM used in B-ISDN networks, usually both admission control and flow control are deployed. If the network provides connectionless service, such as today's best-effort Internet service, only flow control will be applied. However, newly proposed Internet service models that provide more services and better QoS also require admission control mechanisms.

2. OVERVIEW OF ADMISSION CONTROL

An admission control scheme basically consists of signaling messages and admission control units that perform the admission control algorithm or policy. The signaling is usually a part of the call setup/release signaling procedure. The connection to be set up could be either duplex or simplex. Three types of signaling may be involved in this procedure: (1) the signaling between the end user and the network access point, (2) the signaling inside the network, and (3) the signaling between networks.[2] The signaling between the user and network access point is used to send and respond to the call setup request. The signaling inside the network can be used to identify the current resources available for the new calls, inform the switches/routers to prepare for the new call, or inquire as to whether they are able to admit it. If the call needs to trespass several network domains, signaling between networks is required to check whether this new call can be supported by all the networks.

An admission control algorithm or policy is performed by the admission control units in the switch/router or some specified components/devices in the network. Admission control algorithms can be categorized into resource-based and policy-based. *Resource-based algorithms* base their decision on the current resource usage in the network, the resources needed to be allocated to the new call and whether the required QoS can be guaranteed by the network. Moreover, the already admitted calls shall not be affected if a new call is admitted.

Policy-based admission control is needed when different policies are enforced in the network. Its basic purpose is to determine whether a user is qualified to access the network service at a specific time.

3. ADMISSION CONTROL SCHEMES

In the following paragraphs we discuss admission control schemes in two types of networks: B-ISDN and Internet. The admission control for B-ISDN reflects the typical

[1] In today's wired telecommunication networks, especially telephone and data networks, another switching technique, namely the message switching is not widely used.

[2] The signaling protocols for the three cases do not have to be the same.

schemes of current connection-oriented networks that utilize common channel signaling. The schemes designed for the Internet reflect the tendency of future Internet services and models.

3.1. Admission Control in B-ISDN Networks

In B-ISDN networks where ATM is utilized as the transport technology, admission control procedure is implemented through signaling messages and admission control units of the switches. Two types of signaling are involved in the admission control procedure: the signaling at the user–network interface (UNI)[3] and the signaling at the network–network interface (NNI).[4] The ITU-T recommendation for UNI access signaling protocol is Q.2931 [1], which is a modified version of Q.931 — the access signaling protocol for ISDN. In ITU-T recommendations, the signaling protocol for NNI is B-ISUP [2], the B-ISDN user part of Signaling System 7 (SS7) — a widely used signaling protocol in today's telephone networks. ATM forum also defines Private Network-Network Interface (PNNI) protocol to address the issues of interswitch, internetwork, and routing operations, which are not specified in the ITU-T recommendations. In the following paragraphs we describe the signaling procedure that involves admission control based on ITU-T recommendations.

Generally, admission control is coupled with resource management that is involved in two procedures: call setup and call release. An example of a successful point-to-point call setup and release procedure is shown in Fig. 1. In the call setup procedure, the signaling messages SETUP, CALL PROCEEDING, ALERTING, CONNECT, and CONNECT ACK comply with the specification of Q.2931 and the signaling messages IAM, IAA, ACM, and ANM comply with B-ISUP.

When the switch at the UNI (switch A in the example) receives the SETUP message, which includes the calling and called party identities, such as the ongoing traffic descriptor, switch A will reply with CALL PROCEEDING

[3] UNI refers to the interface between the user and the network.
[4] NNI refers to the interfaces between the switches.

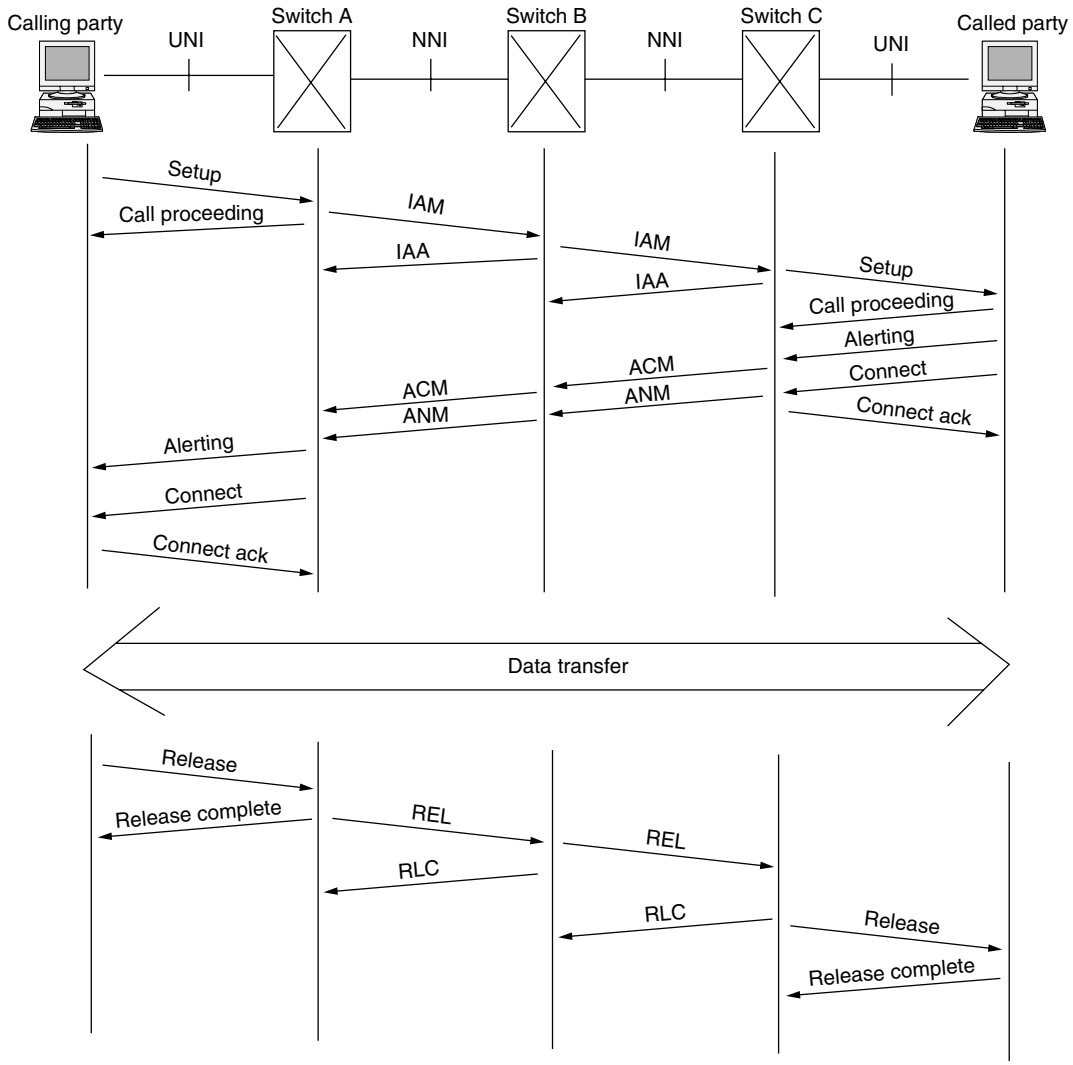


Figure 1. An example of successful point-to-point call setup and release procedures.

message if its admission control algorithm decides that this call can be accepted by the switch. This decision is made based on the identities of calling and called party, the route and resource availability, the traffic descriptor and the corresponding QoS requirements. If the route to the called party trespasses several switches, B-ISUP will be invoked. The switch will send an IAM (initial address message) to its next-hop switch (switch B in the example), which contains all the information in the call SETUP message. An IAA (IAM acknowledgment) will be sent back to switch A. If the next-hop switch can admit this call, and it is not the switch through which the called party accesses the network, it will forward the IAM to its next-hop switch (switch C). The same procedure will repeat until IAM reaches the switch connected to the called party, which is switch C in our example. If switch C is able to admit the call, it will issue a Q.2931 SETUP message through the UNI to the called party. The called party will reply in any of the following three ways: CALL PROCEEDING, ALERTING, and CONNECT if the call can be accepted. If the called party replies with CALL PROCEEDING, the network will wait for the following ALERTING or CONNECT messages from the called party. If the called party replies with ALERTING message, switch C will issue an address complete message (ACM) backward to switch A, where an ALERTING message will be issued to the calling party. Then the network will wait for the CONNECT message from the called party. If the called party replies with CONNECT message, switch C will issue an ANswer Message (ANM) backward to switch A, which indicates the connection is activated. Correspondingly, switch A will issue a CONNECT message to notify the calling party on the activation of the call. The calling party and switch C will also reply with a CONNECT ACK message to switch A and the called party respectively. Thereafter the data transfer phase can start.

The call can be terminated by either the users (calling party or called party) or the network. At the UNI, a RELEASE message will be initiated and a RELEASE COMPLETE message will be responded. Correspondingly, at NNI, a RELease (REL) message and a ReLease Complete (RLC) message will be exchanged between the switches. On receiving the Release message from UNI or REL message from NNI, a switch will release the resource reserved for the connection and these resources become available to the network.

During the setup procedure, if any switch detects that the call cannot be admitted, the call will be rejected. The switch will issue an IAM reject (IAR) message to its previous-hop switch from which it receives the IAM request. If no alternate route can be found, this IAR message will be sent backward until it reaches the edge of the network, where the switch will send a RELEASE message to the calling party.

To support point-to-multipoint connections, ATM Forum defines some supplementary messages for UNI [3]. It also defines PNNI protocol for Private NNI, which includes routing and signaling protocols across private ATM networks. For further details on signaling and its implementation, readers may refer to the corresponding standards and recommendations [1,2,4].

Admission control decisions are made and resource management is performed at each switch by its admission control unit, which can be a centralized module or a decentralized component deployed at each input or output module of the switch. The algorithm performed at the admission control unit is decided by the specific network administrators or equipment vendors and is not standardized. In Section 4 we briefly introduce and discuss some of the proposed approaches.

3.2. Admission Control on the Internet

Traditionally, the Internet only provides connectionless best-effort service; therefore, no admission control at the IP level is required. However, with the wide deployment of the Internet, many applications including many real-time applications may choose the TCP/IP protocol suite as their transport technology. These applications will require better QoS guarantees than what the best-effort service can provide. To address such a demand, new Internet service models have been proposed and are being developed that include the Integrated Service model (Intserv) [5] and the Differentiated service model (Diffserv) [6]. In the Intserv model, traffic is identified by a flow, a concept similar to the virtual connection in ATM; and a signaling protocol, namely, the resource ReSerVation Protocol (RSVP) [7] was proposed so that admission control and resource allocation can be performed in a dynamic manner. In the Diffserv model, traffic is aggregated and classified by service classes, which is significantly different from that in the Intserv model. In the Diffserv model services can be provided based on static or dynamic admission control and resource allocation.

3.2.1. Admission Control via RSVP in the Intserv Model. RSVP is used to setup a simplex path between two hosts. If duplex communication is required, two separate paths in each direction have to be setup via RSVP signaling procedure. A successful point-to-point path setup procedure is shown in Fig. 2. At the beginning of the procedure, the sender (host A) sends a PATH message to the receiver, which sets up a path from the source to the destination (downstream). The PATH message contains the "previous hop" information that the message has trespassed, the information about the sender, the traffic descriptor and the information about the status of the network. It installs a "path state" at each intermediate node. When the receiver (host D) receives the path message, it will send a RESV message back to the sender along the path that was set up by the PATH message (upstream). The RESV message makes resource reservation request at each node, which will decide whether to accept this request, based on both the availability of the resources and the policy enforced in the network. If the request can be admitted, a "reservation state" will be created at the node and the RESV message will continue to be forwarded along the path until it reaches the sender. Otherwise the request will be rejected and an error message will be sent to the receiver. Once the sender has received the RESV message, it can start its data transmission to the receiver. The path state and reservation state at the intermediate routers form a "soft state," which must be refreshed by PATH and

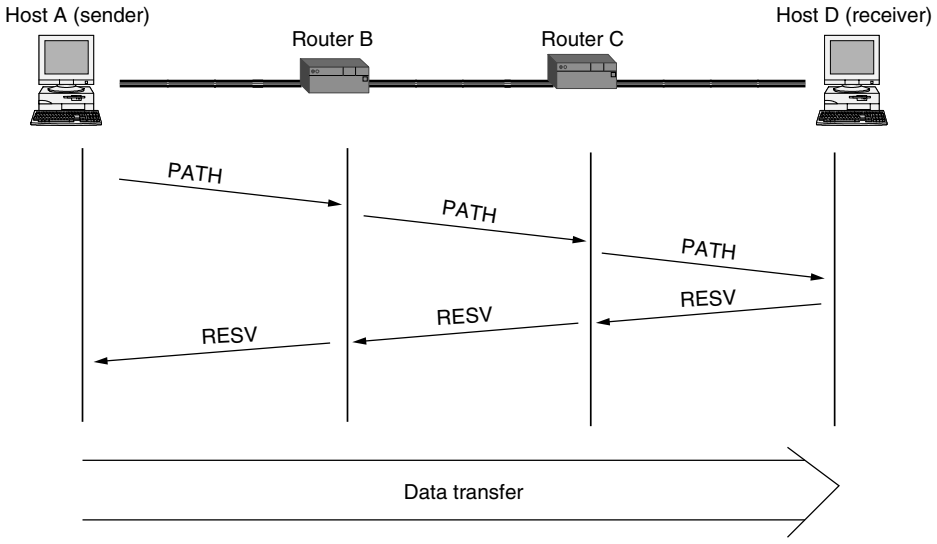


Figure 2. An example of successful point-to-point call setup via RSVP.

RESV messages periodically during data transmission. The state is timed out and deleted if no matching PATH or RESV messages arrive within the timeout interval. In this case the reservation is released. The state is also able to be torn down explicitly by the hosts (sender or receiver) when data transfer finishes. RSVP also supports multicast and flow aggregation.

The resource reservation request should be mapped to the link-layer technology where the resource-based admission control and resource allocation algorithms are actually deployed. The IETF workgroup ISSLL have developed some specifications on how to map an RSVP reservation request onto specific link layer technologies such as ATM [8] and Ethernet [9].

3.2.2. Admission Control in Diffserv Networks. The underlying principle of the Diffserv model is that service is provided to users based on the service-level agreement (SLA) between the user and the service provider. Traffic flows will be marked by the host or leaf router and classified, metered, shaped, and possibly re-marked at the ingress router of the Diffserv network where the flows will be aggregated according to the service class set by SLA and forwarded to the core routers. At core routers, the QoS is provided by the "Per hop behavior" (PHB) associated with service classes. SLA can be static, half-static or dynamic.

In static SLA, the service provided to the user is negotiated in advance and may be manually configured by the network administrator periodically. At the ingress node of the network, the flows of a user will be monitored to ensure its conformance to the SLA. Therefore, admission control can be performed in a static and implicit manner by which there is no signaling procedure to request resource reservation.

In half-static and dynamic SLA, there exists an agent in the Diffserv network, called "bandwidth broker" (BB), to manage the resources in its domain. In the half-static mode, resources have been preallocated to the users according to their SLA; however, it needs a signaling procedure to activate and install states at the border (ingress and egress) routers before transmission, and to deactivate and clear states after transmission. Usually the duration of such a state is in the time-scale of hours. In the dynamic mode, there are no preallocated resources to the users. The user has to use an explicit signaling protocol to request admission to the network. In both modes, there are two options to perform admission control and resource management. First, only boundary nodes are signaling-aware. The resource is managed through resource management agents, for example, BB. The interior routers are not signaling-aware. In this case, an admission request will be forwarded to the BB from the edge node by using signaling messages. The second option is that the interior routers are also signaling-aware and the signaling procedure is hop by hop similar to the procedure described in Intserv. However, the interior nodes will schedule and forward traffic only on the basis of the traffic class, while in the Intserv model traffic is scheduled and forwarded according to the flow specifications. The signaling protocol in Diffserv can be RSVP, extensions of RSVP or any other customized protocol.

3.2.3. Policy-Based Admission Control on the Internet. Policy-based admission control is still in its early stage. Generally, it is complementary to the resource-based admission control to resolve those issues not addressed by resource-based admission control, which may include priority of users and applications, security, or time-of-day traffic. RFC 2753 [10] provides a framework for policy-based admission control in which two basic network elements were proposed: policy enforcement point (PEP) and policy decision point (PDP). PEP is used to enforce policy for admission and PDP is the component that actually makes policy decisions. Usually PEP is deployed as a function unit in the edge routers and PDP is deployed as a centralized server in the network. An optional local decision point (LDP) can be deployed together with PEP to provide local policy decision. PDP may also need to access other servers in the network to reach a policy decision. A typical configuration of a QoS-enabled network with policy-based admission control capability is shown in

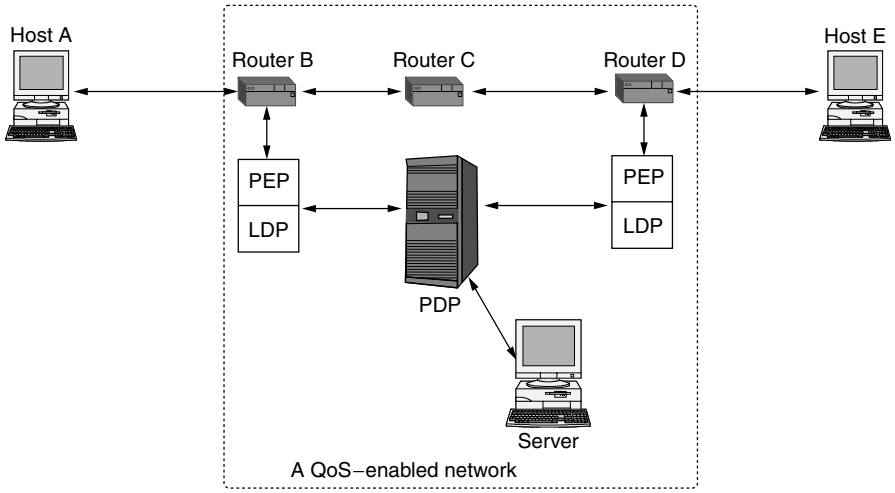


Figure 3. A typical configuration of QoS-enabled network with policy-based admission control.

Fig. 3, which can be applied to both Intserv networks and Diffserv networks.

In a network that enables both resource-based and policy-based admission control, when a signaling message reaches the edge router requesting admission to the network and reserving resources, the router will first check whether there are sufficient resources for the request through the local resource management unit or through the remote resource management components. Then it will check through PEP whether there are policies to be enforced to the request. The PEP will inquire LDP and PDP as to possible policy decisions. This can be referred to as a dynamic mode. PDP and LDP are also able to issue policies to be enforced at PEP simply according to the conditions of the network or the policies of the service provider, which can be referred to as a static mode. Since PDP is usually deployed at a remote server, a protocol between PEP and PDP is needed for them to communicate. A candidate for such a protocol is COPS [11]. COPS employs a client/server model and uses TCP as its transport protocol for reliable communication between PEP and PDP. A PEP can send its request through COPS REQ message and receive decision from PDP through COPS DEC message. An RSVP signaling message is also able to carry policy requests so that end-to-end policies can be enforced. The extension of RSVP to support policy-based admission control is proposed in RFC 2750 [12], in which a new data object, policy data object, is defined for this purpose.

3.2.4. MPLS and Its Admission Control. In connectionless networks such as today's Internet, each router analyzes the network layer header of a packet and makes its routing and forwarding decision independently, which is referred to as "hop by hop." However, a router will handle packets with the same forwarding equivalence class (FEC) in the same manner and these packets are actually indistinguishable to the router [13]. Moreover, it is believed that the network layer header provides much more information than necessary to perform routing and forwarding [13], which also makes it difficult to achieve fast analysis of the header and makes routing of packets a bottleneck in high-performance routers. Therefore, multiprotocol label switching (MPLS) was proposed, in which a small fixed label was bound to the packet to locally identify FEC. Similar to an ATM virtual circuit, a label-switched path (LSP) has to be set up so that the adjacent label−switching routers (LSRs) along the LSP can forward packets by simply looking into the locally bound labels without having to analyze the entire network layer header. A label distribution protocol (LDP) is used by an LSR to inform the other LSR about the label binding to a specific FEC. During the label assignment and distribution procedure, attributes of the FEC can be set up so that QoS and policies of the FEC can be specified. Since an LSP is similar to an ATM virtual circuit, admission control is performed during the LSP setup and LDP is used as the signaling. Although Rosen [13] does not specify the LDP scheme, it requires that label binding is determined by the downstream LSR. MPLS is a technology between layers 2 and 3 (referred to as layer 2.5 technology), whose resource management and admission control algorithm depend on the underlying layer 2 technology.

4. ADMISSION CONTROL ALGORITHMS

As can be seen from the previous discussion, resource-based admission control algorithms are closely coupled with the problem of resource management. Generally, a resource-based admission control algorithm consists of two parts: estimation of resource usage and QoS performance, and optimized decisionmaking on whether to accept the connection request. Resource-based admission control algorithms have been under intensive research and development efforts since the emergence of ATM. This is due to several reasons:

1. Data networks, including ATM and the Internet, support and encourage statistical multiplexing so that the resource utilization can be enhanced, which, however, increases the algorithm's complexity. The estimation of resource usage must be accurate so that the QoS performance will not be harmed as a result of the statistical multiplexing. This, in turn, requires the development and use of a good model for

estimation of resource usage and QoS performance, and an optimized solution to it.

2. There are many different services provided by the data networks that have quite different statistical nature and QoS requirements. This complicates the modeling of the problem and makes the optimization of the solution quite difficult.
3. Admission control must be performed in real time. Therefore, the algorithm needs to be accurate on one hand, and simple enough on the other hand. However, these two requirements are sometimes contradictory to each other.

Compared to resource-based admission control, policy-based admission control is an approach that manages the network resources in a relatively static manner. A connection will be accepted if it matches the policies of the network. When a connection falls into several policies, the network node needs to find which policy fits the connection best and basis its admission decision on that policy.

4.1. Model-Based Admission Control

Resource-based admission control algorithms can be further divided into model-based algorithms and measurement-based algorithms. Some traffic descriptors are used to model the traffic and measure the performance in both types of algorithms. In ATM, these descriptors include peak cell rate, sustainable cell rate (SCR), and maximum burst size (MBS) [3]. Two approaches have been adopted in model-based algorithms: deterministic approach and statistical approach.

The *deterministic* approach allocates resources simply according to peak data rate while assuming no data loss. This approach was adopted by telephone networks and CBR traffic in ATM. The advantage of the deterministic approach is that it is easy to be implemented and can be performed in real time. However, although QoS of each connection will be guaranteed, the resources, such as bandwidth and buffer allocated to each connection, cannot be shared, which in turn may lead to underutilization of resources because usually arriving traffic in data networks is bursty in nature rather than of constant rate; therefore a lot of resources will be wasted. For ATM networks, various statistical approaches have been proposed that take the statistical nature of the traffic into account, so that resources can be shared among different connections efficiently.

4.1.1. Traffic Models. The basic objective of the statistical approach is to accept as many connections as possible so that resources can be efficiently utilized while the QoS of each connection is still guaranteed. To achieve this objective, the arriving traffic needs to be modeled accurately so that resources can be allocated properly. It has been found that traffic from many applications such as voice and video present bursty characteristics that can be described by various "ON/OFF" models. These ON/OFF models consist of two states: a busy state in which data are transmitted from the traffic source, and an idle state in which no data are transmitted from the source. Various on-off models differ from each other in the distribution of the duration at each state, and the arrival process in the busy state. Commonly used ON/OFF models for a virtual connection in ATM and B-ISDN include Interrupted Poisson Process (IPP), Interrupted Bernoulli Process (IBP), and Interrupted Fluid Process (IFP). Further studies on ATM traffic introduce more complex distributions into the modeling such as Markov modulated poisson process (MMPP), Markov modulated bernoulli process (MMBP), and Markov modulated fluid process, each of which consists of several different states with state-dependent arrival rates.

Through measurement on Ethernet traffic, it has been recognized that IP traffic presents the characteristic of self-similarity which can be modeled by fractional Gaussian noise and fractional autoregressive integrated moving-average processes [14]. The self-similar traffic can be obtained by superposition of many on-off sources whose ON and OFF states strictly alternate and have high variability [15].

According to the various traffic model assumptions, such as the ON/OFF traffic model or its Gaussian approximation on traffic aggregation, many model-based admission control algorithms have been proposed that can be categorized into single-link approach and multiple-link approach.

4.1.2. Single-Link Approach. The single-link approach studies the admission control problem on a single link, specifically, the output of a multiplexor at the output port of a switch. The objective is to optimize or maximize the utilization on the link. By using traffic descriptors and traffic models, an admission region can be calculated assuming a target QoS performance [e.g., cell loss probability (CLP)].

A direct approach is to try to find the relationship between the arrival pattern of incoming traffic and CLP. As an example, an upper bound of CLP that is based on the average cell rate and peak cell rate or the rate variance in a fixed interval can be obtained [16]. A new connection is admitted if the resulting upper bound is below a pre-defined threshold.

Since traffic rate fluctuates between the minimum rate and peak rate, equivalent capacity has been proposed to describe the bandwidth needed to accept N connections for a given CLP threshold. In other words, this problem can be rephrased as, given a link capacity C, a predefined CLP threshold ε, and a buffer length K, how many connections can be accepted. For instance, if the connection requests are homogeneous with same SCR and PCR, by using equivalent capacity, the number of connections that can be accepted is between C/PCR and C/SCR depending on the value of ε. Admission control via equivalent capacity of a single service class and multiple service classes has been extensively studied. For a detailed review and comparison of single-link admission control algorithms interested readers may refer to Refs. 17 and 18.

4.1.3. Multiple-Link Approach. From the perspective of the service provider, the concern is how to efficiently utilize the resources not only on a single link but also in the entire network. For instance, when a specific link to a destination node can no longer accept connections, the connection may still be accepted by carefully rerouting the

connection through another path to the same destination node. Therefore, the optimization objective function should also take into account the routing algorithms. Further consideration of this problem leads to the observation that in the multiple service networks, the optimization objective could be the maximization of network revenue or gain, with multiple constraints such as routing, QoS, and policies. Solutions to the problem include decomposition of the multiple-link problem into single-link problems and the use of various techniques such as neurodynamic programming and reinforcement learning.

4.2. Measurement-Based Admission Control

Model-based admission control algorithms rely heavily on the corresponding traffic model assumptions to achieve the desired objective. However, this type of approach presents several problems:

1. The traffic model may not be sufficiently accurate, which may lead to the problem that the admission region is not properly designed. Although more accurate traffic models can be found by in-depth study of the traffic pattern, it may be too complex to be incorporated into the admission control model, or make the admission control algorithms too complicated to be performed in real time.
2. A user may overestimate or underestimate the traffic it will generate, which leads to an inefficient admission decision made by the network node.
3. Model-based admission control is usually a conservative scheme, which allocates resources based on worst-case scenario and may result in waste of the resources. Therefore, more recently measurement-based admission control algorithms have been extensively studied, most of which differ in three aspects: (a) the objective of measurement, (b) the approach to estimate and evaluate the measurement, and (c) the approach to make the admission decision.

Two types of objectives are used to perform measurement: to evaluate the CLP and to evaluate the equivalent bandwidth of aggregate traffic flows. Compared to the evaluation of CLP, measurement to estimate the equivalent bandwidth requires less computational power and is more straight-forward. The basic idea of this approach is as follows. The current equivalent bandwidth of aggregate traffic flows is estimated through measurements. Combined with the parameters declared in the new connection request, the equivalent bandwidth required is calculated assuming that the new connection is accepted. If the result is less than the bandwidth of the link, the new connection can be accepted; otherwise it is rejected. In this approach, CLP or CLR (cell loss ratio) is implicitly taken into account when estimating and evaluating the equivalent bandwidth.

An advantage of measurement of equivalent bandwidth or flow aggregation is that this approach does not need to track per-flow information of existing connections. Therefore, it can be applied not only in ATM but also in Internet where scalability of the algorithm is a big concern. For example, Cetinkaya et al. [19] proposed an admission control architecture in which an admission control algorithm is performed only at egress routers where the aggregate traffic envelopes are measured and estimated by using an adaptive algorithm. In this scheme, there is no involvement of backbone routers or per-flow management in the admission control procedure. As a result, the algorithm can achieve a good scalability.

To further address the scalability concern, a new approach based on measurement was proposed for the Internet. This approach is significantly different from conventional schemes in that it does not use signaling messages to make connection requests and the network node is not responsible for admission decisions any more. It is the host or end system that actually makes the decision as to whether to access the network. This is achieved by sending probe packets through the network to check the congestion level. If the probe packet indicates the current congestion level is low and will not harm the QoS of the existing and new connections, the host will admit the new flow; otherwise it will hold and give up the connection request. This approach is referred to as *endpoint admission control* or *distributed admission control* [20].

5. CONCLUDING REMARKS

In this article we discussed the problem of admission control in wired networks, with reference to the most current and future key networking infrastructures. The main objective of admission control in wired networks is to control the access of the users to the network resources in order to maximize the network utilization while providing the required quality of service and avoiding the occurrences of network congestion. An admission control scheme consists mainly of signaling messages and admission control units that perform the admission control algorithm or policy. In this article we introduced several typical admission control schemes and signaling procedures that can be applied in connection-oriented networks and on the next-generation Internet. We also presented some admission control algorithms that can be used in the admission control schemes, and discussed their characteristics. In general the admission control algorithms and schemes can be categorized into resource-based and policy-based admission control. *Resource-based admission control* algorithms base their decisions on the current resource usage, the resources needed to be allocated to the new calls, and whether the required quality of service can be guaranteed by the network. The basic purpose of *policy-based control* is to determine whether a user is qualified to access the network service at a specific time, and are required when different policies are enforced in the network. The implementation of admission control for Internet is still under development and deployment. As communication infrastructures are evolving into multiplexed and multiple service-class networks, network resource sharing by multiple service-classes correlates the performances of all classes that are supported in logically partitioned networks, and therefore additional scalable, less complicated and of high-efficiency admission control approaches and architectures are required.

BIOGRAPHIES

Symeon Papavassiliou received a diploma in electrical engineering from the National Technical University of Athens, Greece, in 1990 and his M.Sc. and Ph.D. degrees in electrical engineering from Polytechnic University, Brooklyn, New York in 1992 and 1995, respectively. From 1995 to 1996 Dr. Papavassiliou was a technical staff member at AT&T Bell Laboratories in Holmdel, New Jersey, and from 1996 to August 1999 he was a senior technical staff member at AT&T Laboratories in Middletown, New Jersey. From June 1996 till August 1999 he was also an adjunct professor at the Electrical Engineering Department of Polytechnic University, Brooklyn, New York. Since August 1999, he has been an assistant professor at the Electrical and Computer Engineering Department of New Jersey Institute of Technology, Newark, New Jersey. Dr. Papavassiliou was awarded the Best Paper Award in INFOCOM'94 and the AT&T Division Recognition and Achievement Award in 1997. Dr. Papavassiliou has an established record of publications in his filed of expertise, he is the director of the Broadband, Mobile, and Wireless Networking Laboratory at NJIT, and one of the founding members of the New Jersey Center for Wireless Networking and Security (NJWINS). His main research interests lie in the areas of computer and communication networks with emphasis on wireless communications and high-speed networks, network design and management, TCP/IP and internetworking, computer network modeling and performance evaluation and optimization of stochastic systems.

Jie Yang received a B.S. degree in information engineering, and an M.S. degree in communication and information system from Xidian University, P.R.China, in 1996 and 1999, respectively. He is currently a Ph.D. candidate in electrical engineering at the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, and a research assistant in the Broadband, Mobile and Wireless Networking Laboratory at NJIT, as well as a member of the New Jersey Center for Wireless Networking and Internet Security. From September 1999 till December 2001 he was a member of the New Jersey Center for Multimedia Research where he had been working on the design of high-speed networks. His current research interests are high-speed switch/router architectures, admission control, resource allocation and traffic engineering, and Internet security.

BIBLIOGRAPHY

1. ITU-T Recommendation Q.2931, *Digital Subscriber Signalling System No. 2 (DSS 2)—User-Network Interface (UNI) Layer 3 Specification for Basic Call/Connection Control*, 1995.
2. ITU-T Recommendation Q.2761, *Functional Description of the B-ISDN User Part (B-ISUP) of Signaling System No. 7*, 1999.
3. ATM Forum, *ATM User-Network Interface Specification V3.1*, 1994.
4. ATM Forum, *Private Network-Network Interface Specification Version 1.0 (PNNI 1.0)*, March 1996.
5. R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, June 1994.
6. S. Blake et al., *An Architecture for Differentiated Services*, RFC 2475, Dec. 1998.
7. R. Braden et al., *Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification*, RFC 2205, Sept. 1997.
8. E. Crawley et al., *A Framework for Integrated Services and RSVP over ATM*, RFC 2382, Aug. 1998.
9. R. Yavatkar et al., *SBM (Subnet Bandwidth Manager): A Protocol for RSVP-Based Admission Control over IEEE 802-Style Networks*, RFC 2814, May 2000.
10. R. Yavatkar, D. Pendarakis, and R. Guerin, *A Framework for Policy-Based Admission Control*, RFC 2753, Jan. 2000.
11. D. Durham et al., *The COPS (Common Open Policy Service) Protocol*, RFC 2748, Jan. 2000.
12. S. Herzog, *RSVP Extensions for Policy Control*, RFC 2750, Jan. 2000.
13. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, Jan. 2001.
14. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (Extended version), *IEEE/ACM Trans. Network.* **2**(1): 1–15 (Feb. 1994).
15. W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level, *IEEE/ACM Trans. Network.* **5**(1): 71–86 (Feb. 1997).
16. H. Saito, Call admission control in an ATM network Using upper bound of cell loss probability, *IEEE Trans. Commun.* **9**(40): 1512–1521 (Sept. 1992).
17. H. G. Perros and K. M. Elsayed, Call admission control schemes: A review, *IEEE Commun. Mag.* 82–91 (Nov. 1996).
18. E. W. Knightly and N. B. Shroff, Admission control for statistical QoS: Theory and practice, *IEEE Network* 20–29 (March/April 1999).
19. C. Cetinkaya, V. Kanodia, and E. W. Knightly, Scalable services via egress admission control, *IEEE Trans. Multimedia* **3**(1): 69–81 (March 2001).
20. F. P. Kelly, P. B. Key, and S. Zachary, Distributed admission control, *IEEE J. Select. Areas Commun.* **18**(12): 2617–2628 (Dec. 2000).

ADMISSION CONTROL IN WIRELESS NETWORKS

SYMEON PAPAVASSILIOU
JIONGKUAN HOU
New Jersey Institute of Technology
University Heights
Newark, New Jersey

SEBNEM OZER
MeshNetworks, Inc.
Orlando, Florida

1. INTRODUCTION

The goal of wireless communications is to provide users with ubiquitous information access, that is, to allow the

Table 1. Service Classification

	Real-Time		Nonreal-Time	
	Conventional	Streaming	Interactive	Background
Examples	Voice	Videostreaming	Web browsing	Email
Delay	Bounded		Sensitive	Tolerable
Rate	Guaranteed		Not guaranteed	
BER	$\leq 10^{-3}$	$\leq 10^{-6}$	≈ 0	

users to access the capabilities and resources of the global network at any time without regard to their location and mobility. Technological advances such as time and space diversity systems, low noise filters, efficient equalizers, advanced modulation and coding schemes, and the rapid development of handheld wireless terminals have facilitated the rapid growth of wireless communications and mobile computing.

Compared with fixed networks the most salient features of the wireless networks include the users' mobility, the limited bandwidth and power resources, the highly dynamic network (re)configuration, and the higher link bit error rates. In cellular wireless networks, due to the mobility of wireless subscribers, the network configuration is rearranged every time a subscriber moves into the coverage region (cell) of a base station or a new network. Furthermore, for current and future wireless networks designed to support high-data-rate applications, among the major limitations are the propagation conditions, such as fading and multipath, and power consumption that determine the communication range. In general, the tradeoff is between creating a dense infrastructure with high handoff rate and a more sparse deployment at the expense of high power consumption. Therefore in order to maximize the utilization efficiency of the limited radio resources, while meeting the quality-of-service (QoS) requirements of mobile users, efficient admission control and resource management schemes are required in the emerging wireless network architectures.

Wireless network management services can be categorized as call management, radio resource management, and mobility management [1]. Call management, for setting up and terminating communication sessions is necessary for both conventional information networks and the wireless communication systems. The other two categories of network management tasks are new and specific to wireless communications. The distinctive features in terms of mobility, traffic patterns and QoS requirements, and availability of the limited resource are the factors that govern admission control in wireless networks.

QoS is the ability of a network element to provide some level of assurance that its traffic and service requirements can be satisfied. For admission control depending on the QoS requirements, two main service classes can be considered: real-time and non-real-time services. Each class can be further divided into subclasses corresponding to different applications with given traffic characteristics and QoS parameters. For instance, in UMTS (Universal Mobile Telecommunications Systems), the supported traffic types are divided into four different service classes such as conventional class, streaming class, interactive and background class. Table 1 summarizes this classification and highlights the respective traffic characteristics of each class.

Features specific to the mobile environment such as the particular problems of highly variable connection quality, management of data location, the restrictions of battery life and cost, all impact the delivery of the required QoS. Therefore, admission control mechanisms combined with effective resource allocation schemes are crucial for the efficient design and use of wireless systems.

2. OVERVIEW OF ADMISSION CONTROL

The wireless network management techniques perform various processes such as power control, channel allocation, and handoff. The call and radio resource management problem is to assign, at different timescales, to each terminal, a base station, a physical channel and transmitter power levels, for both uplink (from mobile terminal to base station) and downlink (from base station to mobile terminal) communication. Depending on the access technology, the channels may take the form of time slots in time-division multiple access (TDMA) systems, or frequency bands in frequency-division multiple access (FDMA) systems, or different codes in code-division multiple access (CDMA) systems. According to these management techniques, admission control can be implemented at call (circuit switching), packet (packet switching) or burst (burst switching) level, or a common access scheme (random/common packet access) may be deployed. As mentioned before due to the user mobility, a user might move across the cell boundary while a call is in progress. In this case the system automatically transfers the call to a new channel belonging to the new base station. This process is called handoff or handover. In order to provide uninterrupted service to the mobile subscribers, handoffs must be performed successfully and should be imperceptible to the users. Users may have to change their radio cells a number of times during the lifetime of their connections, and as a result, availability of wireless network resources at the connection setup time does not necessarily guarantee that wireless network resources will be available throughout the whole lifetime of the connection. Thus the handoff events make the call admission control process for wireless networks more complicated than those for wired networks.

More specifically the admission control steps in wireless networks can be summarized as follows [2]: (1) assign one or more (e.g., soft handoff) base stations for a new or handoff call if the call has been accepted; (2) assign one

or more channels (e.g., frequency, time slots and codes or a combination of them) according to rate requirements; (3) assign transmitting powers for the base station and mobile nodes (power levels are adjusted according to the channel conditions, user location and required QoS); and (4) allocate resources according to the traffic classes (a time scheduler decides when to use the allocated resources).

The resources are estimated from the measured interference conditions, radio channel characteristics, current load in the cell site, sessions' traffic characteristics, and quality of service requirements. These inputs along with historical values and capacity models are used for the admission control tasks. An important consideration in specifying these functions is the interplay of the granularity of their response, and the load they create on the system. Too little monitoring may cause out of specification performance for a period of time while these measurement and management functions themselves will place a load on the systems they are monitoring. Another measurement required for wireless systems is done in order to estimate the link quality for highly variable air interface and user mobility. Bursty data transmission poses a new problem as the link quality cannot be measured efficiently at long idle times where the distance between transmissions can be considerably changed. A tradeoff between using estimated average link qualities versus keeping the link alive at a minimum idle power level must be taken into account for different ratios of idle rate to mobility rate.

In the following we examine in more detail each one of the main elements involved in the admission control process in wireless networks.

3. POWER CONTROL

The objective of power control is to deliver to each radio receiver a signal that is strong enough to overcome noise and interference from other signals but not so strong as to cause excessive interference to other communications. In general power control guards against changes in the system load, jamming, slow and fast variations in the channel conditions, and sudden improvements or degradations in the links. The gain from power control can be seen in conserving energy for prolonged power supply, in satisfying stable QoS for multimedia services, in efficient handling of mobility (handoffs), in increasing overall capacity, and in other applications.

The carrier-to-interference ratio (CIR) [or signal-to-interference ratio (SIR)] balancing technique for power control purposes has been presented in several early power control schemes [3,4]. The power control algorithms in the literature can be classified as distributed and centralized algorithms. For mainly practical considerations, most efforts have concentrated more on distributed power control schemes than on centralized schemes, because the centralized power control suffers from problems such as large-scale data management, complexity, network vulnerability, and latency, etc.

In any of these algorithms either the path gains are assumed to be known a priori, or measured SIRs on the active links are utilized. In general two main distributed power control approaches have been proposed. In the first approach, the receiver's signal-to-interference ratio (SIR) is measured and the transmission power is adjusted according to whether the SIR is below or above some target value. The drawback of this algorithm is that the local adjustments, without a global consistency, increase the interference to the neighboring areas, which, in turn, results in an increase of power in this area, and finally in an increase of interference. In the second approach, the transmitted power is adjusted in order to balance the SIRs of all links and to maximize the worst SIR in the channel. The drawback of this approach is that during the iterations of power adjustment or after reaching the steady state, the SIRs of the links may fall below the required value.

Besides the ability to compute capacity margin, the desired properties of a power control scheme are stated [5] as to be distributed (at the node or link level) in order to require minimal usage of network resources for control signaling, simple to be suitable for real-time implementation, agile for fast tracking and adaption to the channel changes and mobility, robust to be able to adapt to stressful contingencies, and scalable to perform at various network scales of interest. In predictive call admission control, a predictor is used to predict the future traffic from its present and past values. If the call setup time is longer relative to the traffic variations, the advantage of the predictive algorithm is higher.

Power control in code division multiple access (CDMA) cellular systems is a crucial issue since the capacity of CDMA networks is mainly interference limited [6,7]. Present CDMA cellular systems have been optimized for voice transmission. For voice CDMA systems based on the Interim Standard (IS95) standard, power control is used to combat the near–far problem by maintaining nearly constant received power at the base station. Power control is used as a means of minimizing multiuser interference and improving capacity by adjusting the powers to obtain the same carrier to interference power ratio on all links.

In the IS95 reverse link (uplink), the signal from each mobile unit should arrive at the base station with the minimum signal-to-noise ratio (SNR) needed to maintain the desired quality. In reverse-link–open-loop control, the mobile unit estimates the path loss from the cell site by comparing the received power to the transmitted power. Then the mobile adjusts its power such that the transmitted power is lowered if the signal is determined to be too strong or is increased slightly otherwise. In reverse-link–closed-loop control, the demodulator at each cell site compares the received SNR to the desired value and commands the appropriate adjustments. In the forward (downlink) link, at certain locations, the signal received by a mobile unit may be too weak to accurately decode data due to the excessive shadowing and interference from a neighboring cell. The cell periodically reduces the transmitted power in order not to transmit high power if not necessary. When a mobile detects an increase in its frame error rate, it requests higher power and the cell site increases the power by a predetermined amount.

Several power control algorithms have been proposed to address the problem of admission control in a DS-CDMA (direct-sequence code-division multiple access) network

with integrated services [8,9]. The main objective is to achieve optimality in the sense of maintaining active link quality (QoS of active users) while maximizing free capacity of new admissions. Bursty packet applications can introduce high interference during active periods. Multiaccess interference is regulated by controlling the transmit powers of the users for active link quality protection. This is done by computing the "interference margin," that is, the amount of excess interference that can be tolerated by active users without violating their SNR thresholds.

In many systems, transmission power control and channel allocations according to the traffic classes are managed jointly in order to maintain the SIR's of all links above the required quality factor at all times. For instance, the conventional power control scheme can be used with one power level for each slot if packets with equal or similar bit error rate (BER) requirements are transmitted in the same slots [10]. Otherwise, an optimal power distribution can be computed to provide the required BER of media with high priority and achieve the maximum throughput and minimum BER of media with low priority [11].

4. CHANNEL ALLOCATION AND ADMISSION CONTROL

In general the channel allocation problem can be viewed as a combinatorial optimization problem. Frequency-division multiplexing and time-division multiplexing provide a "channelization" of the spectrum. In code-division multiplexing schemes waveform allocations are permuted in a random fashion [12]. Depending on the dedication of bandwidth, we can categorize the channel allocation schemes into four sets as follows: dedication of channels for call duration (circuit switching), dedication of channels for packet duration (packet switching), dedication of channels for the duration of burst data (burst switching), and random access transmission (common channel packet switching) that do not require a reservation of channels. While circuit switching is a fixed dedicated assignment, packet and burst switching are demand-based assignments. For each one of these sets and switching schemes, different channel allocation and admission control processes are required.

4.1. Circuit Switching

In circuit switching, the users are allocated a dedicated channel and a continuous connection is guaranteed during a session. Hence, circuit switching is a static admission control where the negotiation is done for the call duration in the specific cell. The steps are specification of QoS requirements, negotiation for an agreed specification between all parties, admission control for prediction of the capability to meet the users' requirements, and resource reservation for allocation of resources to connections. These functions are supported by a database of multimedia documents that has information such as historical characteristics of a link, a profile manager that maintains QoS-related information for different classes of users, and a network monitor that monitors the system's state at the new and handoff call arrivals.

The user first sends a request message containing the information for the specification of the QoS requirements. The cell site decides to accept or reject the user according to the active users QoS requirements. If feasible power and rate vectors are found, the new user is accepted to the system. A power vector is feasible if for every active link, a positive transmission power level smaller than the peak transmit power can be found. Similarly, a rate vector is feasible if for every active link a rate level greater than the minimum required rate can be assigned. If the user is accepted, an acknowledgment message with the assigned channel and the required power level is send to the mobile. Each session arrival is either allocated to a dedicated channel or blocked. If a negative acknowledgment is received or no response is received within a predetermined time interval, the user resends the request message after a random delay. A blocked user is lost if the waiting time exceeds the tolerance time or a maximum number of access attempts is reached.

4.2. Packet Switching

Dynamic admission control is more effective for bursty multimedia data and highly variable wireless channel conditions. Its corresponding functions include: monitoring of the QoS parameters, policing for ensuring that all parties satisfy the QoS contracts, maintenance of QoS by modification of some network parameters, renegotiation of a contract, and adaptation to the changes in the system. Depending on whether the renegotiation is done on a packet or burst basis, the packet and burst switching techniques are performed. In packet switching, data terminals must contend for a channel for each packet that must be sent. Therefore the network utilization is maximized while access delay per packet is increased.

4.3. Burst Switching

Circuit and packet switching techniques are insufficient in meeting the quality of service (QoS) requirements of bursty long multimedia messages due to the poor channel utilization and high per-packet delay, respectively. For instance, the proposed burst switching technique in cdma2000 MAC layer [13] attempts to overcome these problems by allocating the dedicated channels to the burst of data and releasing them at the end of the bursts. This ideal burst switching system would immediately release the circuit at the beginning of the idle period following the packet burst, so that the allocation delay constraints would be satisfied, while the channel utilization is maximized [14].

Since the traffic channels are allocated for the duration of a burst, admission control at burst level is considered. Admission control must be dependent not only on the active users but also on the registered users in the inactive state (since they can reaccess the system), in order to foresee the potential to admit a new user. Depending on the traffic and control channels allocation and registration process, a terminal can be in different states where the state transitions are controlled by the base station via "timer" values. The optimal timer that determines the burst length depends on the user traffic characteristics, the timescale of interest, the system load, and the corresponding QoS requirements.

4.4. Common Channel Packet Switching

For short bursty messages, the exchange of resource allocation control information can be avoided by using ALOHA-type random-access methods where terminals compete for radio resources. This approach requires the resolution of collisions and the use of retransmission policies.

The common packet channel (CPCH) mechanism has been shown to be an efficient transfer mechanism of packet data in wireless environments for non-real time applications such as email, HTTP, and FTP [15]. CPCH message transmission typically operates in power controlled CDMA systems. Each message can have variable length where the maximum length is a higher-layer parameter. Since error control via acknowledgments and retransmission in non-real time applications is crucial, especially in the environments where message losses are usually higher, a retransmission scheme is used. The additional delays caused by retransmissions are likely to be tolerable in applications with less stringent delay bounds, while the loss of some of the messages is often intolerable since completeness of information delivery is essential.

There is no guaranteed QoS during the packet transmission. A positive or negative acknowledgment is sent to the user after the reception of each packet according to the packet error at the receiver side. If negative acknowledgment is received or no response is received within a predetermined timeout value, the user retransmits the packet after a random delay. Acknowledgments can be send for each packet or for a burst of packets.

In general, the advantage of burst reservation schemes for data services is minimization of the interference for voice and data packets at the expense of higher overhead to control and measure the channel load. On the other hand, ALOHA-type common packet transmission requires a higher rate of retransmission for data users while a simpler control mechanism is needed.

4.5. Hybrid Schemes

Hybrid channel assignment schemes are used for integrated voice/data services where voice traffic is transmitted in a circuit mode while data traffic is transmitted in packet/burst mode or on a common channel. In this case, data users can use the voice channels during OFF time of voice users without degrading the QoS of the voice users. Some proposed channel assignment techniques are described below.

For Dynamic TDMA/TDD mode, users send transmission requests to the base station that processes them with a schedule table based on the QoS parameters of user traffic. For constant-bit-rate (CBR) and variable-bit-rate (VBR) traffic, slot allocation is performed once during call establishment, as in circuit mode. For available-bit-rate (ABR) and unspecified-bit-rate (UBR) traffic, slot allocation is performed on a burst-by-burst basis via dynamic reservation of ABR/UBR slots and unused CBR/VBR slots as in burst switching.

In packet reservation multiple access (PRMA), each of the slots are classified as being either reserved or available. Reservations are limited to terminals transmitting real-time data such as voice or video. Data terminals must contend for a time slot for each packet that must be sent as in packet switching. Speech activity detectors are used to hold reservations only for the duration of the talk spurt and to release them during quiet spurts so that the channel bandwidth can be used by other terminals with packets to send. One drawback of this algorithm is that while voice terminals are able to minimize collisions by reserving slots, terminals must still contend for initial access and data packets must contend for each slot. Furthermore, the permission probability of the different terminals must be controlled so that terminals with time-sensitive data are able to access the channel without excessive delays. Various improvements to the basic PRMA protocol have been proposed for multimedia cellular systems. For TDMA systems, dynamic and centralized PRMA are proposed where time slots are assigned to users according to the amount of bandwidth required and their priority levels. After a contention period, the base station allocates as much of the user's requested rate as possible. For time-division CDMA (TDCDMA) systems, PRMA-based techniques further divide each time slot into subslots using up to eight spreading codes. These subslots are used for contention and data transmission. The reservation will last until the end of the voice burst or for a certain numbers of data frames. By controlling the contention access and allocation for data services, the protocol is able to track delay requirements and dropping probabilities for different services.

For CDMA systems, the packets are classified according to their traffic rate and queued according to their priority levels. The terminal can be at three states: idle, active, and blocked. Different techniques are studied to meet different rate requirements by using variable processing gain or multiple codes. In hybrid systems [16], short packets are transmitted on an ALOHA basis using random access with no access delay and minimum overhead. In the case of larger packets the terminal will request a dedicated channel (code) on the access channel. The base station will evaluate if the request resources are available to assign a transmission format with the time that the user can start transmitting. Once the transmission is finished the terminal will maintain the link for a certain time. If a packet is generated within that time, the user transmits immediately, but if the packet is very large, the user has to request the channel again.

5. HANDOFF AND ADMISSION CONTROL

According to the call initiation position, in a wireless network two types of calls submit admission requests to a base station: new calls, which are initiated by mobile subscribers in the current cell; and handoff calls, which are initiated in other cells and handed off into the current cell. The function of admission control as mentioned before is to determine whether to grant radio resources to an incoming new/handoff call on the basis of information such as the current channel occupation, the bandwidth and QoS requirements of calls in service, and the characteristics of the call that requests admission.

When a call hands off to a neighboring cell whose admission control process decides to reject its admission

request, the call is forced to be terminated prematurely (dropped). One of the important tasks of call admission control is to limit the probability of such forced termination of ongoing calls, because from the viewpoint of mobile subscribers, having a call abruptly terminated in the middle of the conversation is less desirable than new call attempts being blocked occasionally. Hence most of the wireless admission control schemes are handoff-prioritized schemes, which offer handoff calls higher priorities over new calls to access the limited radio resources.

It should be noted here that a lower handoff call blocking probability is obtained at the cost of an increase in the new-call blocking probability. Therefore the admission control schemes must be carefully designed to balance these two types of blocking in order to achieve a better performance. Many handoff priority-based admission control schemes, that range from static to dynamic, have been proposed in the literature and they can be roughly classified into three categories [17]: guard channel schemes, queuing priority schemes, and channel borrowing schemes.

5.1. Guard Channel Schemes

In guard channel schemes (also called cutoff priority schemes), some of the radio channels are reserved for the exclusive use of handoff calls while the rest of the channels are shared equally by both new calls and handoff calls. One critical factor that influences the performance of guard channel schemes is the number of channels that need to be reserved. If the reservation is low (underreservation), the QoS requirements on handoff call blocking probability cannot be met as shown in Fig. 1a. On the other hand, a higher level of reservation (over-reservation) may result in a large number of new-call attempts being rejected.

Depending on how the number of guard channels is determined, guard channel schemes can be further classified into static schemes and dynamic schemes. For static guard channel schemes the number of channels reserved for handoff purposes is fixed for each cell, and it is calculated based on the knowledge of the traffic pattern of the area and the estimation of channel occupancy time distribution at the system design stage. The major advantage of this static approach is its simplicity since no communication and computation overheads are involved. However, the problems of overreservation and underreservation are unavoidable if the cell traffic does not conform to the prior knowledge. Therefore dynamic reservation schemes are designed to overcome these problems. Through the use of current system information, such as user mobility information and channel occupation information, dynamic schemes can determine the number of guard channels in a real-time fashion, and therefore they can easily adjust to the changing conditions of the system.

It should be noted here that no matter how the reservation is made, the guard channel schemes may result in a reduction of the total carried traffic (as shown in Fig. 1b). In general it is the originating calls and not the ongoing calls that really add to the total traffic [18]. Because fewer channels are available to new calls, the larger the number of the guard channels, the higher the probability the originating calls being blocked and, hence the less the overall traffic carried by the system.

5.2. Queuing Priority Schemes

The basic idea of the queuing priority scheme is that when a new call or a handoff call cannot be granted the required channels at its arrival time, the call is put into a queue waiting for its admission conditions to be met. Queuing of handoff calls is possible due to the fact that there is a finite time interval between the time that the received signal level drops below the handoff threshold and the time the call is terminated due to insufficient signal level. As shown in Fig. 2, handoff can occur at any time during the time interval Δt.

Queuing of new calls is possible because of the use of common channel signaling in digital communication systems. In the standard Public Switched Telephone Network (PSTN) the queuing of new calls is impractical since the signaling needed for the dialing is done on the communication channel itself. Queuing of a new call would therefore result in multiple redials that would unnecessarily occupy some communication channels. In cellular systems, however, the setup of a call is done on a separate control

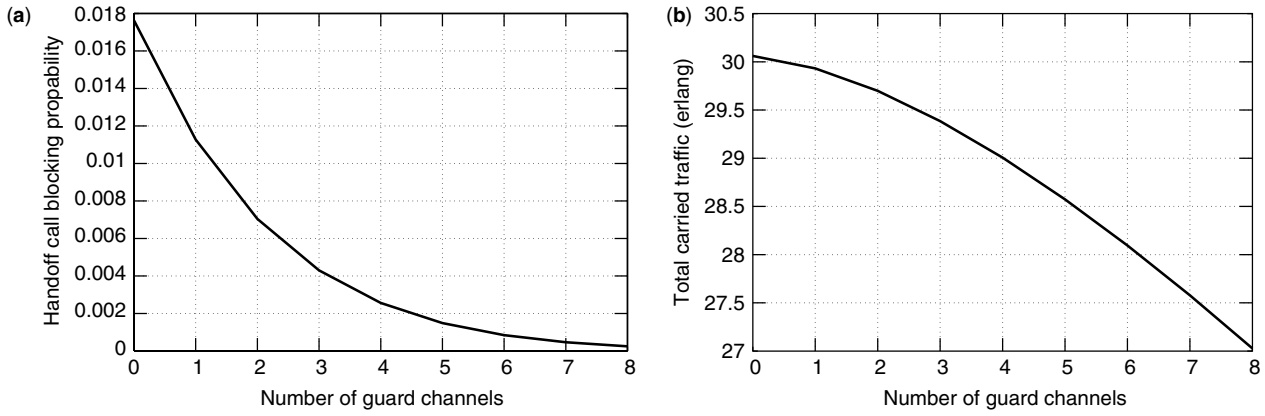


Figure 1. The system performance versus number of guard channels: (**a**) handoff call blocking probability; (**b**) total carried traffic.

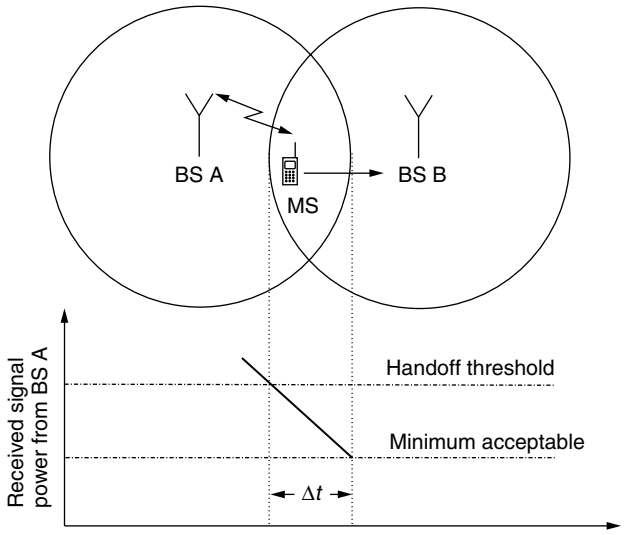


Figure 2. Handoff time interval.

channel, which can provide the system with a way of queuing new calls without affecting the transmission channels.

According to the types of calls that are queued, the queuing priority schemes can be further classified into handoff call queuing, new-call queuing, and new/handoff call queuing. Handoff call queuing schemes put the incoming handoff call in the queue and block new call attempts if there are no available channels. New call queuing is often used in combination with guard channel schemes to increase the carried traffic; if the new-call admission conditions are not met, then the arriving new calls are put into a queue to wait for the channels to be released. In the new/handoff call queuing schemes, both new calls and handoff calls are queued in the same queue and handoff calls are given non-preemptive priorities over new calls.

5.3. Channel Borrowing Schemes

The channel borrowing scheme is a combination of fixed and dynamic channel assignment schemes. The channel borrowing schemes work as follows: when all the channels in a cell are occupied, the cell borrows channels from other cells to accommodate the incoming handoff calls, as long as the borrowed channels do not interfere with the ones used by existing calls. The channel borrowing schemes are more flexible in the sense that by "moving" (borrowing) channels from less busy cells to more busy cells, a balanced performance throughout in the system can be achieved.

One problem associated with the channel borrowing scheme is channel locking. This occurs when cells within the required minimum channel reuse distance from a cell that is using a borrowed channel cannot use the same channel. Reuse distance in a cellular system is defined as the minimum distance between two cells that may use the same channels. Reuse factor is a cell plan parameter equivalent to reuse distance. It is the minimum number of channels needed to establish one call connection at each cell without reusing a channel in cells closer than the reuse distance [19]. Figure 3 shows a part of a wireless network that has frequency reuse factor 7. When cell B

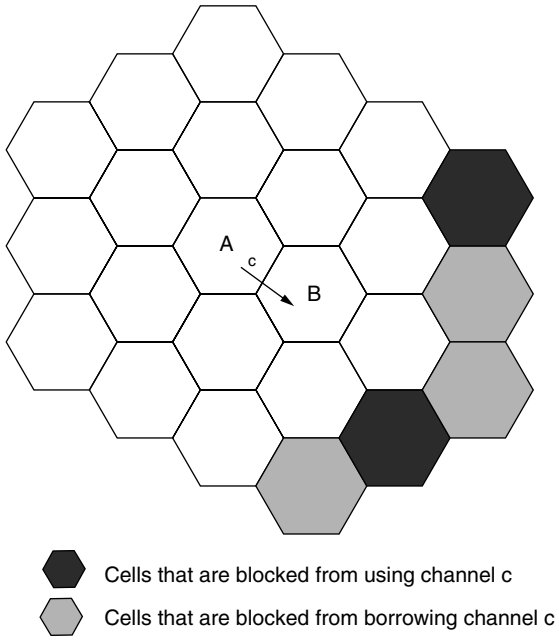


Figure 3. Channel locking caused by borrowing action.

borrows channel c from cell A to serve an incoming handoff call, the predefined frequency reuse pattern is temporarily violated: those darkest shadowed cells are prohibited from using channel c, although channel c is originally assigned to these cells and the darker shadowed cells cannot borrow channel c due to the cochannel interference requirements.

Channel locking problems make the channel borrowing schemes more complicated than the guard channel schemes and queuing priority schemes because the decreased handoff call blocking probability is obtained at the cost of decreasing the capacity of other cells, which, in turn, will cause QoS degradation in these cells. In order to achieve a better performance, a channel borrowing scheme must try to minimize this cost through cell coordination and centralized control. Two commonly used borrowing protocols are the minimum influence borrowing and channel reallocation. The minimum influence borrowing algorithm aims to borrow the channel that has minimum impact on the overall performance of the system. When channel borrowing in necessary, all the borrowable channels are compared in terms of the traffic conditions in the blocked cells of each borrowable channel, and predictions are made accordingly. Then the channel that will cause the least QoS degradation in the expected future is chosen. The channel reallocation process aims at minimizing the time that a borrowed channel is used. Instead of returning the channel when the call that uses the borrowed channel completes or hands off, if there is a channel that is released by another call in the borrower cell, the borrowed channel is returned and the released channel is allocated to this call.

6. ADMISSION CONTROL IN 3G STANDARDS

Second–Generation (2G) wireless systems have focused on the development of mobile networks in order to support

conventional telephony services. The aim of next generation systems, such as 3G technologies wideband CDMA (W-CDMA), cdma2000, and the Universal Mobile Telecommunications System (UMTS), [International Mobile Telecommunications in 2000 (IMT-2000)], is to support a variety of data services while increasing the system capacity. As an interim solution, the global system for mobile communications (GSM) operators are moving toward the general packet radio system (GPRS) technology. At the same time the TDMA (and some GSM) operators are planning for enhanced data rate for global evolution (EDGE). The IS95 CDMA operators are considering 1XRTT, which is the interim step toward CDMA-2000. Two other interesting approaches being developed are higher data rate (HDR) and "1 EXTREME." One industry group, the 3rd Generation Wireless Partnership Project (3GPP), is developing the 3G standards for GSM-based and wideband CDMA (WCDMA) air interface, while the 3rd Generation Partnership Project 2 (3GPP2), is developing 3G standards for cdma2000-based systems and the Universal Wireless Communications Consortium (UWCC) for the evolution of North American-TDMA (NA-TDMA) technology.

For the establishment of a packet data session, a GPRS UE (user equipment) must activate a packet data protocol (PDP) context where QoS parameter values are negotiated according to the availability of resources [20]. The QoS profile consists of five attributes: delay, service precedence, reliability, mean throughput, and peak throughput. However, since there is no per-flow prioritization and only best effort traffic is supported, end-to-end QoS, such as delay attribute, is not implemented. In UMTS, the PDP context mechanism has been improved to support QoS for multiple application flows with enhanced QoS negotiation and setup, and as a result, network QoS for end-to-end services can be realized.

For a CDMA air interface, QoS requirements are satisfied for different traffic classes. For instance, hard QoS guarantees are provided to realtime applications such as voice and video, and best-effort service to non-real-time applications such as packet data. Therefore, the resources are offered in accordance with the specific group characteristics. Group behavior of a class is implemented by power control and spreading control. An extensive research is done for both uplink and downlink cases with class-based bandwidth scheduling schemes to attain differentiated QoS on the CDMA air interface. The admission control in these schemes requires a radio resource allocation framework that characterizes the capacity model of a CDMA air interface and QoS models of various traffic classes.

In WCDMA, the power and rate adaptation algorithms can be implemented by the power/rate scheduler and the transmission time control by the time scheduler. The necessary radio channel characteristics can be provided by a resource estimator and the built-in capacity models in a call admission control module, and/or the resource estimator that can translate the gain from optimal power and rate allocation into better capacity estimation and utilization. In TD-CDMA, the admission control also has a time-slot assignment for uplink or downlink. This flexibility provides better adaptation to different scenarios, such as traffic asymmetry between uplink and downlink. Furthermore, allocation of CDMA codes and TDMA time slots provides higher granularity.

In IS95B, higher-data-rate service is provided through code aggregation. Specifically, up to eight codes can be assigned for the duration of a burst (one fundamental code and seven supplemental codes) requiring a burst-level admission control. A call origination with the packet data service option is used to establish a packet data service level registration. When the user remains idle for a predetermined inactivity time, the air interface resource is deallocated but the packet data registration remains established. The fundamental channel is assigned for the duration of the packet data call, whereas the supplemental channels are assigned for the duration of the packet data burst. Some of the major enhancements of CDMA2000 include the addition of a pilot channel on the reverse link and closed loop power control on the forward link. Furthermore, two additional states are added. In the active state, both a traffic and a control channel are assigned to a mobile. In the control hold state, the traffic channel is released but the control channel is maintained. In the suspended state, no interface channels are assigned, but the radiolink protocol state is remembered to avoid delays during reinitialization. The CDMA2000 physical layer provides a single supplemental channel with variable spreading gain. CDMA2000 defines a multimode enhanced random, and reservation-access scheme [13]. It is a combination of well-known packet reservation multiple access and common channel multiple access concepts. For an efficient high-speed packet data traffic with variable duration and data rates, admission control is enhanced with fast congestion control, fast capture feedback, interference control, and closed-loop power control. Depending on the service type and packet size, the mobile may need to request a dedicated channel or a reserved common channel, or simply include its packet in its random-access probe.

Standardization activities also include defining 3G wireless networks based on a TDMA air interface, namely EGPRS. It uses a TDMA-based packet-switched radio technology and a new air interface, EDGE, and GPRS core network designed for best-effort packet data services. Most of the research and development efforts in supporting QoS for multiple service classes in TDMA networks include packet scheduling schemes and multiple time-slot assignments.

7. CONCLUDING REMARKS

The explosive growth of the Internet and the continued dramatic increase for all wireless services are fueling the demand for increased capacity, data rates, supported multimedia services, and support for different QoS requirements for different classes of services. The scarcity of available radio spectrum limits the obtainable user data rates, and therefore issues associated with the QoS, network management and control, and system adaptability are rapidly gaining critical research and commercial importance. In this article we discussed the problem of admission control in circuit-switched and packet-switched wireless networks, and presented

efficient admission control schemes in order to maximize the utilization of the limited radio resources, while maintaining the QoS requirements of mobile users. The various elements and processes associated with the admission control in wireless networking technologies include the assignment and allocation of limited network resources, such as bandwidth, channels, and transmission powers. Depending on the dedication of bandwidth and channels, we classified the channel allocation mechanisms and the corresponding system operation in the following modes: circuit switching, packet switching, burst switching, and common channel packet switching. For each one of these various modes we described in detail the overall admission control process. We have also provided an overview of the available handoff priority-based admission control schemes, in order to address tradeoffs between the handoff call blocking probability and the new call blocking probability. Finally we discussed the role and operation of admission control in the various standards for the current and next-generation wireless networks.

BIOGRAPHIES

Symeon Papavassiliou received a diploma in electrical engineering from the National Technical University of Athens, Greece, in 1990 and the M.Sc. and Ph.D. degrees in electrical engineering from Polytechnic University, Brooklyn, New York in 1992 and 1995, respectively. From 1995 to 1996 Dr. Papavassiliou was a technical staff member at AT&T Bell Laboratories in Holmdel, New Jersey, and from 1996 to August 1999 he was a senior technical staff member at AT&T Laboratories in Middletown, New Jersey. From June 1996 till August 1999 he was also an adjunct professor at the Electrical Engineering Department of Polytechnic University, Brooklyn, New York. Since August 1999 he has been an assistant professor at the Electrical and Computer Engineering Department of New Jersey Institute of Technology, Newark, New Jersey. Dr. Papavassiliou was awarded the Best Paper Award in INFOCOM'94 and the AT&T Division Recognition and Achievement Award in 1997. Dr. Papavassiliou has an established record of publications in his filed of expertise, he is the Director of the Broadband, Mobile, and Wireless Networking Laboratory at NJIT, and one of the founding members of the New Jersey Center for Wireless Networking and Security (NJWINS). His main research interests lie in the areas of computer and communication networks with emphasis on wireless communications and high-speed networks, network design and management, TCP/IP and internetworking, computer network modeling and performance evaluation and optimization of stochastic systems.

Sebnem Zorlu Ozer received a B.S. degree in electronics and telecommunications engineering with highest honors in 1992 from Istanbul Technical University, Turkey, an M.S. degree in electrical engineering in 1995 from Bogazici University, Turkey, and a Ph.D. degree in electrical engineering in 2001 from New Jersey Institute of Technology. Since 2001, she has been a systems engineer at Meshnetworks Inc, where she has been working on the design and development of ad-hoc networks. Her primary responsibilities are focused on integrating quality of service within a dynamic mobile network. Her areas of interest are design and management of wireless networking systems, performance analysis of computer networks, optimization of stochastic systems and quality of service management in mobile networks.

Jiongkuan Hou received his B.S. and his M.S. degrees in electrical engineering from Northern Jiaotong University, Beijing, China, in 1993 and 1999, respectively. He is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at New Jersey Institute of Technology, and a research assistant in the Broadband, Mobile, and Wireless Networking Laboratory at NJIT, as well as a member of the New Jersey Center for Wireless Telecommunications. His research interests lie in the areas of design and management of mobile cellular networks, with emphasis on resource allocation, call admission control, pricing, and multimedia service support.

BIBLIOGRAPHY

1. D. J. Goodman, *Wireless Personal Communications*, Addison Wesley, Longman, 1997.
2. L. Jorguseski, E. Fledderus, J. Farserotu, and R. Prasad, Radio resource allocation in third-generation mobile communication systems, *IEEE Commun. Mag.* 117–123 (Feb. 2001).
3. R. W. Nettleton and H. Alavi, Power control for a spread spectrum cellular mobile radio system, *Proc. IEEE Vehicular Technology Conf.*, 1983, pp. 242–246.
4. J. Zander, Performance of optimum transmitter power control in cellular radio systems, *IEEE Trans. Vehic. Technol.* **41**(1): 57–62 (Feb. 1992).
5. N. Bambos, Toward power-sensitive network architectures in wireless communications: Concepts, issues, and design aspects, *IEEE Pers. Commun.* 50–59 (June 1998).
6. W. C. Y. Lee, Overview of cellular cdma, *IEEE Trans. Vehic. Technol.* **40**(2): 291–302 (May 1991).
7. K. S. Gilhousen and I. M. Jacobs, On the capacity of a cellular cdma system, *IEEE Trans. Vehic. Technol.* **40**(2): 303–312 (May 1991).
8. S. Ramakrishna and J. M. Holtzman, A scheme for throughput maximization in a dual-class cdma system, *IEEE J. Select. Areas Commun.* **16**(6): 830–844 (Aug. 1998).
9. J. M. Capone and L. F. Merakos, Integrating data traffic into a cdma cellular voice system, *Wireless Networks* **1**(4): 389–401 (1995).
10. I. F. Akyildiz, D. A. Levine, and I. Joe, A slotted cdma protocol with ber scheduling for wireless multimedia networks, *IEEE/ACM Trans. Network.* **7**(2): 146–158 (April 1999).
11. J. Wu and R. Kohno, A wireless multimedia cdma system based on transmission power control, *IEEE J. Select. Areas Commun.* **14**(4): 683–691 (May 1996).
12. J. Zander, Radio resource management in future wireless networks — requirements and limitations, *IEEE Commun. Mag.* **35**: 30–36 (Aug. 1997).
13. Telecommunications Industry Association, *The cdma2000 ITU-R RTT Candidate Submission-TR45-5.5*, 1998.

14. S. Z. Ozer, S. Papavassiliou, and A. Akansu, On performance of switching techniques for integrated services in cdma wireless systems, *Proc. IEEE Vehicular Technology Conf.*, 1973, 2000, Vol. 4, pp. 1967–1973.

15. ETSI TS 125 322 V3.1.2 (2000–01), *Technical Specification Universal Mobile Telecommunications System (UMTS); RLC Protocol Specification*.

16. C. Roobol, P. Beming, J. Lundsjo, and M. Johansson, A proposal for an rlc/mac protocol for wideband cdma capable of handling real time and non real services, *Proc. IEEE Vehicular Technology Conf.*, May 1998, pp. 107–111.

17. J. Hou, J. Yang, and S. Papavassiliou, Integration of pricing with call admission control for wireless networks, *Proc. IEEE Vehicular Technology Conf.*, 2001, Vol. 3, pp. 1344–1348.

18. I. Katzela and M. Naghshineh, Channel assignment schemes for cellular mobile telecommunication system: A comprehensive survey, *IEEE Pers. Commun.* **3**: 10–31 (June 1996).

19. S. Papavassiliou, L. Tassiulas, and P. Tandon, Meeting QoS requirements in a cellular network with reuse partitioning, *IEEE J. Select. Areas Commun.* **12**(8): 1389–1400 (Oct. 1994).

20. R. Koodli and M. Puuskari, Supporting packet-data QoS in next-generation cellular networks, *IEEE Commun. Mag.* **39**(2): 180–188 (Feb. 2001).

ALOHA PROTOCOLS

JOHN J. METZNER
Pennsylvania State University
University Park, Pennsylvania

1. INTRODUCTION

The ALOHA method originated at the University of Hawaii as a means for multiple users to send short data packets via radio transmission over a common channel to a central station, in a largely uncoordinated manner [1]. The central station (see Fig. 1) would rebroadcast all its receptions on a different channel, so that the sending users would be fed back their own transmissions along with others. If the user observes that its data packet has collided with another, it would retransmit some time later. If the sender sees its own packet unaltered, this serves as an acknowledgment.

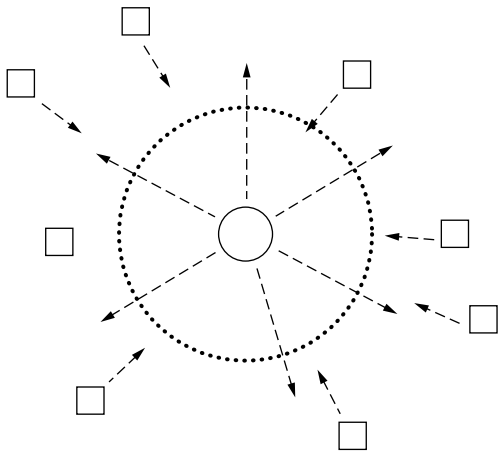


Figure 1. Rebroadcast by the central station (circle) of signals sent by the ALOHA participants (squares).

Two major categories of ALOHA are unslotted ALOHA and slotted ALOHA. In unslotted ALOHA, a transmitted packet can be sent at any starting time and can be of varied duration. In slotted ALOHA, there are fixed-sized synchronized time slots. A sender can send in any time slot, and the data packets should all be slightly smaller than a time slot duration. Collisions can be complete or partial in unslotted ALOHA, but are complete in slotted ALOHA. Even a partial collision usually calls for retransmission of the packet.

If the broadcast method does not provide enough information for the sending user to be certain of successful transmission, another option is for the destination or the central station to send back an individual acknowledgment signal.

ALOHA is very effective in light traffic situations; the sender can use the whole channel, and send very rapidly, usually without collision. With heavier traffic, ALOHA suffers from problems of efficiency and stability. Various protocols and signal processing methods have been invented to alleviate these problems. These include collision resolution algorithms, control of arrival rate and retransmission time, reservation ALOHA, multichannel ALOHA, capture ALOHA, and diversity reception. Also, a modification of unslotted ALOHA using carrier-sense multiple access with collision detection (CSMA/CD) is the basis for the popular Ethernet local-area network protocol [2].

2. QUANTITATIVE ANALYSIS OF IDEALIZED SLOTTED ALOHA COMMUNICATION

Suppose slotted ALOHA is used and all users observe the results of all transmissions. Models have assumed either (1) an infinite number of users having a finite total message arrival rate or (2) a finite number of users. The time delay for retransmission is assumed to be independently randomized for each retransmission; if it were the same for two colliders, they would collide again for certain.

2.1. The Infinite Number of Users Model

Packets (including packet retransmissions) are presumed to be generated by the infinite set of all users at a total finite rate of G packets per time slot. The number of transmissions in a slot is assumed to obey a Poisson distribution:

$$P[k] = \frac{G^k}{k!}e^{-G} \tag{1}$$

There is a successful transmission in a slot if and only if exactly one transmission occurs. Let S be the fraction of successful slots, also called *the throughput per slot*.

$$S = Ge^{-G} \tag{2}$$

S is maximum at $G = 1$, where it equals e^{-1} or 0.368.

G consists of two components: an average new packet arrival rate denoted as λ, and an average retransmission rate denoted as r: ($G = \lambda + r$). This model does not properly

reflect the effect of statistical variations. For a given average arrival rate λ, there is a finite probability that a short-term packet transmission rate G will occur that is large enough to cause the success (departure) rate to fall below λ. This will cause increased r (more retransmissions needed), which reduces S, leading to still higher r, until almost all traffic is retransmission traffic and hardly any are successful. This potential instability problem can be alleviated by blocking new packet transmissions and/or increasing average retransmission delay on observing excessive collision.

2.2. The Finite Number of Users Model

Assume that there is a fixed number M of active users. Suppose that k of these users are backlogged, which means that they have a need to retransmit a previously collided packet. Backlogged users do not send a new packet until after the backlogged packet is successful, at which time it becomes a nonbacklogged user. Assume that each $M - k$ nonbacklogged user has, independently of other users or its own past events, a probability P_A of sending a packet in the next slot, and that each of the k backlogged users has, independently, a probability P_R of sending its retransmission in the next slot.

This model is artificial because there is no fixed number of "active" users in practice, and active users are bursty in their needs for transmission. A large number of users tend to average out the behavior, however, so the model may approximate the real situation.

Suppose at a given time there are k backlogged users. Call this state k. The offered load in state k is

$$G(k) = (M - k)P_A + kP_R \tag{3}$$

For moderately large M, the success rate $S(k)$ can be approximated by [3]

$$S(k) \approx G(k)e^{-G(k)} \tag{4}$$

where $G(k)$ is as given by (3).

In state k, there is an average arrival rate of $(M - k)P_A$. The success rate $S(k)$ can be thought of as a departure rate. When arrivals exceed departures, the backlog increases and G increases, since, normally, $P_R > P_A$. The reverse happens when departures exceed arrivals. Figure 2 illustrates how system behavior tends to drift.

The arrows along the departure curve denote the drift direction. Point A is a desirable stable operating point. However, occurrence of a short-term jump in arrival rate can easily move the system past the unstable equilibrium point B, after which k increases to send the system to the undesired stable operating point C.

For very small P_R there can be a single desirable stable point, but very small P_R means long delay. Ideal operation would be to control P_R as a function of k so as to keep $G(k)$ as close to 1 as possible, but the state is difficult to know exactly. A simple control scheme could increase P_R when an idle slot is observed and decrease it when a collision is observed.

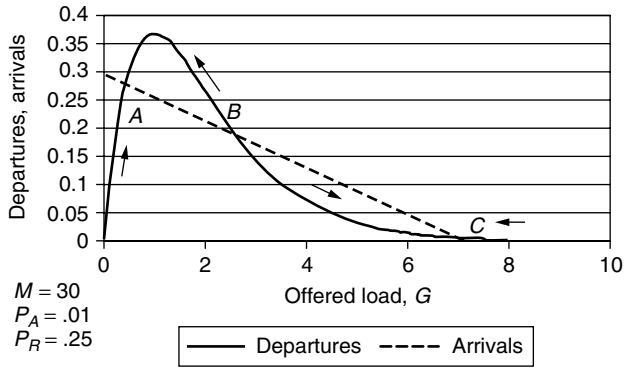


Figure 2. Drift analysis of ALOHA stability.

3. UNSLOTTED ALOHA

Unslotted ALOHA packets can start at any time and can be of variable size. If one assumes that any overlap of two frames causes both to be lost, unslotted ALOHA is considerably less efficient than slotted ALOHA. For equal-size packets the maximum efficiency is $\frac{1}{2}e$, or half that of slotted ALOHA. With unequal-size packets the maximum efficiency is slightly higher than $\frac{1}{2}e$. However, variable size packets often are desired, in which case slotted ALOHA would suffer inconvenience and/or inefficiency if packets had to be broken up or only partially filled slots.

4. IMPROVING THE EFFICIENCY OF ALOHA

There are three avenues for modifying ALOHA communication to improve its efficiency:

1. Use protocols to eliminate or reduce collision frequency.
2. Tolerate collision through signal design, signal processing, and/or diversity reception.
3. Increase the number of bits per slot by using nonbinary transmission.

Techniques in the first avenue include (a) unslotted — send only when no activity is sensed on the channel, called carrier-sense multiple access (CSMA); (b) unslotted — stop sending if a collision is detected, called collision detection (CD); (c) slotted — collision resolution algorithms.

Avenues 1a and 1b are employed as CSMA/CD in the Ethernet protocol for local-area networks. The sensing of the channel does not result in perfect collision avoidance because of the nonzero propagation time between potential senders. Wireless networks can use CSMA but seldom CD, because the locally strong signal of a transmitting station prevents detection of a locally weak signal transmitted remotely in the same timeframe and frequency band. With the central station model the central station broadcasts on a different band than the multiaccess users. The collisions are detected, but the round-trip propagation time may make it too late to benefit from a halt in transmission.

5. COLLISION RESOLUTION ALGORITHMS

These algorithms [4] improve the efficiency of slotted ALOHA by forcing a resolution of a collision. After a collision, only the colliders (the backlogged set) are permitted to send until the collision is resolved by their success. The algorithm ensures successes by the colliders within a minimum average number of slot times, while also allowing all users to know when the collision has been resolved. All potential senders are assumed to be informed of whether the prior slot experienced a success or a collision, or was idle; they are not presumed to know how many have collided.

One may question how the potential senders all get to know of the collision just before the next slot. The slotted ALOHA channel could occupy a slot in a TDM frame whose duration is longer than the round-trip time needed to learn of the collision. Figure 3 illustrates such a slot.

The gist of the algorithms, without going into the details, is as follows. Each collider picks a number in some agreed-on range; the range is split into two subsets, and those in the first subset send. If there is no collision, the first subset is resolved and attention turns to the second subset. If there is a collision, the first subset is split in two, eventually leading to resolution of the first subset.

With these algorithms, the maximum efficiency is increased from 0.368 to close to 0.5, with the exact gain dependent on the algorithm details and assumptions. Perhaps more importantly, the algorithms ensure stability at all arrival rates less than the maximum efficiency, which is the limit as $n \to \infty$ of n divided by the expected number of slots to resolve n colliders. This is because, with n colliders, the average number of arrivers during the average number of slots needed to resolve the colliders remains less than n. Thus n always tends to drift lower. Delay also is reduced, because it is not necessary to use a small P_r for a backlogged packet; the backlogged packet delay is no more than the time to resolve its collision.

6. RESERVATION ALOHA AND HYBRID SYSTEMS

In reservation systems, the information transmitted to make a reservation of data to be sent afterward is normally a minute fraction of the total information flow. Thus, a small portion (subchannel) of the channel can be devoted to reservation traffic, and even that subchannel can be lightly used. ALOHA methods work well in a light traffic channel, because transmissions are fast and collisions are rare. A collision merely becomes a delayed reservation, which is not too troublesome if rare.

ALOHA can also be used to implicitly reserve a slot in a TDM system for an indefinite period of successive frames by successfully transmitting data in that slot. Others could defer attempting to send in that slot until it was observed that the slot had become idle. This method is called *reservation ALOHA* [5].

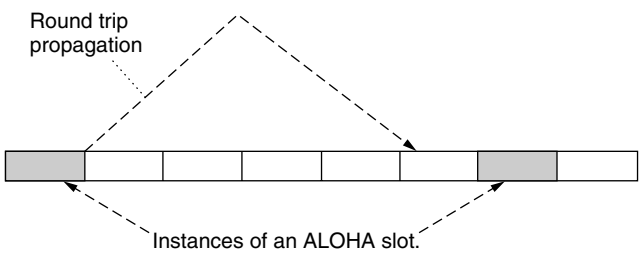


Figure 3. Occurrences of successive instances of an ALOHA slot.

7. MULTICHANNEL ALOHA AND MULTICOPY ALOHA

The ALOHA idea can be extended to multiple channels. These channels might be different frequency bands (FDM), different time slots of a synchronized frame (TDM), or different orthogonal code sequences in CDMA. If there are n frequency channels instead of 1, the time to send a packet increases by a factor of n. This would be a slight disadvantage under light traffic, since it would take much longer to send a burst of packets unless more than one channel was used simultaneously. There is no gain in total capacity by splitting into multiple channels.

Consider TDM with multiple ALOHA channels. One possibility is that if a collision occurs, the sender should resend in a randomly chosen channel. But this is just like picking at random to send in one of n slots, which is not much different than the one-channel strategy of sending a random time later after a collision.

Another thought is to send multiple copies of a packet in the hope that at least one will get through. This may be okay to do for a small subset of important, delay-sensitive messages. However, it is highly inefficient as a general policy. This is because (1) the limit of a fraction 0.368 of total successful packets can't be exceeded and (2) multiple successes of the same packet are included in this fraction, so the proportion of different successful packets will be much lower than 0.368.

8. CAPTURE ALOHA AND ABILITY TO TOLERATE COLLISION

The discussion so far has assumed that all colliding packets are lost. This may be overly pessimistic. There is a capture effect such that if one of two received signals is a certain amount stronger than an interferer, the strongest signal can be decoded. More sophisticated decoding and signal processing techniques may even allow both of two colliding packets to be decoded. If the stronger of two interfering signals is decoded, the effect of this now known signal could be subtracted out to a large extent, possibly allowing successful decoding of the weaker signal.

With slotted ALOHA, the number of simultaneously transmitted signals is rarely greater than 2, so simultaneous decoding may not be that complex. If two could be decoded simultaneously, but not more than two, the throughput per slot would be

$$S = Ge^{-G} + G^2e^{-G} \tag{5}$$

where S is maximized at $G = 1.618$ and is

$$S_{\max} = 0.840 \text{ packets/slot} \tag{6}$$

One simple way of using the capture effect to increase the throughput of slotted ALOHA is to use two different packet transmission power levels. Assume that if exactly one strong signal packet is sent in a slot, along with any number of weak signal packets, the strong signal packet is successfully decoded, but the weak one(s) will be lost. In any other collision event, all packets are assumed to be lost. This allows a higher efficiency than ALOHA without capture, and for an optimum proportion of high-power senders the maximum throughput is increased [6] from $1/e$ or 0.368 to $(e^{-(1-1/e)})$, or about 0.53.

Even without intentionally designing for two different power levels, received signals will come in with different powers. With fading, individual senders will sometimes come in at high power, sometimes with low power. This is an advantage [7]. A receiver could have multiple antennas that experience independent fading on different antennas. With this diversity reception feature, a signal A could be received more strongly than signal B on one antenna, while signal B could be received more strongly than signal A on another antenna, such that both could be captured. In ALOHA systems, fading may actually create greater throughput than without fading.

9. MULTIBASE ALOHA

Extending the idea of multiple antennas for diversity reception in ALOHA systems, it is possible to have a network of cooperating base stations [8]. A mobile sender transmits a packet, and if at least one base station can decode it, the packet could be successful. Duplicates could be recognized by the network, and the base station with the strongest reception could supply the feedback acknowledgment. This way several mobile senders can be successful simultaneously. Also, handoff would not be necessary unless the mobile left the entire base station network.

10. ALOHA WITH A TIME CONSTRAINT

ALOHA systems rely on being able to repeat collided packets. This is a drawback to time-constrained traffic that must meet some deadline. Still, the time constraint can allow several retransmissions if round-trip acknowledgment is much shorter than the delay tolerance. On mobile-base communication the distances are relatively short, allowing short round-trip times. If the deadline for a packet is not met, no further retransmissions are attempted. As long as these losses are tolerable, this helps system stability, since backlogged packets are removed from the offered load by the deadline.

The capture effect with two power levels can be used to give priority to one class of signals. For example, time-constrained traffic such as real-time voice can be transmitted with higher power than non-real-time data transfer. Another option, if there is time for multiple retransmissions, is to use the high power only if the deadline is imminent [9]. Also, in the multichannel case, it has been suggested [10] to send copies of the same packet on multiple channels when the deadline is imminent. Both the higher-power and the multiple-copy techniques use the principle of devoting more signal energy to the transmission of the packet when the deadline nears.

11. NONBINARY TRANSMISSION WITH ALOHA

In systems where collision is a more serious problem than noise, signal-to-noise ratio is relatively high. Channel capacity would then indicate that more bits per hertz could be sent by using nonbinary transmission. Consider the channel capacity formula for white Gaussian noise channels:

$$C = F \times \log\left(1 + \frac{P}{N_0 F}\right) \tag{7}$$

where C is capacity in bits per second, F is the bandwidth in hertz, $N = N_0 F$ is the noise power, and P is the signal power. When P/N is large, C is proportional to P/N in dB (decibels). Thus there would be 5 times the capacity at 40 dB as at 8 dB. If ALOHA sent at 5 times the rate, its same size packets would only be one-fifth as long in duration, and thus would be much less likely to collide if data were generated at the same long-term average rate. If P/N varied widely, however, adaptation to the current channel capacity could be complex and difficult to accomplish.

12. CDMA AND SPREAD ALOHA

Code-division multiple access [11] is a technique whereby many users send simultaneously over a wide frequency band. In the direct sequence version, a sender sends a bit by sending its unique pseudorandom L-bit binary code sequence or its inverse. The individual data rate is lower by a factor of L (spreading factor) than the data rate at which a single user could send. The receiver decodes a particular sender by correlating with the sender's known code. Abramson [12] has suggested an idea of everyone using one code, in a scheme he calls "spread ALOHA." Suppose that two different users start sending a sequence of data at starting times d seconds apart. A correlator will output two interleaved streams of spikes spaced d seconds apart corresponding to the two senders. The approximate output is illustrated in Fig. 4, where the solid pulses represent the bits of one stream and the dashed line pulses represent the bits of the other stream. If the spacing d is not very close to zero or to a multiple of the period, the two senders' data can be readily decoded. A similar effect can be achieved by sending a narrow, low-duty-cycle pulsetrain for a packet [13]; no code or correlator would then be needed. This latter alternative could be called pulse time-spread ALOHA. Spread ALOHA is simpler than everyone having their own code, but it suffers from the usual ALOHA problem of collisions when the pulsetrains are too closely spaced. With unequal powers, the capture effect could result in one of two or more colliders being successfully decoded. For example, if the two pulsetrains in Fig. 4 were in step, the net sum of the two responses would always have the correct sign of the larger pulse, in the absence of noise.

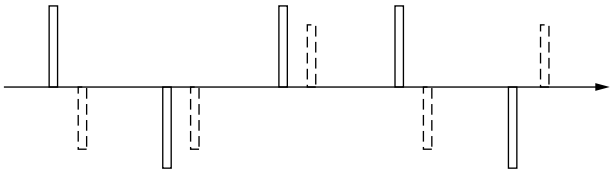


Figure 4. Two decorrelated streams or two user pulsetrains.

BIOGRAPHY

John J. Metzner is a professor of computer engineering, with appointments in both the Department of Electrical Engineering and the Department of Computer Science and engineering. He received his B.E.E., M.E.E., and Eng.Sc.D. degrees from New York University in 1953, 1954, and 1958, respectively. He has held faculty and research appointments at New York University, Polytechnic University, Brooklyn, New York, Wayne State University, Detroit, Michigan, Oakland University, Rochester, Michigan, and, since 1986, The Pennsylvania State University, University Park, Pennsylvania. He served a year as acting dean of the School of Engineering and Computer Science at Oakland University, and two years as acting director of the computer engineering program at Penn State. In research, Dr. Metzner has devised various ARQ protocols for reliable and efficient data communication, techniques for efficient comparison of remote replicated data files, efficient acknowledgement protocols for slotted ring networks, improved broadcast retransmission protocols, methods for improved utilization of ALOHA and spread spectrum multiaccess, and techniques for simpler and more effective error correction.

BIBLIOGRAPHY

1. N. Abramson, The ALOHA system, in N. Abramson and F. Kuo, eds., *Computer Communication Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
2. R. Metcalf and D. Boggs, Ethernet: Distributed packet switching for local computer networks, *Commun. ACM* **19**(7): 395–403 (July 1976).
3. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
4. J. I. Capetanakis, Tree algorithms for packet broadcast channels, *IEEE Trans. Inform. Theory* **IT-25**(5): 505–515 (Sept. 1979).
5. D. J. Goodman, Cellular packet communications, *IEEE Trans. Commun.* **38**(8): 1272–1280 (Aug. 1990).
6. J. Metzner, On improving utilization in ALOHA networks, *IEEE Trans. Commun.* **COM-24**(4): 447–448 (April 1976).
7. J. Arnbak and W. van Blitterswijk, Capacity of slotted ALOHA in Rayleigh-fading channels, *IEEE J. Select. Areas Commun.* **SAC-5**: 261–269 (Feb. 1987).
8. M. Sidi and I. Cidon, A multi-station packet radio network, *Performance Evaluation*, Vol. 8, no. 1, North-Holland, Feb. 1988, pp. 65–72.
9. J. Metzner and J.-M. Chung, Efficient energy utilization with a time constraint and time-varying channels, *IEEE Trans. Vehic. Technol.* **48**(12): 2005–2013 (Dec. 2000).
10. Y. Birk and Y. Keren, Judicious use of redundant transmissions in multichannel ALOHA networks with deadlines, *IEEE J. Select. Areas Commun.* **17**(2): 257–269 (Feb. 1999).
11. S. Tantaratana and K. Ahmed, eds., *Wireless Applications of Spread Spectrum Systems: Selected Readings*, IEEE Press, Piscataway, NJ, 1998.
12. N. Abramson, Multiple access in wireless digital networks, *Proc. IEEE* **82**(9): 1360–1369 (Sept. 1994).
13. J. Metzner, *Reliable Data Communications*, Academic Press, San Diego, CA, 1998.

AMPLITUDE MODULATION

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Speech, images, and video are examples of analog signals that are transmitted routinely over wireline and wireless (radio-) communication channels. In spite of the general trend toward digital transmission of these types of analog signals, there is still today a significant amount of analog signal transmission, specifically, audio and video broadcast. In this article, we describe the transmission of analog signals by amplitude modulation of a sinusoidal carrier. Methods for demodulation of the amplitude modulated sinusoidal carrier to recover the analog signal are also described.

The analog signal to be transmitted is generally characterized as an information-bearing message signal, which is denoted as $m(t)$. The message signal $m(t)$ is an electrical signal that may represent either an audio signal, or a still image, or a video signal. Such a signal is assumed to be a lowpass signal with frequency content that extends from $f = 0$ to some upper frequency limit, say, B Hz. Hence, if the voltage spectrum (Fourier transform) of $m(t)$ is denoted as $M(f)$, then $M(f) = 0$ for $|f| > B$. The bandwidth B of the message signal depends on the type of analog signal. For example, the bandwidth of an audio signal is typically approximately 4 kHz and that of an analog video signal is approximately 6 MHz.

The sinusoidal carrier which is modulated by $m(t)$ is expressed as

$$c(t) = A_c \cos 2\pi f_c t$$

where A_c is the (unmodulated) carrier amplitude and f_c is the carrier frequency. Basically, the modulation of the carrier $c(t)$ by the message signal $m(t)$ converts the message signal from lowpass to bandpass, in the neighborhood of the carrier f_c. This frequency translation resulting from the modulation process is performed in order to achieve one or both of the following objectives: (1) to translate lowpass signal in frequency to the passband of the channel so that the spectrum of the frequency-translated message signal matches the passband characteristics of the channel and (2) to accommodate for the simultaneous transmission of

signals from several message sources, where each message signal modulates a different carrier and, thus, occupies a different frequency band, as in frequency division multiplexing.

2. AMPLITUDE MODULATION

In amplitude modulation, the message signal $m(t)$ is impressed on the amplitude of the carrier signal $c(t)$. There are several different ways to modulate the amplitude of the carrier by the message signal $m(t)$, each of which results in different spectral characteristics for the transmitted signal. Specifically, these methods are called (1) *double-sideband*, *suppressed-carrier AM*, (2) *conventional double-sideband AM*, (3) *single-sideband AM*, and (4) *vestigial-sideband AM*.

2.1. Double-Sideband Suppressed-Carrier AM

A double-sideband, suppressed-carrier (DSB-SC) AM signal is obtained by multiplying the message signal $m(t)$ with the carrier signal $c(t)$. Thus, we have the amplitude modulated signal

$$\begin{aligned} u(t) &= m(t)c(t) \\ &= A_c m(t) \cos 2\pi f_c t \end{aligned} \tag{1}$$

The voltage spectrum of the modulated signal can be obtained by computing the Fourier transform of $u(t)$. The result of this computation is

$$U(f) = \frac{A_c}{2}[M(f - f_c) + M(f + f_c)] \tag{2}$$

Figure 1 illustrates the magnitude and phase spectra for $M(f)$ and $U(f)$.

We observe that the magnitude of the spectrum of the message signal $m(t)$ has been translated or shifted in frequency by an amount f_c. The phase of the message signal has been translated in frequency the same amount. Furthermore, the bandwidth occupancy of the amplitude-modulated signal is $2B$, whereas the bandwidth of the message signal $m(t)$ is B. Therefore, the channel bandwidth required to transmit the modulated signal $u(t)$ is $B_c = 2B$.

The frequency content of the modulated signal $u(t)$ in the frequency band $|f| > f_c$ is called the *upper sideband* of $U(f)$, and the frequency content in the frequency band $|f| < f_c$ is called the *lower sideband* of $U(f)$. It is important to note that either one of the sidebands of $U(f)$ contains all the frequencies that are in $M(f)$. Thus, the frequency content of $U(f)$ for $f > f_c$ corresponds to the frequency content of $M(f)$ for $f > 0$, and the frequency content of $U(f)$ for $f < -f_c$ corresponds to the frequency content of $M(f)$ for $f < 0$. Hence, the upper sideband of $U(f)$ contains all the frequencies in $M(f)$. A similar statement applies to the lower sideband of $U(f)$. Therefore, the lower sideband of $U(f)$ contains all the frequency content of the message signal $M(f)$. Since $U(f)$ contains both the upper and the lower sidebands, it is called a *double-sideband* (DSB *AM signal*.

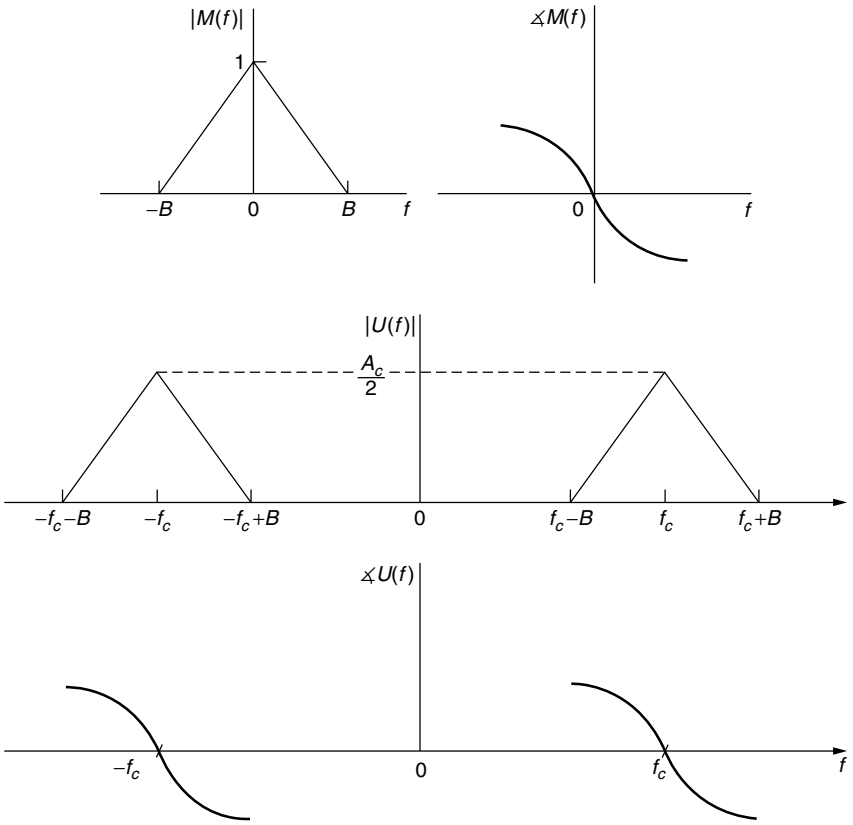


Figure 1. Magnitude and phase spectra of the message signal $m(t)$ and the DSB AM modulated signal $u(t)$.

The other characteristic of the modulated signal $u(t)$ is that it does not contain a carrier component; that is, all the transmitted power is contained in the modulating (message) signal $m(t)$. This is evident from observing the spectrum of $U(f)$. We note that, as long as $m(t)$ does not have any DC component, there is no impulse in $U(f)$ at $f = f_c$, which would be the case if a carrier component was contained in the modulated signal $u(t)$. For this reason, $u(t)$ is called a *suppressed-carrier signal*. Therefore, $u(t)$ is a DSB-SC AM signal.

In the propagation of the modulated signal through the communication channel, the signal encounters a propagation time delay, which depends on the characteristics of the propagation medium (channel). Generally, this time delay is not precisely known to the signal receiver. Such a propagation delay results in a received signal, in the absence of any channel distortion or additive noise, of the form

$$r(t) = A_c m(t) \cos(2\pi f_c t + \phi_c)$$

where ϕ_c is a carrier phase manifested by the propagation delay.

Suppose that we demodulate the received signal by first multiplying $r(t)$ by a locally generated sinusoid $\cos 2\pi f_c t + \phi)$, where ϕ is the phase of the sinusoid, and then passing the product signal through an ideal lowpass filter having a bandwidth B. The multiplication of $r(t)$ with $\cos(2\pi f_c t + \phi)$ yields

$$\begin{aligned} r(t)\cos(2\pi f_c t + \phi) &= A_c m(t)\cos(2\pi f_c t + \phi_c)\cos(2\pi f_c t + \phi) \\ &= \tfrac{1}{2}A_c m(t)\cos(\phi_c - \phi) \\ &\quad + \tfrac{1}{2}A_c m(t)\cos(4\pi f_c t + \phi + \phi_c) \end{aligned} \tag{3}$$

A lowpass filter rejects the double frequency components and passes only the lowpass components. Hence, its output is

$$y_o(t) = \tfrac{1}{2}A_c m(t)\cos(\phi_c - \phi) \tag{4}$$

Note that $m(t)$ is multiplied by $\cos(\phi_c - \phi)$. Thus, the desired signal is scaled in amplitude by a factor that depends on the phase difference between the phase ϕ_c of the carrier in the received signal and the phase ϕ of the locally generated sinusoid. When $\phi_c \neq \phi$, the amplitude of the desired signal is reduced by the factor $\cos(\phi_c - \phi)$. If $\phi_c - \phi = 45^\circ$, the amplitude of the desired signal is reduced by $\sqrt{2}$ and the signal power is reduced by a factor of 2. If $\phi_c - \phi = 90^\circ$, the desired signal component vanishes.

The discussion above demonstrates the need for a *phase-coherent or synchronous demodulator* for recovering the message signal $m(t)$ from the received signal. Thus, the phase for the locally generated sinusoid should ideally be equal to the phase ϕ_c of the received carrier signal.

A sinusoid that is phase-locked to the phase of the received carrier can be generated by use of a phase-locked loop (PLL), which is described in Refs. 1–3.

2.2. Conventional Double-Sideband AM

A conventional AM signal consists of a large carrier component in addition to the double-sideband AM modulated signal. The transmitted signal is expressed mathematically as

$$u(t) = A_c[1 + m(t)]\cos 2\pi f_c t \tag{5}$$

where the message waveform is constrained to satisfy the condition that $|m(t)| \leq 1$. We observe that $A_c m(t)\cos 2\pi f_c t$ is a double-sideband AM signal and $A_c \cos 2\pi f_c t$ is the carrier component. Figure 2 illustrates an AM signal in the time domain.

As long as $|m(t)| \leq 1$, the amplitude $A_c[1 + m(t)]$ is always positive. This is the desired condition for conventional DSB AM that makes it easy to demodulate, as described next. On the other hand, if $m(t) < -1$ for some t, the AM signal is said to be *overmodulated* and its demodulation is rendered more complex. In practice, $m(t)$ is scaled so that its magnitude is always less than unity.

The voltage spectrum of $u(t)$ given by (5) is obtained by computing the Fourier transform of $u(t)$. The result of this computation is

$$U(f) = \frac{A_c}{2}[M(f - f_c) + M(f + f_c) + \delta(f - f_c) + \delta(f + f_c)] \tag{6}$$

This spectrum is sketched in Fig. 3. We observe that spectrum of the conventional AM signal occupies a bandwidth twice the bandwidth of the message signal. As in the case of DSB-SC carrier, conventional AM consists of both an upper sideband and a lower sideband. In addition, the spectrum of a conventional AM signal contains impulses at $f = f_c$ and $f = -f_c$, which correspond to the presence of the carrier component in the modulated signal.

The major advantage of conventional AM signal transmission is the ease with which the signal can be demodulated. There is no need for a synchronous demodulator. Since the message signal $m(t)$ satisfies the condition $|m(t)| < 1$, the envelope (amplitude) $1 + m(t) > 0$. If we rectify the received signal, we eliminate the negative values without affecting the message signal as shown in Fig. 4. The rectified signal is equal to $u(t)$ when $u(t) > 0$ and zero when $u(t) < 0$. The message signal is recovered by passing the rectified signal through a lowpass filter whose bandwidth matches that of the message signal. The combination of the rectifier and the lowpass filter is called an *envelope detector*.

Ideally, the output of the envelope detector is of the form

$$d(t) = g_1 + g_2 m(t)$$

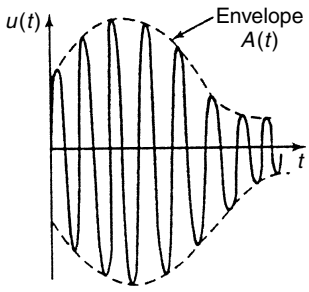


Figure 2. A conventional AM signal in the time domain.

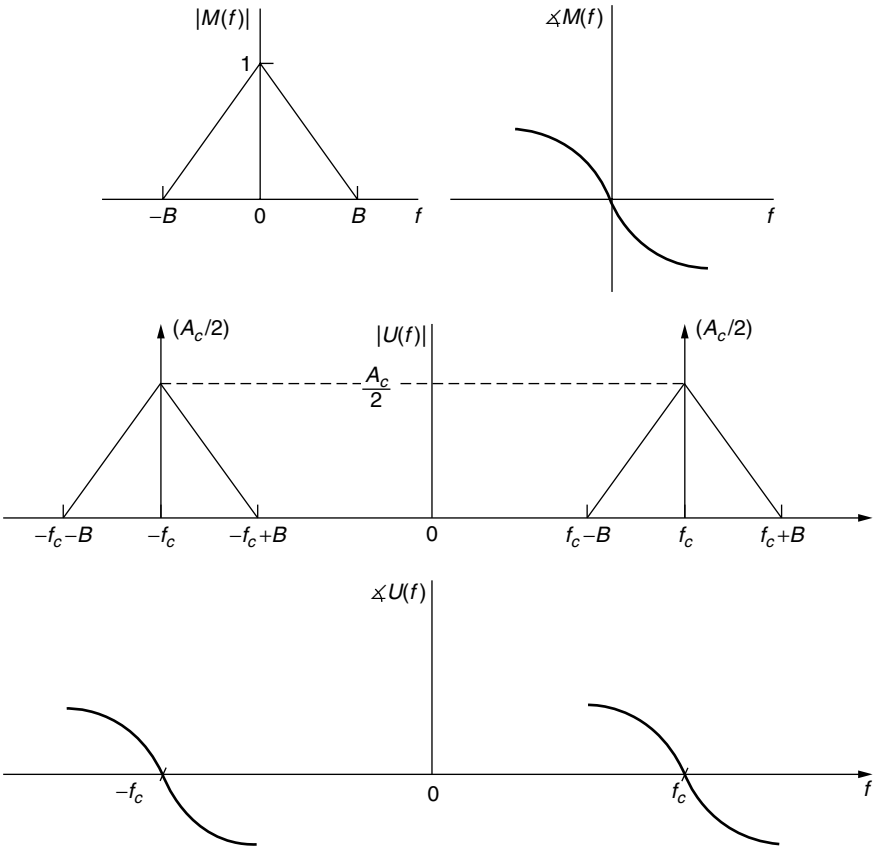


Figure 3. Magnitude and phase spectra of the message signal $m(t)$ and the conventional AM signal.

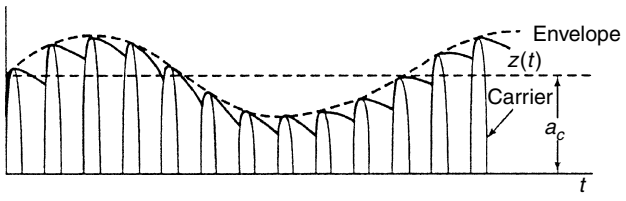


Figure 4. Envelope detection of conventional AM signal.

which g_1 represents a dc component and g_2 is a gain factor due to the signal demodulator. The dc component can be eliminated by passing $d(t)$ through a transformer, whose output is $g_2 m(t)$.

The simplicity of the demodulator has made conventional DSB AM a practical choice for AM radiobroadcasting. Since there are literally billions of radio receivers, an inexpensive implementation of the demodulator is extremely important. The power inefficiency of conventional AM is justified by the fact that there are few broadcast transmitters relative to the number of receivers. Consequently, it is cost-effective to construct powerful transmitters and sacrifice power efficiency in order to simplify the signal demodulation at the receivers.

2.3. Single-Sideband AM

In Section 2.1 it was observed that a DSB-SC AM signal required a channel bandwidth of $B_c = 2B$ for transmission, where B is the bandwidth of the message signal $m(t)$. However, the two sidebands are redundant. In this section, it is demonstrated that the transmission of either sideband is sufficient to reconstruct the message signal $m(t)$ at the receiver. Thus, the bandwidth of the transmitted signal is reduced to that of the message signal $m(t)$.

It can be demonstrated by the use of the Fourier transform that a single-sideband (SB) AM signal can be represented mathematically as

$$u(t) = A_c m(t) \cos 2\pi f_c t \mp A_c \hat{m}(t) \sin 2\pi f_c t \tag{7}$$

where $\hat{m}(t)$ is the Hilbert transform of $m(t)$ and the plus-or-minus sign determines which sideband we obtain (+ for the lower sideband and − for the upper sideband). The Hilbert transform may be viewed as a linear filter with impulse response $h(t) = 1/\pi t$ and frequency response

$$H(f) = \begin{cases} -j, & f > 0 \\ j, & f < 0 \\ 0, & f = 0 \end{cases} \tag{8}$$

Therefore, the SSB AM signal $u(t)$ may be generated by using the system configuration shown in Fig. 5.

The method shown in Fig. 5 for generating a SSB AM signal is one that employs a Hilbert transform filter. Another method, illustrated in Fig. 6, generates a DSB-SC AM signal and then employs a filter that selects either the upper sideband or the lower sideband of the double-sideband AM signal.

To recover the message signal $m(t)$ in the received SSB AM signal, a phase coherent or synchronous demodulator

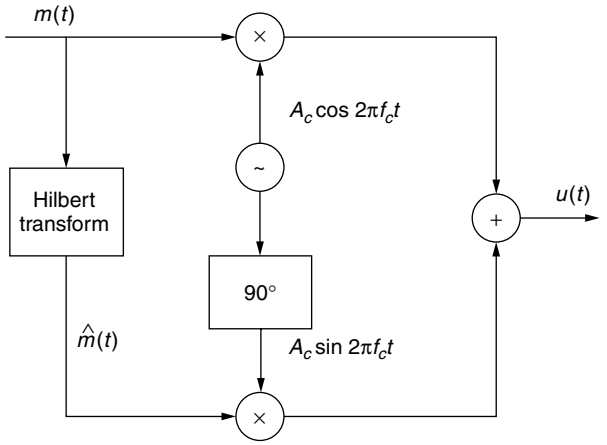


Figure 5. Generation of a single-sideband signal.

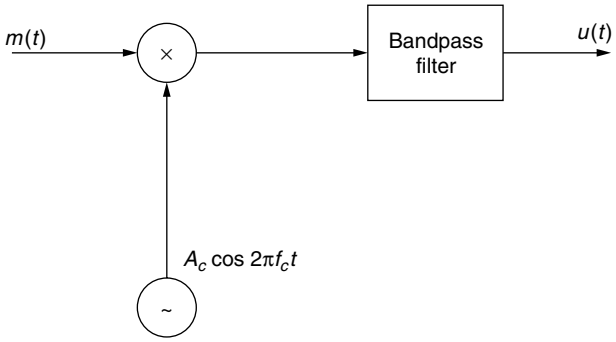


Figure 6. Generation of a single-sideband AM signal by filtering one of the sidebands of a DSB-SC AM signal.

is required, as was the case of DSB-SC AM signals. Thus, for the received SSB signal as given in (7), when demodulated by multiplying $r(t)$ with a sinusoid that has a phase offset ϕ, we obtain

$$\begin{aligned} r(t)\cos(2\pi f_c t + \phi) &= u(t)\cos(2\pi f_c t + \phi) \\ &= \tfrac{1}{2}A_c m(t)\cos\phi \pm \tfrac{1}{2}A_c \hat{m}(t)\sin\phi \\ &\quad + \text{double-frequency terms} \end{aligned} \tag{9}$$

By passing the product signal in (9) through an ideal lowpass filter, the double-frequency components are eliminated, leaving us with

$$y(t) = \tfrac{1}{2}A_c m(t)\cos\phi \pm \tfrac{1}{2}A_c \hat{m}(t)\sin\phi \tag{10}$$

Note that the effect of the phase offset not only is to reduce the amplitude of the desired signal $m(t)$ by $\cos\phi$ but also results in an undesirable sideband signal due to the presence of $\hat{m}(t)$ in $y(t)$. The latter component was not present in a DSB-SC signal and, hence, it was not a factor. However, it is an important element that contributes to the distortion of the demodulated SSB signal.

The transmission of a pilot tone at the carrier frequency is a very effective method for providing a phase-coherent reference signal for performing synchronous demodulation at the receiver. Thus, the undesirable sideband signal component is eliminated. However, this means that a portion of the transmitted power must be allocated to the transmission of the carrier.

The spectral efficiency of SSB AM makes this modulation method very attractive for use in voice communications over telephone channels (wirelines and cables). In this application, a pilot tone is transmitted for synchronous demodulation and shared among several channels.

The filter method shown in Fig. 6 for selecting one of the two signal sidebands for transmission is particularly difficult to implement when the message signal $m(t)$ has a large power concentrated in the vicinity of $f = 0$. In such a case, the sideband filter must have an extremely sharp cutoff in the vicinity of the carrier in order to reject the second sideband. Such filter characteristics are very difficult to implement in practice.

2.4. Vestigial-Sideband AM

The stringent frequency-response requirements on the sideband filter in a SSB AM system can be relaxed by allowing a part, called a vestige, of the unwanted sideband to appear at the output of the modulator. Thus, the design of the sideband filter is simplified at the cost of a modest increase in the channel bandwidth required to transmit the signal. The resulting signal is called *vestigial-sideband* (VSB) *AM*.

To generate a VSB AM signal we begin by generating a DSB-SC AM signal and passing it through a sideband filter with frequency response $H(f)$ as shown in Fig. 7. In the time domain the VSB signal may be expressed as

$$u(t) = [A_c m(t)\cos 2\pi f_c t] * h(t) \tag{11}$$

where $h(t)$ is the impulse response of the VSB filter and the asterisk denotes convolution. In the frequency domain, the corresponding expression is

$$U(f) = \frac{A_c}{2}[M(f - f_c) + M(f + f_c)]H(f) \tag{12}$$

To determine the frequency-response characteristics of the filter, consider the demodulation of the VSB signal $u(t)$. Multiply $u(t)$ by the carrier component $\cos 2\pi f_c t$ and pass the result through an ideal lowpass filter, as shown in Fig. 8. Thus, the product signal is

$$v(t) = u(t)\cos 2\pi f_c t$$

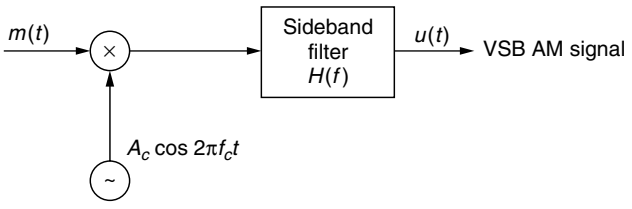


Figure 7. Generation of a VSB AM signal.

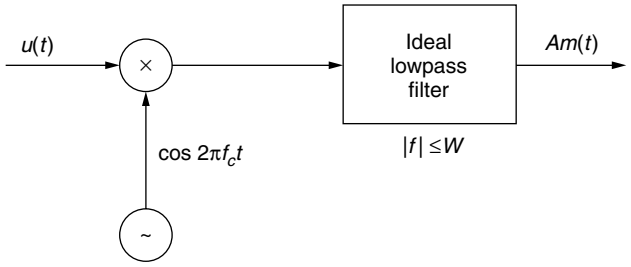


Figure 8. Demodulation of VSB signal.

or, equivalently

$$V(f) = \tfrac{1}{2}[U(f - f_c) + U(f + f_c)] \tag{13}$$

When $U(f)$ is substituted from (12) into (13), the result is

$$\begin{aligned} V(f) = {} & \frac{A_c}{4}[M(f - 2f_c) + M(f)]H(f - f_c) \\ & + \frac{A_c}{4}[M(f) + M(f + 2f_c)]H(f + f_c) \end{aligned} \tag{14}$$

The lowpass filter rejects the double-frequency terms and passes only the components in the frequency range $|f| \leq B$. Hence, the signal spectrum at the output of the ideal lowpass filter is

$$V_o(f) = \frac{A_c}{4}M(f)[H(f - f_c) + H(f + f_c)] \tag{15}$$

We require that the message signal at the output of the lowpass filter be undistorted. Hence, the VSB filter characteristic must satisfy the condition

$$H(f - f_c) + H(f + f_c) = \text{constant}, \quad |f| \leq B \tag{16}$$

This condition is satisfied by a filter that has the frequency-response characteristic shown in Fig. 9. We note that $H(f)$ selects the upper sideband and a vestige of the lower sideband. It has odd symmetry about the carrier frequency f_c, in the frequency range $f_c - f_a < f < f_c + f_a$, where f_a is a conveniently selected frequency that is some small fraction of B; that is, $f_a \ll B$. Thus, we obtain an undistorted version of the transmitted signal. Figure 10 illustrates the frequency response of a VSB filter that selects the lower sideband and a vestige of the upper sideband.

In practice, the VSB filter is designed to have some specified phase characteristic. To avoid distortion of the message signal, the VSB filter should be designed to have linear phase over its passband $f_c - f_a \leq |f| \leq f_c + B$.

3. IMPLEMENTATION OF AM MODULATORS AND DEMODULATORS

There are several different methods for generating AM modulated signals. We shall describe the methods most commonly used in practice. Since the process of modulation involves the generation of new frequency components, modulators are generally characterized as nonlinear and, or, time-variant systems.

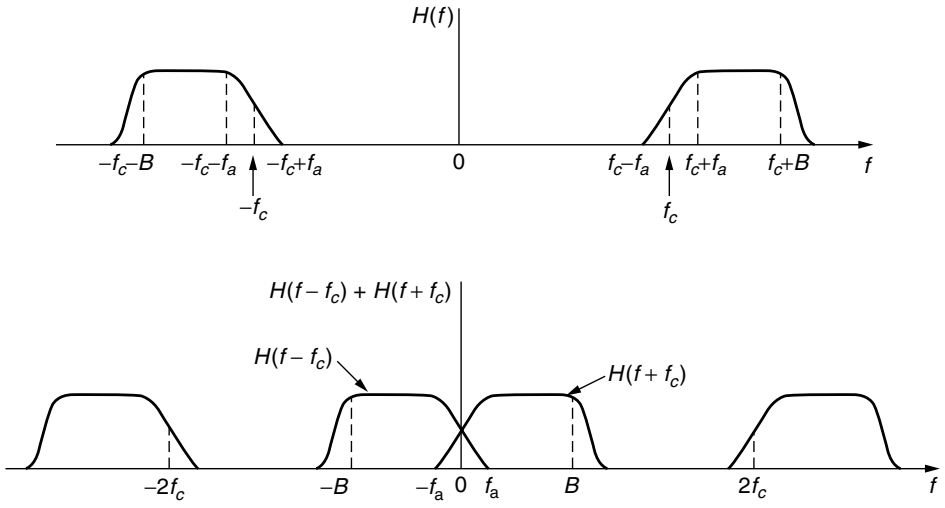


Figure 9. VSB filter characteristics.

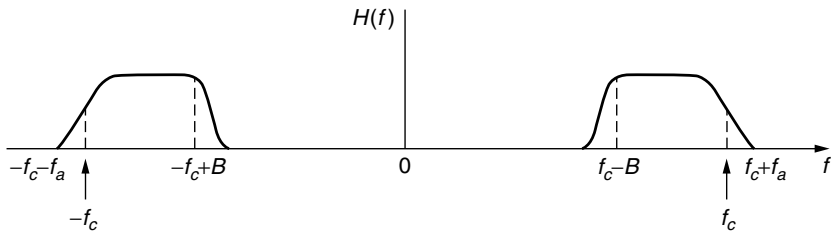


Figure 10. Frequency response of VSB filter for selecting the lower sideband of the message signal.

3.1. Power-Law Modulation

Consider the use of a nonlinear device such as a p-n diode that has a voltage–current characteristic as shown in Fig. 11. Suppose that the voltage input to such a device is the sum of the message signal $m(t)$ and the carrier $A_c \cos 2\pi f_c t$, as illustrated in Fig. 12. The nonlinearity will generate a product of the message $m(t)$ with the carrier, plus additional terms. The desired modulated signal can be filtered out by passing the output of the nonlinear device through a bandpass filter.

To elaborate on this method, suppose that the nonlinear device has an input–output (square-law) characteristic of the form

$$v_0(t) = a_1 v_i(t) + a_2 v_i^2(t) \tag{17}$$

where $v_i(t)$ is the input signal $v_0(t)$ is the output signal, and the parameters (a_1, a_2) are constants. Then, if the input to the nonlinear device is

$$v_i(t) = m(t) + A_c \cos 2\pi f_c t \tag{18}$$

its output is

$$\begin{aligned} v_0(t) &= a_1[m(t) + A_c \cos 2\pi f_c t] \\ &\quad + a_2[m(t) + A_c \cos\cos 2\pi f_c t]^2 \\ &= a_1 m(t) + a_2 m^2(t) + a_2 A_c^2 \cos 2\pi f_c t \\ &\quad + A_c a_1 \left[1 + \frac{2a_2}{a_1} m(t)\right] \cos 2\pi f_c t \end{aligned} \tag{19}$$

The output of the bandpass filter with bandwidth $2B$ centered at $f = f_c$ yields

$$u(t) = A_c a_1 \left[1 + \frac{2a_2}{a_1} m(t)\right] \cos 2\pi f_c t \tag{20}$$

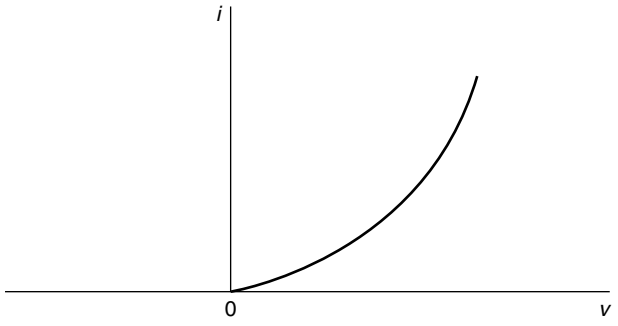


Figure 11. Voltage–current characteristic of p-n diode.

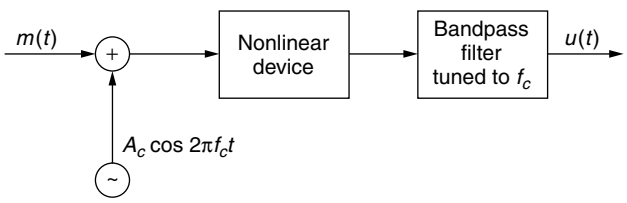


Figure 12. Block diagram of power-law AM modulator.

where $2a_2|m(t)|/a_1 < 1$ by design. Thus, the signal generated by this method is a conventional DSB AM signal.

3.2. Switching Modulator

Another method for generating an AM modulated signal is by means of a switching modulator. Such a modulator can be implemented by the system illustrated in Fig. 13a. The sum of the message signal and the carrier; i.e., $v_i(t)$ given by (18), are applied to a diode that has the input–output voltage characteristic shown in Fig. 13b, where $A_c \gg m(t)$. The output across the load resistor is simply

$$v_0(t) = \begin{cases} v_i(t), & c(t) > 0 \\ 0, & c(t) < 0 \end{cases} \tag{21}$$

This switching operation may be viewed mathematically as a multiplication of the input $v_i(t)$ with the switching function $s(t)$

$$v_0(t) = [m(t) + A_c \cos 2\pi f_c t] s(t) \tag{22}$$

where $s(t)$ is as shown in Fig. 13c.

Since $s(t)$ is a periodic function, it is represented in the Fourier series as

$$s(t) = \frac{1}{2} + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{2n-1} \cos[2\pi f_c t(2n-1)] \tag{23}$$

Hence

$$\begin{aligned} v_0(t) &= [m(t) + A_c \cos 2\pi f_c t] s(t) \\ &= \frac{A_c}{2} \left[1 + \frac{4}{\pi A_c} m(t)\right] \cos 2\pi f_c t + \text{other terms} \end{aligned} \tag{24}$$

The desired AM modulated signal is obtained by passing $v_0(t)$ through a bandpass filter with center frequency $f = f_c$

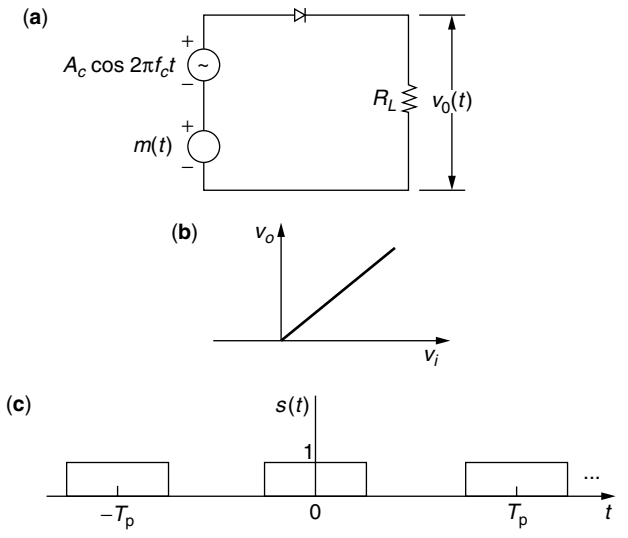


Figure 13. Switching modulator and periodic switching signal.

and bandwidth $2B$. At its output, we have the desired conventional DSB AM signal:

$$u(t) = \frac{A_c}{2}\left[1 + \frac{4}{\pi A_c} m(t)\right] \cos 2\pi f_c t \tag{25}$$

3.3. Balanced Modulator

A relatively simple method for generating a DSB-SC AM signal is to use two conventional AM modulators arranged in the configuration illustrated in Fig. 14. For example, we may use two square-law AM modulators as described above. Care must be taken to select modulators with approximately identical characteristics so that the carrier component cancels out at the summing junction.

3.4. Ring Modulator

Another type of modulator for generating a DSB-SC AM signal is the ring modulator illustrated in Fig. 15. The switching of the diodes is controlled by a square wave of frequency f_c, denoted as $c(t)$, which is applied to the center taps of the two transformers. When $c(t) > 0$, the top and bottom diodes conduct, while the two diodes in the crossarms are off. In this case, the message signal $m(t)$ is multiplied by $+1$. When $c(t) < 0$, the diodes in the crossarms of the ring conduct, while the other two are switched off. In this case, the message signal $m(t)$ is multiplied by -1. Consequently, the operation of the ring modulator may be described mathematically as a multiplier of $m(t)$ by the square-wave carrier $c(t)$:

$$v_0(t) = m(t)c(t) \tag{26}$$

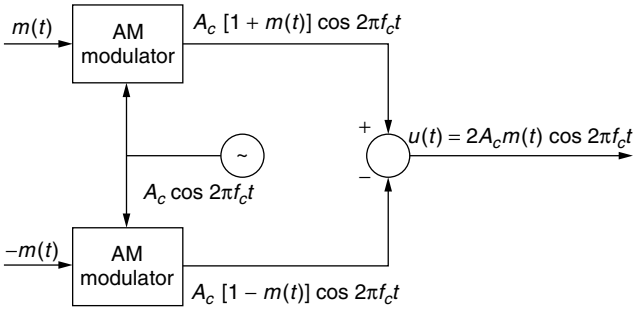


Figure 14. Block diagram of a balanced modulator.

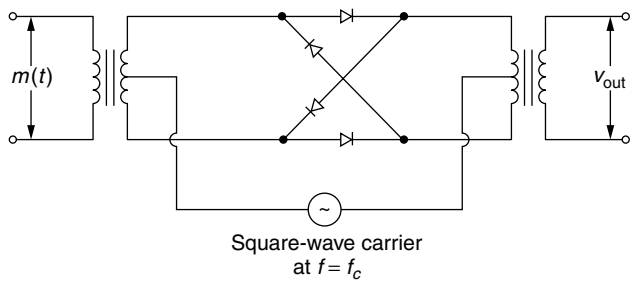


Figure 15. Ring modulator for generating DSB-SC AM signals.

Since $c(t)$ is a periodic function, it is represented by the Fourier series

$$c(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{(-1)n^{-1}}{2n-1} \cos[2\pi f_c(2n-1)t] \tag{27}$$

Hence, the desired DSB-SC AM signal $u(t)$ is obtained by passing $v_0(t)$ through a bandpass filter with center frequency f_c and bandwidth $2B$.

From the discussion above, we observe that the balanced modulator and the ring modulator systems, in effect, multiply the message signal $m(t)$ with the carrier to produce a DSB-SC AM signal. The multiplication of $m(t)$ with $A_c \cos w_c t$ is called a *mixing operation*. Hence, a mixer is basically a balanced modulator.

The method shown in Fig. 5 for generating a SSB signal requires two mixers, specifically, two balanced modulators, in addition to the Hilbert transformer. On the other hand, the filter method illustrated in Fig. 6 for generating a SSB signal requires a single balanced modulator and a sideband filter.

Let us now consider the demodulation of AM signals. We begin with a description of the envelope detector.

3.5. Envelope Detector

As indicated previously, conventional DSB AM signals are easily demodulated by means of an envelope detector. A circuit diagram for an envelope detector is shown in Fig. 16. It consists of a diode and an RC circuit, which is basically a simple lowpass filter.

During the positive half-cycle of the input signal, the diode is conducting and the capacitor charges up to the peak value of the input signal. When the input falls below the voltage on the capacitor, the diode becomes reverse-biased and the input becomes disconnected from the output. During this period, the capacitor discharges slowly through the load resistor R. On the next cycle of the carrier, the diode conducts again when the input signal exceeds the voltage across the capacitor. The capacitor charges up again to the peak value of the input signal and the process is repeated again.

The time-constant RC must be selected so as to follow the variations in the envelope of the carrier-modulated signal. In effect

$$\frac{1}{f_c} \ll RC \ll \frac{1}{B}$$

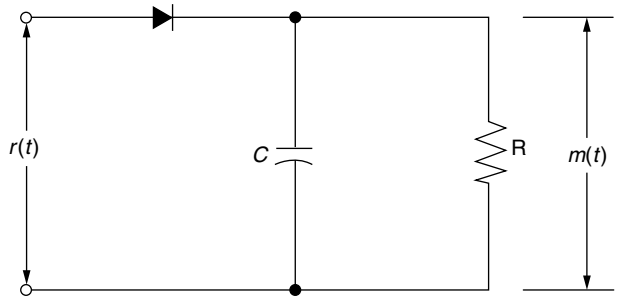


Figure 16. An envelope detector.

In such a case, the capacitor discharges slowly through the resistor, and thus, the output of the envelope detector closely follows the message signal.

3.6. Demodulation of DSB-SC AM Signals

As indicated earlier, the demodulation of a DSB-SC AM signal requires a synchronous demodulator. Thus, the demodulator must use a coherent phase reference, which is usually generated by means of a phase-locked loop (PLL) to demodulate the received signal.

The general configuration is shown in Fig. 17. A PLL is used to generate a phase-coherent carrier signal that is mixed with the received signal in a balanced modulator. The output of the balanced modulator is passed through a lowpass filter of bandwidth B that passes the desired signal and rejects all signal and noise components above B Hz. The characteristics and operation of the PLL are described in Refs. 1–3.

3.7. Demodulation of SSB Signals

The demodulation of SSB AM signals also requires the use of a phase-coherent reference. In the case of signals such as speech, that have relatively little or no power content at d_c, it is straightforward to generate the SSB signal, as shown in Fig. 6, and then to insert a small carrier component that is transmitted along with the message. In such a case we may use the configuration shown in Fig. 18 to demonstrate the SSB signal. We observe that a balanced modulator is used for the purpose of frequency conversion of the bandpass signal to lowpass or baseband.

3.8. Demodulation of VSB Signals

In VSB a carrier component is generally transmitted along with the message sidebands. The existence of the carrier component makes it possible to extract a phase-coherent reference for demodulation in a balanced modulator, as shown in Fig. 18.

In some applications such as TV broadcasting, a large carrier component is transmitted along with the message in the VSB signal. In such a case, it is possible to recover the message by passing the received VSB signal through an envelope detector.

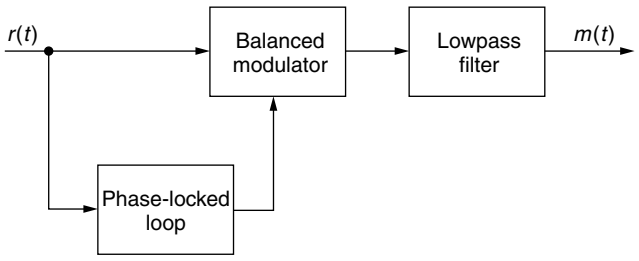


Figure 17. Block diagram of demodulator for DSB-SC AM signals.

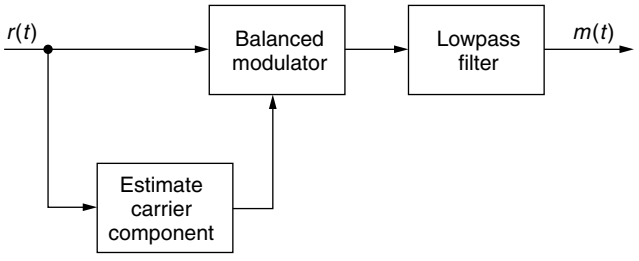


Figure 18. Block diagram of demodulator for SSB AM signal with a carrier component.

4. CONCLUDING REMARKS

This article has covered the different types of amplitude modulation techniques that are used in the transmission of analog signals. Conventional AM is most widely used in radiobroadcasting. It is anticipated that conventional AM in radiobroadcasting will be phased out eventually and replaced by a digital modulation method that provides better signal fidelity. Current analog television broadcasting will also undergo a similar transformation. Single-sideband AM was widely used in telephone systems for audio signal transmission over many decades. However, today, audio signal transmission in the telephone systems is performed by digital modulation after the voice signals are converted to digital form by use of pulse code modulation (PCM) or differential PCM (DPCM).

Treatments of AM modulation can be found in many undergraduate-level textbooks in communication systems. References 4–6 are cited as typical.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the

author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. W. C. Lindsey, *Synchronization Systems in Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1972.
2. F. M. Gardner, *Phaselock Techniques*, Wiley, New York, 1979.
3. W. C. Lindsey and C. M. Chie, A survey of digital phase-locked loops, *Proc. IEEE* **69**: 410–432 (1981).
4. H. Taub and D. L. Schilling, *Principles of Communication Systems*, McGraw-Hill, New York, 1971.
5. J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
6. H. Stark, F. B. Tuteur and J. B. Anderson, *Modern Electrical Communications*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1988.

ANTENNA ARRAYS

JOHN N. SAHALOS
Radiocommunications Laboratory
Aristotle University of Thessaloniki
Thessaloniki, Greece

1. INTRODUCTION

Antennas occupy a palmary position in radiocommunication systems. It is not an overemphasis to say that antennas have become ubiquitous devices. This has occurred because radio and TV as well satellite and mobile communications have experienced the largest growth among the industry systems. The strongest economic and social impact nowadays is coming from cellular telephony, personal communications, and satellite navigation systems. All these applications have served on the motivation for engineers to achieve elegant antennas to be incorporated into handy and portable systems.

Many textbooks provide in-depth resources on antennas. Especially on antenna arrays there are digests and books containing extensive data and techniques. The references cited here [1–20] include some of the most well known and recommended books.

A device able to receive or transmit the electromagnetic energy is called an antenna, which consists of one or more elements. A single-element antenna is usually not enough to cover the technical needs. That happens because its performance is limited. A set of discrete elements made up of an antenna array offers the solution to the transceiving of electromagnetic energy. An antenna array is characterized by the geometry and the type of the elements. A major role in the antenna array is played by the mutual coupling between the elements and their input impedance. For simplicity reasons in both the fabrication and the synthesis, the elements are chosen, if it is possible, to be identical and parallel. For the same reasons uniformly spaced linear arrays are mostly encountered in practice.

In the following paragraphs the properties of various antenna arrays will be presented and the synthesis method will be discussed.

2. ANTENNA ARRAY FACTOR

The radiation characteristics of antennas have to do with the far-field region. In this region the field is separated in two parts: one containing the distance r of the observation point (receiver location) and the other, its spherical coordinate angles θ and φ. The far electric field of a typical antenna element (see Fig. 1) can be expressed as

$$\boldsymbol{E}_n(r) \cong -j\omega\mu \frac{e^{-j\beta r}}{4\pi r} \boldsymbol{f}_n(\theta, \varphi) \tag{1}$$

The angular — dependent vector $\boldsymbol{f}_n(\theta, \varphi)$ gives the directional characteristics of the nth-element electric field [11]:

$$\boldsymbol{f}_n(\theta, \varphi) = (\hat{\theta}\hat{\theta} + \hat{\varphi}\hat{\varphi}) \cdot \int_{\text{element}} \boldsymbol{J}_n(\mathbf{r'_n}) e^{j\beta \hat{r} \cdot (\mathbf{r_n} - \mathbf{r'_n})} dv' \tag{2}$$

where $\boldsymbol{J}_n(r'_n)$ = electric current density of the nth element
r'_n = distance of a source point from the origin

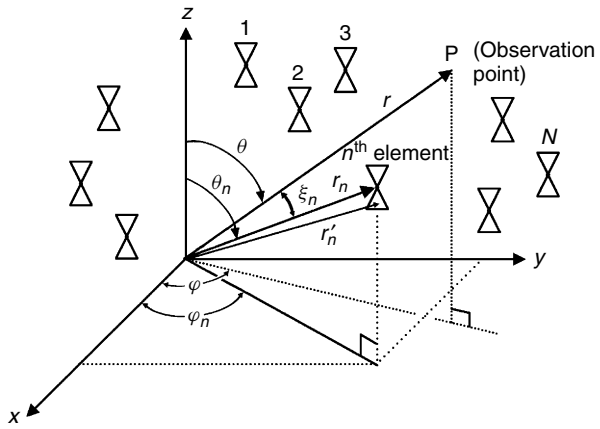


Figure 1. Geometry of a general antenna array.

r = distance of the observation point from the origin
$\beta = \dfrac{2\pi}{\lambda}$, the free-space wavenumber
ω = angular frequency
μ = magnetic permeability of the space

The total electric field of the N-element antenna array is

$$\boldsymbol{E}(r) = \sum_{n=1}^{N} \boldsymbol{E}_n(r) \tag{3}$$

Also the total magnetic field is [6]

$$\boldsymbol{H}(r) = \frac{1}{\eta}\hat{\boldsymbol{r}} \times \boldsymbol{E}(r) \tag{4}$$

where $\eta = \sqrt{\mu/\varepsilon}$ (ε is the electric permeability).

For identical — and, if possible, identically oriented — elements, the current distribution of these elements is approximately the same except for a constant complex multiplier. In Eq. (1), $\boldsymbol{f}_n(\theta, \varphi)$ can be expressed as

$$\boldsymbol{f}_n(\theta, \varphi) = I_n \boldsymbol{f}(\theta, \varphi) \tag{5}$$

$\boldsymbol{f}(\theta, \varphi)$ is called the pattern function of the element and I_n is the complex excitation of the nth element of the array.

Combining Eqs. (1), (2), and (5), we have

$$\boldsymbol{E}(r) = -j\omega\mu \frac{e^{-j\beta r}}{4\pi r} \boldsymbol{f}(\theta, \varphi) \sum_{n=1}^{N} I_n e^{j\beta r_n \cos \xi_n} \tag{6}$$

where $(r_n,\ \theta_n,\ \varphi_n)$ are the spherical coordinates of a convenient reference point of the nth element and $\cos \xi_n = \sin\theta \sin\theta_n \cos(\varphi - \varphi_n) + \cos\theta \cos\theta_n$. The last term of (6) is expressed separately as

$$\mathrm{AF}(\theta, \varphi) = \sum_{n=1}^{N} I_n e^{j\beta r_n \cos \xi_n} \tag{7}$$

$\mathrm{AF}(\theta, \varphi)$ is the array factor. This factor is actually the array pattern of N isotropic point sources positioned at the reference points of the elements of the original array.

From Eqs. (6) and (7) we know that

$$\boldsymbol{E}(r) = -j\omega\mu \frac{e^{-j\beta r}}{4\pi r} \boldsymbol{f}(\theta, \varphi)\mathrm{AF}(\theta, \varphi) \tag{8}$$

This expression states the following pattern multiplication principle: "An array consisting of identically oriented similar elements has a pattern which can be expressed as the product of the element pattern and the array factor."

An antenna engineer must select the element according to the technical requirements. Since the element pattern is known, the design effort is mainly directed to the array factor.

3. ELEMENTS AND ARRAY TYPES

Element types of antenna arrays can be found in the literature [1–14]. Monopoles, dipoles, loops, slots, microstrip patches, and horns are the most common types. More recent studies and innovations have resulted in new types of elements, such as the monolithic, the superconducting, the active, and the electronically and functionally small elements.

In parallel with the development of elements, antenna arrays have experienced a tremendous growth. Their list starts from the linear broadside and endfire arrays, the planar, the circular, and the conformal and goes up to the adaptive arrays. Some of the more recent types are flat-plate slot arrays, digital beamforming, dichroic, slotted, and fractal arrays.

As mentioned above, antenna analysis and synthesis focus mostly on the array factor. Consequently, in the following paragraphs we devote our analysis mainly to this factor.

4. ANTENNA CHARACTERISTICS AND INDICES

One of the main characteristics of an antenna is its radiation pattern. This characteristic graphically presents the radiation properties and can be measured by moving a probe antenna around the antenna under test at a constant distance in the far field (see Fig. 2a). The response as a function of the angular coordinates constitutes the radiation pattern. Depending on the probe type and orientation, the appropriate component of the electric or the magnetic field can be measured. If the probe is moved over the spherical surface, its terminal voltage will present the 3D radiation pattern. A pattern taken in one plane is known as the *plane pattern*. The pattern that contains the electric field vector is the E-plane pattern, while the pattern that contains the magnetic field vector is the H plane. The above two are referred as the *principal plane patterns*. As an example, Fig. 2b presents the 3D radiation pattern of an electric dipole while Figs. 2c and 2d show the E- and H-plane pattern.

Polarization of an antenna is the polarization of the wave transmitted by the antenna. Polarization has to do with a certain direction. If the direction is not stated, then it is assumed that it corresponds in the direction of maximum. The polarization is characterized by the curve traced by the endpoint of the arrow representing the instantaneous electric field. The field is observed in the direction of propagation. Polarization is classified as linear, circular, or elliptical.

For the sake of convenience, some of the antenna indices will be defined as follows:

1. The directive gain $D(\theta_0,\ \varphi_0)$ is a dimensionless quantity that is defined by

$$D(\theta_0, \varphi_0) = \frac{\text{radiation intensity for the direction } (\theta_0, \varphi_0)}{\dfrac{1}{4\pi}(\text{radiation power of the antenna})} \tag{9}$$

Radiation intensity is the power radiated in a given direction per unit solid angle. The maximum $D(\theta_0, \varphi_0)$ is the directivity D of the antenna.

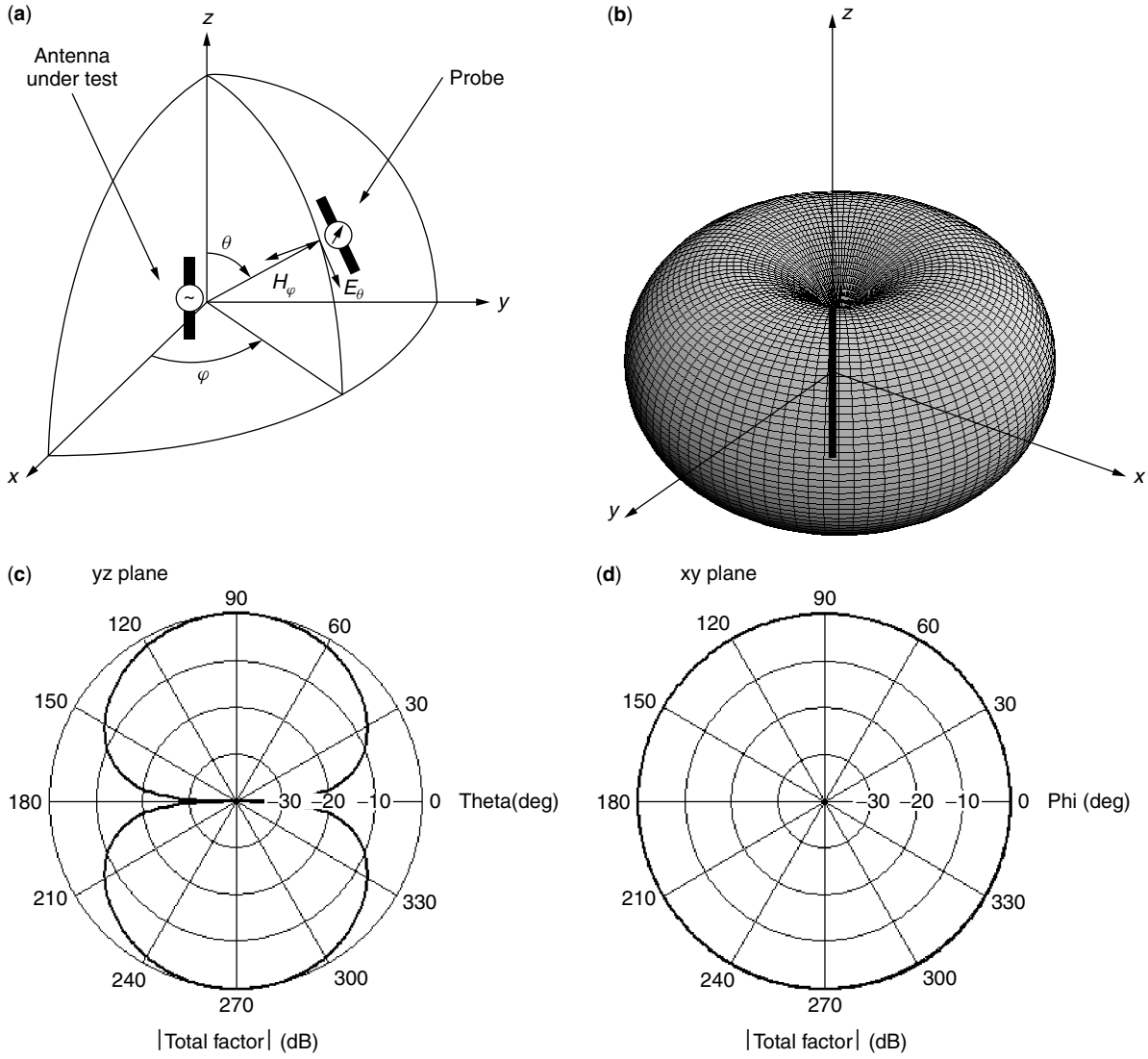


Figure 2. Radiation pattern of an electric $\lambda/2$ dipole: (**a**) pattern measurement scheme; (**b**) three-dimensional plot; (**c**) E-plane radiation pattern; (**d**) H-plane radiation pattern.

2. The power gain $G(\theta_0, \varphi_0)$ is defined by

$$G(\theta_0, \varphi_0) = \frac{\text{radiation intensity for the direction } (\theta_0, \varphi_0)}{\frac{1}{4\pi}(\text{power input to the antenna})} \quad (10)$$

3. The signal-to-noise ratio SNR applies for the receiving antennas and is defined by

$$\text{SNR} = \frac{\text{received power of the desired signal}}{\frac{1}{4\pi}(\text{noise plus interference power})} \quad (11)$$

4. The radiation efficiency η of an N-element array antenna is defined by

$$\eta = \frac{\text{radiation intensity to the direction of maximum}}{N\,(\text{sum of the excitation current magnitude squared})} \quad (12)$$

5. The quality factor Q of an N-element array antenna is defined by

$$Q = \frac{\text{sum of the excitation current magnitude squared}}{\frac{1}{4\pi}\,(\text{power input to the array})} \quad (13)$$

Combining (13) and (12), we find that

$$G_{\max} = N\eta Q \quad (14)$$

6. The half-power (or 3-dB power) beamwidth (HPBW) is the angular width between the angular points half-power (3 dB) below that of the main beam maximum of the antenna.
7. The first null beamwidth (BW_{null}) is defined as the angular width between the first zero crossing of either side of the main-beam maximum of the antenna.

8. The bandwidth of an antenna is defined by the frequency limits at which the maximum gain is reduced to half-power (3 dB). The fractional bandwidth is given by $\Delta f/f$.
9. The sidelobe level (SLL) is the ratio of the pattern value of a sidelobe peak to the corresponding value of the mainlobe. Usually SLL in an antenna is defined as the largest sidelobe level for the whole pattern.

5. LINEAR ARRAYS

One method to obtain directive antennas is to use several individual antennas that add their contributions in preferred directions and cancel in others. This arrangement is known as an *array*, and the individual antennas are called *elements*. The most practical array that consists of a number of elements set up along a straight line is the linear array.

Consider a typical linear array placed along the z axis as shown in Fig. 3. The array factor $\mathrm{AF}(\theta, \varphi)$ depends only on the angle θ and is written as

$$\mathrm{AF}(\theta) = \sum_{n=0}^{N} I_n e^{j\beta d_n \cos\theta} \tag{15}$$

If the elements are placed in the same interelement distance d, then Eq. (15) yields

$$\mathrm{AF}(\theta) = \sum_{n=0}^{N} I_n e^{j\beta n d \cos\theta} = \sum_{n=0}^{N} I_n z^n \tag{16}$$

where

$$z = e^{j\beta d \cos\theta} \tag{17}$$

For $0 \le \theta \le \pi$, AF (θ) is a polynomial of z that moves on a unit circle with a phase bounded between $-\beta d$ and $+\beta d$. The bounded region is called the *visible region*. The unit circle approach, proposed by Schelkunoff [1] visually indicates how the element contributions combine.

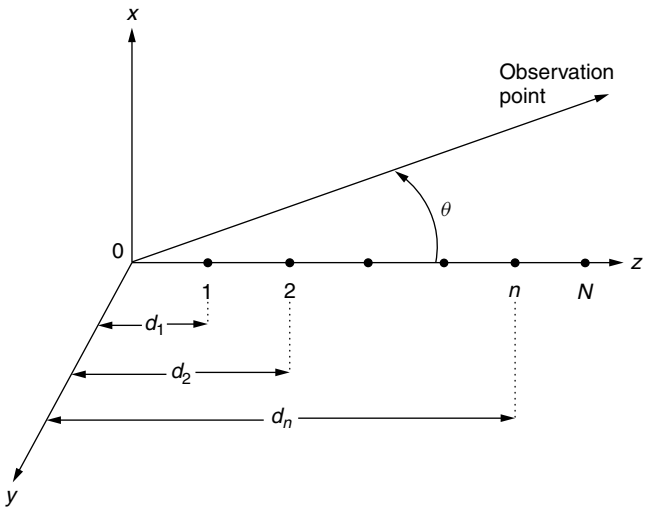


Figure 3. A linear N-element array.

A linear array that cancels noise from N different directions has a pattern of the following form:

$$\mathrm{AF}(\theta) = C\prod_{n=1}^{N}(z - z_n) = C\sum_{k=0}^{N} I_k z^k \tag{18}$$

C is the normalization factor such that $|\mathrm{AF}(\theta)|_{\max} = 1; z_n = e^{j\beta d \cos\theta_n}$, where θ_n is the direction of the nth interference, and I_k is the required excitation of the kth element.

The type and orientation of the elements of the array are selected for receiving the maximum of the desired signal. By varying z_n, it is possible to steer the nulls. This will alter the element excitations. Also some roots can be outside the visible region or the unit circle.

A linear array with all roots equal is known as a *binomial array*. $\mathrm{AF}(\theta)$ is of the form

$$\mathrm{AF}(\theta) = C(z - z_1)^N = C\sum_{k=0}^{N}\binom{N}{k}(-z_1)^k z^{N-k} \tag{19}$$

where

$$\binom{N}{k} = \frac{N!}{(N-k)!k!} \tag{20}$$

The binomial array has low minor lobes. However a wide and difficult to be realized variation between the amplitudes of the elements is present. This variation increases as the number of elements increases.

5.1. Uniform Arrays

Linear arrays with equally spaced elements, identical magnitude and progressive phase, are referred to as *uniform arrays*. For N elements we have $I_k = (e^{i\alpha})^k$ and Eq. (18) becomes

$$\mathrm{AF}(\theta) = C\sum_{n=0}^{N-1}(ze^{j\alpha})^n = C\frac{(ze^{j\alpha})^N - 1}{(ze^{j\alpha}) - 1} \tag{21}$$

According to [6], Eq. (21) may be transformed to

$$\mathrm{AF}(\theta) = e^{j(N-1)\psi/2}\frac{\sin(N\psi/2)}{N\sin(\psi/2)} \tag{22}$$

where

$$\psi = \beta d\cos\theta + \alpha \tag{23}$$

$|\mathrm{AF}(\theta)|$ as a function of ψ, known also as $|F(\psi)|$, is presented for a few values of N in Fig. 4.

The main characteristics of $F(\psi)$ are the following:

1. Maximum values occur at $\psi = \pm 2k\pi$, where $k = 0, 1, 2, \ldots$.
2. Nulls of the array are at $\psi = \pm 2k\pi/N$ where $k = 1, 2, 3, \ldots$ and $k \ne N, 2N, 3N, \ldots$.
3. The 3-dB points of the array factor are at ψ, which gives

$$F(\psi) = \pm\frac{\sqrt{2}}{2} \Rightarrow |F(\psi)|^2 = \frac{1}{2}$$
$$\Rightarrow 10\log|F(\psi)|^2 = 3 \text{ dB}$$

Therefore, $N\psi/2 = \pm 1.391$ rad.

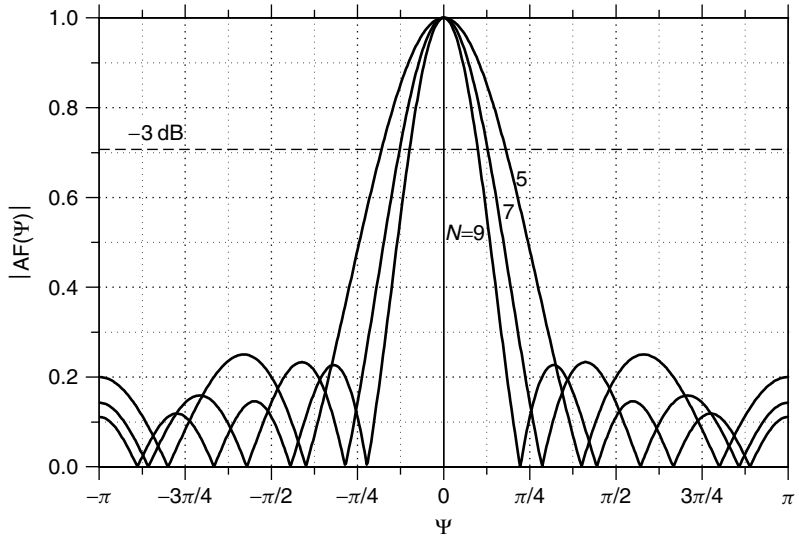


Figure 4. The magnitude of the array factor as a function of ψ.

4. Secondary maxima (minor lobes) occur when $\psi = (2k+1)\pi/N$, where $k = 1, 2, 3, \ldots$.

It is noticed that the first minor lobe is at $\psi = \pm 3\pi/N$, which gives

$$|F(\psi)| = \frac{1}{N \sin \dfrac{3\pi}{2N}} = \text{SLL}$$

When the beam maximum appears at $\theta = \pi/2$, the array is called *broadside*. *Endfire* is an array with the beam maximum at $\theta = 0$ or π. The array beam can be steered at any direction θ_0 if the phase shift is $\alpha = -\beta d \cos\theta_0$. For the endfire array it is $\alpha = \mp\beta d$. In addition to the ordinary endfire, there is the Hansen–Woodyard (HW) endfire array, which is more directive [Eq. (6)]. In HW the progressive phase shift is

$$\alpha = \mp\left(\beta d + \frac{2.94}{N}\right) \cong \mp\left(\beta d + \frac{\pi}{N}\right) \tag{24}$$

Also

$$|\psi| = \pi \begin{cases} \text{for } \theta = \pi \text{ if maximum occurs at } \theta = 0 \\ \text{for } \theta = 0 \text{ if maximum occurs at } \theta = \pi \end{cases} \tag{25}$$

Condition (25) gives

$$d = \frac{\lambda}{4}\left(1 - \frac{2.94}{\pi N}\right) \cong \frac{\lambda}{4}\left(1 - \frac{1}{N}\right) \tag{26}$$

HW is useful for very long arrays with small interelement distance. Useful formulas for the prescribed linear arrays are given in Table 1.

5.2. Chebyshev Arrays

5.2.1. Chebyshev Polynomials. Chebyshev arrays are uniformly spaced linear arrays with nonuniform excitation. They make use of the Chebyshev polynomials [21].

A Chebyshev polynomial $T_m(x)$ of an independent variable x is an orthogonal one of mth order. It contains equal ripples in the region $-1 \le x \le 1$ and the amplitude varies between $+1$ and -1. The polynomial outside this region rises exponentially:

$$T_m(x) = \begin{cases} \cos(m \cos^{-1} x) & |x| \le 1 \\ \left(\dfrac{x}{|x|}\right)^m \cosh(m \cosh^{-1}|x|) & |x| > 1 \end{cases} \tag{27}$$

and

$$\left.\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \end{aligned}\right\} \tag{28}$$

Equation (28) and the recursion relation

$$T_m(x) = 2xT_{m-1}(x) - T_{m-2}(x) \tag{29}$$

create the Chebyshev polynomials of any order.

5.2.2. Dolph–Chebyshev Arrays. Dolph [22] recognized that Chebyshev polynomials could be used to have arrays with maximum directivity for a given sidelobe level. The equal ripples of the polynomials present the sidelobes, while the main beam comes from the exponential increase beyond $|x| = 1$.

The linear array is fed symmetrically about the centerline (see Fig. 5). The array factor can be expressed in terms of $\cos(\psi/2)$, where $\psi = \beta d \cos\theta + \alpha$. The independent variable of the Chebyshev polynomial is

$$x = x_0 \cos\left(\frac{\psi}{2}\right) \tag{30}$$

At $x = x_0$ the polynomial has its maximum value R:

$$T_m(x_0) = R \text{ or } x_0 = \cosh\left(\frac{1}{m}\cosh^{-1} R\right) \tag{31}$$

The zeros of $T_m(x)$ are at

$$x_k = \pm \cos\frac{(2k-1)\pi}{2m} \qquad k = 1, 2, \ldots, m \tag{32}$$

Table 1. Useful Formulas for Uniform Linear Arrays

	Broadside	Endfire	Hansen–Woodyard Endfire	Intermediate with Maximum at $\theta = \theta_0$
Directivity	$\sim 2N\mathrm{d}/\lambda$	$\sim 4N\mathrm{d}/\lambda$	$\sim 1.789\,(4N\mathrm{d}/\lambda)$	Depending on θ_0
HPBW	$\pi - 2\cos^{-1}\left(\dfrac{2.782}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{2.782}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{0.2796\pi}{N\beta d}\right)$	$\left\lvert\cos^{-1}\left(\cos\theta_0 + \dfrac{2.782}{N\beta d}\right) - \cos^{-1}\left(\cos\theta_0 - \dfrac{2.782}{N\beta d}\right)\right\rvert$
Beamwidth between nulls	$\pi - 2\cos^{-1}\left(\dfrac{2\pi}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{2\pi}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{\pi}{N\beta d}\right)$	$\left\lvert\cos^{-1}\left(\cos\theta_0 + \dfrac{2\pi}{N\beta d}\right) - \cos^{-1}\left(\cos\theta_0 - \dfrac{2\pi}{N\beta d}\right)\right\rvert$
Null angular positions	$\cos^{-1}\left(\pm\dfrac{2k\pi}{N\beta d}\right)$	$\cos^{-1}\left(1 - \dfrac{2k\pi}{N\beta d}\right)$	$\cos^{-1}\left[1 + (1-2k)\dfrac{\pi}{N\beta d}\right]$ $k = 1, 2, 3, \ldots, \quad k \neq N, 2N, 3N, \ldots$	$\cos^{-1}\left(\cos\theta_0 \pm \dfrac{2k\pi}{N\beta d}\right)$
Sidelobe maximum positions	$\cos^{-1}\left(\pm\dfrac{(2k+1)\pi}{N\beta d}\right)$	$\cos^{-1}\left(1 - \dfrac{(2k+1)\pi}{N\beta d}\right)$	$\cos^{-1}\left(1 - \dfrac{2k\pi}{N\beta d}\right)$ $k = 1, 2, 3, \ldots, \quad k \neq N, 2N, 3N, \ldots$	$\cos^{-1}\left(\cos\theta_0 \pm \dfrac{(2\ k+1)\pi}{N\beta d}\right)$

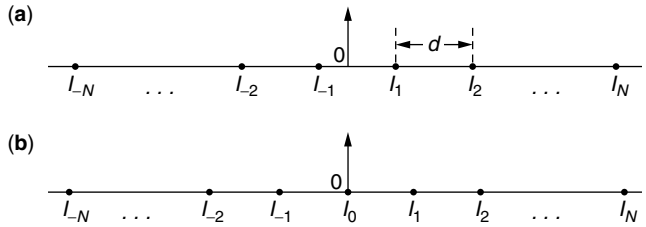


Figure 5. Linear array with (**a**) even and (**b**) odd number of elements, uniformly spaced and symmetrically excited ($I_k = I_{-k}$).

which correspond to

$$\psi_k = \pm 2 \cos\left(\frac{x_k}{x_0}\right) \tag{33}$$

The excitations I_k are found from (18) by using $z_k = e^{j\psi_k}$. The order of the Chebyshev polynomial is one less than the total number of elements of the array.

An example of a nine-element array with SLL $= -25$ dB is given. The polynomial is the $T_8(x)$. For $R = 25$ dB it is $R = 10^{25/20} = 17.7828$ and from (31), $x_0 = 1.1013$. A broadside array has the excitation coefficients presented in Table 2.

An intermediate array with maximum at $\theta = \theta_0$ has the same amplitude as before with a phase shift $\alpha = -\beta d \cos\theta_0$. Figure 6 shows the patterns for $d/\lambda = 0.5$ of a broadside and an intermediate array with $\theta_0 = \pi/3$.

5.2.3. Riblet Arrays. Dolph–Chebyshev arrays are suitable for $d \geq \lambda/2$ and fail to give the optimum design for $d < \lambda/2$. Riblet [23] devised a method to overcome the problem. He used $(2m + 1)$ elements and an independent variable of the form

$$x = a \cos\psi + b \tag{34}$$

The polynomial is $T_m(x)$ with a visible region $-1 \leq x \leq x_0$:

$$\left.\begin{aligned} x_0 &= a + b \\ -1 &= a \cos\beta d + b \end{aligned}\right\} \tag{35}$$

Solving (35), we obtain

$$a = \frac{1 + x_0}{1 - \cos\beta d} \quad \text{and} \quad b = -\frac{1 + x_0 \cos\beta d}{1 - \cos\beta d} \tag{36}$$

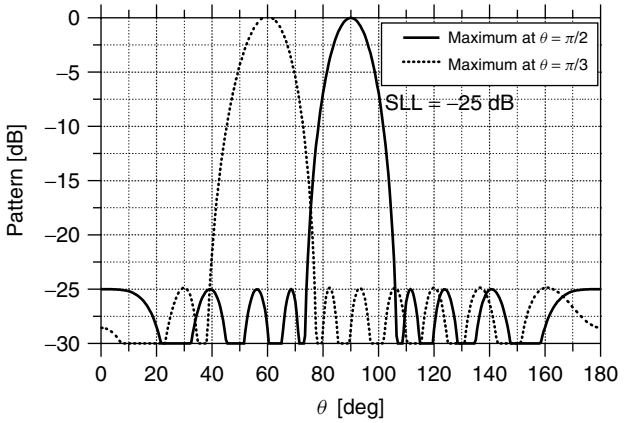


Figure 6. Radiation patterns of a Dolph–Chebyshev array with nine elements in $d/\lambda = 0.5$ with max. at $\theta = \pi/2$ and $\theta = \pi/3$.

Zeros of $T_m(x)$ are given by (32) and the corresponding ψ_k are

$$\psi_k = \pm \cos^{-1}\left(\frac{x_k - b}{a}\right) \tag{37}$$

A nine-element array with $d/\lambda = 0.4$ and SLL $= -20$ dB is presented. The excitation is given in Table 3 and the pattern, in Fig. 7.

5.2.4. Other Chebyshev Arrays. Equal-sidelobe designs with the help of Chebyshev polynomials can be made by

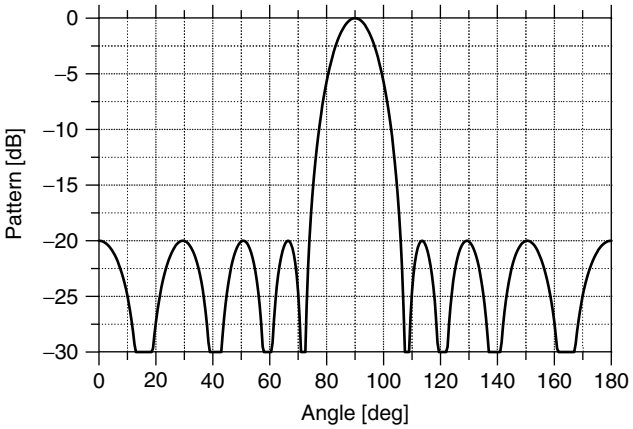


Figure 7. Pattern of a nine-element Riblet array with $d/\lambda = 0.4$ and SLL $= -20$ dB.

Table 2. Normalized Excitation Coefficients for a Nine-Element Dolph–Chebyshev Array

Element number	1, 9	2, 8	3, 7	4, 6	5
Excitation coefficient	3.783×10^{-1}	5.310×10^{-1}	7.639×10^{-1}	9.363×10^{-1}	1

Table 3. Excitation Coefficient of Nine-Element Riblet Array with $d/\lambda = 0.4$

Element number	1, 9	2, 8	3, 7	4, 6	5
Excitation coefficient	5.519×10^{-1}	8.913×10^{-2}	8.393×10^{-1}	1.566×10^{-1}	1

extending the Riblet technique, where x is written as

$$x = a\cos(\beta d\cos\theta + \alpha) + b \tag{38}$$

To find a, b, and α, three characteristic angles θ_1, θ_2 and θ_3, which correspond to x_1, x_2, and x_3, respectively, must be defined.

$$\left.\begin{aligned} x_1 &= a\cos(\beta d\cos\theta_1 + \alpha) + b \\ x_2 &= a\cos(\beta d\cos\theta_2 + \alpha) + b \\ x_3 &= a\cos(\beta d\cos\theta_3 + \alpha) + b \end{aligned}\right\} \tag{39}$$

Solving this equation, one can find

$$\tan\alpha = \frac{\sin y_{21} - \lambda\sin y_{31}}{\lambda\cos y_{31} - \cos y_{21}} \tag{40}$$

where

$$\left.\begin{aligned} y_{21} &= \beta\frac{d}{2}(\cos\theta_2 + \cos\theta_1) \\ y_{31} &= \beta\frac{d}{2}(\cos\theta_3 + \cos\theta_1) \\ \lambda &= \frac{(x_2 - x_1)\sin\left[\dfrac{\beta d}{2}(\cos\theta_3 - \cos\theta_1)\right]}{(x_3 - x_1)\sin\left[\dfrac{\beta d}{2}(\cos\theta_2 - \cos\theta_1)\right]} \end{aligned}\right\} \tag{41}$$

$$a = \frac{x_2 - x_1}{\cos(\beta d\cos\theta_2 + \alpha) - \cos(\beta d\cos\theta_1 + \alpha)} \tag{42}$$

$$b = x_1 - a\cos(\beta d\cos\theta_1 + \alpha) \tag{43}$$

A general approach not only for Chebyshev but also for Legendre arrays can be found in Ref. 24. Table 4 and Figs. 8–11 present four common endfire cases expressed in Eqs. (39)–(43).

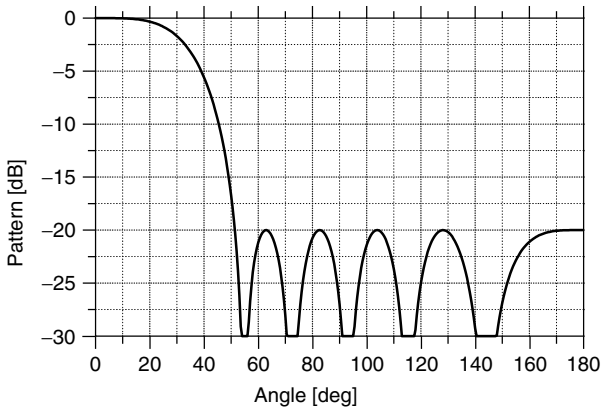


Figure 8. Pattern of the endfire case 1 array for $N = 11$, $d/\lambda = 0.25$, SLL $= -20$ dB.

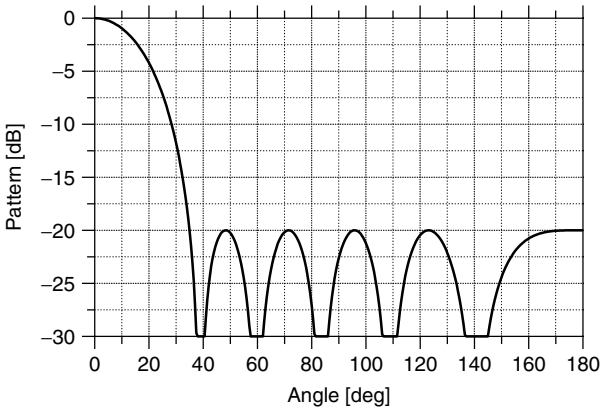


Figure 9. Pattern of the endfire case 2 array for $N = 11$, $d/\lambda = 0.2$, SLL $= -20$ dB.

6. PLANAR ARRAYS

Individual radiators positioned on a plane constitute a planar array. The usual planar array is rectangular. The elements are positioned along a rectangular grid. The grid can be constructed from the combination of two perpendicular linear arrays. Because of the increase in variables, the pattern can be controlled and scanned at any point in space. The more symmetrical patterns and the lower

Table 4. Coefficients of Chebyshev Endfire Arrays

Endfire Case 1	Endfire Case 2	Endfire Case 3 (Optimum)	Endfire Case 4
$x_1 \to \theta_1 = 0$	$x_1 \to \theta_1 = 0$	$x_1 \to \theta_1 = 0$	$x_1 \to \theta_1 = 0$
$x_2 \to \theta_2 = \pi$	$x_2 \to \theta_2 = \pi$	$x_2 \to \theta_2 = \pi$	$-x_2 \to \theta_2 = \pi$
x_3 #	x_3 #	$x_3 = -(a + b)$	$x_3 \to \theta_3 = \theta_{HP}$
$\mathrm{T}_m(x_1) = R$	$\mathrm{T}_m(x_1) = R$	$\mathrm{T}_m(x_1) = R$	$\mathrm{T}_m(x_1) = R, \mathrm{T}_m(x_3) = R\sqrt{2}/2$
$\alpha = -\beta d$	$\alpha = \beta d$	$\alpha = 2\tan^{-1}\left[\sin(\beta d)\dfrac{x_1 + x_2 + 2x_3 - 2\sqrt{(x_1 + x_3)(x_2 + x_3)}}{(x_1 - x_2)[1 + \cos(\beta d)]}\right]$	$\alpha = \cot^{-1}\left[\dfrac{(2\lambda + 1)\sin\beta d - \sin(\beta d\cos\theta_3)}{\cos\beta d - \cos(\beta d\cos\theta_3)}\right]$
			$\lambda = \dfrac{x_3 - x_1}{x_1 + x_2}$
$a = \dfrac{x_1 + x_2}{1 - \cos 2\beta d}$	$a = \dfrac{x_1 - x_2}{\cos 2\beta d - 1}$	$a = \dfrac{x_2 - x_1}{2\sin\beta d\cdot\sin\alpha}$	$a = \dfrac{x_1 + x_2}{2\sin\beta d\cdot\sin\alpha}$
$b = x_1 - a$	$b = x_2 - a$	$b = -x_3 - a$	$b = x_1 - \cos(\beta d + \alpha)$
$d_{\max} = \lambda/4$	$d_{\max} = \lambda/4$	$d_{\max} = \dfrac{\lambda}{2} - \dfrac{\lambda}{2\pi}\sin^{-1}\left(\sqrt{\dfrac{x_1 + x_3}{x_1 + 1}}\right)$	$d_{\max} = \dfrac{\lambda}{2\pi} \times \left[\alpha - \sin^{-1}\left(\sqrt{\dfrac{x_1 + x_2}{x_1 + 1}}\right)\right]$

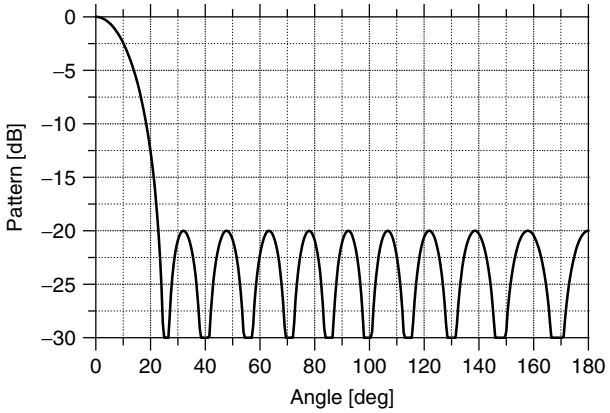


Figure 10. Pattern of the endfire case 3 array for $N = 11$, $d/\lambda = 0.35$, SLL $= -20$ dB.

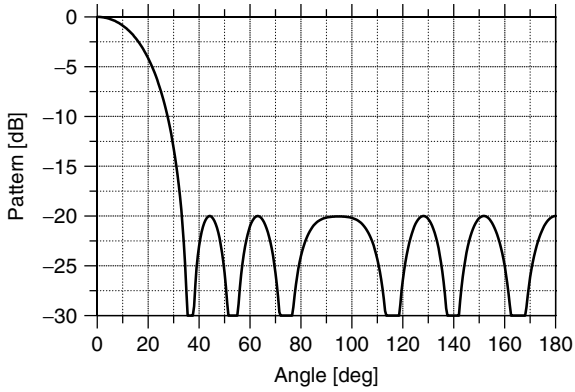


Figure 11. Pattern of the endfire case 4 array for $N = 11$, $d/\lambda = 0.4$, HPBW $= 35°$, SLL $= -20$ dB.

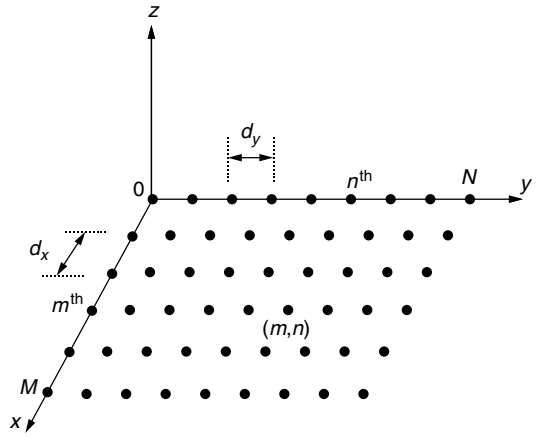


Figure 12. Planar array geometry.

sidelobes are the two main advantages of the planar over the linear arrays. Let us refer to Fig. 12.

The element corresponding to the mth row and the nth column is the mnth element with excitation I_{mn}. The array factor is

$$\mathrm{AF}(\theta, \varphi) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{mn} z_x^m z_y^n \tag{44}$$

where

$$z_x = e^{j\beta d_x \sin\theta \cos\varphi} \text{ and } z_y = e^{j\beta d_y \sin\theta \sin\varphi}$$

For uniform excitation and progressive phase α_x along x axis and α_y along y axis, the array factor becomes

$$\mathrm{AF}(\theta, \varphi) = I_0 \left(\sum_{m=1}^{M} e^{j(m-1)(\beta d_x \sin\theta \cos\varphi + \alpha_x)} \right) \times \left(\sum_{n=1}^{N} e^{j(n-1)(\beta d_y \sin\theta \sin\varphi + \alpha_y)} \right) \tag{45}$$

According to Eqs. (22) and (23), expression (45) gives

$$|\mathrm{AF}(\theta, \varphi)| = \left| \frac{\sin(M\Psi_x/2)}{M \sin(\Psi_x/2)} \right| \left| \frac{\sin(N\Psi_y/2)}{N \sin(\Psi_y/2)} \right| \tag{46}$$

where

$$\left.\begin{aligned} \Psi_x &= \beta d_x \sin\theta \cos\varphi + \alpha_x \\ \Psi_y &= \beta d_y \sin\theta \sin\varphi + \alpha_y \end{aligned}\right\} \tag{47}$$

For d_x and/or $d_y \geq \lambda$, the in-phase addition of the radiated field is made in more than one direction and grating lobes are produced. The grating lobes are located at

$$\left.\begin{aligned} \Psi_x &= \pm 2m\pi \\ \Psi_y &= \pm 2n\pi \quad m \text{ and } n = 1, 2, \ldots \end{aligned}\right\} \tag{48}$$

If the mainlobe direction is at (θ_0, φ_0), then

$$\left.\begin{aligned} \alpha_x &= -\beta d_x \sin\theta_0 \cos\varphi_0 \\ \alpha_y &= -\beta d_y \sin\theta_0 \sin\varphi_0 \end{aligned}\right\} \tag{49}$$

Solving (48) for the direction $(\theta_{mn}, \varphi_{mn})$ where grating lobes occur, we obtain

$$\varphi_{mn} = \tan^{-1}\left[\frac{\sin\theta_0 \sin\varphi_0 \pm n\lambda/d_y}{\sin\theta_0 \cos\varphi_0 \pm m\lambda/d_x} \right] \tag{50}$$

$$\theta_{mn} = \sin^{-1}\left[\frac{\sin\theta_0 \sin\varphi_0 \pm n\lambda/d_y}{\sin\varphi_{mn}} \right] = \sin^{-1}\left[\frac{\sin\theta_0 \cos\varphi_0 \pm m\lambda/d_x}{\cos\varphi_{mn}} \right] \tag{51}$$

A 6×6-element array with $\alpha_x = \alpha_y = 0$ will be given. Figures 13 and 14 show the corresponding patterns for $d_x = d_y = \lambda/2$ and $d_x = d_y = \lambda$. It is obvious that for the large spacing there are grating lobes at $\theta = \pi/2$ and $\varphi = 0$, $\pi/2$, π, $3\pi/2$.

Finally, the pattern of a 6×6 rectangular array that combines two different Dolph–Chebyshev arrays with SLL $= -20$ dB for $d_x = 0.5\lambda$ and $d_y = 0.82\lambda$ is shown in Fig. 15.

7. CIRCULAR ARRAYS

A *circular array* is a planar array with the elements positioned on a circular ring. A circular array with N isotropic elements (see Fig. 16) produces the array factor:

$$\mathrm{AF}(\theta, \varphi) = \sum_{n=1}^{N} I_n e^{j[\beta R \sin\theta \cos(\varphi - \varphi_n) + \alpha_n]} \tag{52}$$

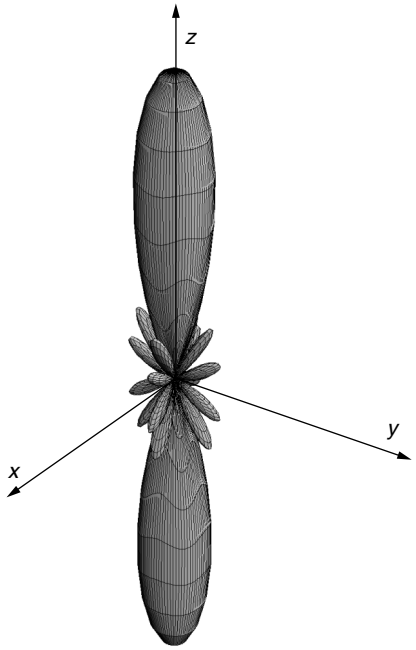


Figure 13. Three-dimensional pattern of a 6×6 uniform planar array with $a_x = a_y = 0$ and $d_x = d_y = 0.5\lambda$.

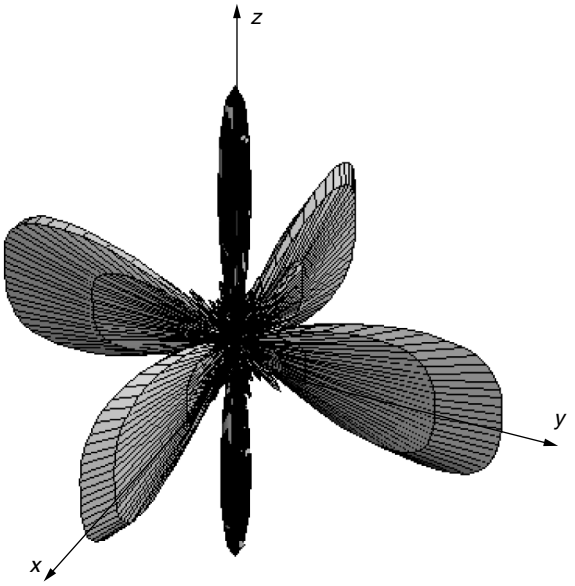


Figure 14. Three-dimensional pattern of a 6×6 uniform planar array with $a_x = a_y = 0$ and $d_x = d_y = 1\lambda$.

where I_n is the amplitude of the excitation of the nth element and α_n is the corresponding phase. To have a peak at (θ_0, φ_0), it must be

$$\alpha_n = -\beta R \sin\theta_0 \cos(\varphi_0 - \varphi_n) \tag{53}$$

and

$$\mathrm{AF}(\theta, \varphi) = \sum_{n=1}^{N} I_n e^{j\beta R[\sin\theta\cos(\varphi-\varphi_n) - \sin\theta_0\cos(\varphi_0-\varphi_n)]} \tag{54}$$

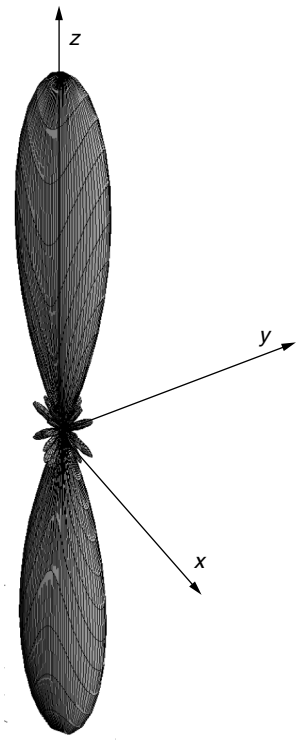


Figure 15. Three-dimensional pattern of a 6×6 Chebyshev planar array with SLL $= -20$ dB, $a_x = a_y = 0$, $d_x = 0.5\lambda$, and $d_y = 0.82\lambda$.

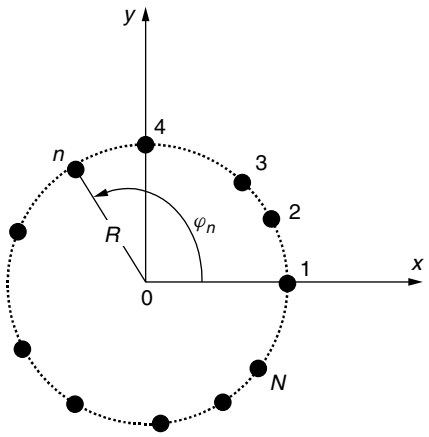


Figure 16. Geometry of a circular array with N elements.

The pattern of a 12-element uniform circular array with $\beta R = 10$, $\theta_0 = 0$, $\varphi_0 = 0$ is shown in Fig. 17.

An interesting array comes from a dipole positioned at the bisector of a corner reflector with angle $\omega_N = 2\pi/\mathrm{N}$. The reflector creates a circular array with the $N - 1$ images (see Fig. 18). Figure 19 presents the corresponding H pattern. It is noticed that the pattern outside the reflectors angle does not exist.

M circular arrays in concentric rings produce an array factor

$$\mathrm{AF}(\theta, \varphi) = \sum_{m=1}^{M} \sum_{n=1}^{N} I_{mn} e^{j[\beta R_m \sin\theta\cos(\varphi-\varphi_{mn}) + \alpha_{mn}]} \tag{55}$$

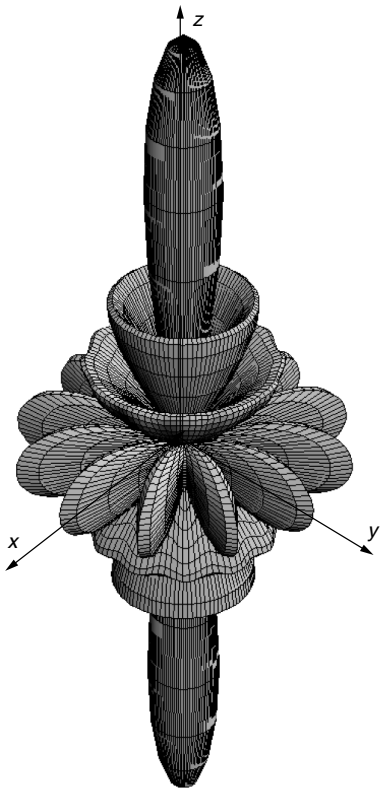


Figure 17. Three-dimensional pattern of a circular array with $N = 12$, $\beta R = 10$, $\theta_0 = 0^\circ$, and $\varphi_0 = 0^\circ$.

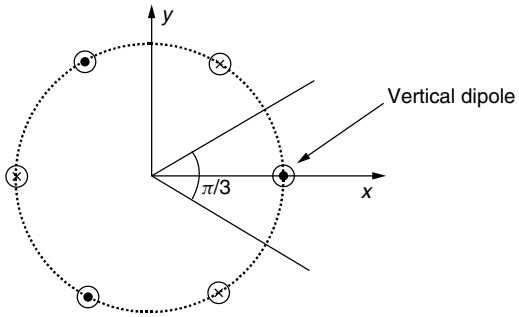


Figure 18. Vertical dipole in front of a corner reflector with its images.

where $I_{mn}e^{j\alpha_{mn}}$ is the excitation of the nth element of the mth ring.

A corner reflector with a linear array positioned in front of it (see Fig. 20) creates concentric rings. For a 2-dipole uniform linear array the pattern is as presented in Fig. 21.

8. 3D ARRAYS

N elements positioned in 3D space constitute a three-dimensional array. The array factor is

$$\mathrm{AF}(\theta,\varphi) = \sum_{n=1}^{N} I_n e^{j[\alpha_n + \beta r_n[\sin\theta\sin\theta_n\cos(\varphi-\varphi_n)+\cos\theta\cos\theta_n]]} \qquad (56)$$

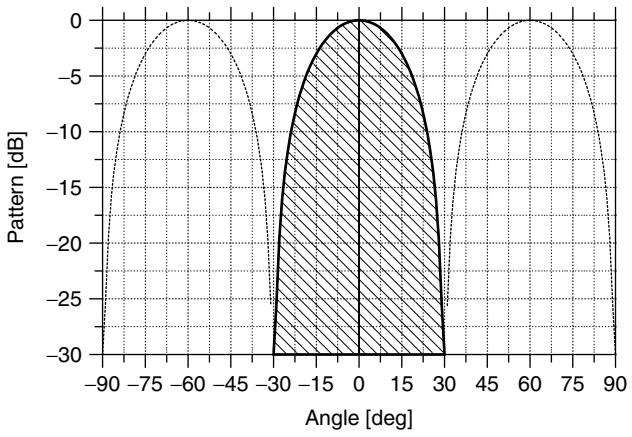


Figure 19. H-Pattern of a dipole in front of a $\pi/3$ corner reflector.

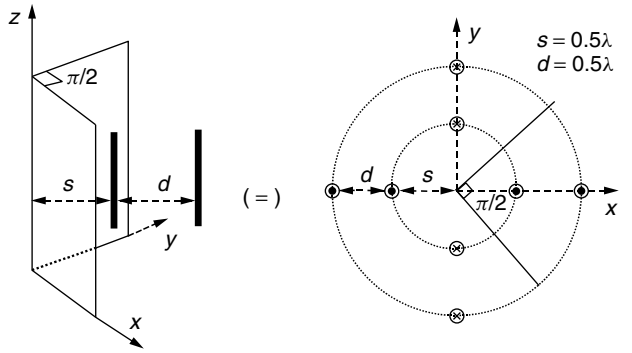


Figure 20. A parallel dipole linear array in front of a $\pi/2$ corner reflector.

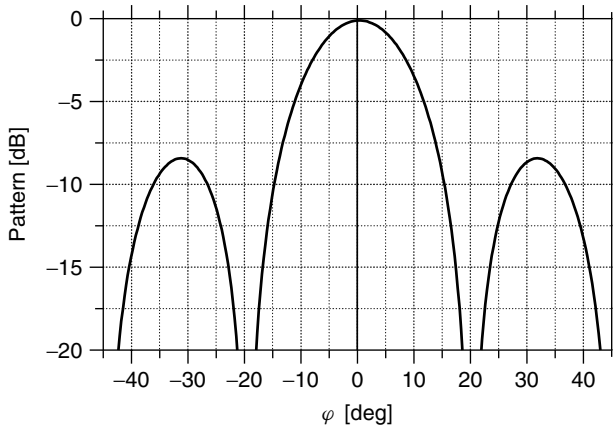


Figure 21. Pattern of a 2-dipole array (see Fig. 20) in front of a $\pi/2$ reflector.

where $(r_n, \theta_n, \varphi_n)$ are the spherical coordinates of the nth element and $I_n e^{j\alpha_n}$ is the corresponding excitation. Parallel circular arrays with their centers on the same axis constitute a cylindrical array. The array factor is simplified to

$$\mathrm{AF}(\theta,\varphi) = \sum_{m=1}^{M}\sum_{I=1}^{L} I_{mI} e^{j[\alpha_{mI} + \beta R_0 \sin\theta\cos(\varphi-\varphi_{mI})+\beta z_m\cos\theta]} \qquad (57)$$

where $I_{mI}e^{j\alpha_{mI}}$ is the excitation of the Ith element of the mth circular array, R_0 is the radius of the circular arrays, and z_m is the position of the mth array on the z axis. A cylindrical array can be made from an array of collinear dipoles in front of a corner reflector (Fig. 22).

To have a maximum at $\theta = \pi/2$, $\varphi = \pi/4$, the array can be uniform. If there are additional constraints on the SLL, then the excitations must be nonuniform. Figure 23 shows the E pattern of a Chebyshev array with $N = 9$ collinear dipoles in front of a $\pi/2$ corner reflector. $d = 0.7\lambda$, SLL ≤ -20 dB and maximum occurs at $\theta = \pi/2$, $\varphi = \pi/4$.

9. CONFORMAL ARRAYS

Arrays with requirements in conformality to a shaped surface are known as *conformal*. Conformal arrays are used in mobile platforms for aerodynamic reasons. Also for specified angles of coverage, arrays can be conformal to stationary shaped surfaces.

Analysis of conformal arrays differs from that of planar ones. The pattern of the array can not be given by multiplying the element pattern and the array factor. These are not separable, and, of course, the pattern is not a simple polynomial. Also, the illumination around the radiating surface as well as the polarization and the pattern of each radiating element must be taken into account separately.

Practical communication and surveillance systems with scanning requirements use conformal arrays of cylindrical shape. In this case one part of the array is illuminated by means of a switching network. For the commutation of a given illumination region around the cylinder, one can use either mechanical rotation, or switch networks, or lens scanning, or matrix beam formers, or digital beam formers [25]. If the array compared to the radius R takes up a small sector and the radius is large ($R \gg \lambda$), then the element pattern is approximated by that of a planar array. If the sector is large compared to the radius R or if the element is in an illuminated region of an array fully wrapped around the cylinder, then the pattern must be carefully calculated and is much different than that of a planar array.

An example of a cylindrical array of 16 microstrip patches is shown in Fig. 24. For a uniform excitation and element phase given by Eq. (53) for a maximum at ($\theta_0 = \pi/2$, $\varphi_0 = 0$) and ($\theta_0 = \pi/2$, $\varphi_0 = \pi/3$), the patterns of the array are presented in Fig. 25.

Conical arrays are used mainly for missiles and aircrafts. The design follows the corresponding one of the cylindrical arrays. Spherical, hemispherical, and conical

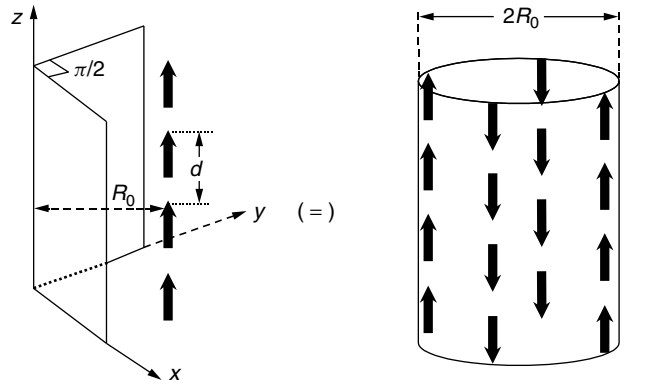


Figure 22. Array of collinear dipoles in front of a $\pi/2$ reflector.

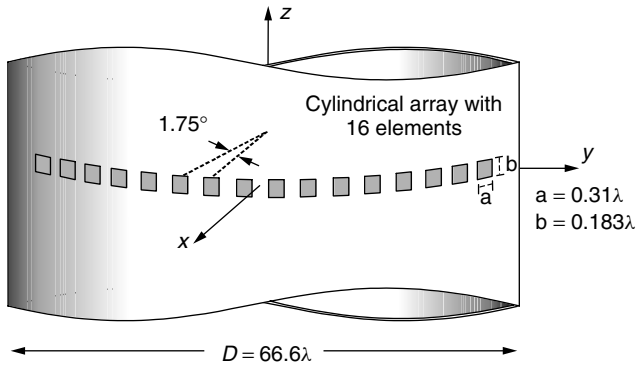


Figure 24. A cylindrical array of 16 rectangular microstrip patches.

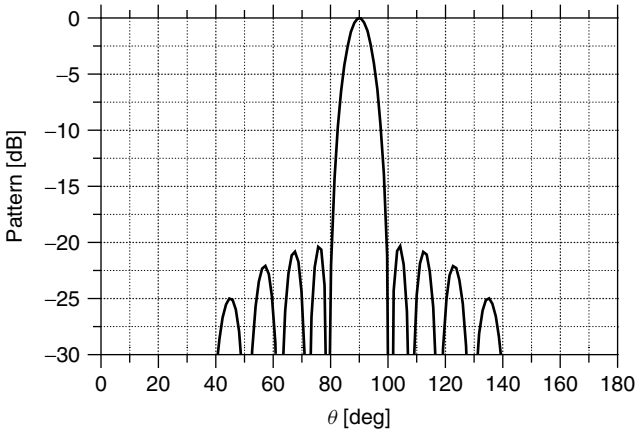


Figure 23. E pattern of a Chebyshev linear array of $N = 9$ collinear dipoles with $d = 0.7\lambda$ and SLL ≤ -20 dB.

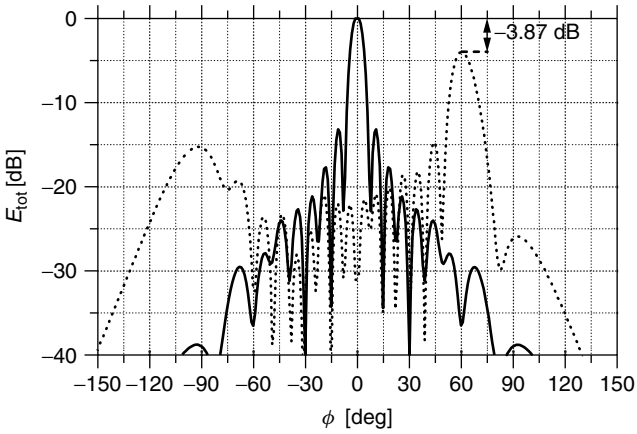


Figure 25. Radiation patterns for a uniform array of 16 axially polarized rectangular patches given in Fig. 24.

are conformal arrays, which are excited in groups of subarrays. Usually they are fed by switch matrices in the same way as the cylindrical arrays.

10. PATTERN SYNTHESIS FOR ARRAYS

The main advantage of arrays is that they can produce accurate approximations of desired radiation patterns. Several techniques have been given in the past for synthesizing array factors. Most of them relate to the synthesis of narrowbeam and low-sidelobe patterns.

Synthesis is based primarily on the antenna engineer experience. The meaningful method, which will result in a realizable solution, must approach the desired property and not the exact requirement. Several synthesis procedures will be described in the following paragraphs.

10.1. Uniform Linear Array Synthesis

Uniform arrays can be used for SLL ≥ -13.3 dB. A linear scanning array with maximum at $\theta = \theta_0$ and half power beamwidth θ_H must have (see Table 1)

$$\theta_0 > \cos^{-1}\left(\cos^2\frac{\theta_H}{2}\right) \tag{58}$$

and

$$N\frac{d}{\lambda} = 0.4428\left[\frac{2(1+\cos\theta_H)}{\sin^4\theta_0 - (\cos\theta_H - \cos^2\theta_0)^2}\right]^{1/2} \tag{59}$$

For example

$$\left.\begin{aligned} &\text{For } \theta_H = 10^\circ \text{ and } \theta_0 = \frac{\pi}{2} \Rightarrow N\frac{d}{\lambda} \\ &\quad = 5.08 \text{ (broadside array)} \\ &\text{For } \theta_H = 10^\circ \text{ and } \theta_0 = \frac{\pi}{6} \Rightarrow N\frac{d}{\lambda} \\ &\quad = 10.28 \text{ (intermediate array)} \end{aligned}\right\} \tag{60}$$

An endfire array where $\theta_0 = 0^\circ$ (Table 1) needs

$$N\frac{d}{\lambda} = \frac{0.4428}{1-\cos\dfrac{\theta_H}{2}} \text{ (ordinary)} \tag{61}$$

$$N\frac{d}{\lambda} = \frac{0.1398}{1-\cos\dfrac{\theta_H}{2}} \text{ (Hansen–Woodyard)} \tag{62}$$

Uniform arrays are useful in practice. For example, an array for mobile communications or with $\theta_0 = 96^\circ$ and $\theta_H = 8^\circ$ must have

$$N\frac{d}{\lambda} = 6.383 \tag{63}$$

Eight $\lambda/2$ collinear dipoles with phase difference of 30° in $d/\lambda = 0.798$ can be used. The antenna E pattern (see Fig. 26) is found by multiplying the array factor by the element pattern.

10.2. Chebyshev Arrays Synthesis

Arrays with constraints on the SLL are useful in communications and radar systems. The Dolph method can be used for $d/\lambda \geq 0.5$. To avoid grating lobes we must have $d \leq d_{\max}$, where

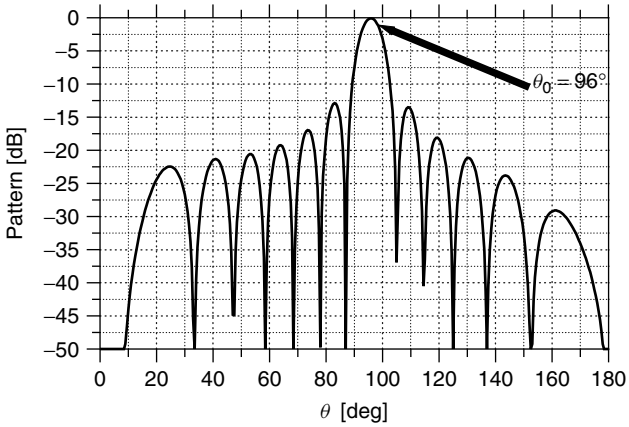


Figure 26. E pattern of a uniform array of eight collinear dipoles in $d/\lambda = 0.798$, $\theta_0 = 96^\circ$, and HPBW $= 8^\circ$.

$$\frac{d_{\max}}{\lambda} = 1 - \frac{\cos^{-1}(1/x_0)}{\pi} \tag{64}$$

x_0 is the distance where the maximum of the array occurs.

For $d/\lambda < 0.5$ the Riblet method must be used. For the Dolph array the HPBW has been related with that of the uniform one (HPBW$_u$) of the same length. The so-called broadening factor f was found to be [7]

$$f = \frac{(\text{HPBW})}{(\text{HPBW}_u)} = 1 + 0.632\left[\frac{2}{R}\cosh\sqrt{(\cosh^{-1}R)^2 - \pi^2}\right]^2 \tag{65}$$

f is valid in the range of -60 dB $\leq$ SLL ≤ -20 dB and for scanning near broadside.

An estimate to the directivity D with the help of f is possible:

$$D = \frac{2R^2}{1 + (R^2 - 1)f\dfrac{\lambda}{Nd}} \tag{66}$$

For the Riblet case the HPBW is given approximately by

$$\text{HPBW} \cong 10.3^\circ\frac{\lambda}{Nd}\sqrt{s + 4.52}\csc\theta_0 \tag{67}$$

where s is the SLL in dB and θ_0 is the scan angle. Also the directivity D [7] is

$$D = \frac{2R^2}{1 + R^2\dfrac{\lambda}{Nd}\sqrt{\dfrac{\ln(2R)}{\pi}}\cdot\sin\left(\beta\dfrac{d}{2}\right)} \tag{68}$$

Array factors of the form [24]

$$\text{AF}(\theta) = T_m(x)T_1^n(x) \tag{69}$$

are able to give either equal or unequal sidelobes. The number of nulls depends on m and n. If we compare (69) with an array factor

$$\text{AF}_1(\theta) = T_{m+n}(x) \tag{70}$$

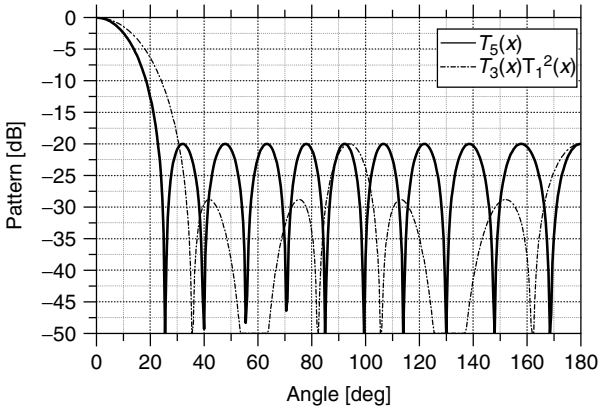


Figure 27. Patterns of case 3 endfire Chebyshev array with array factor $T_5(x)$ and $T_3(x)T_1^2(x)$ for SLL $= -20$ dB.

we see that $\mathrm{AF}_1(\theta)$ has $(m+n)$ roots while $\mathrm{AF}(\theta)$ has either $(m+1)$ roots for $m = 2k$ or m roots for $m = 2k+1$.

Figure 27 presents for comparison the factors $T_5(x)$ and $T_3(x)T_1^2(x)$ of the case 3 endfire array for $N = 11$, $d/\lambda = 0.35$, and SLL $= -20$ dB.

10.3. Synthesis by Sampling or by Root Matching

Continuous distributions create excellent patterns with low sidelobes. Discrete arrays coming from sampling them can give similar patterns. For large-element spacing, the patterns of the array and the line source do not match well. To avoid this problem the method of root matching is used. In other words, the nulls of the continuous distribution pattern appear in the pattern of the discrete array. If the pattern does not yield the desired accuracy, a perturbation technique [26] can be applied. In this case the distribution of the discrete-element array varies to improve the accuracy.

10.3.1. Simple Distributions. A discrete-element array of a fixed length is transformed to a continuous distribution as the number of elements approaches to infinity (Fig. 28).

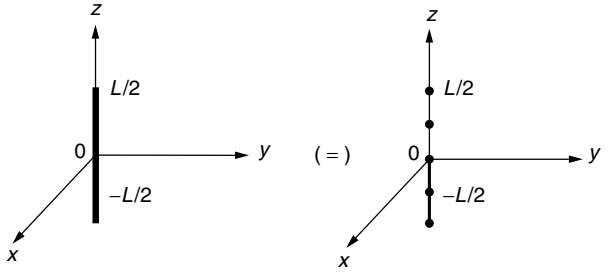


Figure 28. Line source and its equivalent discrete-element array.

The array factor reduces to an integral and is called the space factor (SF):

$$\mathrm{SF}(\theta) = \int_{-L/2}^{L/2} I(z')e^{j(\beta\cos\theta-\alpha)z'}dz' = \int_{-L/2}^{L/2} I(z')e^{j\xi z'}dz' \qquad (71)$$

where $I(z')$ and α are the amplitude distribution and phase progress along the source. This equation is the finite one-dimensional Fourier transform relating the far field to the excitation.

Changing the bounds of integration, Eq. (71) becomes

$$\mathrm{SF}(\theta) = \int_{-\infty}^{\infty} I(z')e^{j\xi z'}dz' \qquad (72)$$

$I(z')$ is zero outside of $-L/2 \le z' \le L/2$. Using the Fourier transform we have

$$I(z') = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{SF}(\theta)e^{-jz'\xi}d\xi \qquad (73)$$

If L is large enough, Eq. (71) gives the desired pattern within a certain error. In the discrete-element array $I(z')$ is sampled in appropriate intervals. It is observed that a small difference between the two patterns appears. Useful distributions with their characteristics are presented in Table 5.

10.3.2. Taylor Distribution (Chebyshev Error). Chebyshev arrays provide an optimum relation between the SLL

Table 5. Radiation Characteristics for Line Sources and Linear Arrays with Uniform, Triangular, Cosine, and Cosine-Squared Distributions

Distribution	Uniform	Triangular	Cosine	Cosine-Squared
Distribution I_n	I_0	$I_1\left(1-\frac{2}{L}\mid z'\mid\right)$	$I_2\cos\left(\frac{\pi}{L}z'\right)$	$I_3\cos^2\left(\frac{\pi}{L}z'\right)$
Space factor (SF) $u=\left(\frac{\pi L}{\lambda}\right)\sin\theta$	$I_0L\frac{\sin u}{u}$	$I_1\frac{L}{2}\left[\frac{\sin\left(\frac{u}{2}\right)}{\frac{u}{2}}\right]^2$	$I_2L\frac{\pi}{2}\frac{\cos(u)}{(\pi/2)^2-u^2}$	$I_3\frac{L}{2}\frac{\sin(u)}{u}\left[\frac{\pi^2}{\pi^2-u^2}\right]$
Half-power beamwidth (degrees) $L\lambda$	$\frac{50.6}{(L\lambda)}$	$\frac{73.4}{(L\lambda)}$	$\frac{68.8}{(L\lambda)}$	$\frac{83.2}{(L\lambda)}$
First null beamwidth (degrees) $L\lambda$	$\frac{114.6}{(L\lambda)}$	$\frac{229.2}{(L\lambda)}$	$\frac{171.9}{(L\lambda)}$	$\frac{229.2}{(L\lambda)}$
First side lobe max. (to main max.) (dB)	-13.2	-26.4	-23.2	-31.5
Directivity factor (Llarge)	$2\left(\frac{L}{\lambda}\right)$	$0.75\left[2\left(\frac{L}{\lambda}\right)\right]$	$0.810\left[2\left(\frac{L}{\lambda}\right)\right]$	$0.667\left[2\left(\frac{L}{\lambda}\right)\right]$

and the HPBW. Another distribution characterized by low SLL of the first N sidelobes next to the main beam is the Taylor distribution. The other sidelobes gradually fall off in value. The space factor of the Taylor distribution comes from Dolph–Chebyshev if the elements of the array become infinite [27]:

$$\mathrm{SF}(\theta) = \frac{\cosh\left[\sqrt{(\pi A)^2 - u^2}\right]}{\cosh(\pi A)} \tag{74}$$

where

$$\left.\begin{aligned} u &= \pi \frac{L}{\lambda}(\cos\theta - \cos\theta_0) \\ \cosh(\pi A) &= R \end{aligned}\right\} \tag{75}$$

Since Eq. (74) cannot be realized physically, Taylor [27] presented a space factor whose roots are the zeros of SF (θ). Because the factor is the approximation of the ideal Chebyshev, it is also known as the *Chebyshev error*. The space factor is

$$\mathrm{SF}(u, A, \overline{n}) = \frac{\sin u}{u} \frac{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1 - \left(\frac{u}{u_n}\right)^2\right]}{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1 - \left(\frac{u}{n}\right)^2\right]} \tag{76}$$

where $(\overline{n} - 1)$ is a parameter that defines the number of pairs of inner nulls.

$$u_n = \overline{n} \frac{\sqrt{A^2 + \left(n - \frac{1}{2}\right)^2}}{\sqrt{A^2 + \left(\overline{n} - \frac{1}{2}\right)^2}} \quad n = 1, 2, \ldots, \overline{n} - 1 \tag{77}$$

The Taylor distribution can be found by the Fourier transform:

$$I(z') = \mathrm{SF}(0, A, \overline{n}) + 2\sum_{m=1}^{\overline{n}-1} \mathrm{SF}(m, A, \overline{n}) \cos\left(m\pi \frac{z'}{L}\right) \tag{78}$$

Sampling $I(z')$, we create the discrete array. Problems that arise and cause inaccuracies, even for large arrays, were addressed earlier [28].

Figure 29 presents the Taylor pattern for SLL $= -25$ dB, $L = 7\lambda$, and $\overline{n} = 5$. The amplitudes of the elements at $d/\lambda = 0.5$ are given in Fig. 30.

10.3.3. Taylor One-Parameter Distribution. In low-noise systems it is desirable to have the first sidelobes at a certain level while the others decay as the angle increases. Taylor [6] developed a procedure for synthesizing such a pattern. The distribution is referred as the *Taylor one-parameter* and is of the following form:

$$I_n(z') = I_0\left[\pi B \sqrt{1 - \left(\frac{2z'}{L}\right)^2}\right] \tag{79}$$

where z' = distance from the center of the line source
L = length of the line source

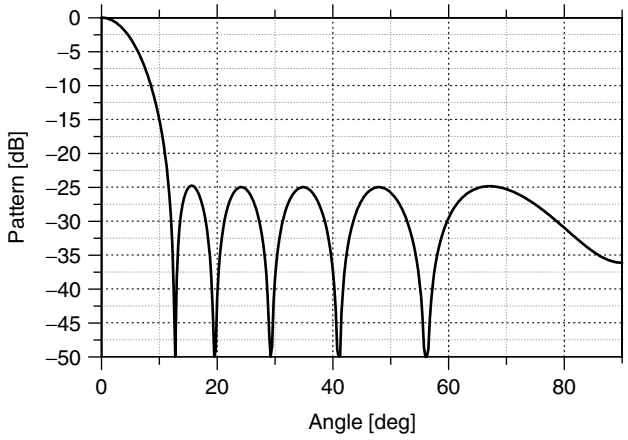


Figure 29. Pattern of a Taylor distribution with SLL $= -25$ dB, $\overline{n} = 5$, $d/\lambda = 0.5$, and $N = 14$.

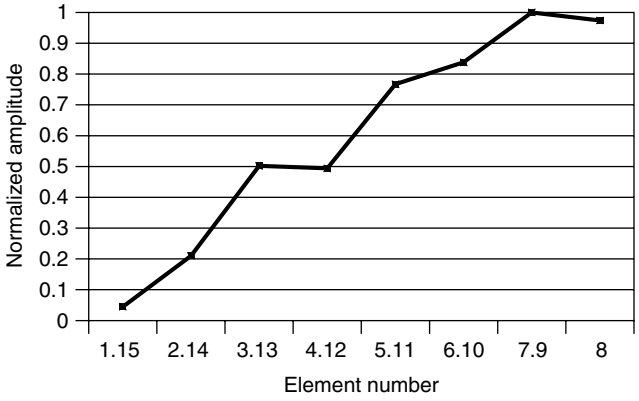


Figure 30. Amplitudes of the elements at $d/\lambda = 0.5$ of the Taylor distribution with the pattern given in Fig. 29.

B = parameter that determines the sidelobe
I_0 = modified Bessel function of the first kind and zero order.

The space factor associated with (79) can be obtained by the Fourier transform:

$$\mathrm{SF}(\theta) = \begin{cases} \dfrac{\sin[\sqrt{(\pi B)^2 - u^2}]}{\sqrt{(\pi B)^2 - u^2}}, & u^2 < (\pi B)^2 \\ \dfrac{\sinh[\sqrt{u^2 - (\pi B)^2}]}{\sqrt{u^2 - (\pi B)^2}}, & u^2 > (\pi B)^2 \end{cases} \tag{80}$$

where

$$u = \beta \frac{L}{2}(\cos\theta - \cos\theta_0) \tag{81}$$

Parameter B is found from [29]

$$R = 4.60333 \frac{\sinh \pi B}{\pi B} \tag{82}$$

10.3.4. Bayliss Line Source. A pattern null on boresight with the appropriate sidelobe level was developed by Bayliss [30]. Monopulse tracking systems use an auxiliary pattern of the form of Bayliss in coincidence with a beam

peak of the main pattern. The Bayliss pattern is described in terms of two parameters, A and $\overline{n}$. The pattern is

$$\mathrm{SF}(\theta) = u\cos(\pi u)\frac{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{u}{u_n}\right)^2\right]}{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{u}{n+\frac{1}{2}}\right)^2\right]} \tag{83}$$

$(\overline{n}-1)$ is the parameter that defines the number of inner nulls:

$$u_n = \begin{cases} \left(\overline{n}+\frac{1}{2}\right)\left(\frac{\xi_n^2}{A^2+\overline{n}^2}\right)^{1/2} & n = 1,2,3,4 \\ \left(\overline{n}+\frac{1}{2}\right)\left(\frac{A^2+n^2}{A^2+\overline{n}^2}\right)^{1/2} & n = 5,6,\ldots,\overline{n}-1 \end{cases} \tag{84}$$

A, ξ_n, and the location $u_{\max}$ where SF is maximized are found as a function of $S = |\text{sidelobe level (dB)}|$ (see Table 6):

$$x = \alpha_1 + S\{\alpha_2 + S[\alpha_3 + S(\alpha_4 + S\cdot\alpha_5)]\} \tag{85}$$

The aperture distribution by the sine Fourier series with $\overline{n}$ terms is

$$I(z') = \sum_{m=0}^{\overline{n}-1} B_m \sin\left[(2m+1)\frac{\pi z'}{L}\right] \tag{86}$$

where

$$B_m = \frac{(-1)^m\left(m+\frac{1}{2}\right)^2}{2j}\frac{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{m+\frac{1}{2}}{u_n}\right)^2\right]}{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{m+\frac{1}{2}}{n+\frac{1}{2}}\right)^2\right]} \tag{87}$$

A Bayliss and a Taylor pattern ($\overline{n}=5$) for SLL $= -30$ dB, $N = 14$ and $d/\lambda = 0.5$ are shown in Fig. 31.

10.3.5. Modified Patterns by Iteration. Taylor and Bayliss patterns with individual different sidelobes can be made by using a perturbation procedure. According to Elliot [5], we express the patterns in more general forms:

$$\text{Taylor:}\quad \mathrm{SF}(u) = C_0\frac{\sin\pi u}{u}\frac{\displaystyle\prod_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1}\left(1-\frac{u}{u_n}\right)}{\displaystyle\prod_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1}\left(1-\frac{u}{n}\right)} \tag{88}$$

Table 6. Coefficients of the Parameters A, ξ_n, and $u_{\max}$

x	α_1	α_2	α_3	$\alpha_4\cdot 10^5$	$\alpha_5\cdot 10^7$
A	0.3038753	0.05042922	−0.00027989	0.343	−0.2
ξ_1	0.9858302	0.0333885	0.00014064	−0.19	0.1
ξ_2	2.00337487	0.01141548	0.0004159	−0.373	0.1
ξ_3	3.00636321	0.00683394	0.00029281	−0.161	0
ξ_4	4.00518423	0.00501795	0.00021735	−0.088	0
$u_{\max}$	0.4797212	0.01456692	−0.00018739	0.218	−0.1

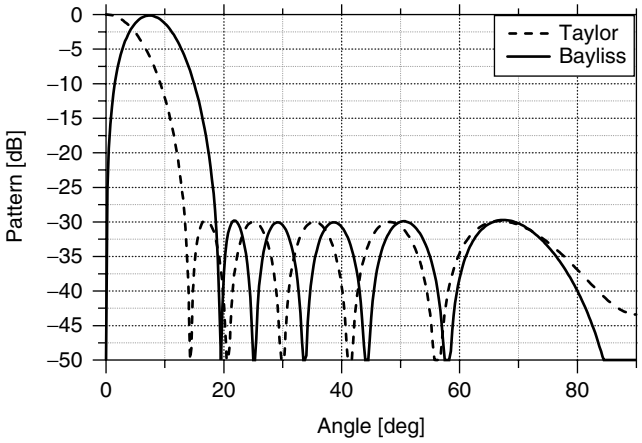


Figure 31. Pattern of a Taylor and a Bayliss array to give SLL $= -30$ dB ($\overline{n}=5$) for $N = 14$ and $d/\lambda = 0.5$.

where

$$\left.\begin{aligned} u_n &= \overline{n}_R\frac{\sqrt{A_R^2+\left(n-\frac{1}{2}\right)^2}}{\sqrt{A_R^2+\left(\overline{n}_R-\frac{1}{2}\right)^2}} \\ u_n &= -\overline{n}_L\frac{\sqrt{A_L^2+\left(n+\frac{1}{2}\right)^2}}{\sqrt{A_L^2+\left(\overline{n}_L-\frac{1}{2}\right)^2}} \end{aligned}\right\} \tag{89}$$

where R and L identify the right and left side of the pattern. $\overline{n}_R$ and $\overline{n}_L$ denote the transition roots of the two sides, and A_R and A_L are the corresponding SLL parameters.

$$\text{Bayliss:}\quad \mathrm{SF}(u) = C_0 u\cos\pi u\frac{\displaystyle\prod_{n=-(\overline{n}_{L-1})}^{\overline{n}_R-1}\left(1-\frac{u}{u_n}\right)}{\displaystyle\prod_{n=-\overline{n}_L}^{\overline{n}_R-1}\left(1-\frac{u}{n+\frac{1}{2}}\right)} \tag{90}$$

We start from a pattern $\mathrm{SF}_0(u)$ with the SLL on both sides to be the average of the desired ones. All the roots of Eqs. (88) and (90) u_n^0 are known.

We assume that the roots of the desired pattern are

$$u_n = u_n^0 + \delta u_n \tag{91}$$

with a small perturbation δu_n. Then if

$$C = C_0 + \delta\, C$$

$\mathrm{SF}(u)$ becomes

$$\frac{\mathrm{SF}(u)}{\mathrm{SF}_0(u)} - 1 = \frac{\delta C}{C_0} + \sum_{n=-(\overline{n}_{L-1})}^{\overline{n}_{R-1}}\frac{\dfrac{u}{(u_n^0)^2}}{1-\dfrac{u}{u_n^0}}\delta u_n \tag{92}$$

The peak positions u_m^p give

$$\frac{\mathrm{SF}(u_m^p)}{\mathrm{SF}_0(u_m^p)} - 1 = \frac{\delta C}{C_0} + \sum_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1} \frac{\dfrac{u_m^p}{(u_n^0)^2}}{1 - \dfrac{u_m^p}{u_n^0}} \delta u_n \tag{93}$$

For the $\overline{n}_R + \overline{n}_L - 1$ lobes we have an equal number of linear equations of the form (93). The system is solved for $\delta C/C_0$ and the $\overline{n}_R + \overline{n}_L - 2$ values δu_n since $\delta u_0 = 0$. The new values of u_n are substituted in (88) and the new pattern is checked. The process is repeated until the new pattern differs from the desired by a minimum predefined amount.

The same procedure is applied for the Bayliss distribution. Equation (92) is modified to

$$\frac{\mathrm{SF}(u_m^p)}{\mathrm{SF}_0(u)} - 1 = \frac{\delta C}{C_0} - \frac{\delta u_0}{u_m^p} + \sum_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1} \frac{\dfrac{u_m^p}{(u_n^0)^2}}{1 - \dfrac{u_m^p}{u_n^0}} \delta u_n \tag{94}$$

which gives a system of $\overline{n}_R + \overline{n}_L$ unknowns, which is solved. The perturbation process is repeated in the same way as before.

Figures 32 and 33 show the pattern of two modified distributions for $\overline{n} = 6$, SLL $= -20$ dB and three intermost pairs of lobes -30 dB.

In the preceding cases, an iterative procedure can be applied for power pattern synthesis where we have additional degrees of freedom. Orchard et al. [13] proposed a technique by dividing the pattern in the shaped beam region and the sidelobe region.

The array factor in general is

$$\mathrm{AF}(\theta) = \prod_{n=1}^{N} (z - z_n) = \sum_{n=0}^{N} I_n z^n \tag{95}$$

It is assumed that the zero locations are complex of the form

$$z_n = \exp(a_n + jb_n) \tag{96}$$

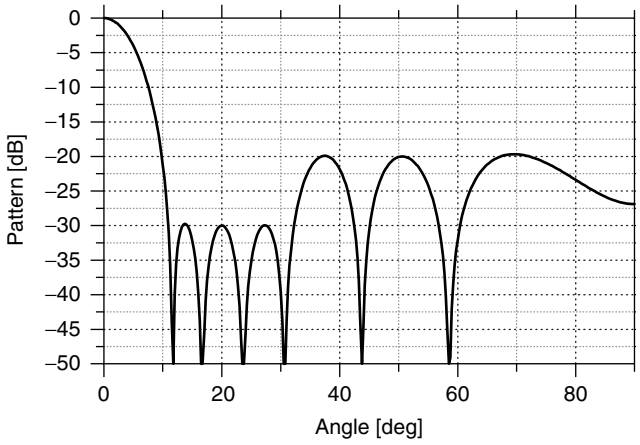


Figure 32. Modified Taylor pattern for $\overline{n} = 6$, three intermost pairs of lobes with -30 dB level and the other lobes with -20 dB level ($N = 14$, $d/\lambda = 0.5$).

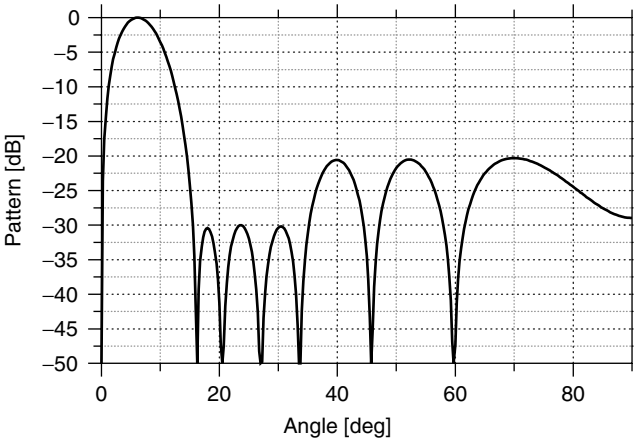


Figure 33. Modified Bayliss pattern for $\overline{n} = 6$, three intermost pairs of lobes with -30 dB level and the other lobes with -20 dB level ($N = 14$, $d/\lambda = 0.5$).

and z is written as

$$z = \exp(j\varphi) \tag{97}$$

Orchard sets the Nth root $z_N = 1$ and expresses the power pattern in decibels:

$$G = \sum_{n=1}^{N-1} 10 \log[1 - 2e^{a_n} \cos(\varphi - b_n) + e^{2a_n}] + 10 \log[2(1 + \cos\varphi)] + C_1 \tag{98}$$

C_1 is a constant that allows G to have at the main beam a given value.

The unknown coefficients a_n, b_n, and φ are found by using an iterative scheme. This scheme uses the derivatives of G and the difference between the existing and the desired power pattern. The procedure does not produce an optimum result. However it offers flexibility and control to the ripple and the sidelobe level as well as to the entire radiation pattern.

11. FOURIER TRANSFORM AND THE ORTHOGONAL METHOD

A linear uniformly spaced array with nonuniform excitation has an array factor of the form

$$\mathrm{AF}(\psi) = \sum_{n=1}^{N} I_n e^{jn\psi} \tag{99}$$

We expand a desired $\mathrm{AF}_d(\psi)$ in a Fourier series with infinite terms of the form (99). The first N coefficients of the two series are equated to approximate the desired pattern. The coefficients are found by using the orthogonality of the expansion functions:

$$I_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{AF}_d(\psi) e^{-jn\psi} d\psi \tag{100}$$

The Fourier method is adequate for spacing $d = 0.5\lambda$. For $d > 0.5\lambda$ it fails, while for $d < 0.5\lambda$ and a sufficient

number of elements the pattern more closely matches the desired one.

The general expression of (99) comes from nonuniformly spaced arrays, [33–37]:

$$\mathrm{AF}(u) = \sum_{n=1}^{N} I_n e^{jx_n u} \tag{101}$$

where

$$\left.\begin{aligned} u &= \pi \cos\theta \\ x_n &= \frac{d_n}{\lambda/2} \end{aligned}\right\} \tag{102}$$

The basis functions of (101) are not orthogonal. Their inner product is

$$k_{in} = \int_{-\pi}^{\pi} e^{j(x_i - x_n)u} du = \frac{\sin(x_i - x_n)\pi}{(x_i - x_n)\pi} \tag{103}$$

$\mathrm{AF}(u)$ can be expressed by the Gram–Schmidt procedure [33] in an orthogonal basis $\{\Psi_n(u)\}$:

$$\Psi_n(u) = \sum_{i=1}^{n} c_i^{(n)} e^{jx_i u} \tag{104}$$

and

$$\mathrm{AF}(u) = \sum_{n=1}^{N} B_n \Psi_n(u) \tag{105}$$

With the aid of the orthogonality, we have

$$B_n = \int_{-\pi}^{\pi} \mathrm{AF}_d(u) \cdot \Psi_n^*(u)\, du \tag{106}$$

$\Psi_n^*(u)$ is the conjugate of $\Psi_n(u)$.

Combining (101), (104), and (105), we have

$$I_n = \sum_{i=n}^{N} c_n^{(i)} B_i \tag{107}$$

For comparison purposes, an 7-element array with 0.85λ spacing and a constant beam between 85° and 95° is designed. Figure 34 presents the pattern obtained by the orthogonal method and the Fourier transform.

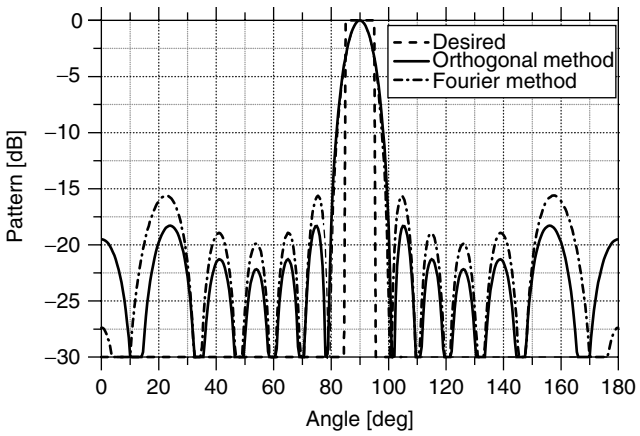


Figure 34. Pattern of an 7-element array with $d/\lambda = 0.85$ by the orthogonal and Fourier methods.

12. WOODWARD–LAWSON (WL) METHOD AND ORTHOSYNTHESIS

A uniform linear array with an array factor of the form $\sin(N\psi/2)/N\sin(\psi/2)$ has the narrowest pattern that can be achieved with an array. The uniform pattern is a useful tool for synthesis because it can be a member of an orthogonal set of beams. By devising lossless networks one can superimpose groups of beams in order to synthesize a desired pattern. A uniform array with N elements in equal distance d/λ produces a normalized beam pattern:

$$f_m(\theta) = b_m \frac{\sin(N\psi_m/2)}{N\sin(\psi_m/2)} \tag{108}$$

where

$$\psi_m = \beta d(\cos\theta - \cos\theta_m) \tag{109}$$

If we assume that a desired factor is the superposition of terms of the form (108), then

$$\mathrm{AF}(\theta) = \sum_{m=-M}^{M} b_m \frac{\sin(N\psi_m/2)}{N\sin(\psi_m/2)} \tag{110}$$

where

$$b_m = \mathrm{AF}(\theta_m) \tag{111}$$

and

$$\theta_m = \cos^{-1}\left(m\frac{\lambda}{Nd}\right) \tag{112}$$

The excitation of each element becomes

$$I_n = \frac{1}{N}\sum_{m=-M}^{M} \mathrm{AF}(\theta_m) e^{-j\beta d_n \cos\theta_m} \tag{113}$$

For a line source, we again can superimpose groups of beams of the form [6]

$$f_m(\theta) = b_m \frac{\sin[\beta L/2(\cos\theta - \cos\theta_m)]}{\beta L/2(\cos\theta - \cos\theta_m)} \tag{114}$$

The space factor is

$$\mathrm{SF}(\theta) = \sum_{m=-M}^{M} b_m \frac{\sin[\beta L/2(\cos\theta - \cos\theta_m)]}{\beta L/2(\cos\theta - \cos\theta_m)} \tag{115}$$

where

$$b_m = \mathrm{SF}(\theta_m) \tag{116}$$

and

$$\theta_m = \cos^{-1}\left(m\frac{\lambda}{L}\right) \tag{117}$$

The excitation distribution is

$$I(z') = \sum_{m=-M}^{M} b_m e^{-j\beta z' \cos\theta_m} \tag{118}$$

Instead of sampling $\mathrm{AF}(\theta)$ and $\mathrm{SF}(\theta)$ in θ_m we could apply the orthogonal method termed *orthosynthesis*. In this case

θ_m can be different from that in Eq. (112) or (117) and can have values that optimize the solution.

Figure 35 presents a cosecant-squared power pattern of a line source with $L = 10\lambda$. The same pattern with discrete elements ($N = 20$ and $N = 30$) is presented in Fig. 36. Figure 37 presents the desired of a modified cosecant-squared pattern and the pattern by orthosynthesis and WL for $N = 16$ and $d/\lambda = 0.5$. From the value of the mean-square error it appears that orthosynthesis is better than the WL.

13. ORTHOGONAL PERTURBATION METHOD

In the shape design of an array, an adjustment of the spacing between the elements can be made. A procedure that combines the adjustment with the orthogonal method is known as the *orthogonal perturbation method* [38,39]. A linear array has an array factor of the form

$$\mathrm{AF}_0(\theta) = \sum_{i=1}^{N} I_i e^{j\beta d_i \cos\theta} = \sum_{i=1}^{N} I_i \Phi_i(\theta) \tag{119}$$

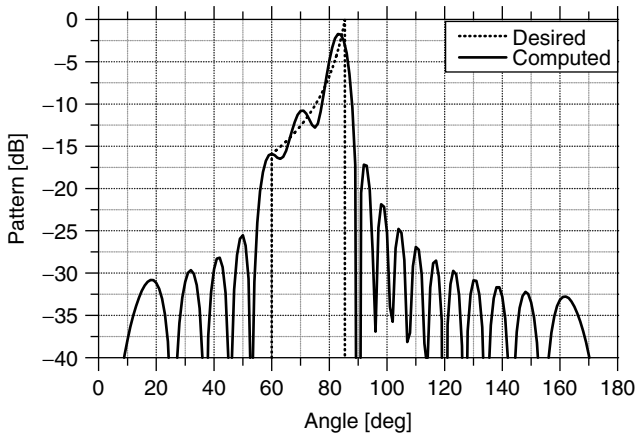


Figure 35. A cosecant-squared power pattern of a line source with $L = 10\lambda$ taken by WL method.

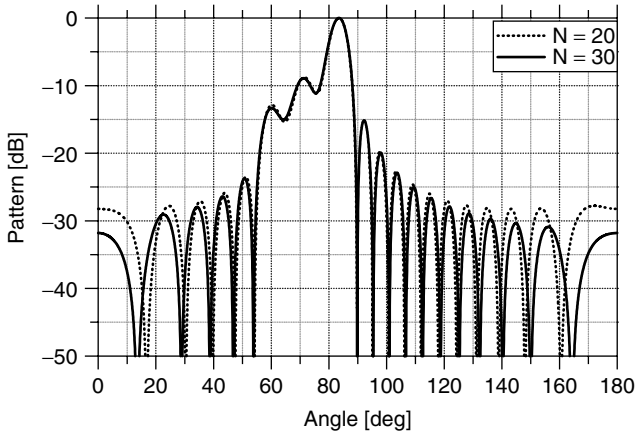


Figure 36. A cosecant-squared power pattern by sampling the line source with pattern of Fig. 35 to have $N = 20$ and 30 elements.

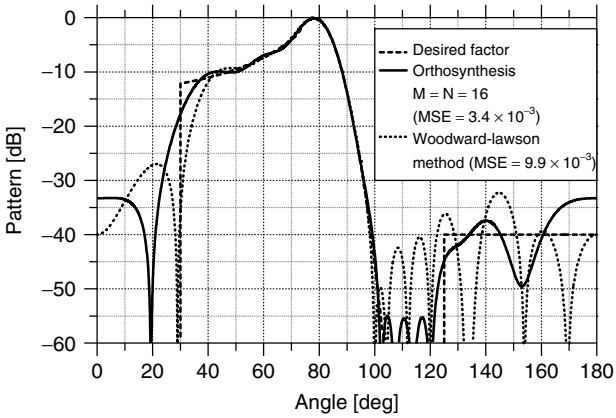


Figure 37. Cosecant-squared power pattern for $N = 16$ elements and $d/\lambda = 0.5$ by orthosynthesis and WL.

If we perturb the position d_i of each element such that $\beta(\delta d_i) \gg 1$, then the array factor becomes

$$\mathrm{AF}_1(\theta) \cong \sum_{i=1}^{N} [1 + j\beta(\delta d_i)\cos\theta] I_i e^{j\beta d_i \cos\theta} \tag{120}$$

Substituting (119) into (120) and dividing by $\cos\theta$ we have

$$F(\theta) = \frac{\mathrm{AF}_1(\theta) - \mathrm{AF}_0(\theta)}{\cos\theta} = \sum_{i=1}^{N} A_i \Phi_i(\theta) \tag{121}$$

where

$$A_i = j\beta(\delta d_i) I_i \tag{122}$$

It is clarified that for $\theta = \pi/2$, $F(\theta)$ is already kept equal to zero. By the orthogonal method we have

$$\left.\begin{aligned} \mathrm{AF}_0(\theta) &= \sum_{i=1}^{N} B_i^0 \Psi_i(\theta) \\ F(\theta) &= \sum_{i=1}^{N} B_i \Psi_i(\theta) \end{aligned}\right\} \tag{123}$$

I_i and A_i are

$$I_i = \sum_{j=i}^{N} B_j^0 c_i^{(j)} \tag{124}$$

$$A_i = \sum_{j=i}^{N} B_j c_i^{(j)} \tag{125}$$

Instead of I_i, we can use quantized approximate values for the initial array. After the quantization the array is perturbed and from $F(\theta)$ we take A_i, which gives δd_i. The perturbation continues by an iterative procedure until the desired approximation is achieved. If the result is not the expected, the procedure is repeated for a larger number of amplitudes.

An example with a Chebyshev pattern $T_5(x)$ with $\mathrm{SLL} = -30$ dB and $\mathrm{HPBW} = 15^\circ$ is presented. Three quantized amplitudes and 11 elements are used. The results are shown in Fig. 38 and Table 7.

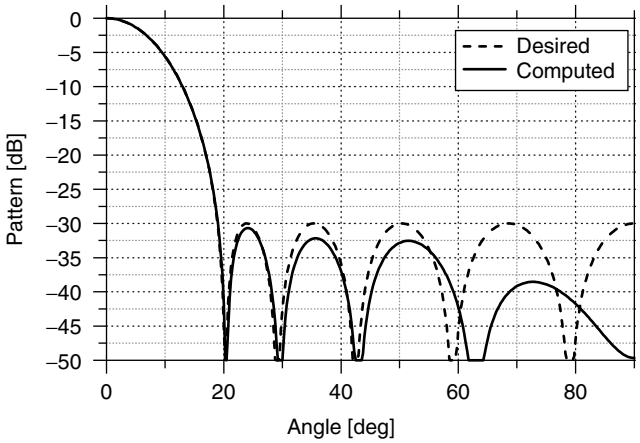


Figure 38. Chebyshev pattern with SLL = −30 dB and HPBW = 15° taken by the orthogonal perturbation method.

Table 7. Amplitudes and Positions of an 11-Element Array that Produces a Chebyshev Pattern

Element Number (i)	Distance (λ)	Quantized Current $I(i)$
1,11	±1.975	2
2,10	±1.603	2
3,9	±1.204	5
4,8	±0.800	5
5,7	±0.378	8
6	0	5

14. SYNTHESIS AS AN OPTIMIZATION PROBLEM

The antenna synthesis is mainly a nonlinear optimization procedure. In this procedure a convenient real function, which takes an optimum value at the reached properties of the desired antenna, is constructed. More than one function can be used to fit several antenna properties [40]:

1. Radiation pattern at a single frequency or at a number of frequencies.
2. Antenna impedance at a single frequency or at a number of frequencies.
3. Antenna index without or under constraint on another index.
4. Antenna impedance and/or radiation pattern in a given frequency range.
5. Coupling of antennas.

The optimization parameters may characterize the excitation, the shape, the size, the loadings, and the current distribution of the antenna elements. Any variation of some parameters requires completely new solutions.

Most of the optimization methods are divided into two categories. The first makes use of the values of the optimization function itself. The second looks at the gradient of the above function. The optimization function in some cases is not an explicit function but it is simply computed numerically.

Except for the abovementioned methods, procedures based on random search are available. These are based on the use of a random-number generator by which the successive points are determined. Finally, the simulated annealing and the genetic algorithms are two global optimizers.

15. OPTIMIZATION OF AN INDEX

An antenna index, I, such as directivity, gain, or quality factor, can be written as

$$I = \frac{[\tilde{a}]^*[A][a]}{[\tilde{a}]^*[M][a]} \tag{126}$$

where $[\tilde{a}]^* = [a_1, a_2, \ldots, a_N]^*$ is the conjugate transpose of $[a]$. By $[a]$ one can represent the current or the voltage excitation vector of the array. $[A]$ and $[M]$ with

$$\left.\begin{aligned}[A] &= [\alpha_{ij}] \\ [M] &= [m_{ij}]\end{aligned}\right\} \tag{127}$$

are both Hermitian $N \times N$ square matrices. Also $[M]$ is positive-definite.

An index I of the form (126) will be optimized under the constraint that another index I_1 is

$$I_1 = \frac{[\tilde{a}]^*[M_2][a]}{[\tilde{a}]^*[M_3][a]} = \gamma \tag{128}$$

According to several authors [41,42], a solution can be found by using the Lagrange multiplier and setting the quantity L

$$L = \frac{[\tilde{a}]^*[A][a]}{[\tilde{a}]^*[M][a]} + \lambda \left\{ \frac{[\tilde{a}]^*[M_2][a]}{[\tilde{a}]^*[M_3][a]} - \gamma \right\} \tag{129}$$

stationary with respect to $[a]$ and λ.

Zeroing the first variation of L, we have

$$[a] = q[K]^{-1}[\tilde{B}]^* \tag{130}$$

where q is a constant and

$$[K] = [M] + p\{\gamma[M_3] - [M_2]\} \tag{131}$$

where p is found from (128) by solving an eigenvalue equation [43,44]. If there is no constraint on I_1, $p = 0$. In the optimization procedure, pattern values and index constraints can also be combined.

An example of a wire dipole array shown in Fig. 39 with maximum gain G_1 in $\theta_1 = 0^\circ$ at the frequency f_1 under the constraint that at $f_2 = f_1/2$ the gain is G_2 in the direction θ_2 is presented in Fig. 40.

An interesting case is the optimization of the directivity of a 21-element array of 1λ length, which gives $D = 48.77$, $Q = 2.258875 \times 10^5$, and $\eta = 1.028 \times 10^{-13}\%$. The minimum to maximum excitation is 4.2044×10^{-4}. A five-element uniform array with the same length has maximum $D = 5$, $Q = 1$, and $\eta = 100\%$. The first array is superdirective, usually known as *supergain*. In supergain arrays the ohmic losses are extremely large. This is the

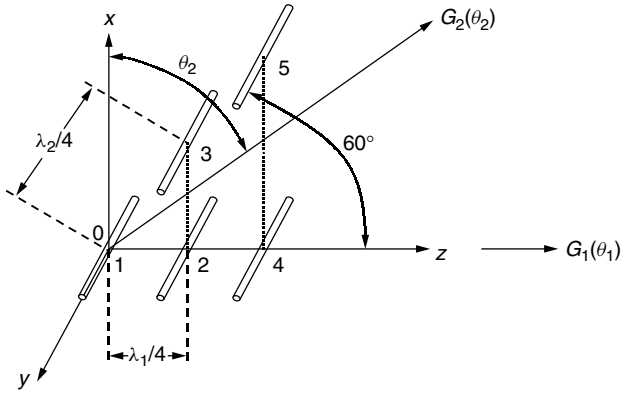


Figure 39. A 5-parallel-wire dipole triangular array.

penalty than one must pay in loss of efficiency if reduction of length is important.

16. OPTIMIZATION BY SIMPLEX AND GRADIENT METHODS

Simplex and gradient [45–47], are both local optimizer methods. A *simplex* is a body in multidimensional space. The optimization function at the vertices of a simplex is computed. On this basis, a new smaller simplex is chosen within which an optimum should be situated. The optimum depends on the initial simplex [45].

Gradient methods are known as steepest-descent methods. A starting point is chosen and the direction where the optimization function decreases most rapidly is found. Adopting a new point in that direction at a desired distance and repeating the process, a minimum of the optimization function is achieved.

It is noticed that it is difficult to judge if the minimum is the global one or a local one. From an antenna engineering point of view, we are usually interested in a suitable solution and not necessarily the global optimum.

The following example illustrates the optimization of the radiation pattern and the antenna impedance of a four-element Yagi–Uda array. For the pattern, the antenna gain obtained was larger than a prescribed value. For the impedance, the mutual coupling was taken into account. Use of initial values for the optimization parameters was based on experience. After 12 simplex iterations, an antenna was obtained with (see Fig. 41) $L_R = 0.50\lambda$, $L_F = 0.468\lambda$, $L_D = 0.45\lambda$, $d_r = 0.25\lambda$, $d = 0.30\lambda$, $a = 0.002\lambda$.

It was found that $G = 9.15$ dB and $Z_{in} = (36 + j0)\Omega$. The front-to-back-ratio was 14.1 dB and the HPBW = 65.2°. Figure 42 shows the pattern of the antenna. A similar optimization can be applied for log-periodic dipole arrays [9].

17. OPTIMIZATION BY SIMULATED ANNEALING METHODS

The basic idea in simulated annealing (SA) is to combine local search with Monte Carlo techniques in analogy

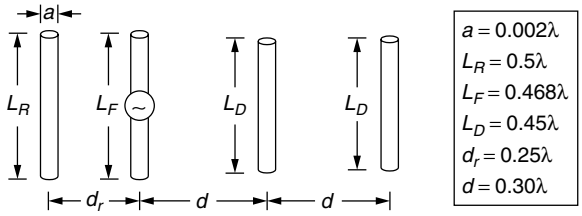


Figure 41. Yagi–Uda antenna with four elements.

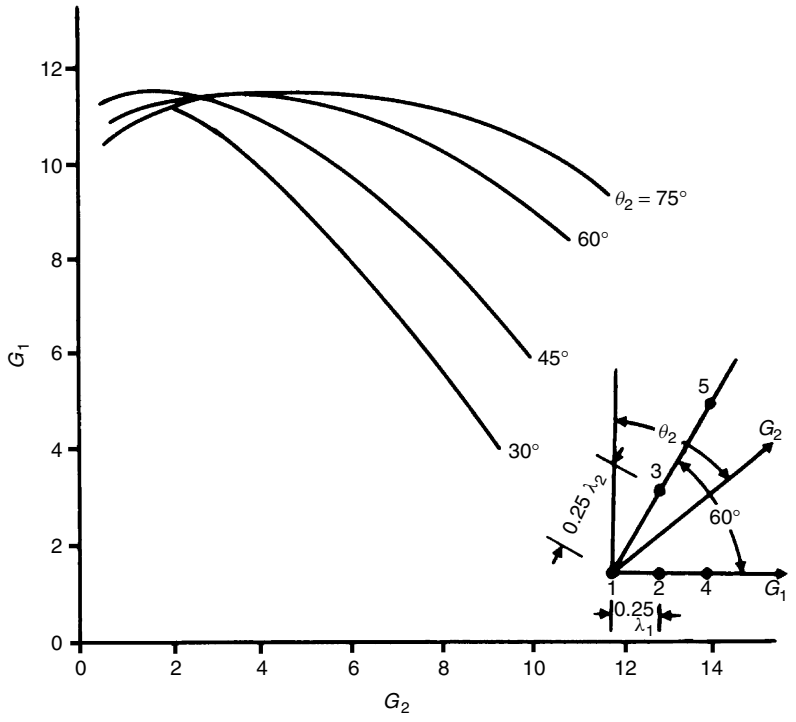


Figure 40. Maximum gain G_1 in frequency f_1 versus G_2 in frequency $f_1/2$. $\theta_1 = 0°$ and θ_2 is 30°, 45°, 60°, and 75°.

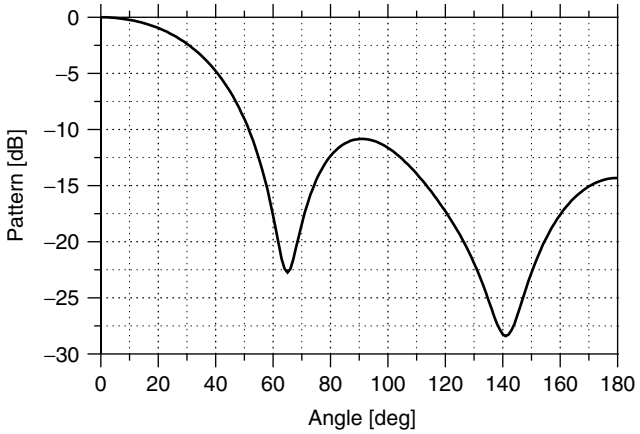


Figure 42. H pattern of the optimized four-element Yagi antenna with $G = 9.15$ dB, $F/B = 14.1$ dB and $Z_{\text{in}} = 36\Omega$.

to cooling processes in thermodynamics. *Simulated annealing* [48] refers to a process used to reveal the low-temperature state of some material. At high temperatures the molecules of a liquid move freely with respect to one another. If the liquid is cooled slowly, the thermal mobility is lost. The atoms are able to line up in a crystal, which represents the minimum-energy state for the system. The time spent at each temperature must be sufficiently long to allow a thermal equilibrium to be realized. If the system is cooled quickly, it does not reach the minimum energy state but one having higher energy.

In optimization by SA we simulate the annealing process by a Monte Carlo method where the global minimum of the objective function represents the low-energy configuration [49,50].

Variations of the simulated annealing process can include parallelization techniques with the use of multiple CPUs [51]. SA has been used in various combinatorial optimization problems.

An example of an array of eight $\lambda/2$ collinear dipoles is presented. The initial array is a uniform array with $d/\lambda = 0.93$ and phase shift $\alpha = 52^\circ$. The main-beam maximum is at $\theta = 99^\circ$. It is observed that in the area $\theta \le \pi/2$, SLL ≥ -10 dB occurs. In practice, this is not desirable. Mobile and radio stations aim at lower upward of horizon. By using the appropriate cost function with the geometry constraints for SLL ≤ -18 dB in $\theta \le \pi/2$, a new array is found. Table 8 and Fig. 43 present the array and the pattern. Applications for wire antenna arrays as well as slot arrays can be found in the literature [52,53].

Table 8. Antenna Array with Collinear Dipoles by Simulated Annealing

Dipole Number	Dipole Position	Dipole Phase (degree)
1	0 λ	0
2	0.70 λ	73
3	1.53 λ	116
4	2.37 λ	155
5	3.23 λ	−172
6	4.05 λ	−130
7	4.92 λ	−93
8	5.62 λ	−19

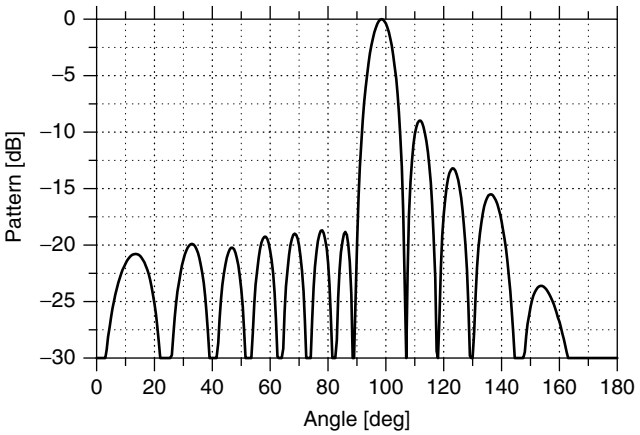


Figure 43. E pattern of $N = 8$ collinear dipole for SLL ≤ -18 dB at $\theta \le \pi/2$ and maximum at $\theta = 99^\circ$.

18. OPTIMIZATION BY GENETIC ALGORITHMS (GAs)

Genetic algorithms (GAs) are global optimizers. GAs [54] follow two main principles: the ability to encode complex structures and the use of simple transformations to improve such structures. GAs are well suited for a wide range of problems in electromagnetics [54]. They have the advantage of quick and easy programming and implementation. They are also suitable for constrained optimization. GAs are based on Darwin's principle [55]: survival of the fittest. The basic idea is an analogy between an individual and a solution on one hand and between an environment and a given problem on the other. The function to be minimized or maximized represents the fitness. This is computed for a given individual and determines how that person "fits" or, in other words, how good this solution for the given problem is.

Many categories of GAs have been designed. A simple GA [56] has nonoverlapping populations. Very popular for electromagnetics are the steady-state GAs with overlapping populations. The best individuals survive to the next generation. Another approach is the deme GA [57], which involves parallel evolving populations with migration.

GAs have been successfully applied in many engineering problems [58–61]. GAs can be applied to thinned arrays. A *thinned array* is a subset of aperiodic arrays. Thinning an array means turning off some elements in a uniformly spaced or periodic array. The OFF elements remain in the array, so the mutual coupling for the interior elements remains the same.

A thinned array offers essentially the same beamwidth with less directivity and fewer elements than does a uniform array of the same size [13,16,20]. The most realistic applications of GAs to array thinning have to do with optimizing the SLL of a large number of elements.

Concluding GAs, an example of an 11-element endfire (case 1) array with SLL $= -20$ dB and HPBW $\cong 72^\circ$ is presented. The elements have the same amplitude. Figure 44

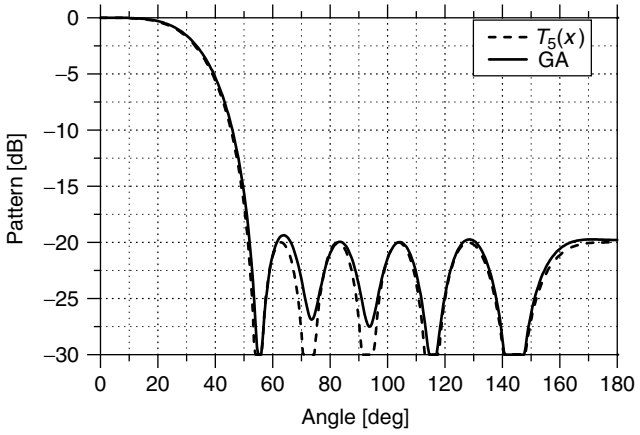


Figure 44. Pattern of 11-element Chebyshev endfire array with SLL = −20 dB, HPBW = 72° (case 1).

Table 9. Position and Phase of an 11-Element Array Able to Create a $T_5(x)$ End-Fire Pattern

Element Position	Distance (λ)	Phase (degree)
1	0.000	0
2	0.400	−140.1
3	0.719	111.8
4	0.855	47.4
5	1.080	−36.4
6	1.300	−108.4
7	1.611	156.8
8	1.860	88.5
9	2.025	10.8
10	2.131	−63.7
11	2.410	−156.5

presents the pattern, and Table 9 gives the position and the phase of the elements.

19. SPACE AND TIME OPTIMIZATION AND SMART ANTENNAS

Antenna arrays combined with signal processing in space and real time are known as "smart" antennas. The low-cost and fast digital processors now available have made possible the implementation of smart antennas. Smart antennas can be used with great success in cellular and satellite mobile communications. They improve the system performance by increasing the spectrum efficiency and the channel capacity. They also extend the range of coverage by multiple-beam steering and electronic compensation of the distortion. Smart antennas can reduce propagation problems such as multipath fading, cochannel interference, and delay spread as well as communication indices such as bit error rate (BER) and outage probability. Their main advantage is the capability to provide a certain channel at a certain direction. This results in *spatial-division multiple access* (SDMA), which performs differently from the frequency (FDMA), the time (TDMA), and the code (CDMA) division multiple accesses.

Smart antennas are known as *adaptive arrays*, *intelligent antennas*, *spatial processing*, *digital beamforming antennas*, and by other terms, [62]. They direct their main-beam maximum to the user while the pattern nulls are in the direction of possible interference [17]. Two main types of beam patterns are available: (1) the switched-beam and (2) the adaptive system. The switched beam divides the communication sector in microsectors. Each microsector contains a predetermined fixed beam pattern. The adaptive systems dynamically alter the patterns to optimize the communications performance. They utilize sophisticated signal processing algorithms [17,63], which update the beam patterns on the basis of changes in both the desired and the interfering signal directions.

Adaptive array theory is based on the optimization methods given before and on the real-time response in a transient environment.

In the literature one can find a lot of special journal issues, books, and specialized research papers in the area of smart antennas [64–69].

Consider a uniform linear array immersed in a homogeneous medium in the far field of M uncorrelated sinusoidal point sources of frequency f_0 (see Fig. 45).

The time difference taken of a plane wave, coming from the ith source in the direction (θ_i, φ_i), to arrive from the kth element to the origin, is

$$r_k = \frac{d}{c}(k-1)\cos\theta_i \tag{132}$$

The signal induced on the first element due to the same source is $m_i(t)e^{j2\pi f_0 t}$. The function $m_i(t)$ depends on the type of the access used:

$$m_i(t) = A_i e^{j\xi_i(t)} \quad \text{(frequency modulation for FDMA)} \tag{133}$$

$$m_i(t) = \sum_n d_i(n)p(t-n\Delta) \quad \text{(TDMA)} \tag{134}$$

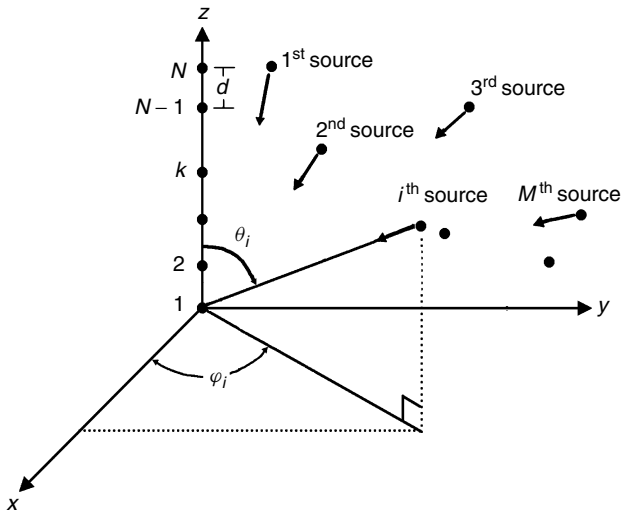


Figure 45. Signal model of a linear array.

where $p(t)$ is the sampling pulse, $d_i(n)$ is the message symbol, and Δ is the sampling interval:

$$m_i(t) = d_i(t)g(t) \quad \text{(CDMA)} \tag{135}$$

where $d_i(t)$ is the message sequence and $g(t)$ is a pseudorandom binary sequence.

The signal induced at the kth element is $m_i(t)e^{j2\pi f_0(t+r_k)}$. It is assumed that the bandwidth of the signal is narrow enough and the array dimensions are small enough for the modulating function $m_i(t)$ to stay almost constant during r_k.

The total signal induced at the kth element due to all sources plus the noise $n_k(t)$ is

$$x_k = \sum_{i=1}^{M} m_i(t)e^{j2pf_0(t+r_k)} + n_k(t) \tag{136}$$

Let us now consider a narrowband beamformer where signals from each element are multiplied by a complex weight and summed to form the array output $y(t)$ (see Fig. 46):

$$y(t) = \sum_{k=1}^{N} w_k^* x_k(t) = [\tilde{w}^*][x(t)] \tag{137}$$

where $[w]$ and $[x(t)]$ are column vectors containing the weights and the inputs of the elements of the array. The values of $[w]$ are determined by using one of the optimization methods.

Smart antennas are analyzed for different network topologies and mobility scenarios. The array geometries

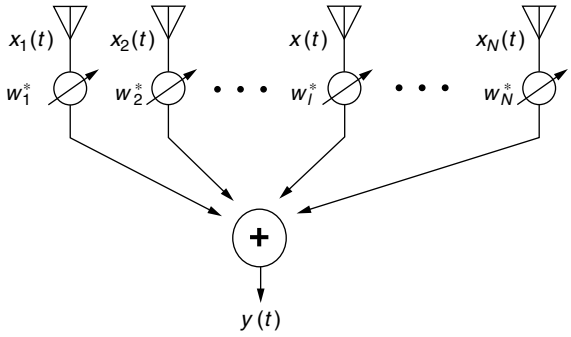


Figure 46. Narrowband beamformer structure.

will be realized with the feed networks and the algorithms for fast beamforming and direction of arrival. A smart antenna system is presented in Fig. 47.

Smart antennas have undergone significant progress since the early 1990s, and their future looks bright. Cost will continue to be the most critical point. It is believed that an explosive development of array processing algorithms within communication systems will appear.

20. ELEMENT PATTERN AND MUTUAL COUPLING

Analysis and synthesis of antenna arrays are given for array elements with known current or aperture field characteristics. It was assumed that these characteristics are proportional to the excitations, the same for similar elements, and unchanged as the array is scanned. In general, all the currents and fields differ in magnitude,

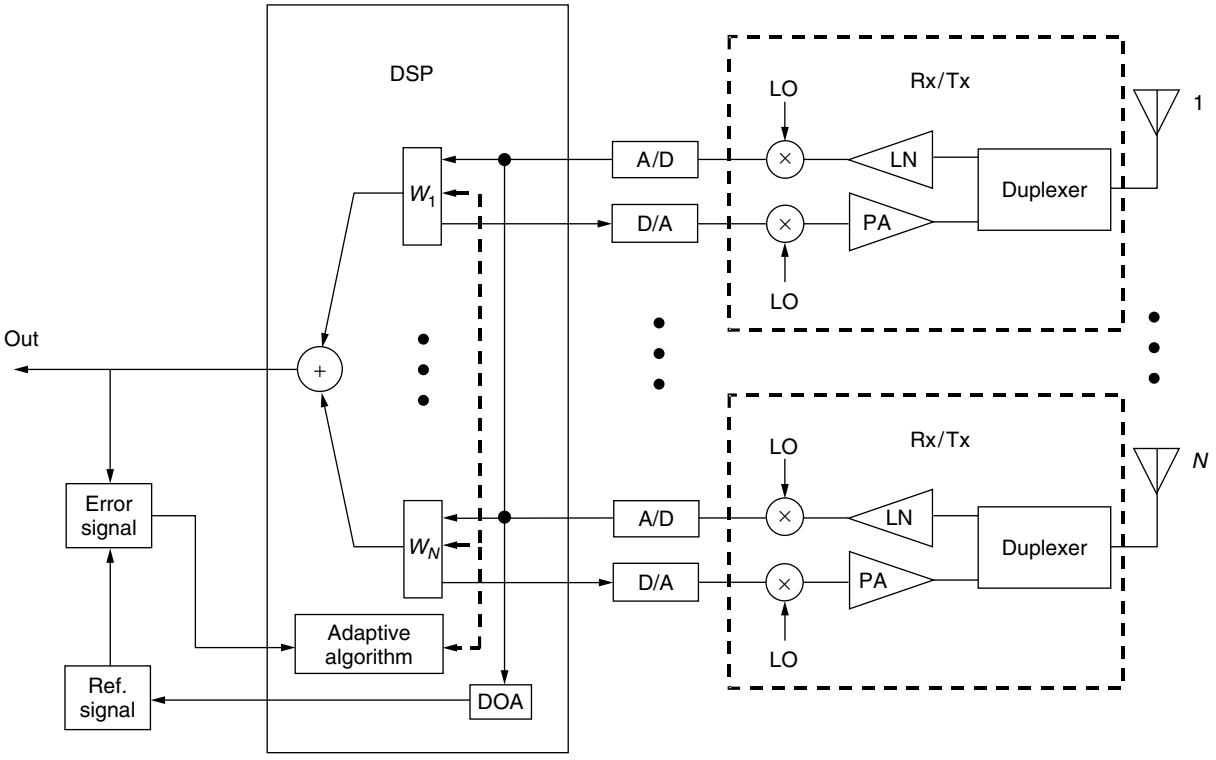


Figure 47. Diagram of a smart antenna system.

phase, and distribution from element to element. The differences depend on the frequency and the scan angle as well as on the geometry of the array. That happens because mutual coupling plays an important role in the behavior of the elements. Actually the radiated field can be expressed as generalized integrals that include the appropriate distributions over the radiating antenna elements and nearby diffracting bodies. The array characteristics are dominated by the mutual coupling between the elements.

Mutual coupling alters mainly the amplitudes and phases between various elements while the currents or aperture distributions remain very similar. In the antenna array synthesis the required distributions can be found by using different methods depending on the problem. Among these, the boundary-value, the transmission line, and the Poynting vector methods are the main ones [1,3,6]. In the late 1960s the integral equations with suitable numerical solutions were successively applied. The numerical techniques are collectively referred to as the *method of moments* (MoM), [6,41,70]. This method is simple and versatile and requires fast and large amounts of computation. The speed and storage capacity of the computer characterize the limitation of the method.

There are several forms of integral equations. For the time-harmonic EM fields, the electric (EFIE) and the magnetic field (MFIE) integral equations are popular [71]. The EFIE enforces the electric field, while the MFIE enforces the magnetic field boundary condition. MoM reduces the integral equations to a system of simultaneous linear algebraic ones in terms of the unknown current or aperture distribution. For radiation problems, especially for wire antennas, there are popular integral equations as the Pocklington, the Hallen, and the reaction integral equations. There are computer codes for the evaluation of the radiation characteristics of antennas. They make use of the abovementioned equations and compute the appropriate quantities in the near and far fields.

20.1. Finite and Infinite Arrays

Let a wire structure (Fig. 48) be composed of straight segments of circular cross section. For electrically thin wires it was found that the total current (conduction plus displacement) on the structure can be found by using one of the electric field integral equations [70,71]. The current on the wires is expanded in a finite series as follows:

$$I(\ell) = \sum_{n=1}^{N} I_n F_n(\ell) \tag{138}$$

where $F_n(\ell)$ are the current expansion functions.

Substituting (138) into the integral equation used, the following system of simultaneous linear algebraic equations yields

$$\sum_{n=1}^{N} I_n Z_{mn} = V_m \quad m = 1, 2, 3, \ldots, N \tag{139}$$

where V_m are the applied voltages and I_n are the complex amplitudes of the current distribution. The elements Z_{mn} are the mutual impedance elements. Equation (139) can be expressed in matrix form:

$$[Z][I] = [V] \tag{140}$$

The only nonzero elements V_m of $[V]$ are these where a generator is at the terminals of the mth segment.

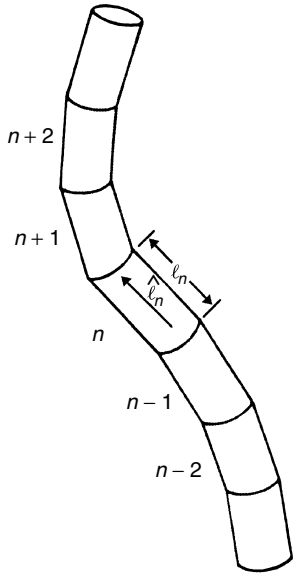


Figure 48. Wire structure of straight segments.

Let us assume that a −20-dB Chebyshev broadside array with five parallel $\lambda/2$ dipoles is desired. Table 10 shows the required currents and the corresponding voltages at the main ports for equal spacing $d = 0.25\lambda$ and $d = 0.5\lambda$. The resulting pattern for $d = 0.25\lambda$ is shown in Fig. 49. If we suppose that there is no coupling between the elements, then the currents have the same relative values as the voltages. In this case the resulting pattern is also given in Fig. 49 and is very different from the desired.

In addition to the numerical methods, one could use measured data to evaluate the mutual coupling effects. In this case the currents and voltages can be found by using one of the classical methods of synthesis [72].

The prediction of element impedance as a function of scan and element patterns in an infinite array is very different in a finite one. Elements away from the edge of large finite arrays have approximately the same characteristics to these of infinite arrays. So, the study of infinite arrays has a practical aim.

Table 10. Relative Input Currents and Voltages for a Five-Element Chebyshev Broadside Array with SLL = −20 dB

	I_i		V_i	
Element	$d/\lambda = 0.25$	$d/\lambda = 0.5$	$d/\lambda = 0.25$	$d/\lambda = 0.5$
1, 5	1	1	$1\angle 0°$	$1\angle 0°$
2, 4	−1.194	1.608	$1.902\angle 251°$	$1.383\angle 341°$
3	2.178	1.932	$3.027\angle 32°$	$1.804\angle -6°$

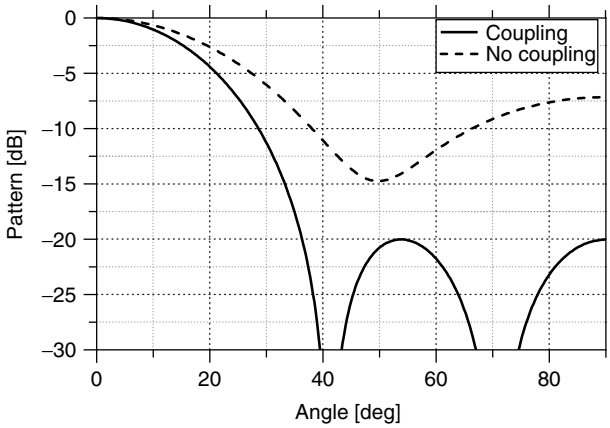


Figure 49. Pattern of 5 parallel $\lambda/2$ dipole linear array with $d = 0.25\lambda$.

In infinite arrays a wave-type formulation or a mode-matching approach with a direct solution of the differential equations can be used. Any of the abovementioned approaches is based on the periodic nature of the fields.

Infinite array theory is a good approximation of the impedance behavior of central elements in large arrays [6,13,16].

21. ARRAY FEEDS

Linear arrays or assemblies thereof making planar arrays are the most usual fixed-beam arrays. Using linear arrays as the building blocks, appropriate feed networks are developed. Two kinds of feeds are more usual: the "series" and the "shunt" ones.

In the series feed the elements of the array are in series along the transmission line. Similarly, in the shunt feed the elements are in parallel with the line or the network. Feeds must offer an acceptable in-band performance in relatively modest cost. The feed choice depends on the application as well as on its physical, processing, and electrical properties. The weight with the conformity and the material used characterize the physical properties. The fabrication and the availability of the materials define the processing properties. Finally the losses, the shielding, the design ability, and the performance over a specified bandwidth characterize the electrical properties. The ability of the array feed to control the power distribution allows the antenna engineer to meet the appropriate requirements.

A critical function of a feed network is that of impedance matching. By matching the impedances as closely as possible at each portion of the network, the reflection coefficients and therefore the VSWR of the feed is kept to with certain levels.

The feed network must keep the isolation between outputs. This means that any energy entering in the ith output port should not reappear at any other via the network.

An array feed should have the ability to steer its main beam. This is accomplished by using discrete phase shifters and attenuators located between the outputs of the feed network and the elements of the array.

The most common shunt feed is the corporate one (Fig. 50). A series feed can be constructed by using the transmission line junctions (Fig. 51). Multiple-beam feeds are made by series-fed beamforming networks (Fig. 52) or by parallel feed as the Butler matrix (Fig. 53). Planar arrays use arrangement of series-series or series–parallel topologies [73–79].

Finally, optical hardware for the care and feeding of an array can be used. An extended analysis of photonic feed systems can be found in the literature [80,81].

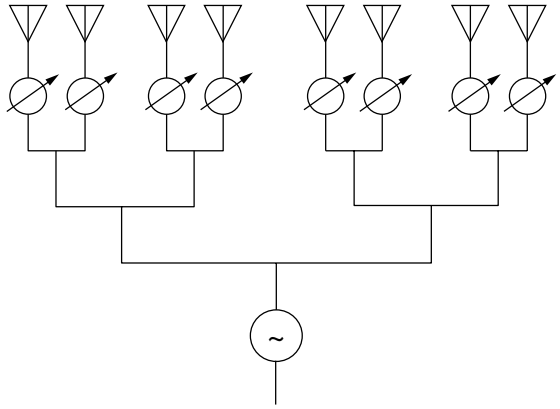


Figure 50. Parallel corporate feed.

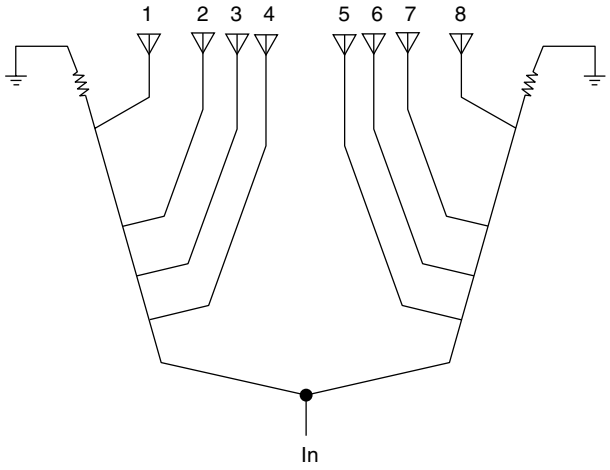


Figure 51. Series feed with transmission line junctions.

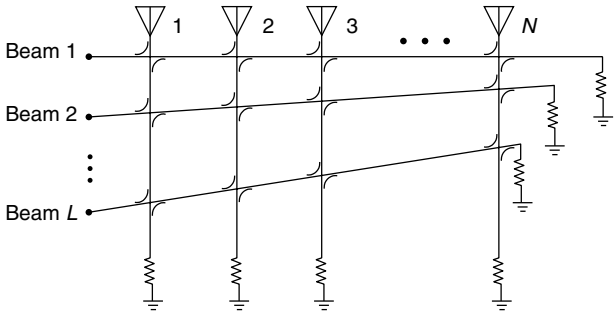


Figure 52. Series-fed beamforming network.

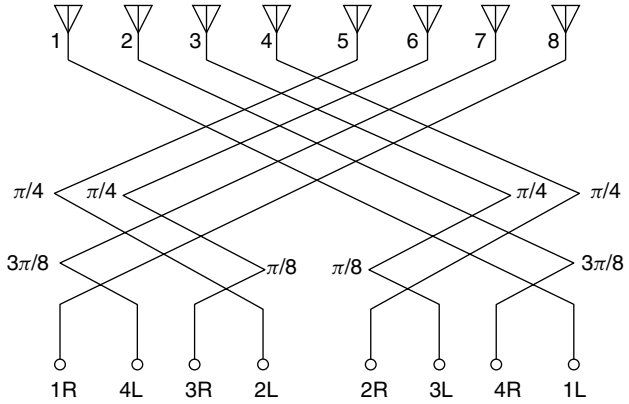


Figure 53. Eight beams and elements Butler matrix.

BIOGRAPHY

John N. Sahalos received his B.Sc. degree in physics and the Diploma in civil engineering from the University of Thessaloniki, Greece, in 1967 and 1975, respectively. He also received the Diploma of postgraduate studies in electronics in 1975 and a Ph.D. in electromagnetics in 1974. During 1976 he was with Electroscience Laboratory, the Ohio State University, Columbus, as a postdoctoral university fellow. From 1977–1985 he was a professor in the Electrical Engineering Department, University of Thrace, Greece. Since 1985 he has been a professor at the School of Science, University of Thessaloniki, Greece. During 1982 and 1989, he was a visiting professor at the University of Colorado, Boulder, and at the Universidad Politecnica de Madrid, Spain, correspondingly. He is the author of three books and more than 200 articles published in the scientific literature. His research interests are in the area of applied electromagnetics, antennas, high-frequency methods, communications, microwaves, and biomedical engineering.

Dr. Sahalos is a professional engineer and a consultant to industry. He is on the editorial board of two scientific journals. Since 1999 he has been the president of the Greek URSI committees, is a member of five IEEE Societies, and a member of both the New York Academy of Science and the Technical Chamber of Greece.

BIBLIOGRAPHY

1. S. A. Schelkunoff and H. T. Friis, *Antenna Theory and Practice*, Wiley, New York, 1952.
2. R. W. P. King, *The Theory of Linear Antennas*, Harvard Univ. Press, Cambridge, MA, 1956.
3. J. D. Kraus, *Antennas*, McGraw-Hill, New York, 1988.
4. R. C. Hansen, ed., *Microwave Scanning Antennas*, Academic Press, New York, Vol. I, 1964; Vols. II, III, 1966. (Peninsula Publishing, 1985).
5. R. S. Elliot, *Antenna Theory and Design*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
6. C. A. Balanis, *Antenna Theory*, Analysis and Design, Wiley, New York, 1997.
7. T. Milligan, *Modern Antenna Design*, McGraw-Hill, New York, 1985.
8. N. Amitay, V. Galindo, and C. P. Wu, *Theory and Analysis of Phased Arrays*, Wiley-Interscience, New York, 1972.
9. M. T. Ma, *Theory and Applications of Antenna Arrays*, Wiley-Interscience, New York, 1974.
10. A. W. Rudge, K. Milne, A. D. Olver, and P. Knight, eds., *The Handbook of Antenna Design*, IEE/Peregrinus, London, 1983.
11. Y. T. Lo and S. W. Lee, *Antenna Handbook*, Van Nostrand Reinhold, New York, 1988.
12. R. C. Johnson and H. Jasik, *Antenna Engineering Handbook*, McGraw-Hill, New York, 1993.
13. R. C. Mailloux, *Phased Array Antenna Handbook*, Artech House, Norwood, MA, 1994.
14. J. R. James and P. S. Hall, eds., *Handbook of Microstrip Antennas*, Vols. I, II, IEE/Peter Peregrinus, London, 1989.
15. N. Fourikis, *Phased Array-Based Systems and Applications*, Wiley-Interscience, New York, 1997.
16. R. C. Hansen, *Phased Array Antennas*, Wiley-Interscience, New York, 1998.
17. R. T. Compton, Jr., *Adaptive Antennas*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
18. T. S. Rappaport, ed., *Smart Antennas*, IEEE Press, 1998.
19. G. V. Tsoulos, ed., *Adaptive Antennas for Wireless Communications*, IEEE Press, 2001.
20. Y. Rahmat-Samii and E. Michielssen, *Electromagnetic Optimization by Genetic Algorithms*, Wiley-Interscience, New York, 1999.
21. M. Abramowitz and L. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.
22. C. L. Dolph, A current distribution for broadside arrays which optimizes the relationship between beam width and side-lobe level, *Proc. IRE* **34**: 335–338 (1946).
23. H. J. Riblet, Discussion on a current distribution for broadside arrays which optimizes the relationship between beam width and side-lobe level, *Proc. IRE* **35**: 489–492 (1947).
24. G. Miaris, M. Chryssomalis, E. Vafiadis, and J. N. Sahalos, A unified formulation for Chebyshev and Legendre superdirective end-fire array design, *Archiv Elektrotechnik* **78**(4): 271–280 (1995).
25. W. H. Kummer, general ed., *IEEE Trans. Antennas Propag.* (Special Issue on Conformal Arrays), **AP-22**(1) (1974).
26. R. S. Elliot, On discretizing continuous aperture distributions, *IEEE Trans. Antennas Propag.* **AP-25**(5): 617–621 (1977).
27. T. T. Taylor, Design of line source antennas for narrow beamwidth and low sidelobes, *IEEE Trans. Antennas Propag.* **AP-3**: 16–28 (1955).
28. A. T. Villeneuve, Taylor patterns for discrete arrays, *IEEE Trans. Antennas Propag.* **AP-32**(10): 1089–1093 (1984).
29. R. C. Hansen, Linear arrays, in A. Rudge, ed., *Handbook of Antenna Design*, Vol. 2, Peter Peregrinus, London, 1983, Chap. 9.
30. E. T. Bayliss, Design of monopulse antenna difference patterns with low sidelobes, *Bell Syst. Tech. J.* **47**: 623–650 (1968).
31. D. K. Cheng, *Analysis of Linear Systems*, Addison-Wesley, Reading, MA, 1959.

32. S. R. Laxpatti, Planar array synthesis with prescribed pattern nulls, *IEEE Trans. Antennas Propag.* **AP-30**(6): 1176–1183 (1982).

33. J. N. Sahalos, The orthogonal method of nonuniformly spaced arrays, *Proc. IEEE* **62**: 281 (1974).

34. H. Unz, Nonuniformly spaced arrays: The orthogonal method, *Proc. IEEE* **54**: 53–54 (1966).

35. J. N. Sahalos, A solution of nonuniformly linear array with the help of the Chebyshev polynomials, *IEEE Trans. Antennas Propag.* **AP-24**: 109–112 (1976).

36. J. N. Sahalos, K. Melidis, and S. Lampou, On the optimum directivity of general nonuniformly spaced broadside arrays of dipoles, *Proc. IEEE* **64**: 1706–1709 (1974).

37. J. N. Sahalos, A solution of the general nonuniformly spaced antenna array, *Proc. IEEE* **64**: 1292–1294 (1976).

38. G. Miaris and J. N. Sahalos, The orthogonal method for the geometry synthesis of a linear antenna array, *IEEE-AP Mag.* **41**(1): 96–99 (1999).

39. S. Goudos, G. Miaris, and J. N. Sahalos, On the quantized excitation and the geometry synthesis of a linear array by the orthogonal method, *IEEE Trans. Antennas Propag.* **AP-49**(2): 298–305 (2001).

40. B. D. Popovic, M. B. Dragovic, and A. R. Djordjevic, *Analysis and Synthesis of Wire Antennas*, Research Studies Press, Wiley, New York, 1982.

41. R. F. Harrington, *Field Computations by Moment Method*, IEEE Press, New York, 1993.

42. Y. T. Lo, S. W. Lee, and Q. H. Lee, Optimization of directivity and SNR of an arbitrary antenna array, *Proc. IEEE* **54**(8): 1033–1045 (1966).

43. L. P. Winkler and M. Schwartz, A fast numerical method for determining the optimum SNR of an array subject to a Q factor constraint, *IEEE Trans. Antennas Propag.* **AP-20**(4): 503–505 (1972).

44. P. Zimourtopoulos and J. N. Sahalos, On the gain maximization of the dual frequency and direction array consisting of wire antennas, *IEEE Trans. Antennas Propag.* **AP-33**: 874–880 (1985).

45. J. A. Nelder and R. Mead, A simplex method for function minimization, *Comput. J.* **7**: 308–313 (1965).

46. S. L. S. Jacoby, J. S. Kowalik, and J. T. Pizzo, *Iterative Methods for Nonlinear Optimization Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

47. P. R. Abdy and M. A. H. Dempster, *Introduction to Optimization Methods*, Chapman & Hall, London, 1974.

48. N. Metropolis, A. W. Rosenbluth, A. H. Teller, and E. Teller, Equations of state calculations by fast computing machines, *J. Chem. Phy.* **21**: 1087–1091 (1953).

49. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge Univ. Press, 1992.

50. A. Torn and A. Zilinskas, *Global Optimization, Lecture Notes in Computer Science*, Springer-Verlag, 1987.

51. R. Azencott, *Simulated Annealing Parallelization Techniques*, Willey, New York, 1992.

52. Z. Zaharis, E. Vafiadis, and J. N. Sahalos, On the design of a dual-band base station wire antenna, *IEEE Antennas Propag. Mag.* **42**(6): 144–151 (2000).

53. K. Kechagias, E. Vafiadis, and J. N. Sahalos, On the RLSA antenna optimum design for DBS reception, *IEEE Trans. Broadcast.* **44**(4): 460–469 (1998).

54. Y. Rahmat-Samii and E. Michielssen, *Electromagnetic Optimization by Genetic Algorithms*, Willey, New York, 1999.

55. C. Darwin, *On the Origin of Species*, John Murray, London, 1859.

56. D. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989.

57. M. Wall, *GAlib: A C++ Library of Genetic Algorithm Components*, Version 2.4, Document Revision B, MIT, 1996.

58. J. H. Holland, *Adaptation in Natural and Artificial Systems*, Univ. Michigan Press, 1975.

59. M. Mitchell, *An Introduction to Genetic Algorithms*, 2nd Pr., MIT Press, 1996.

60. L. Chambers, *Practical Handbook of Genetic Algorithms*, Vol. I, *Applications*, CRC Press, Boca Raton, FL, 1995.

61. K. F. Man, K. S. Tang, and S. Kwong, *Genetic Algorithms*, Springer-Verlag, London, 1999.

62. M. Chryssomallis, Smart antennas, *IEEE Antennas Propag. Mag.* **42**(3): 129–136 (2000).

63. J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, 2nd edn., Macmillan, New York, 1992.

64. R. Schreiber, Implementation of adaptive array algorithms, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**: 1038–1045 (1986).

65. S. Choi and T. K. Sarkar, Adaptive antenna array utilizing the conjugate gradient method for multipath mobile communications, *Signal Process.* **29**: 319–333 (1992).

66. A. El Zooghby, C. G. Christodoulou, and M. Georgiopoulos, Neural network-based adaptive beamforming for one and two dimensional antenna arrays, *IEEE Trans. Antennas Propag.* **AP-46** (1998).

67. Th. S. Rappaport, ed., *Smart Antennas: Adaptive Arrays, Algorithms and Wireless Position Location*, IEEE Press, 1998.

68. L. C. Godara, Application of antenna arrays to mobile communications, Part I: Performance improvement, feasibility and system considerations, *Proc. IEEE* **85**(7): 1031–1060 (1997).

69. L. C. Godara, Application of antenna arrays to mobile communications, Part II: Beam-forming and direction-of-arrival considerations, *Proc. IEEE* **85**(8): 1195–1245 (1998).

70. W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, Wiley, New York, 1998.

71. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.

72. H. Steyskal and J. S. Herd, Mutual coupling compensation in small array antennas, *IEEE Trans. Antennas Propag.* **AP-38**(12): 1971–1975 (1990).

73. L. Young, *Parallel Coupled Lines and Directional Couplers*, Artech House, Norwood, MA, 1992.

74. R. S. Elliot, An improved design procedure for small arrays of shunt slots, *IEEE Trans. Antennas Propag.* **AP-32**: 48–53 (1983).

75. R. C. Hansen and G. Brunner, Dipole mutual impedance for design of slot arrays, *Microwave J.* **22**: 54–56 (1979).

76. N. A. Begovich, Frequency scanning, in R. C. Hansen, ed., *Microwave Scanning Antennas*, Vol. III, Peninsula Publishing, 1983, Chap. 2.

77. J. S. Ajioka, Frequency-scan antennas, in R. C. Johnson, ed., *Antenna Engineering Handbook*, McGraw-Hill, New York, 1993, Chap. 19.

78. J. L. Butler, Digital, Matrix and intermediate frequency scanning, in R. C. Hansen, ed., *Microwave Scanning Antennas*, Peninsula Publishing, 1985, Chap. 3.

79. J. R. James and P. S. Hall, *Handbook of Microstrip Antennas*, Vols. 1, 2, IEE, Peter Peregrinus, London, 1989.

80. H. Zmuda and E. N. Toughlian, eds., *Photonic Aspects of Modern Radar*, Artech House, Norwood, MA, 1994.

81. A. Kumar, *Antenna Design with Fiber Optics*, Artech House, Norwood, MA, 1996.

ANTENNA MODELING TECHNIQUES

JOHN L. VOLAKIS
University of Michigan
Ann Arbor, Michigan

THOMAS F. EIBERT
T-Systems Nova GmbH
Technologiezentrum
Darmstadt, Germany

1. INTRODUCTION

Antennas are key components in any wireless communication system [1,2]. They are the devices that allow for the transfer of a signal (in a wired system) to waves that in turn propagate through space and can be received by another antenna. The receiving antenna is responsible for the reciprocal process: that of turning an electromagnetic wave into a signal or voltage at its terminals that can subsequently be processed by the receiver. The receiving and transmitting functionalities of the antenna structure itself are fully characterized by Maxwell's equations [3] and are fairly well understood. The dipole antenna (a straight wire fed at the center by a 2-wire transmission line) was the first antenna ever used and is also one of the best understood [1,2]. For effective reception and transmission, it must be approximately $\lambda/2$ long (λ = wavelength) at the frequency of operation (or multiples of this length). Thus, it must be fairly long when used at low frequencies ($\lambda = 1$ m at 300 MHz), and even at higher frequencies (UHF and above), its protruding nature makes it quite undesirable. Further, its low gain (2.15 dB), lack of directionality, and extremely narrow bandwidth make it even less attractive. Not surprisingly, the Yagi–Uda antenna (typically seen on the roof of most houses for television reception) was considered a breakthrough in antenna technology when introduced in the early 1920s because of its much higher gain of 8–14 dB. Log periodic wire antennas introduced in the late 1950s and 1960s and wire spirals allowed for both gain and bandwidth increases. On the other hand, high-gain antennas even today rely on large reflectors (dish antennas) and waveguide arrays [used for airborne/warning and control system (AWACS) radar] that are expensive and cumbersome to deploy.

Until the late 1970s, antenna design was based primarily on practical approaches using off-the-shelf antennas such as various wire geometries (dipoles, Yagi–Uda, log periodics, spirals), horns, reflectors and slots/apertures as well as arrays of some of these. The antenna engineer could choose or modify one of them based on design requirements that characterize antennas such as gain, input impedance, bandwidth, pattern beamwidth, and sidelobe levels (see, e.g., Refs. 1 and 2 or any of the several antenna textbooks for a description of these quantities). Antenna development required extensive testing and experimentation and was therefore funded primarily by the governments. However, more recently, dramatic growth in computing speeds and development of effective computational techniques [4–6] for realistic antenna geometries has allowed for low-cost virtual antenna design. Undoubtedly the explosive growth of wireless communications and microwave sensors, microwave imaging needs and radars has been the catalyst for introducing a multitude of new antenna designs since 1990 and an insatiable desire for using modern computational techniques for low cost designs. Requirements for conformal (nonprotruding) antennas for airborne systems, increased bandwidth requirements, and multifunctionality have led to heavy exploitation of printed (patch) or other slot-type antennas [7] and the use of powerful computational tools (commercial and noncommercial) for designing such antennas (see Fig. 1) [8]. The accuracy of these techniques is also remarkable, as seen by the results shown in Fig. 1 [9] for a cavity-backed slot spiral antenna. Needless to mention, the commercial mobile communications industry has been the catalyst for the recent explosive growth in antenna design needs. Certainly, the 1990s have seen an extensive use of antennas by the public for cellular, GPS, satellite, wireless LAN for computers, upcoming Bluetooth technology, and so on. However, future needs will be even greater when a multitude of antennas will be integrated into automobiles for all sorts of communication needs. Such antennas must be designed with the platform in mind (see Fig. 2) and must therefore satisfy gain and pattern requirements in the presence of the platform. Concurrent modeling of the large structure is therefore needed, resulting in a substantial increase of computational requirements for analysis and design purposes. For military applications, there is an increasing need for multifunctional antennas that can satisfy a plethora of communications needs using a single aperture as small as possible. Such apertures are also intended for unmanned airborne vehicles (UAVs) and small general aviation vehicles where real estate is even more limited. The multitude of design requirements for such antennas implies use of fast computational tools as well as optimization methods to arrive at designs that satisfy the specific mission or product needs.

In this article we summarize the most popular antenna analysis methods. These can be subdivided into time-domain and frequency-domain methods. *Time-domain methods* are appropriate for broadband analysis (many frequencies), and among them the finite-difference–time domain method [5,10] is the most popular technique. Time-domain–integral equation methods [11–13] are gaining attention by combining them with fast algorithms. However, in general, time-domain approaches are still slow and seldom attractive for narrowband analysis and design.

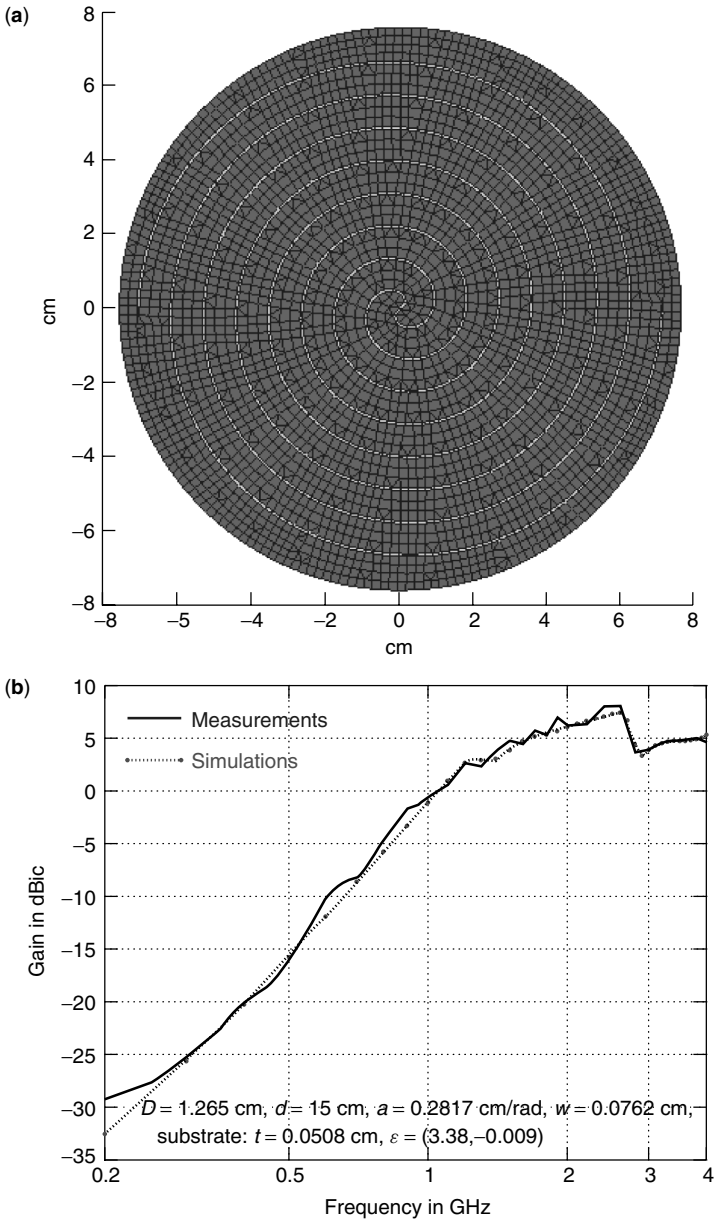


Figure 1. (**a**) Computational grid for a slot spiral antenna. (**b**) Comparison of gain calculations with measurements for slot spiral antennas [9]. The slot spiral is situated on the aperture of a circular cavity of diameter $d = 15$ cm and depth $D = 1.265$ cm. The slot spiral is a metal surface residing on a dielectric substrate ($\varepsilon_r = 3.38 - j0.009$ and of thickness $t = 0.0508$ cm) with the shown slot imprint of width $w = 0.0762$ cm.

In this article, we will discuss *frequency-domain methods* for antenna analysis. Like the time-domain methods, frequency domain approaches [14] can be categorized under (1) integral, (2) differential, and (3) hybrid techniques. The popular finite-element (FE) method [6,15] used in most branches of engineering belongs to the second category, and the ensuing procedure entails a direct solution of Maxwell's equations. Differential or FE methods are the choice modeling techniques for finite or inhomogeneous dielectric regions. In contrast, integral methods [4] are the choice techniques for modeling metallic structures situated in free space or on thin substrates (layers of dielectric). As can be understood, hybrid techniques involve a suitable combination of finite-element and integral or other modeling methods, including high-frequency techniques. The latter were actually the first to be used for accurate analysis of reflector and horn antennas [16] and for predicting antenna interactions on complex platforms such as aircraft [17], and work in this area continues to be explored [18]. More recently the combination of integral and FE methods [referred to as *finite-element–boundary integral* (FEBI) *methods*] has been successful for modeling complex antenna geometries constructed of metallic and nonmetallic materials [19]. The recent introduction of fast methods [20–22] has played an important role in the use of hybrid FEBI methods for design [23].

Below we proceed to discuss details associated with the implementation of integral and FE methods after we first present some basic electromagnetic concepts.

3. SOME BASIC EQUATIONS

Electromagnetic phenomena are governed by Maxwell's equations, a system of coupled time space partial differential equations established in the nineteenth century [3].

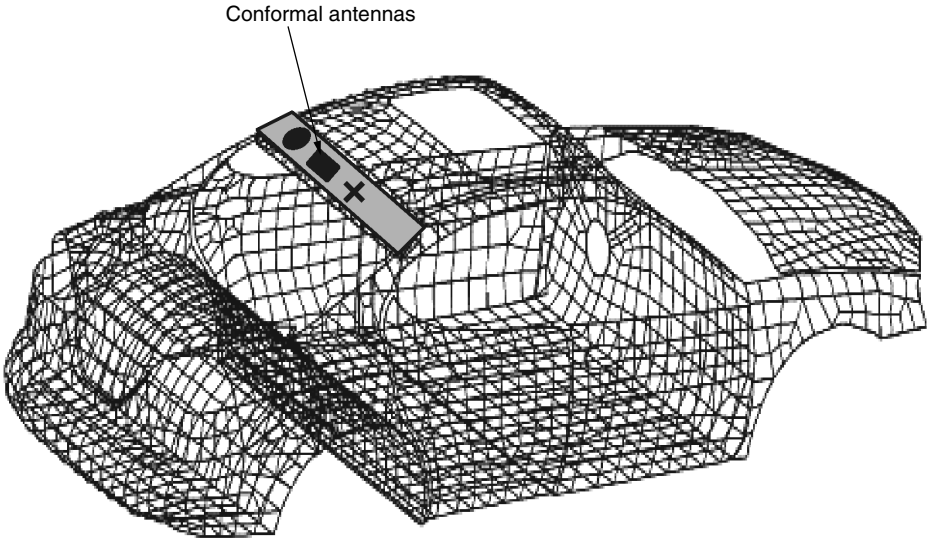


Figure 2. Quadrilateral grid used for modeling the automobile surface for onboard antenna evaluations. Antenna radiation can be introduced using the antenna aperture fields or currents computed for the isolated antenna.

Although Maxwell's equations allow for very general field variations in both space and time, for simplicity we shall consider only time-harmonic fields where the $e^{+j\omega t}$ time dependence is assumed and suppressed. In addition, we shall consider only fields in a linear and isotropic medium. With these assumptions, Maxwell's equations in differential (point) form can be written as

$$\nabla \times \mathbf{E} = -\boldsymbol{M}_i - jkZ\boldsymbol{H} \tag{1}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}_i + jkY\boldsymbol{E} \tag{2}$$

where $\boldsymbol{H}$ is the magnetic field in amperes per meter (A/m), $\boldsymbol{E}$ is the electric field in volts per meter (V/m), $\boldsymbol{J}_i$ is the impressed electric current (source antenna radiating current), and $\boldsymbol{M}_i$ is a fictitious magnetic current (source) often used for mathematical convenience. The radiating medium is completely described by its intrinsic impedance $Z = 1/Y = (\mu/\varepsilon)^{1/2}$, where ε and μ denote the medium's permittivity and permeability, respectively. The permittivity characterizes the medium's response in the presence of an electric field whereas the permeability is associated with the magnetic field. The wavenumber is denoted by $k = \dfrac{2\pi}{\lambda} = \omega(\mu\varepsilon)^{1/2}$, where λ is the wavelength and ω is the corresponding angular frequency. The faraday (1) and Ampère–Maxwell laws (2) are independent first-order vector equations. To solve for $\boldsymbol{E}$ and $\boldsymbol{H}$, we typically combine (1) and (2) to obtain the vector wave equation

$$\nabla \times \nabla \times \begin{Bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{Bmatrix} - k^2 \begin{Bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{Bmatrix} = -j\omega \begin{Bmatrix} \mu \boldsymbol{J}_i \\ \varepsilon \boldsymbol{M}_i \end{Bmatrix} \mp \nabla \times \begin{Bmatrix} \boldsymbol{M}_i \\ \boldsymbol{J}_i \end{Bmatrix} \tag{3}$$

The wave nature of $\boldsymbol{E}$ and $\boldsymbol{H}$ is easily surmised when we introduce the identity $\nabla \times \nabla \times \boldsymbol{E} = -\nabla^2 \boldsymbol{E} + \nabla(\nabla \cdot \boldsymbol{E}) = -\nabla^2 \boldsymbol{E}$, where we have assumed that $\nabla \cdot \boldsymbol{E} = 0$, true away from the source region (away from the antenna). With this replacement, the vector wave equation reduces to scalar wave or rather Helmholtz equations of the form $\nabla^2 E_\xi + k^2 E_\xi = f_i$, where E_ξ implies the ξth component of the vector field and f_i is the appropriate right-hand side reduced from (3). A solution of the wave equation can be accomplished provided the boundary conditions are enforced on canonical surfaces (spheres, cubes, infinite cylinders, and planes). For practical problems, however, boundary conditions must be enforced on noncanonical surfaces, and consequently a closed-form solution is not possible. This is the reason for resorting to a numerical solution of Maxwell's equations.

When dealing with arbitrary antenna structures, it is customary to invoke the surface equivalence principle [24], as illustrated in Fig. 3. The antenna itself is enclosed in a mathematical surface S, and equivalent or mathematical surface currents $\boldsymbol{J}_s$ and $\boldsymbol{M}_s$ are introduced on that surface. When $(\boldsymbol{J}_i, \boldsymbol{M}_i)$ in (1) and (2) are replaced by $(\boldsymbol{J}_s, \boldsymbol{M}_s)$, it follows that an integral representation of the fields everywhere is

$$\boldsymbol{E}(\boldsymbol{r}) = -jkZ \iint_S \left[\boldsymbol{J}_s(\boldsymbol{r}') G(\boldsymbol{r}, \boldsymbol{r}') + \frac{1}{k^2} (\nabla_s' \cdot \boldsymbol{J}_s(\boldsymbol{r}')) \nabla G(\boldsymbol{r}, \boldsymbol{r}') \right] ds' + \iint_S \left[\boldsymbol{M}_s(\boldsymbol{r}') \times \nabla G(\boldsymbol{r}, \boldsymbol{r}') \right] ds' + \boldsymbol{E}^{\text{inc}} = \boldsymbol{E}^{\text{rad}} + \boldsymbol{E}^{\text{inc}} \tag{4}$$

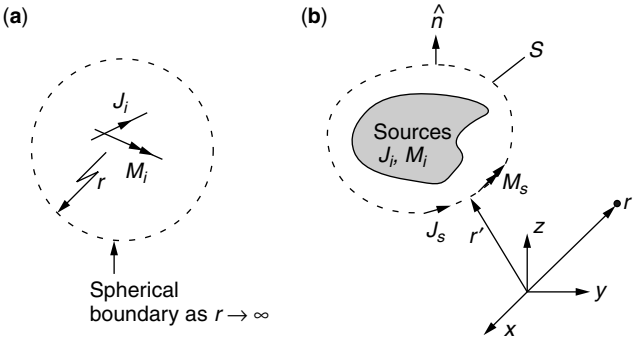


Figure 3. (**a**) Source currents enclosed by a spherical boundary. (**b**) Illustration of surface equivalence where the source current or field is replaced by equivalent surface currents located on a surface enclosing the antenna/source; $(\boldsymbol{J}_s, \boldsymbol{M}_s)$ can be found by enforcing the boundary conditions on the antenna structure.

where $\boldsymbol{r}$ and $\boldsymbol{r}'$ refer to the observation and source (i.e., $\boldsymbol{J}_s$ and $\boldsymbol{M}_s$) position vectors, respectively. Here, $\nabla_s \cdot$ denotes the surface divergence operator [25] and the equivalent surface current densities ($\boldsymbol{J}_s, \boldsymbol{M}_s$) are related to the surface fields via the relations

$$\begin{aligned} \boldsymbol{J}_s &= \hat{n} \times \boldsymbol{H} \\ \boldsymbol{M}_s &= \boldsymbol{E} \times \hat{n} \end{aligned} \tag{5}$$

where $\hat{n}$ denotes the unit normal to the integration surface S. With this identification, (4) becomes the Stratton–Chu equation [26]. Also, $\boldsymbol{E}^{\text{inc}}$ is the incident or excitation field intensity and is nonzero for radar scattering problems, but typically set to zero for antenna analysis. Further, G is the scalar Green function[1] given by

$$G(\boldsymbol{r}, \boldsymbol{r}') = \frac{1}{4\pi} \frac{e^{-jk|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|} \tag{6}$$

This Green function also incorporates the radiation condition stating that outgoing waves are of the form e^{-jkr}/r as $r \to \infty$. In mathematical form, the corresponding boundary condition is

$$\lim_{r\to\infty} r[\nabla \times \boldsymbol{E} + jk\hat{r} \times \boldsymbol{E}] = 0$$

a necessary condition for the unique solution of (3).

The analysis of antennas using integral equation methods amounts to finding ($\boldsymbol{J}_s, \boldsymbol{M}_s$) or some other predefined currents subject to boundary conditions satisfied by the antenna structure. To illustrate the implementation of integral equation methods, we refer to Fig. 4, showing a reflector and a patch antenna [24]. For the reflector antenna, on application of the equivalence principle (here the closed surface S is collapsed onto the reflector surface and $\boldsymbol{J}_s = \boldsymbol{J}_s^+ - \boldsymbol{J}_s^-$ is the net current), the reflector is removed and replaced by the equivalent current $\boldsymbol{J}_s$. There is no need for magnetic currents since $\hat{n} \times \boldsymbol{E} = 0$ on metallic surfaces in accordance with (5). The unknown $\boldsymbol{J}_s$ can subsequently be found by solving the integral equation

$$\hat{n} \times [\boldsymbol{E}^{\text{rad}}(\boldsymbol{r}) + \boldsymbol{E}^{\text{inc}}(\boldsymbol{r})] = 0, \qquad \boldsymbol{r} \in S \tag{7}$$

valid for $\boldsymbol{r}$ on the reflector's surface, where $\boldsymbol{E}^{\text{inc}}$ is the excitation field from the reflector feed. In the case of the patch antenna structure, the $\boldsymbol{M}_s$ current is also introduced across the antenna aperture where $\boldsymbol{E}$ is nonzero. Referring to Fig. 4, the corresponding integral equations for $\boldsymbol{J}_s$ and $\boldsymbol{M}_s$ are

$$\begin{aligned} &\hat{n} \times \boldsymbol{E}_1(\boldsymbol{r}) = 0, && \boldsymbol{r} \in \text{ exterior metallic surfaces} \\ &\hat{n} \times \boldsymbol{E}_1(\boldsymbol{r}) = \hat{n} \times \boldsymbol{E}_2(\boldsymbol{r}) && \boldsymbol{r} \in \text{ cavity aperture} \\ &\hat{n} \times \boldsymbol{H}_1(\boldsymbol{r}) = \hat{n} \times \boldsymbol{H}_2(\boldsymbol{r}) && \boldsymbol{r} \in \text{ cavity aperture} \\ &\hat{n} \times \boldsymbol{E}_2(\boldsymbol{r}) + \boldsymbol{E}^{\text{probe}}(\boldsymbol{r}) = 0, && \boldsymbol{r} \in \text{ interior metallic surfaces} \end{aligned} \tag{8}$$

[1] The Green function of a given solution domain characterizes the fields caused by a unit point source.

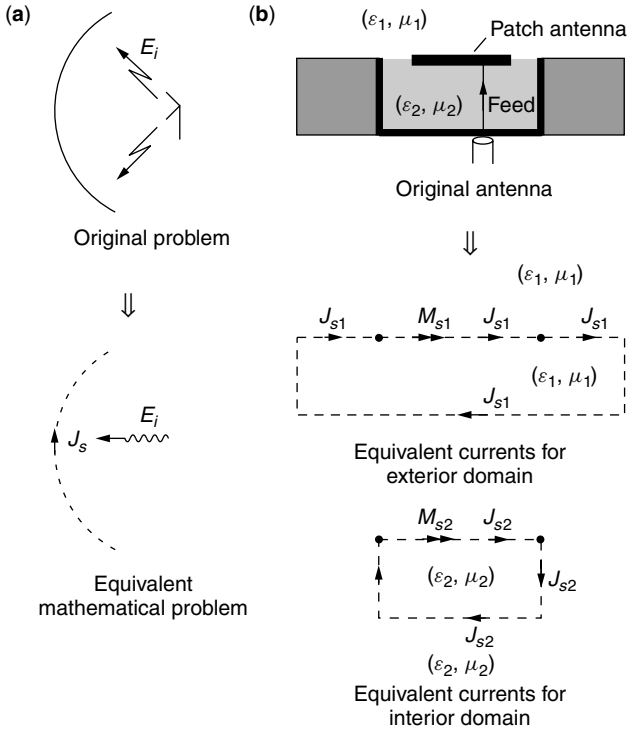


Figure 4. Surface equivalence principle applied to a reflector (left) and a patch (right) antenna.

where $\boldsymbol{E}_1$ is due to $\boldsymbol{J}_{s1}$ and $\boldsymbol{M}_{s1}$ radiating in a homogeneous medium (ε_1, μ_1) and $\boldsymbol{E}_2$ is the corresponding field due to $\boldsymbol{J}_{s2}$ and $\boldsymbol{M}_{s2}$ radiating in the homogeneous medium (ε_2, μ_2); that is, $(\boldsymbol{J}_{s1}, \boldsymbol{M}_{s1})$ and $(\boldsymbol{J}_{s2}, \boldsymbol{M}_{s2})$ are conveniently introduced to satisfy the boundary conditions on the boundary of the piecewise homogeneous region. We can readily conclude that the continuity conditions across the aperture (excluding the patch) can be satisfied a priori by setting $\boldsymbol{J}_{s1} = -\boldsymbol{J}_{s2}$ and $\boldsymbol{M}_{s1} = -\boldsymbol{M}_{s2}$ across that section of the surface. The equivalent currents for the other surfaces must be found via a numerical solution of the integral equation enforcing the boundary conditions.

4. INTEGRAL EQUATION TECHNIQUES

Integral equation (IE) techniques are among the oldest and most successful computational antenna modeling approaches. Their basic idea is to replace an antenna or scattering object by equivalent sources (currents) such that those sources radiate in a domain whose Green function is known. An IE is then derived by enforcing the boundary or continuity conditions for the fields such as those in (8) or (7). This is the first step (*step* 1) in any simulation test and typically involves use of the equivalence principle illustrated in Fig. 3 and 4. The equivalence principle relies on the uniqueness theorem, stating that the resulting solution is unique provided it satisfies Maxwell's equations and the boundary conditions. Use of the representation (4) guarantees the first, whereas the boundary conditions are enforced with the appropriate choice of the equivalent current. It is important to note the introduction of the equivalent currents implies removal of the

actual structure whose presence is only implied through the boundary conditions to be enforced.

The second step (*step* 2) in a numerical simulation is the discretization of the geometry (as done in Fig. 5 and 2) and the unknown equivalent sources using an appropriate set of basis or expansion functions. The coefficients of the expansion then become the unknowns or degrees of freedom (DoFs) used for setting up the linear system of equations. Such a linear system is formed by enforcing the boundary conditions associated with the original geometry using point matching, the method of moments (MoM) [27] procedure or Nyström's method [28]. For example, at metallic surfaces the boundary condition to be enforced is that the tangential electric fields vanish on that surface (viz., $\hat{n} \times \boldsymbol{E} = 0$ on perfect conductors). On a dielectric surface, the pertinent boundary conditions must enforce continuity of tangential $\boldsymbol{E}$ and $\boldsymbol{H}$ across the interface. To do so, it is necessary that both electric and magnetic equivalent currents be introduced at dielectric interfaces as done, for example, in Fig. 4(b) for the cavity aperture. On the other hand, for metallic surfaces only one equivalent current is required to satisfy the boundary conditions.

The third and final step (*Step* 3) in a numerical simulation is the solution of the linear system generated in step 2. For small linear systems involving less than

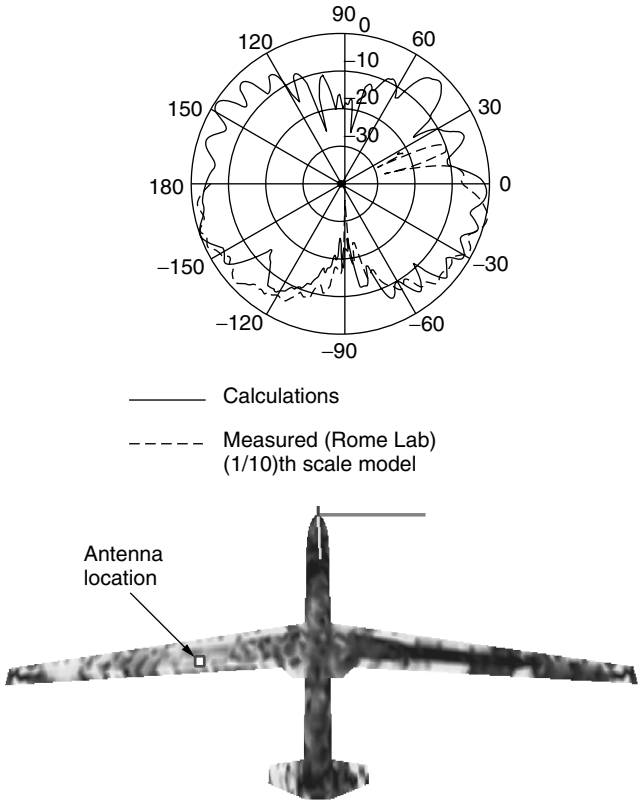


Figure 5. Comparison of measured and calculated patterns for a UHF antenna radiating on a Global Hawk unmanned aerial vehicle (UAV). The surface field plot shows the currents or surface magnetic field on the UAV (red implies highest strength). Measurements are courtesy of the Air Force Research Laboratory (Rome, New York, USA).

1000–5000 unknowns, direct inversion (Gauss elimination) and Lower-Upper (LU) decomposition [29] methods are typically the choice approaches. However, these methods require $O(N^3)$ central processing unit (CPU) solution time and $O(N^2)$ for storing the matrix (N: number of unknowns), and typically over a million discrete elements (and unknowns) are needed to discretize a simple full-scale aircraft at radar frequencies. Furthermore, design for antennas and microwave circuits can be realized only by using extremely fast algorithms that provide results in seconds or minutes rather than hours or days. In the mid-1990s, fast algorithms such as the fast multipole method (FMM) and its multilevel version [20] and adaptive integral method (AIM) [21,22] were introduced to alleviate the CPU bottleneck associated with realistic EM simulations. Basically, FMM and AIM overcome the $O(N^3)$ "curse of dimensionality," as it is often called, in solving dense matrix systems using direct solution methods. Their CPU and memory requirements reduce down to $O(N \log N)$ or so. The basic idea of AIM and FMM is the same as that used in the highly efficient fast Fourier transform (FFT) method. AIM does even make direct utilization of the FFT in its implementation, and the multilevel FMM can be considered as a fast FFT for data with unequal separation intervals.

Let us now proceed to implement the three steps mentioned above. For a metallic antenna structure as in Fig. 4, only electric equivalent currents are needed to express the radiated fields, and in this case the condition to be enforced on the metallic surface of the reflector or some other structure is $\hat{n} \times [\boldsymbol{E}^{\text{rad}} + \boldsymbol{E}^{\text{inc}}] = 0$, where $\boldsymbol{E}^{\text{inc}}$ denotes the field from the horn feed (see Fig. 5) and $\boldsymbol{E}^{\text{rad}}$ is the integral expression in (4). The enforcement of this condition gives the integral equation

$$jkZ\hat{n} \times \iint_S \left[\boldsymbol{J}_s(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}') + \frac{1}{k^2}(\nabla_s' \cdot \boldsymbol{J}_s(\boldsymbol{r}'))\nabla G(\boldsymbol{r},\boldsymbol{r}') \right] ds' \Bigg|_{\boldsymbol{r}\in S} = \hat{n} \times \boldsymbol{E}^{\text{inc}}(\boldsymbol{r})|_{\boldsymbol{r}\in S} \tag{9}$$

referred to as the *electric field integral equation* (EFIE) since it enforces the boundary condition on $\boldsymbol{E}$. The unknown here is the current density $\boldsymbol{J}_s$ and to solve for it, we proceed to step 2 of the analysis.

Step 2 of the analysis amounts to discretizing the current density and the associated geometry as displayed in Figs. 1 and 2 using triangles, quads, or other surface patches. Specifically, we introduce the expansion

$$\boldsymbol{J}_s(\boldsymbol{r}) = \sum_{n=1}^{N} I_n \boldsymbol{f}_n(\boldsymbol{r}) \tag{10}$$

where $\boldsymbol{f}_n$ are the expansion or basis functions defined on S. This current expansion is substituted into (9) and application of the MoM means testing the resulting equation with N different testing or weighting functions $\boldsymbol{w}_m$. Since testing is equivalent to multiplication with the individual testing functions and subsequent integration over S, N linear algebraic equations for the unique determination of

the N expansion coefficients I_n are obtained:[2]

$$jkZ\sum_{n=1}^{N} I_n \iint_S \iint_S \boldsymbol{w}_m(\boldsymbol{r})\cdot\left[\boldsymbol{f}_n(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}') + \frac{1}{k^2}(\nabla_s'\cdot\boldsymbol{f}_n(\boldsymbol{r}'))\nabla G(\boldsymbol{r},\boldsymbol{r}')\right] ds'\,ds = \iint_S \boldsymbol{w}_m(\boldsymbol{r})\cdot\boldsymbol{E}^{\text{inc}}(\boldsymbol{r})\,ds, \qquad m=1,\ldots,N. \quad (11)$$

For each mth weighting function we obtain a single equation for a total of N equations, leading to the matrix system

$$[Z_{mn}]\{I_n\} = \{V_m\} \quad (12)$$

where $[Z_{mn}]$ is an $N\times N$ fully populated matrix containing the coupling integrals on the left-hand side of (11), $\{I_n\}$ is a column vector containing the unknown coefficients I_n, and $\{V_n\}$ is a column vector containing the weights (moments) of the incident field $\boldsymbol{E}^{\text{inc}}$ on the right-hand side of (11). Depending on the choices for expansion and testing functions, an enormous number of different IE techniques have been developed. If the testing functions are the delta functions, the resulting IE method is called a *point-matching* or collocation method since the IE is enforced only at distinct observation points. The so-called *Galerkin method* performs testing using a weighting function that is identical to the expansion function in (10). If the expansion functions have a domain that spans the entire structure, they are referred to as *entire-domain basis functions*. Entire domain basis functions have only been used for specific cases where the structure shape can be of advantage. *Subdomain basis functions* are typically used and have been the most successful. Subdomain bases have their domain restricted to a single or a pair of surface discretization elements (triangles or quads as shown in Figs. 1 and 2). Typically, they approximate the surface current density as a constant, linear, or quadratic function over the element.

A very crucial point related to all integral equation techniques in electromagnetics is the evaluation of the coupling integrals Z_{mn}. An analytic evaluation of these integrals is usually not possible and numerical integration is plagued by the singular behavior of the Green function $G(\boldsymbol{r},\boldsymbol{r}')$ near $\boldsymbol{r}=\boldsymbol{r}'$. To overcome this difficulty, a series of specialized integrations has been developed over the years that combine analytic integration techniques for the extracted singularity [30,31] with numerical quadrature rules for the remainder integrands. Another possibility is the application of integration variable transformations such as the Duffy's transform [32] to allow for robust numerical integration.

MoM techniques have been very successful in modeling antenna structures and antennas on platforms (see Fig. 5). To better understand the moment method procedure, next we consider the solution of a rather simple integral equation, that associated with radiation by a wire dipole antenna.

[2] $\hat{n}\times$ can be omitted after testing.

4.1. Wire Modeling

A wire antenna is simply a cylindrical conductor whose diameter (a) is much smaller than the length (L) of the wire and also $a \ll \lambda$, where λ is the wavelength of operation. Since the wire is very thin (see Fig. 6), we can practically model the radiation from such a cylindrical antenna using a filamentary (equivalent) current flowing through the center of the wire instead of considering a surface current $\boldsymbol{J}_s$. The use of this filamentary current also implies that no currents exist that are transverse to the wire axis (z), a reasonable approximation for very thin wires. Thus, the integral equation can be rewritten using line rather than surface integrals. We have

$$jkZ\sum_{n=1}^{N} I_n \int_{-L/2}^{L/2}\int_{-L/2}^{L/2} w_m(z)f_n(z')\left[1+\frac{1}{k^2}\frac{\partial^2}{\partial z^2}\right]G_w(z-z')\,dz'\,dz = \int_{-L/2}^{L/2} w_m(z)E_z^{\text{inc}}(z)\,dz, \qquad m=1,\ldots,N \quad (13)$$

and by enforcing the boundary condition on the surface of the wire, $\rho = a$ (see Fig. 6b), the Green function takes the form

$$G_w(z-z') = \frac{1}{4\pi}\frac{e^{-jk[a^2+(z-z')^2]^{1/2}}}{[a^2+(z-z')^2]^{1/2}} \quad (14)$$

We remark that since the current $I(z)$ is not at the same location where the boundary condition is enforced, the Green function is nonsingular at $z=z'$. Another important advantage of thin-wire models is the very small number of expansion functions needed for modeling real-world configurations achieved by the one-dimensional (line current) representation of the originally two-dimensional equivalent surface current densities [33,34] (when away from the feed, usually a single wire of radius $a=w/4$, where w is the strip width, can be used to model a surface as a wire grid).

The wire antenna excitation is performed by an appropriate specification of E_z^{inc} in (13). For a voltage source excitation at the wire center, a good approximation is $E_z^{\text{inc}} = V_0\delta(z)$ as illustrated in Fig. 6d, where V_0 is the voltage source applied at $z=0$. An alternative magnetic frill–current excitation has been considered [35] and is appropriate for the wire monopole antenna seen on most

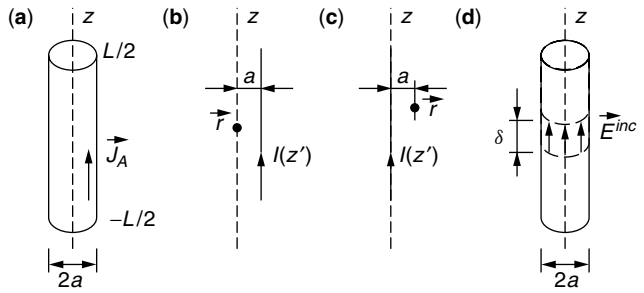


Figure 6. (**a**) Straight wire along z axis; (**b**, **c**) interpretations of thin-wire model; (**d**) δ-gap excitation of wire antenna.

automobiles. The input impedance of the antenna can then be calculated from $Z_{\text{in}} = V_0/I(z_{\text{feed}})$ once the equation system is solved, where the feed location z_{feed} is adjusted for impedance control.

To solve for $\{I_n\}$ using (13), we must choose the basis and weighting functions. The simplest functions useful for the line current approximation are the piecewise constant or pulse basis functions displayed in Fig. 7, but the triangular bases displayed in Fig. 8 provide a much improved approximation without discontinuities at the junctions of the wire segments. Testing is often done using point matching or Galerkin's methods. Figure 9 shows the complex input impedance Z_{in} of a center-fed wire antenna computed using pulse bases with point matching for testing. About 100 basis functions were equally distributed along the wire, and the excitation frequency was varied to investigate the L/λ dependence of the wire antenna. Wire resonances were found at $L/\lambda = 0.48$ and $L/\lambda = 0.84$. For this reason, the optimal operation of the dipole antenna occurs when the dipole is 0.48λ long (see Fig. 10).

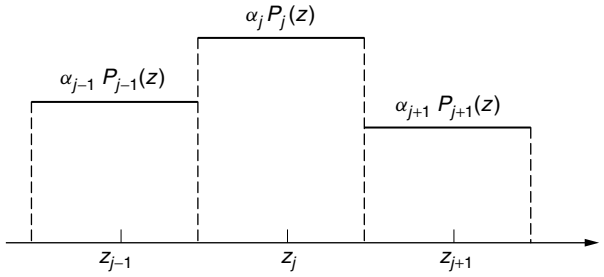


Figure 7. Pulse basis functions for line current modeling.

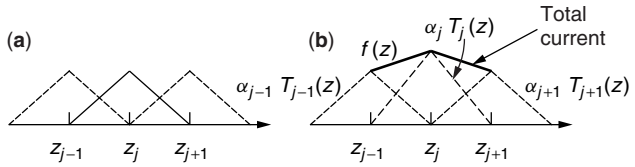


Figure 8. Triangular basis functions (**a**) and (**b**) the resulting piecewise linear current approximation.

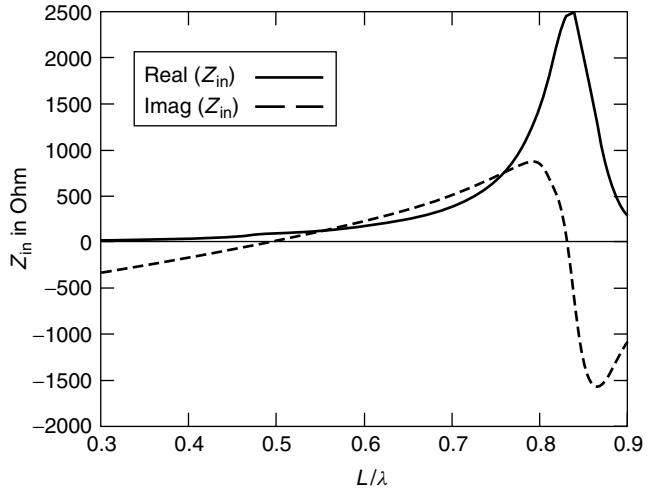


Figure 9. Complex input impedance as a function of length for the wire antenna in Fig. 6 ($a = 0.001\lambda$).

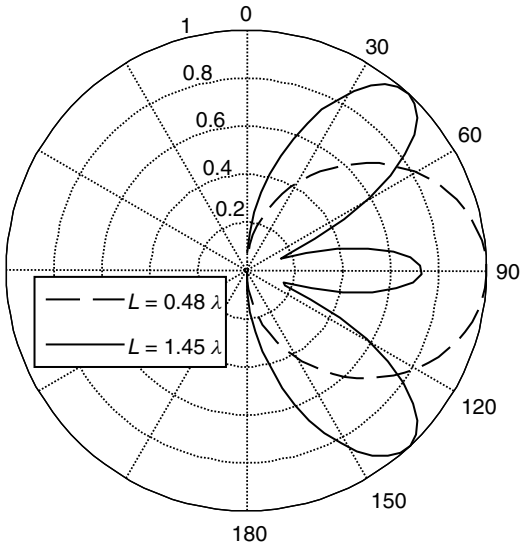


Figure 10. Normalized E-plane radiation patterns for two different wire lengths as shown in Fig. 6 ($a = 0.001\lambda$).

The wire integral equation (13) can be readily generalized to curved wires. In any case, once the line currents are computed along the wire contour, the far field [35] is found from

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{jkZ}{4\pi r}\int_C \hat{r} \times [\hat{r} \times \hat{l}]I(l')e^{-jk|\boldsymbol{r}-\boldsymbol{r}'|}\,dl' \qquad (15)$$

where $\hat{r}$ is the unit vector in the radial direction and $\hat{l}$ is the tangential unit vector along the wire contour.

Thin-wire IE models are useful not only for wire antennas and scatterers but also for computing the radiation of metallic antenna structures (reflectors, horns, etc.) when sufficiently dense wire-grid models are employed to represent the equivalent surface current densities on the structure's surface [33]. For this purpose, wire junctions must be included into the thin-wire theory. This is not an issue for pulse basis approximations, since no current continuity is enforced between wire segments. However, for triangular or higher-order basis functions, special junction conditions must be introduced to fulfill Kirchhoff's current continuity law at wire junctions.

4.2. Surface Modeling

For a variety of antenna problems, wire-grid models are accurate enough to produce useful simulation with high computational efficiency. For other applications such as accurate shielding or planar antenna analyses, the thin-wire approximations cannot produce acceptable results and a direct evaluation of the surface currents as given in (11) is necessary. The formal MoM IE solution is already given by (11), and a numerical implementation can be realized with appropriately assigned surface basis functions for the current representation. Again, popular surface current IE models are based on subdomain bases, where triangular and quadrilateral surface subdomains

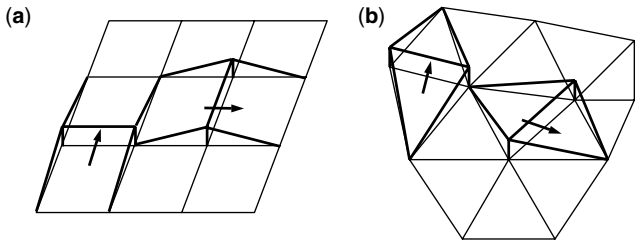


Figure 11. Illustration of mixed linear/constant (rooftop) surface current expansions on (**a**) rectangular and (**b**) triangular subdomains.

as displayed in Fig. 11 are often used. Such techniques are also known as *boundary element* or *boundary integral methods*. The surface current distribution on the individual subdomains is typically chosen to fulfill the necessary physical constraints (e.g., zero normal current components at the edge of a plate) for surface currents. Edge-based surface current basis functions with mixed interpolations are illustrated in Fig. 11 [36]. The degrees of freedom are assigned to the edges of the surface mesh and each basis function has constant normal surface current density components over exactly one edge. Perpendicular to the edge, the current density variation is linear (mixed linear/constant). The basis functions shown in Fig. 11(a) are the *rooftop* basis functions on rectangular subdomains, and those in Fig. 11(b) are the well-known *Rao–Wilton–Glisson* functions introduced in 1982 [36]. On applying coordinate transformations, these basis functions can be transformed to fit to curved surface patches, allowing for more accurate surface current representations [37–39]. Even higher accuracy can be achieved by higher-order basis functions that typically include high-order polynomials subject to constraints at the element junctions for current continuity [34].

The linear system in (11) was derived under the assumption of a metallic surface S. However, surface IE techniques can also be applied for the analysis of antennas involving dielectric and even lossy material objects. As stated earlier, for this case, two current densities must be introduced on S, such as the electric $\boldsymbol{J}_s$ and magnetic $\boldsymbol{M}_s$ surface currents, so that field continuity can be enforced as illustrated in Fig. 4 [24,36]. When dealing with antennas that include dielectrics, volume IEs can be combined with surface IEs for the analysis of the antenna volumetric sections [40,41].

5. FINITE-ELEMENT METHODS

As noted above, IE or boundary element methods result in fully populated matrix systems. For volumetric integral equations where materials must be handled, their storage and computational requirements become prohibitive as the size of the structure increases. This is because the Green function couples all boundary elements regardless of their separation distance. To avoid the introduction of a Green function, one can instead pursue a direct solution of Maxwell's equations in their differential form. Specifically, one could replace the continuous derivatives by finite differences (FDs) to construct a set of equations that can then be solved iteratively in conjunction with specified or natural boundary conditions. This simple, yet very powerful approach became especially popular for time-domain electromagnetic field analysis [e.g., 5,10].

The standard FE method can be derived by applying Galerkin's testing to the differential form of Maxwell's equation, that is, by weighting Maxwell's equations with a suitable (testing) function whose domain is usually restricted over a small region or element (test element) of the computational volume [6,38]. Similar elements can also be used for discretizing the volume region where simple geometric shapes such as quads, tetrahedrons, or triangular prisms (see Fig. 12) are often used. The tetrahedron provides the most flexible element for discretizing volumetric regions. Distorted hexahedra have also been successful for volumetric modeling and may lead to fewer unknowns. On the other hand, shapes with higher symmetry are associated with simple discretization algorithms and may result in better conditioned linear systems.

Although, as discussed above, the finite-element method is based on a direct discretization and solution of the wave equation (3), the weighted or weak form of (3) is used for discretization and solution [6]. The latter enforces Maxwell's equations on an average sense over the discrete element (tetrahedron, quadrilateral, etc.) and is obtained by first dotting (3) with the weighting function/basis $\boldsymbol{W}$ and then making use of the divergence theorem [3] to obtain

$$\iiint_V \left[\frac{1}{\mu_r}(\nabla \times \boldsymbol{E}) \cdot (\nabla \times \boldsymbol{W}) - k^2 \varepsilon_r \boldsymbol{E} \cdot \boldsymbol{W}\right] dv + jkZ \oiint_{S_V} \boldsymbol{W} \cdot (\boldsymbol{H} \times \hat{n})\, ds = 0 \tag{16}$$

Here V denotes the volume domain of interest, S_V is the surface enclosing V, and $\hat{n}$ is the outward normal to S_V. As in (11), $\boldsymbol{W}$ is some weighting function spanning the domain of, say, the eth discrete element. By choosing an expansion for the electric field such as

$$\boldsymbol{E} = \sum_{e=1}^{N} \sum_{i=1}^{N_e} E_i^e N_i^e(\boldsymbol{r}) \tag{17}$$

we can then follow the same steps done for integral equation methods to construct the linear system to find the unknown E_i^e. Although (17) is in principle the same as (10), its form is different but rather convenient in addition to

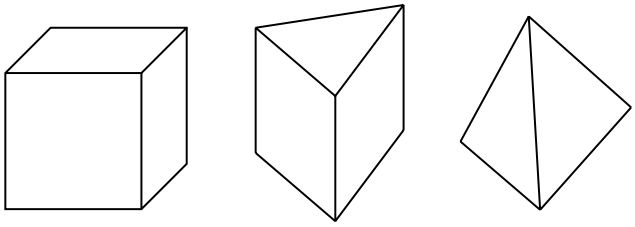


Figure 12. Popular finite-element subdomains: quad, triangular prism, tetrahedron.

having a physical meaning (see Ref. 6 for details). To explain it, let's assume that tetrahedra are used to model the volume of interest. Since the tetrahedron has six edges, we will then set $N_e = 6$, and if N represents the number of tetrahedra used to discretize the entire volume, then the sums in (17) run through all the edges of the tetrahedra constituting the volume. However, if an edge is common to, say, three tetrahedra, then it will appear as many times in the sum. Of particular interest for the expansion (17) is the choice of the basis functions $N_i^e(\boldsymbol{r})$, whose domain based on our notation is only within the eth element and is associated with the ith edge of that element. If we choose $N_i^e(\boldsymbol{r})$ so that

$$N_i^e(\boldsymbol{r}) = \begin{cases} 1 & \boldsymbol{r} = \boldsymbol{r}_i \\ 0 & \boldsymbol{r} = \boldsymbol{r}_j, \quad j \neq i \end{cases}$$

that is, if it is unity at the ith edge located along $\boldsymbol{r}_i$ (similar to Fig. 11) and goes to zero at all other edges, then the coefficient E_i^e is simply the field along the ith edge of the eth element.

The linear system for solving E_i^e is constructed via Galerkin's method by setting $\boldsymbol{W} = \boldsymbol{N}_i^e(\boldsymbol{r})$; then, for $i = 1, \ldots, N_e$ and $e = 1, \ldots, N$, we will obtain NN_e equations. Clearly, these are more than required ($N_{\text{total edges}} < NN_e$) because most of the edges are shared by multiple tetrahedra. For example, an edge shared by three tetrahedra will generate three equations, but actually only one is needed for the unknown field at that edge. Thus, the NN_e equations must be reduced down to $N_{\text{total edges}}$, and this is done via the so-called assembly process. The latter is nothing more than taking the average of the equations generated by testing with $N_j^e(\boldsymbol{r})$, which is unity at the same edge.

As we observe, (16) also includes the unknown magnetic field $\boldsymbol{H}$ at the boundary surface S_V of the domain. Because it is on the boundary, the unique solution of the wave equation requires that it be specified with an external condition unrelated to Maxwell's equations. Of course, if the surface S_V is far away from the radiating source, then the radiation condition [26] can be employed to relate $\boldsymbol{E}$ and $\boldsymbol{H}$ on S_V and thus obtain a deterministic system for the solution of E_i^e. Since the radiation condition can be used only for S_V far away from the source, the enclosed volume is enlarged significantly, requiring many elements for its discretization. Therefore, the number of unknowns becomes unmanageable when the classic radiation condition must be used [25,26], and this plagued the practical application of finite element methods to electromagnetics until the late 1980s (nearly two decades since the method was used by Silvester [42] for waveguide propagation).

The introduction of absorbing boundary conditions (ABCs) such as [6,43]

$$-jkZ\hat{n} \times \boldsymbol{H} = jk\boldsymbol{E}_t + \frac{1}{2jk}\{[\nabla \times [\hat{n}(\hat{n} \cdot \nabla \times \boldsymbol{E})] + \nabla_t(\nabla \cdot \boldsymbol{E}_t)]\} \tag{18}$$

where the subscript "t" denotes the tangential components of $\boldsymbol{E}$ or the nabla operator when evaluated on S_V, provided the means for practical implementation of the finite-element method. This ABC can be enforced with S_V placed on only a fraction of a wavelength from the source. As an alternative to the ABC, one could also use an integral equation to relate $\boldsymbol{E}$ and $\boldsymbol{H}$ on S_V. Such integral equations are of the same form as (4) with $(\boldsymbol{J}_s, \boldsymbol{M}_s)$ replaced by $(\hat{n} \times \boldsymbol{H}, \boldsymbol{E} \times \hat{n})$. Clearly, such an integral equation will lead to a dense submatrix for relating the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields on the surface S_V. Since the finite-element method used for discretizing the interior volume fields is a sparse matrix (usually having a bandwidth of 40 elements or so), the resulting overall system will be partly sparse and partly dense [44]. This is referred to as a *hybrid system*, and the associated methodology is the successful hybrid FE-BI method [44]. It is attractive because S_V can be placed as close to the radiator as desired without restrictions, thus leading to the least number of unknowns. The drawback of this advantage is the increased complexity due to the partly dense system, but more recent use of fast integral methods has alleviated the related issues of CPU and memory complexity [20,45].

The hybrid FE-BI method has been extensively used for the analysis of cavity-backed antennas (see Fig. 4) [44]. Here, the FE volume V includes the possibly inhomogeneous dielectric below the patch only and a BI is applied only at the aperture. Utilizing a half-space Green function, it is sufficient to place magnetic surface current densities $\boldsymbol{M}_s$ on the dielectric portion of the aperture only, and thus a very efficient implementation of the method is possible [6,44]. Fig. 13 shows the input impedance of a cavity-backed and coaxially fed rectangular microstrip patch antenna that has been computed with the hybrid FE-BI technique.

6. CONCLUDING REMARKS

During the 1990s several integral, differential, and hybrid methods matured, and a number of commercial level antenna simulation packages were developed for small

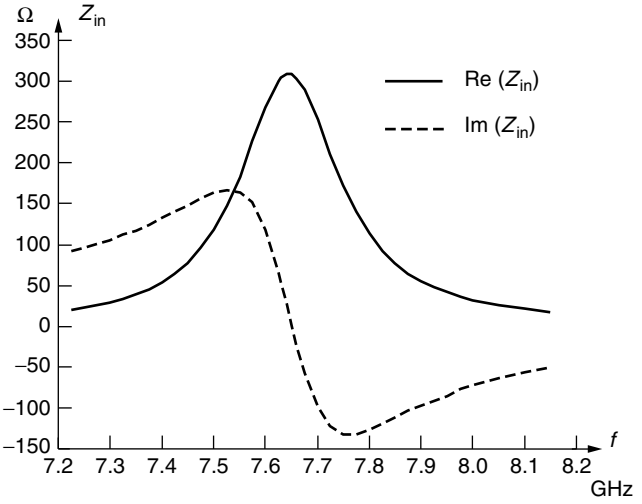


Figure 13. Input impedance of rectangular coaxially fed cavity-backed microstrip patch antenna computed using the hybrid finite-element–boundary-integral technique. Patch: length $L = 1.2$ cm, width $W = 0.8$ cm, feed distance to edge $d = 0.2$ cm, dielectric: thickness $t = 0.1$ cm, $\varepsilon_r = 2.2$.

size but practical antenna models. One can say that the introduction of fast methods during the second half of the 1990s has truly provided the community with computational tools that can allow for both geometric adaptability and material generality, as well as practical size simulations. As an example, using the fast multipole method, a problem involving as many as 170,000 unknowns and resulting in a dense matrix can now be solved in 3 h on a desktop PC using 700 MB (megabytes) of memory, whereas in the mid-1990s the same problem was unsolvable. However, further research on fast methods and related iterative solvers is required prior to their commercialization. Topics such as solver convergence, material modeling, robust and higher-order basis functions, and methods and solver hybridizations are all issues of current research for a variety of specific applications.

BIOGRAPHIES

John L. Volakis obtained his B.E. degree, summa cum laude, in 1978 from Youngstown State University, Ohio, his M.Sc. in 1979 from the Ohio State University, Columbus, Ohio, and a Ph.D. degree in 1982, also from the Ohio State University. Before joining the University of Michigan, Ann Arbor, faculty he worked at the Ohio State University. ElectroScience Laboratory and at Rockwell International. He is currently a professor in the Department of Electrical Engineering and Computer Science (EECS). His primary research deals with the development and application of computational and design techniques to scattering, antennas, and bioelectromagnetics. Dr. Volakis has published over 180 articles in major refereed journal articles, more than 220 conference papers, and 9 book chapters. In addition, he coauthored two books: *Approximate Boundary Conditions in Electromagnetics* (Institution of Electrical Engineers, 1995) and *Finite Element Method for Electromagnetics* (IEEE Press, 1998). In 1998, he received the University of Michigan College of Engineering Research Excellence award, and in 2001 he received the Department of Electrical Engineering and Computer Science Service Excellence Award. Dr. Volakis is a fellow of the IEEE and has served on the editorial boards of several journals. He is also listed in several *Who's Who* directories.

Thomas F. Eibert received the Dipl.-Ing.(FH) degree in 1989 from Fachhochschule Nuernberg, Germany, the Dipl.-Ing. degree in 1992 from the University of Bochum, Germany, and the Dr.-Ing. Degree in 1997 from the University of Wuppertal, Germany, all in electrical engineering. From 1997 to 1998 he was with the Radiation Laboratory at the Electrical Engineering and Computer Science (EECS) Department of the University of Michigan, Ann Arbor, and in 1998 he joined Deutsche Telekom, Darmstadt, Germany, as a research engineer where he has been working on wave propagation and coverage predictions for terrestrial mobile communications and radio broadcasting networks. His major areas of interest are numerical and analytical techniques for electromagnetic and terrestrial wave propagation problems from low frequencies up to millimeter waves.

BIBLIOGRAPHY

1. J. D. Kraus, *Antennas*, McGraw-Hill, 1988.
2. C. A. Balanis, *Antenna Theory*, 2nd ed., Wiley, 1997.
3. J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Dover Publications, New York, 1981 (republication of the 3rd Clarendon Press ed. of 1891).
4. E. K. Miller, L. Medgyesi-Mitchang, and E. H. Newman, eds., *Computational Electromagnetics, Frequency-Domain Method of Moments*, IEEE Press, 1992.
5. A. Taflove, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, Artech House, Boston, 1995.
6. J. L. Volakis, A. Chatterjee, and L. C. Kempel, *Finite Element Method for Electromagnetics*, IEEE Press, USA, 1998.
7. D. M. Pozar and D. H. Schaubert, eds., *Microstrip Antennas*, IEEE Press, 1995.
8. Special issue on advanced numerical techniques in electromagnetics, *IEEE Trans. Antennas and Propagation*, March 1997 issue.
9. D. Filipović and J. L. Volakis, Design and demonstration of a novel conformal slot spiral antenna for VHF to L-band operation, *2001 IEEE Antennas and Propagation Symp. Digest*, 2001.
10. K. S. Kunz and R. J. Luebbers, *The Finite Difference Time Domain Method for Electromagnetics*, CRC Press, Boca Raton, FL, 1993.
11. A. Arif Ergin, B. Shanker, and E. Michielssen, The plane-wave time-domain algorithm for the fast analysis of transient wave phenomena, *IEEE Antennas Propag. Mag.* **41**(4): 39–52 (Aug. 1999).
12. T. K. Sarkar, W. Lee, and S. M. Rao, Analysis of transient scattering from composite arbitrarily shaped complex structures, *IEEE Trans. Antennas Propag.* **48**(10): 1625–1634 (Oct. 2000).
13. T. Abboud, J.-C. Nedeléc, and J. L. Volakis, Stable solution of the retarded potential equations, *Proc. 17th Annual Progress Review in Applied Electromagnetism* (*ACES*), *Digest*, Monterey CA, March 2001, pp. 146–151.
14. A. Peterson, S. Ray, and R. Mittra, *Computational Methods for Electromagnetics*, IEEE Press, 1998.
15. P. Silvester and R. Ferrari, *Finite Elements for Electrical Engineers*, 2nd ed., Cambridge Univ. Press, 1990.
16. R. C. Hansen, ed., *Geometrical Theory of Diffraction*, IEEE Press, 1981.
17. R. J. Marhefka and W. D. Burnside, Antennas on complex platforms, *Proc. IEEE* **80**: 204–208 (Jan. 1992).
18. U. Jakobus and F. M. Landstorfer, Improvement of the PO-MoM hybrid method by accounting for effects of perfectly conducting wedges, *IEEE Trans. Antennas Propag.* **43** (1995).
19. J. L. Volakis, T. Özdemir, and J. Gong, Hybrid finite element methodologies for antennas and scattering, *IEEE Antennas Propag.* 493–507 (March 1997).
20. W. C. Chew, J.-M. Jin, E. Michielssen, and J. M. Song, *Fast and Efficient Algorithms in Computational Electromagnetics*, Artech House, Boston, 2001.
21. E. Bleszynski, M. Bleszynski, and T. Jaroszewicz, AIM: Adaptive integral method compression algorithm for solving large-scale electromagnetic scattering and radiation problems, *Radio Sci.* **31**(5): 1225–1251 (Sept./Oct. 1996).

22. T. F. Eibert and J. L. Volakis, Adaptive integral method for hybrid FE/BI modelling of 3D doubly periodic structures, *IEE Proc. Microwaves, Antennas Propag.* **146**(1): 17–22 (Feb. 1999).

23. Z. Li, Y. E. Erdemli, J. L. Volakis, and P. Y. Papalambros, Design optimization of conformal antennas with the hybrid finite element method, *IEEE Antennas Propag.*

24. E. Arvas, A. Rahhal-Arabi, A. Sadigh, and S. M. Rao, Scattering from multiple conducting and dielectric bodies of arbitrary shape, *IEEE AP Mag.* **33**(2): 29–36 (April 1991).

25. J. van Bladel, *Electromagnetic Fields*, Hemisphere Publishing, New York, 1985.

26. J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, New York, 1941.

27. R. F. Harrington, *Field Computation by Moment Methods*, Macmillan, New York, 1968.

28. L. S. Canino et al., Numerical solution of the Helmholtz equation in 2D and 3D using a high-order Nyström discretization, *J. Comput. Phys.* **146**: 627–633 (1998).

29. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Fannery, *Numerical Recipes in C*, Cambridge Univ. Press, Cambridge, UK, 1992.

30. D. R. Wilton et al., Potential integrals for uniform and linear source distributions on polygonal and polyhedral domains, *IEEE Trans. AP* **32**(3): 276–281 (March 1984).

31. T. F. Eibert and V. Hansen, On the calculation of potential integrals for linear source distributions on triangular domains, *IEEE Trans. Antennas Propag.* **43**(12): 1499–1502 (Dec. 1995).

32. M. G.Duffy, Quadrature over a pyramid or cube of integrands with a singularity at a vertex, *SIAM J. Num. Anal.* **19**(6): 1260–1262 (Dec. 1982).

33. E. K. Miller and E. J. Deadrick, Some computational aspects of thin wire modeling, in R. Mittra, ed., *Numerical and Asymptotic Techniques in Electromagnetics*, Springer-Verlag, 1975.

34. R. Mittra, *Computer Techniques for Electromagnetics*, Pergamon Press, New York, 1973.

35. W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, 2nd ed., Wiley, New York, 1998.

36. S. M. Rao, D. R. Wilton, and A. W. Glisson, Electromagnetic scattering by surfaces of arbitrary shape, *IEEE Trans. AP* **30**(3): 409–418 (May 1982).

37. S. Wandzura, Electric current basis functions for curved surfaces, *Electromagnetics* **12**: 77–91 (1992).

38. G. E. Antilla and N. G. Alexopoulos, Scattering from complex three-dimensional geometries by a curvilinear hybrid fine element–integral equation approach, *J. Opt. Soc. Am. A* **11**: 1445–1457 (April 1994).

39. R. D. Graglia, D. R. Wilton, and A. F. Peterson, Higher-order interpolatory vector bases for computational electromagnetics, *IEEE Trans. Antennas Propag.* **45**: 329–342 (March 1997).

40. D. H. Schaubert, D. R. Wilton, and A. W. Glisson, A tetrahedral modeling method for electromagnetic scattering by arbitrarily shaped inhomogeneous dielectric bodies, *IEEE Trans. Antennas Propag.* **32**(1): 77–85 (Jan. 1984).

41. T. J. Peters and J. L. Volakis, Application of a conjugate gradient FFT method to scattering by thin material plates, *IEEE Trans. Antennas Propag.* **AP-36**: 518–526 (April 1988).

42. P. Silvester, Fine element solution of homogeneous waveguide problems, *Alta Freq.* **38**: 313–317 (1969).

43. T. B. A. Senior and J. L. Volakis, *Approximate Boundary Conditions in Electromagnetics*, IEE Press, 1995.

44. J. M. Jin and J. L. Volakis, A hybrid finite element method for scattering and radiation from microstrip patch antennas and arrays residing in a cavity, *IEEE Trans. Antennas Propag.* **A-39**: 1598–1604 (1991).

45. J. L. Volakis, T. F. Eibert, and K. Sertel, Fast integral methods for conformal antenna and array modeling in conjunction with hybrid finite element formulations, *Radio Sci.* **35**(2): 537–546 (March–April 2000).

ANTENNAS

CONSTANTINE A. BALANIS
ANASTASIS C. POLYCARPOU
Arizona State University
Tempe, Arizona

1. INTRODUCTION

An antenna is the system component that is designed to radiate or receive electromagnetic waves. In other words, the antenna is the electromagnetic transducer that is used to convert, in the receiving mode, free-space waves to guided waves. In a modern wireless system, the antenna must also act as a directional device to optimize or accentuate the transmitted or received energy in some directions while suppressing it in the others. The antenna serves to the communication system the same purpose that eyes and eyeglasses serve to a human.

The history of antennas dates back to James Clerk Maxwell, who unified the theories of electricity and magnetism and eloquently represented their relations through a set of profound equations best known as *Maxwell's equation*. His work was first published in 1873. He also showed that light was electromagnetic, and that both light and electromagnetic waves travel by wave disturbances of the same speed. In 1886, Professor Heinrich Rudolph Hertz demonstrated the first wireless electromagnetic system. He was able to produce in his laboratory at a wavelength of 4 m a spark in the gap of a transmitting $\lambda/2$ dipole that was then detected as a spark in the gap of a nearby loop. It was not until 1901 that Guglielmo Marconi was able to send signals over large distances. He performed, in 1901, the first transatlantic transmission from Poldhu in Cornwall, England, to St. John's, Newfoundland [1].

From Marconi's inception through the 1940s, antenna technology was centered primarily on wire-related radiating elements and frequencies up to about UHF. It was not until World War II that modern antenna technology was launched and new elements (waveguide apertures, horns, reflectors, etc.) were primarily introduced. A contributing factor to this new era was the invention of microwave sources (such as the klystron and magnetron) with frequencies of 1 GHz and above.

While World War II launched a new era in antennas, advances made in computer architecture and wireless

communications technology during the 1960s–1990s have had a major impact on the advance of modern antenna technology, and they are expected to have an even greater influence on antenna engineering in the new millennium. Beginning primarily in the early 1960s, advanced numerical and computational methods were introduced that allowed previously intractable complex antenna system configurations to be analyzed and designed very accurately. Antenna design plays a critical role in overall system design since the success of a system strongly relies on the performance of the antenna. Detailed analysis, design, and measurements of antennas can be found in Ref. 2. A tutorial on antennas is described in Ref. 3.

2. ANTENNA ELEMENTS

Prior to World War II, most antenna elements were of the wire type (long wires, dipoles, helices, rhombuses, fans, etc.), and they were used either as single elements or in arrays. During and after World War II, many other radiators, some of which may have been known for some time and others of which were relatively new, were put into service. This created a need for better understanding and optimization of their radiation characteristics. Many of these antennas were of the aperture type (e.g., open-ended waveguides, slots, horns, reflectors, lenses), and they have been used for communication, radar, remote sensing, and deep-space applications on both airborne and earth-based platforms. Many of these operate in the microwave region.

Prior to the 1950s, antennas with broadband pattern and impedance characteristics had bandwidths not much greater than about 2 : 1. In the 1950s, a breakthrough in antenna evolution was created that extended the maximum bandwidth to as great as 40 : 1 or more. Because the geometries of these antennas are specified by angles instead of linear dimensions, they have ideally an infinite bandwidth. Therefore, they are referred to as *frequency-independent* [2]. These antennas are primarily used in the 10–10,000 MHz region in a variety of applications, including TV, point-to-point communications, and feeds for reflectors and lenses.

It was not until almost 20 years later that a fundamental new radiating element, which has received a lot of attention and many applications since its inception, was introduced. This occurred in the early 1970s when the *microstrip* or *patch* antenna was reported [2]. This element is simple, lightweight, inexpensive, low-profile, and conformal to the surface. Microstrip antennas and arrays can be flush-mounted to metallic and other existing surfaces. Operational disadvantages of microstrip antennas include low efficiency, narrow bandwidth, and low power handling capabilities. Major advances in millimeter wave antennas have been made, including integrated antennas where active and passive circuits are combined with radiating elements in one compact unit (monolithic form).

The unparalleled advances in telecommunications have brought a dramatic increased interest and activity in antenna design. This has resulted in many new elements and design concepts [4], including increased interest in adaptive arrays and "smart" antennas [5,6].

3. THEORY

To analyze an antenna system, the sources of excitation are specified, and the objective is to find the electric and magnetic fields radiated by the elements. Once this is accomplished, a number of parameters and figures of merit (directivity, input impedance, effective area, polarization, etc.) that characterize the performance of the antenna system can be found. To design an antenna system, the characteristics of performance are specified, and the sources to satisfy the requirements are sought.

3.1. Maxwell's Equations

An antenna system is an electromagnetic boundary problem. Therefore, the fields radiated must satisfy Maxwell's equations, which, for lossless medium ($\sigma = 0$) and time-harmonic fields (assuming an $e^{j\omega t}$ time convention), can be written as

$$\nabla \times \vec{E} = -\vec{M}_i - j\omega\mu\vec{H} \tag{1a}$$

$$\nabla \times \vec{H} = \vec{J}_i + j\omega\mu\vec{E} \tag{1b}$$

$$\nabla \cdot (\varepsilon\vec{E}) = q_{ve} \tag{1c}$$

$$\nabla \cdot (\mu\vec{H}) = q_{vm} \tag{1d}$$

In Eqs. (1a)–(1d) both electric $\vec{J}_i$ and magnetic $\vec{M}_i$ current densities, and electric q_{ve} and magnetic q_{vm} charge densities, are allowed to represent the sources of excitation. The respective current and charge densities are related by the continuity equations

$$\nabla \cdot \vec{J}_i = -j\omega q_{ve} \tag{2a}$$

$$\nabla \cdot \vec{M}_i = -j\omega q_{vm} \tag{2b}$$

Although magnetic sources are not physical, they are often introduced as *electrical equivalents* to facilitate solutions of physical boundary-value problems [2,7]. In fact, for some configurations, both electric and magnetic equivalent current densities are used to represent actual antenna systems. For a metallic wire antenna, such as a dipole, an electric current density is used to represent the antenna. However, an aperture antenna, such as a waveguide or horn, can be represented by either an equivalent magnetic current density or by an equivalent electric current density or both. For a radiation problem, the first step is to represent the antenna excitation by its source, represented by the current density $\vec{J}_i$ or $\vec{M}_i$ or both. The next step is to solve the Maxwell equations, subject to a given set of boundary conditions, for $\vec{E}$ and $\vec{H}$. This is a difficult step, and it usually involves an integral with a complicated integrand. This procedure is represented in Fig. 1 as path 1.

To reduce the complexity of the problem, it is common practice to break the procedure into two steps. This is represented in Fig. 1 by path 2. The first step involves an integration while the second involves a differentiation. To accomplish this, auxiliary vector potentials are introduced. The most commonly used potentials are $\vec{A}$ (magnetic vector potential) and $\vec{F}$ (electric vector potential). Although the electric and magnetic field intensities ($\vec{E}$ and $\vec{H}$) represent

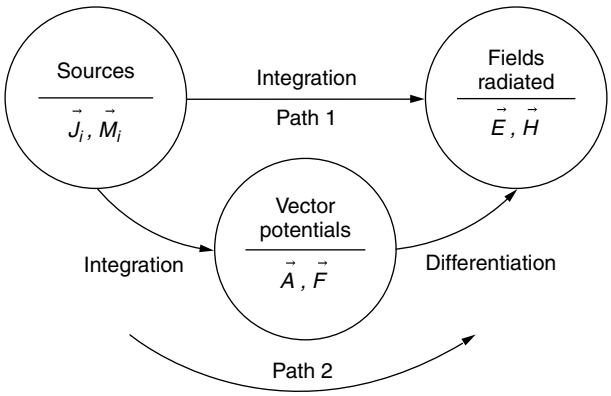


Figure 1. Procedure to solve antenna radiation.

physically measurable quantities, for most engineers the vector potentials are strictly mathematical tools [7]. The procedure along with the appropriate analytical formulations are detailed in Refs. 2 and 7.

3.2. Field Regions

The space surrounding an antenna is usually subdivided into three regions: the *reactive near-field* region, the *radiating near-field* (*Fresnel*) region, and the *far-field* (*Fraunhofer*) region. These regions are so designated to identify the field structure in each. Although no abrupt changes in the field configurations are noted as the boundaries are crossed, there are distinct differences among them. The boundaries separating these regions are not unique, although various criteria have been established and are commonly used to identify the regions [2]. The following definitions in quotations are from an IEEE standard [8].

1. The *reactive near-field region* is defined as "that region of the field immediately surrounding the antenna wherein the reactive field predominates." For most antennas, the outer boundary of this region is commonly taken to exist at a distance $R < 0.62\sqrt{D^3/\lambda}$ from the antenna, where λ is the wavelength and D is the largest dimension of the antenna.
2. The *radiating near-field* (*Fresnel*) *region* is defined as "that region of the field of an antenna between the reactive near-field region and the far-field region wherein radiation fields predominate and wherein the angular field distribution is dependent on the distance from the antenna." The radial distance R over which this region exists is $0.62\sqrt{D^3/\lambda} \leq R < 2D^2/\lambda$ (provided D is large compared to the wavelength). This criterion is based on the maximum phase error of $\pi/8$ radians (22.5°). In this region the field pattern is, in general, a function of the radial distance and the radial field component may be appreciable.
3. The *far-field* (*Fraunhofer*) *region* is defined as "that region of the field of an antenna where the angular field distribution is essentially independent of the distance from the antenna." In this region, the real part of the power density is dominant. The radial distance of this region is $R \geq 2D^2/\lambda$ (provided D is large compared to the wavelength). The outer boundary is ideally at infinity. The criterion is also based on the maximum phase error of $\pi/8$ radians (22.5°). In this region, the field components are essentially transverse to the radial direction, and the angular distribution is independent of the radial distance.

3.2.1. Far Field. The analytical formulation followed to find the field radiated by an antenna at any point, near field or far field, is in general complex and is outlined in detail in Ref. 2. Since antennas are primarily used to communicate over long distances, only the procedure used to find the fields in the far zone will be summarized; it is also less complex.

The following procedure can be followed to determine the electric and magnetic fields radiated by an antenna at an observation point in the far-field region. Once the current densities $\vec{J}_s$ and/or $\vec{M}_s$, either physical or equivalent, are selected to represent the physical antenna, then the vector potentials $\vec{A}$ and $\vec{F}$ are found according to

$$\vec{A} = \frac{\mu}{4\pi} \iint_S \vec{J}_s \frac{e^{-j\beta R}}{R}\, ds' \tag{3a}$$

$$\vec{F} = \frac{\varepsilon}{4\pi} \iint_S \vec{M}_s \frac{e^{-j\beta R}}{R}\, ds' \tag{3b}$$

where R is the distance from any point on the source to the observation point and β is the phase constant ($\beta = 2\pi/\lambda$). The surface current densities $\vec{J}_s$ and $\vec{M}_s$ have the units of A/m and V/m (amperes and volts per meter), respectively. If the current densities are distributed over a volume, the surface integrals of Eqs. (3a) and (3b) are replaced by volume integrals; line integrals are used for thin wire elements.

In the far-field (Fraunhofer) region, the radial distance R of Fig. 2a can be approximated by

$$R \cong \begin{cases} r - r' \cos\psi & \text{for phase terms} \quad (4a) \\ r & \text{for amplitude terms} \quad (4b) \end{cases}$$

Graphically, the approximation of (4a) is illustrated in Fig. 2b, where the radial vectors R and r are parallel to each other. Although such relation is strictly valid only at infinity, it becomes more accurate as the observation point is moved outward at radial distances exceeding $2D^2/\lambda$. Since the far-field region extends at radial distances of at least $2D^2/\lambda$, the approximation of (4a) for the radial distances R leads to phase errors that do not exceed $\pi/8$ radians (22.5°). It has been shown that such phase errors do not have a pronounced effect on the variations of the far-field amplitude patterns.

Using the approximations of (4a) and (4b) for observations in the far-field region, the integrals in (3a) and (3b)

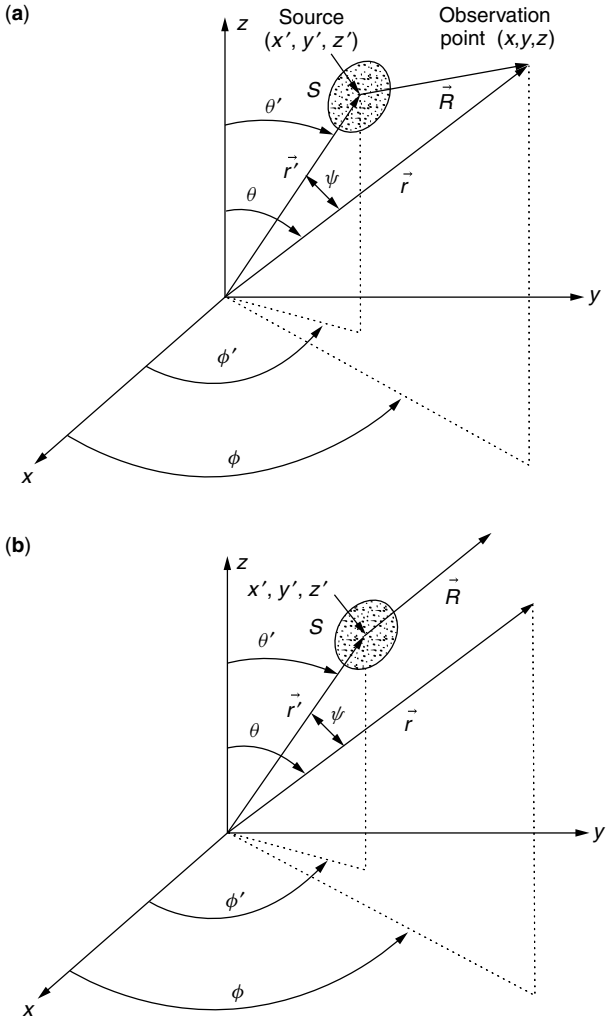


Figure 2. Coordinate system arrangements for (**a**) near-field and (**b**) far-field radiation.

can be reduced to

$$\vec{A} \cong \frac{\mu}{4\pi}\frac{e^{-j\beta r}}{r}\iint_S \vec{J}_s e^{j\beta r\cos\psi}\,ds' = \frac{\mu}{4\pi}\frac{e^{-j\beta r}}{r}\vec{N} \tag{5a}$$

$$\vec{N} = \iint_S \vec{J}_s e^{j\beta r'\cos\psi}\,ds' \tag{5b}$$

$$\vec{F} \cong \frac{\varepsilon}{4\pi}\frac{e^{-j\beta r}}{r}\iint_S \vec{M}_s e^{j\beta r\cos\psi}\,ds' = \frac{\varepsilon}{4\pi}\frac{e^{-j\beta r}}{r}\vec{L} \tag{6a}$$

$$\vec{L} = \iint_S \vec{M}_s e^{j\beta r'\cos\psi}\,ds' \tag{6b}$$

Once the vector potentials are determined, the corresponding spherical components of the electric and magnetic fields in the far-field region can be found in scalar form using [2]

$$E_r \cong 0 \tag{7a}$$

$$E_\theta \cong -j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}[L_\phi + \eta N_\theta] \tag{7b}$$

$$E_\phi \cong j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}[L_\theta - \eta N_\phi] \tag{7c}$$

$$H_r \cong 0 \tag{8a}$$

$$H_\theta \cong j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}\left[N_\phi - \frac{L_\theta}{\eta}\right] \tag{8b}$$

$$H_\phi \cong -j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}\left[N_\theta + \frac{L_\phi}{\eta}\right] \tag{8c}$$

where η is the intrinsic impedance of the medium ($\eta = \sqrt{\mu/\varepsilon}$) while N_θ, N_ϕ and L_θ, L_ϕ are the spherical θ and ϕ components of $\vec{N}$ and $\vec{L}$ from (5b) and (6b), respectively. In antenna theory, the spherical coordinate system is the one most widely used.

By examining (7a)–(8c), it is apparent that

$$E_\theta \cong \eta H_\phi \tag{9a}$$

$$E_\phi \cong -\eta H_\theta \tag{9b}$$

The relations of (9a) and (9b) indicate that in the far-field region the fields radiated by an antenna, and observed in a small neighborhood on the surface of a large-radius sphere, have all the attributes of a transverse electromagnetic (TEM) wave whereby the corresponding electric and magnetic fields are orthogonal to each other and to the radial direction.

To use the procedure described above, the sources representing the physical antenna structure must radiate into an infinite homogeneous medium. If that is not the case, then the problem must be reduced further (e.g., through the use of a theorem, such as the image theorem) until the sources radiate into an infinite homogeneous medium.

4. ANTENNA SOURCE MODELING

The first step in the analysis of the fields radiated by an antenna is the specification of the sources to represent the antenna. Here we will present two examples of source modeling: one for thin-wire modeling (such as a dipole) and the other for an aperture antenna (such as a waveguide). These are two distinct examples each with a different source modeling; the wire requires an electric current density while the aperture is represented by an equivalent magnetic current density.

4.1. Wire Source Modeling

Let us assume that the wire antenna is a dipole, as shown in Fig. 3. If the wire has a circular cross section with radius a, the electric current density induced on the surface of the wire will be symmetric about the circumference (no ϕ variations). If the wire is also very thin ($a \ll \lambda$), it is common to assume that the excitation source representing the antenna is a current along the axis of the wire. This current must vanish at the ends of the wire. For a center-fed resonant dipole, the excitation is often represented by [2]

$$\vec{I}_e = \begin{cases} \hat{a}_z I_0 \sin\left[\beta\left(\dfrac{l}{2} - z'\right)\right] & 0 \le z' \le \dfrac{l}{2} \\ \hat{a}_z I_0 \sin\left[\beta\left(\dfrac{l}{2} + z'\right)\right] & -\dfrac{l}{2} \le z' \le 0 \end{cases} \tag{10}$$

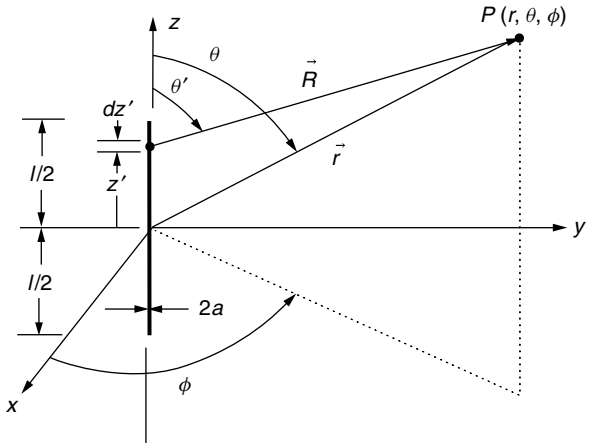


Figure 3. Dipole geometry for electromagnetic wave radiation.

No magnetic source representation is necessary for this type of antenna. For observations in the far-field region, the fields radiated by the antenna can be found using (5a)–(8c) where the surface integrals are replaced by line integrals. On the basis of these far-field expressions, the radiated electric and magnetic fields can be written as

$$E_\theta \cong j\eta \frac{I_0 e^{-j\beta r}}{2\pi r}\left[\frac{[\cos(\beta l/2)\cos\theta] - \cos(\beta l/2)}{\sin\theta}\right] \quad (11a)$$

$$H_\phi \cong \frac{E_\theta}{\eta} \quad (11b)$$

To illustrate the field variation of (11a), a three-dimensional graph of the normalized field amplitude pattern for a half-wavelength ($l = \lambda/2$) dipole is plotted in Fig. 4. A 90° angular section of the pattern has been omitted to illustrate the figure-eight elevation plane pattern variation. As the length of the wire increases, the pattern becomes narrower. When the length exceeds one wavelength ($l = \lambda$), sidelobes are introduced into the elevation plane pattern.

Wire type radiating elements include dipoles, monopoles, loops, and helices. Arrays of dipoles are very popular for wireless communication base stations. Monopoles and very thin helices are widely used in cellular phones and mobiles, and in many wireless communication systems. Loop antennas are used in pagers and also being suggested for cellular phones [9].

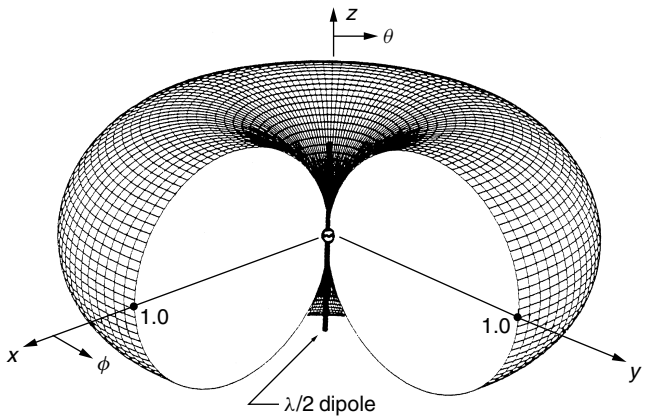


Figure 4. Three-dimensional amplitude pattern of a $\lambda/2$ dipole.

4.2. Aperture Source Modeling

To analyze aperture antennas, the most often used procedure is to model the source representing the actual antenna by the *field equivalence principle* (FEP), also referred to as *Huygen's principle* [2,7]. With this procedure, the actual antenna is replaced by equivalent sources (electric or magnetic or both) that, externally to a closed surface enclosing the actual antenna, produce the same fields as those radiated by the actual antenna. This procedure is analogous to the Thévenin equivalent of circuit analysis, which produces the same response, to an external load, as the actual circuit.

To demonstrate the use of FEP to calculate the fields radiated by an antenna, consider an open-ended rectangular waveguide aperture mounted on an infinite planar PEC (Perfect Electric Conductor) radiating in a semiinfinite homogeneous medium, as shown in Fig. 5a. Let us assume that the fields in the waveguide aperture are those of the dominant TE_{10} mode. Hence, the tangential electric field over the x–y plane is

$$\vec{E}_s = \begin{cases} \hat{a}_y E_0 \cos\left(\dfrac{\pi}{a}x'\right) & -\dfrac{a}{2} \le x' \le \dfrac{a}{2}, -\dfrac{b}{2} \le y' \le \dfrac{b}{2} \\ 0 & \text{elsewhere over the PEC} \end{cases} \quad (12)$$

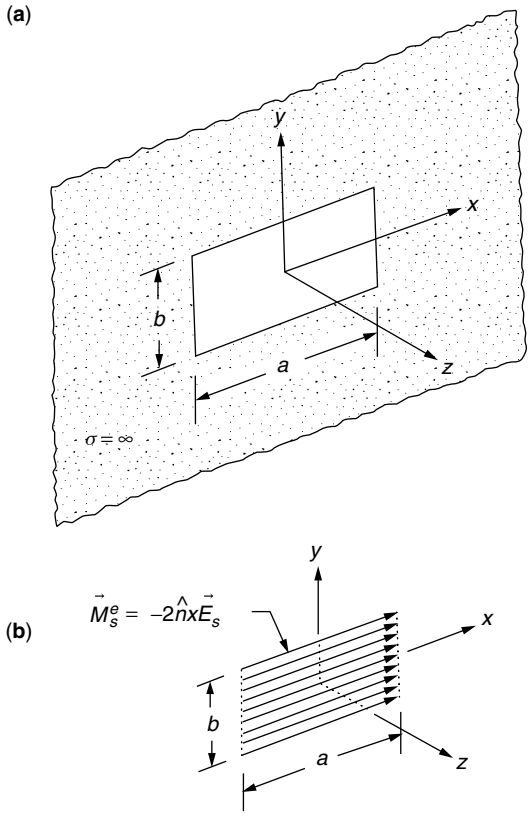


Figure 5. (**a**) Waveguide aperture on an infinite ground plane and (**b**) its equivalent.

By adopting the FEP, an imaginary closed surface is chosen to coincide with an infinite PEC (x–y plane) and covers also the waveguide aperture. The imaginary closed surface is chosen to coincide with the x–y plane because the tangential component of the electric field, and thus the equivalent magnetic current density, are nonzero only in the aperture. Using (12), we can write that

$$\vec{M}_s = -\hat{n} \times \vec{E}_s = \begin{cases} \hat{a}_x E_0 \cos\left(\frac{\pi}{a}x'\right) & -\frac{a}{2} \le x' \le \frac{a}{2}, -\frac{b}{2} \le y' \le \frac{b}{2} \\ 0 & \text{elsewhere over the PEC} \end{cases} \tag{13}$$

Using image theory, the infinite ground plane can be removed by replacing $\vec{M}_s$ with a magnetic current density of twice the strength of (13). The new $\vec{M}_s$ is now radiating into an infinite homogeneous medium, as shown in Fig. 5b. Using (5a)–(8c), the far-field electric and magnetic fields radiated by the waveguide can be written by

$$E_r \cong H_r \cong 0 \tag{14a}$$

$$E_\theta \cong -\frac{\pi}{2} C \sin\phi \frac{\cos(X)}{X^2 - \left(\frac{\pi}{2}\right)^2} \frac{\sin(Y)}{Y} \tag{14b}$$

$$E_\phi \cong -\frac{\pi}{2} C \sin\theta \cos\phi \frac{\cos(X)}{X^2 - \left(\frac{\pi}{2}\right)^2} \frac{\sin(Y)}{Y} \tag{14c}$$

$$H_\theta \cong -\frac{E_\phi}{\eta} \tag{14d}$$

$$H_\phi \cong +\frac{E_\theta}{\eta} \tag{14e}$$

$$X = \frac{\beta a}{2} \sin\theta \cos\phi \tag{14f}$$

$$Y = \frac{\beta b}{2} \sin\theta \sin\phi \tag{14g}$$

$$C = j\frac{ab\beta E_0 e^{-j\beta r}}{2\pi r} \tag{14h}$$

The three-dimensional normalized field pattern of a rectangular aperture with dimensions of $a = 3\lambda$ and $b = 3\lambda$ is shown in Fig. 6. The minor lobes formed throughout the space are clearly shown.

Aperture antennas include waveguides, horns, reflectors, and microstrips. Horns mounted on tall towers are used by telephone companies as transmitting and receiving antennas. Reflectors, because of their high gain, are utilized as ground-based antennas for spaceborne missions and at gateway stations of wireless communication systems. Microstrip antennas, because of their light weight, conformal shapes, low profile, versatility, and other attractive radiation characteristics, are excellent candidates for adaptive and smart antennas.

5. ANTENNA PARAMETERS AND FIGURES OF MERIT

Many different parameters and figures of merit characterize the performance of an antenna system. Some of the most important are included here.

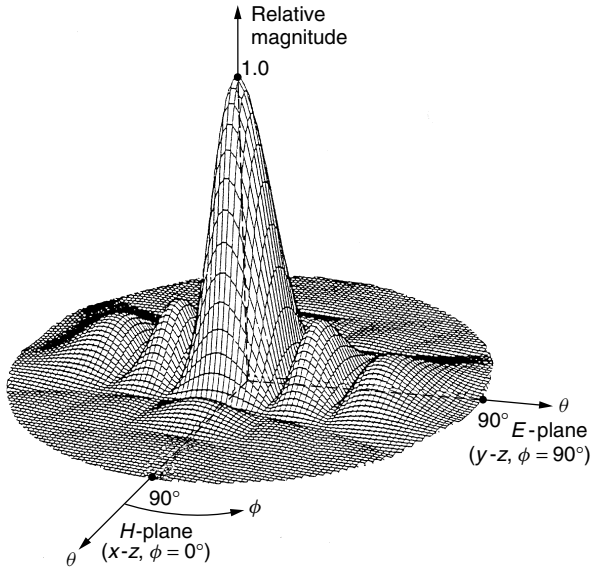


Figure 6. Three-dimensional amplitude radiation pattern for an aperture ($a = 3\lambda$, $b = 3\lambda$) on an infinite ground plane.

An *antenna pattern* is defined as a "graphical representation, usually in the far-field region, of one of the antenna's parameters. For a complete description, the parameters of interest are usually plotted as a function of the spherical angles θ, ϕ." Parameters of interest include amplitude, phase, polarization, and directivity. An amplitude pattern is usually comprised of a number of lobes.

A *main* (*major*) *lobe* is defined as "the radiation lobe containing the direction of maximum radiation. In certain antennas, such as multilobed or split-beam antennas, there may exist more than one major lobe." A *sidelobe* is defined as "a radiation lobe in any direction other than that of the major lobe." The amplitude level of a sidelobe relative to the main lobe (usually expressed in decibels) is referred to as *sidelobe level*.

An antenna, in the transmitting and receiving modes, is often represented by a Thévenin equivalent circuit with an antenna impedance Z_A, as shown in Fig. 7. The antenna impedance Z_A consists of the *radiation resistance* R_r, the *loss resistance* R_L, and an imaginary part $X_A [Z_A = R_A + jX_A = (R_r + R_L) + jX_A]$. The radiation resistance is the resistance that represents antenna radiation or scattering. The loss resistance is the resistance that accounts for the conductive and dielectric losses of the antenna. Expression for R_r and R_L for dipoles and small circular loops can be found, respectively, in Chaps. 4 and 5 of Ref. 2.

Input impedance is defined as "the impedance presented by an antenna at its terminals." It is expressed at the terminals as the ratio of the voltage to current or the ratio of the appropriate components of the electric to magnetic fields, and it is usually complex. When the antenna impedance Z_A is referred to the input terminals of the antenna, it reduces to the input impedance.

Radiation efficiency is defined as "the ratio of the total power radiated by an antenna to the net power accepted by an antenna from the connected transmitter." Using the

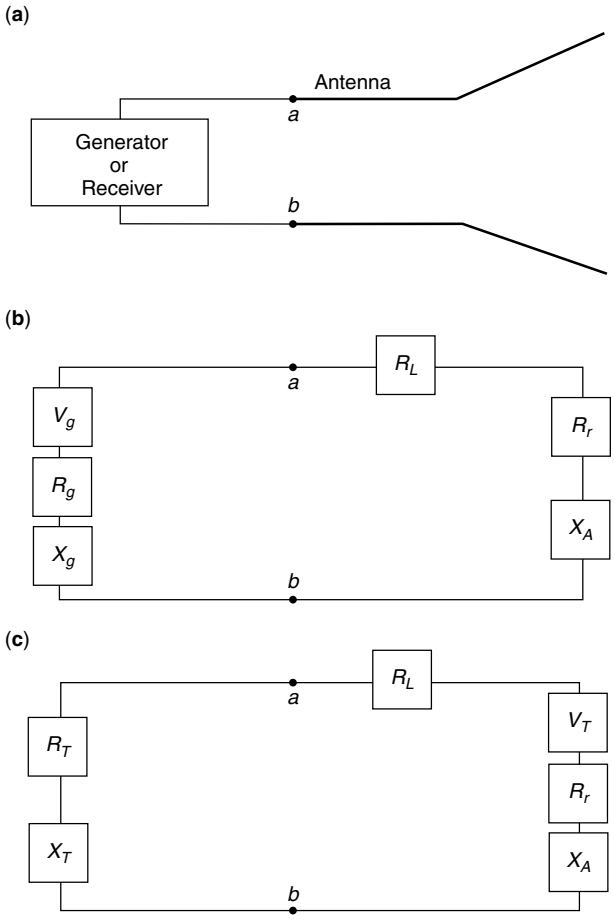


Figure 7. Thévenin equivalents for transmitting and receiving antennas: (**a**) antenna system; (**b**) Thévenin equivalent — transmitting; (**c**) Thévenin equivalent — receiving.

equivalent-circuit representation of an antenna of Fig. 7, the radiation efficiency can be written as

$$e_r = \frac{R_r}{R_r + R_L} \tag{15}$$

Power density S is defined as the power density [in watts per square meter (W/m^2)] of the fields radiated by the antenna. In general, the power density is complex. In the reactive near field, the imaginary component is dominant. In the far field, the real part is dominant. In equation form, the power density $\vec{S}$ is expressed as

$$\vec{S} = \tfrac{1}{2}\vec{E} \times \vec{H}^* = \vec{S}_r + j\vec{S}_i \tag{16}$$

where $\vec{E}$ and $\vec{H}$ are the fields radiated by the antenna (*indicates complex conjugate). The real part of (16) is usually referred to as *radiation density*.

Radiation intensity U is defined as "the power radiated from an antenna per unit solid angle (steredian)." The radiation intensity is usually defined in the far field and is related to the real part of the power density by

$$U = r^2 S_r \tag{17}$$

where r is the spherical radial distance.

Beamwidth is defined as the angular separation between two directions in which the radiation intensity is identical, with no other intermediate points of the same value. When the intensity is one-half of the maximum, it is referred to as *half-power beamwidth*.

An *isotropic radiator* is defined as "a hypothetical, lossless antenna having equal radiation intensity in all directions." Although such an antenna is an idealization, it is often used as a convenient reference to express the directive properties of actual antennas. The radiation intensity S_{r0} of an isotropic radiator and intensity U_0 are defined, respectively, as

$$S_{r0} = \frac{P_r}{4\pi r^2} \tag{18a}$$

$$U_0 = \frac{P_r}{4\pi} \tag{18b}$$

where P_r represents the power radiated by the antenna.

Directivity is one of the most important figures of merit that describes the performance of an antenna. It is defined as "the ratio of the radiation intensity in a given direction from the antenna to the radiation intensity averaged over all direction." Using (18b), it can be written as

$$D = \frac{U(\theta, \phi)}{U_0} = \frac{4\pi U(\theta, \phi)}{P_r} \tag{19}$$

where $U(\theta, \phi)$ is the radiation intensity in the direction θ, ϕ and P_r is the radiated power. For antennas radiating both electric field components (E_θ and E_ϕ), partial directivities D_θ and D_ϕ can be defined as associated, respectively, with E_θ and E_ϕ [2]. The total directivity is then the sum of the two ($D = D_\theta + D_\phi$). If the direction of observation is not specified, it implies the direction of maximum radiation intensity (maximum directivity) expressed as

$$D_0 = \frac{U_m(\theta, \phi)}{U_0} = \frac{4\pi U_m(\theta, \phi)}{P_r} \tag{20}$$

The directivity is an indicator of the relative directional properties of the antenna. As defined by Eqs. (19) and (20), the directional properties of the antenna in question are compared to those of an isotropic radiator. Figure 8 displays the directivity pattern of a $\lambda/2$ dipole and an isotropic source. In each angular direction, only the greater directivity between the two radiators is shown. This allows us to relate the directivity of the element in question to that of an isotropic radiator by simply adding (if expressed in decibels) the relative directivities of one element to another. This procedure is analogous to that used to determine the overall gain of cascaded amplifiers.

Gain is probably the most important figure of merit of an antenna. It is defined as "the ratio of the radiation intensity in a given direction, to the radiation intensity that would be obtained if the power accepted by the antenna were radiated isotropically." Antenna gain is expressed as

$$G = \frac{4\pi U(\theta, \phi)}{P_a} \tag{21}$$

where P_a is the accepted (input) power of the antenna. If the direction is not specified, it implies the direction of

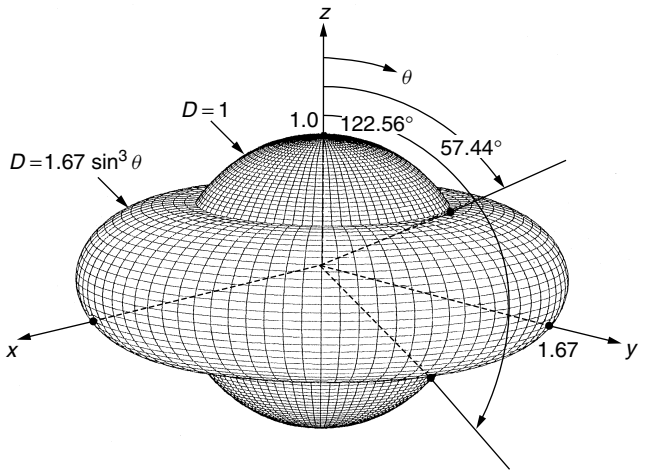


Figure 8. Three-dimensional directivity patterns of a $\lambda/2$ dipole and isotropic radiator.

maximum radiation (maximum gain). In simplest terms, the main difference between the definitions of directivity and gain is that the directivity is based on the radiated power while the gain is based on the accepted (input) power. Since all of the accepted (input) power is not radiated (because of losses), the two are related by

$$P_r = e_r P_a \tag{22}$$

where e_r is the radiation efficiency of the antenna as defined by Eq. (15). By using (19), (21), and (22) the gain can be expressed as

$$G = e_r \frac{4\pi U(\theta, \phi)}{P_r} = e_r D \tag{23}$$

For a lossless antenna, its gain is equal to its directivity.

Antenna polarization in a given direction is determined by the polarization of the fields radiated by the antenna. In general, the polarization of an antenna is classified as linear, circular, or elliptical. Although linear and circular polarizations are special cases of elliptical polarization, in practice they are usually treated separately. Circular and elliptical polarizations also are classified according to the rotation of the transmitted field vectors; the rotation can be either clockwise (right-hand) or counterclockwise (left-hand) as viewed in the direction of propagation.

Polarization efficiency (polarization mismatch or loss factor) is defined as "the ratio of the power received by an antenna from a given plane wave of arbitrary polarization to the power that would be received by the same antenna from a plane wave of the same power flux density and direction of propagation, whose state of polarization has been adjusted for a maximum received power." This is an important factor that must be included in the power budget of communications systems and is one that sometimes is neglected.

Effective area in a given direction is defined as "the ratio of the available power at the terminals of a receiving antenna to the power flux density of a plane wave incident on the antenna from that direction, *the wave being polarization-matched to the antenna*. If the direction is not specified, the direction of maximum radiation intensity is implied." The maximum effective area is related to the antenna gain by

$$A_{em} = pq \left(\frac{\lambda^2}{4\pi} \right) G_0 \tag{24}$$

where p is the polarization efficiency or polarization loss factor, G_0 is the maximum gain of the antenna, and q is the impedance matching efficiency between the transmission line and the antenna defined as

$$q = (1 - |\Gamma_{\text{in}}|^2) \tag{25}$$

where Γ_{in} is the reflection coefficient at the input terminals of the antenna. When multiplied by the power density of the incident wave that impinges on the antenna, the maximum effective area determines the maximum power that is delivered to a matched load connected to the antenna.

Aperture efficiency, usually expressed in percent, is defined as the ratio of antenna's maximum effective aperture to its physical aperture, which can also be expressed on the ratio of the maximum directivity of the aperture to the standard directivity, or

$$\varepsilon_{ap} = \frac{A_{em}}{A_p} = \frac{D_0}{D_s} \tag{26}$$

where A_p is the physical area of the antenna and D_s is the standard directivity of antenna ($4\pi A_p/\lambda^2$ when $A_p \gg \lambda^2$ and with radiation confined to a half-space).

For a rectangular aperture mounted on an infinite ground plane and with a triangular aperture distribution, its aperture efficiency is 75%. However, for an aperture with a sinusoidal aperture distribution, its aperture efficiency is 81%. Again, we see that the aperture distribution, which satisfies the wave equation and the boundary conditions of the structure, determines its aperture efficiency. If an aperture could support a uniform field distribution, its aperture efficiency would be 100%.

6. ARRAYS

Specific radiation pattern requirements usually cannot be achieved by a single antenna element, because single elements usually have relatively wide radiation patterns and low directivities. To design antennas with very large directivities, it is usually necessary to increase the electrical size of the antenna. This can be accomplished by enlarging the electrical dimensions of the chosen single element. However, mechanical problems are usually associated with very large elements.

An alternative way to achieve large directivities, without increasing the size of the individual elements, is to use multiple single elements to form an array. An array is really a sampled version of a very large single element. In an array, the mechanical problems of large single elements are traded for the electrical problems associated with the feed networks of arrays. However, with today's solid-state

technology, very efficient and low-cost feed networks can be designed.

Arrays are the most versatile antenna systems. They find wide applications not only in many spaceborne systems but also in many earth-bound missions. In most cases, the elements of an array are identical; this is not necessary, but it is often more convenient, simpler, and more practical. In general, the radiation characteristics of an array depend on many factors, some of which are:

1. The geometric configuration of the overall array (linear, circular, rectangular, spherical, etc.)
2. The relative displacement between the elements
3. The excitation amplitude of the individual elements
4. The excitation phase of the individual elements
5. The relative pattern of the individual elements

Therefore the designer has many controls or degrees of freedom that can be exercised in order to make the antenna very versatile and meet the specifications of the design.

With arrays, it is practical to not only *synthesize* almost any desired amplitude radiation pattern, but the main lobe can be scanned, resulting in a *scanning* array, by controlling the relative phase excitation between the elements. This is most convenient for applications where the antenna system is not readily accessible, especially for spaceborne missions. The beamwidth of the main lobe along with the sidelobe level can be controlled by the relative amplitude excitation (distribution) between the elements of the array. In fact, there is a tradeoff between the beamwidth and the sidelobe level, based on the amplitude distribution of the elements [2]. The spacing between the elements can be used to control many characteristics of an array, including the pattern, beamwidth, bandwidth, input impedance, and sidelobe level.

There are a plethora of array designs. Two classic array configurations include the Yagi–Uda and log-periodic arrays [2]. The Yagi–Uda is a popular antenna used by amateur radio enthusiasts and for TV. The log-periodic array, because of its large and attractive bandwidth, is probably the most widely used home TV antenna. Arrays of waveguides, horns, reflectors, and microstrips are also very popular [10]. Microstrip arrays will play a key role in the realization of unique designs of adaptive and smart antennas for wireless communications.

Designs of uniform distribution arrays include the broadside, end-fire, and scanning arrays. Classic nonuniform distribution arrays include the binomial, Dolph–Tschebyscheff, Woodward–Lawson, Fourier transform, and Taylor (Chebyshev error and line source) [2]. There are many other array designs, too numerous to name here.

7. ADAPTIVE ARRAYS AND SMART ANTENNAS

In *adaptive antenna arrays* [11,12], the amplitude and phase distribution between the elements are adaptively chosen to improve signal reception or transmission in certain directions and reduce noise and interference in all other directions. Adaptive signal processing algorithms are often used in conjunction with the array architecture to obtain an optimum set of weights that maximizes *signal-to-noise ratio* (SNR) and minimizes *mean-square error* (MSE). In this context, such adaptive arrays are commonly referred to as *smart antennas* [5]. In *code-division multiple-access* (CDMA) applications, such as cellular and mobile communications, smart antennas at the base station can form a main beam toward the subscriber and low-level sidelobes or nulls toward interfering signals. Through adaptive beamforming, smart antennas can penetrate through buildings and cover areas that are otherwise unattainable by a single-element antenna at the base station. In addition, smart antennas can more effectively alleviate problems due to multipath, fading, and time dispersion. This results in an improved system performance compared to an isotropic antenna. Also, dynamic beamforming in smart antennas enhances antenna gain in the direction of the subscriber, thus extending signal coverage. Coverage enhancement can reduce manufacturing cost in cellular and mobile communications by requiring a smaller number of base stations within a given area.

8. CONCLUSIONS

Antenna engineering has enjoyed a very successful period of development since 1950. Responsible for its success have been the introduction and technological advances of some new elements of radiation, such as aperture antennas, horns, reflectors, frequency-independent antennas, and microstrip/patch antennas. Excitement has been created by the advancement of numerical methods that have been instrumental in analyzing many previously intractable problems. Another major factor in the success of antenna technology has been the advances in the computer architecture and wireless communications. Today antenna engineering is considered a truly fine engineering art.

Although a certain level of maturity has been attained, many challenging opportunities and problems remain to be solved. Unique and innovative adaptive and smart antenna designs for wireless communication are creating new enthusiasm and interest in the exploding wireless communication technology. Phased array architecture integrating monolithic MIC (Monolithic Integrated Circuits) technology is still a challenging problem. Integration of new materials into antenna technology offers many advantages, and numerical methods will play key roles in their incorporation and system performance. Computational efficiency in numerical methods will allow modeling, design, and optimization of antennas on complex platforms without the need of supercomputing capabilities. Innovating antenna designs to perform complex and demanding system functions always remain a challenge. New basic elements are always welcomed and offer refreshing opportunities.

BIOGRAPHIES

Constantine A. Balanis received the B.S.E.E. degree from Virginia Tech, Blacksburg, Virginia, in 1964, the

MEE degree from the University of Virginia, Charlottesville, Virginia, in 1966, and the Ph.D. degree in Electrical Engineering from Ohio State University, Columbus, in 1969.

From 1964 to 1970 he was with NASA Langley Research Center, Hampton VA, and from 1970 to 1983 he was with the Department of Electrical Engineering, West Virginia University, Morgantown, West Virginia. Since 1983, he has been with the Department of Electrical Engineering, Arizona State University, Tempe, where he is now regents' professor. His research interests are in low- and high-frequency computational methods for antennas and scattering, smart antennas for wireless communication, and high-intensity radiated fields (HIRF). He received the 2000 IEEE Third Millennium Medal, 1996–97 Arizona State University Outstanding Graduate Mentor Award, 1992 Special Professionalism Award from the IEEE Phoenix Section, the 1989 IEEE Region 6 Individual Achievement Award, and the 1987–1988 Graduate Teaching Excellence Award, School of Engineering, Arizona State University.

Dr. Balanis is a fellow of the IEEE, and he has served as associate editor of the *IEEE Transactions on Antennas and Propagation* and the *IEEE Transactions on Geoscience and Remote Sensing*, and as editor of the Newsletter for the IEEE Geoscience and Remote Sensing Society. He is the author of *Antenna Theory: Analysis and Design* (Wiley, 1997, 1982) and *Advanced Engineering Electromagnetics* (Wiley, 1989).

Anastasis C. Polycarpou received his B.S., M.S., and Ph.D. degrees in electrical engineering from Arizona State University in 1992, 1994, and 1998, respectively. He then joined the Department of Electrical Engineering as an associate research faculty where he worked on various funded research projects. At Arizona State University, he worked on the development and enhancement of numerical methods, in particular the finite element method and the method of moments, for the analysis of complex electromagnetic problems such as microwave circuits, interconnects and electronic packaging, cavity-backed slot antennas in the presence of magnetized ferrites, and helicopter electromagnetics. He has published ten journal papers and 25 conference proceedings. He is currently an associate professor at a small private college in Cyprus. His areas of interest are antennas, electromagnetic theory, and numerical methods.

BIBLIOGRAPHY

1. J. D. Kraus, Antennas since Hertz and Marconi, *IEEE Trans. Antennas Propag.* **AP-33**: 131–137 (Feb. 1985).
2. C. A. Balanis, *Antenna Theory: Analysis and Design*, Wiley, New York, 1997, 1982.
3. C. A. Balanis, Antenna theory: A review, *Proc. IEEE* **80**: 7–23 (Jan. 1992).
4. *Special Issue on Wireless Communications*, *IEEE Trans. Antennas Propag.* **AP-46**: (June 1998).
5. J. C. Liberti, Jr. and T. S. Rappaport, *Smart Antennas for Wireless Communications: IS-95 and Third Generation CDMA Applications*, Prentice-Hall PTR, Englewood Cliffs, NJ, 1999.
6. T. S. Rappaport, ed., *Smart Antennas: Adaptive Arrays, Algorithms, & Wireless Position Location*, IEEE, 1998.
7. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.
8. IEEE, *IEEE Standard Definitions of Terms for Antennas*, IEEE Standard 145-1983, reprinted in *IEEE Trans. Antennas Propag.* **AP-31**(Pt. II of two parts): 5–29 (Nov. 1983).
9. K. D. Katsibas, C. A. Balanis, P. A. Tirkas, and C. R. Birtcher, Folded loop antenna for mobile hand-held units, *IEEE Trans. Antennas Propag.* **AP-46**: 260–266 (Feb. 1998).
10. Special Issue on Phased Arrays, *IEEE Trans. Antennas Propag.* **47**(3): (March 1999).
11. Special Issue on Adaptive Antennas, *IEEE Trans. Antennas Propag.* **AP-24**: (Sept. 1976).
12. Special Issue on Adaptive Processing Antenna Systems, *IEEE Trans. Antennas Propag.* **AP-34**: (Sept. 1986).

ANTENNAS FOR MOBILE COMMUNICATIONS

MICHAEL T. CHRYSSOMALLIS
Democritus University of Thrace
Xanthi, Greece

CHRISTOS G. CHRISTODOULOU
University of New Mexico
Albuquerque, New Mexico

1. ANTENNAS AND MOBILE COMMUNICATION SYSTEM REQUIREMENTS

Although some mobile communication systems are alleged to have originated in 1885, broadly recognized first real mobile services started around 1900 with wireless telegraph on ships, introduced by G. Marconi. He used long vertical wire antennas in various forms; wire antennas were the main type used in mobile systems up to 1970, when integrated antennas appeared as a consequence of the rapid progress in semiconductor integrated circuits. Printed antenna technology introduced the possibility of producing lightweight, less bulky, low-cost, easy-to-manufacture radiating structures, fully compatible with the newly integrated electronic packages [1].

Antennas are used to transmit or receive electromagnetic waves, and also serve as transducers converting guided waves into free-space waves and vice versa. Although all antennas, regardless of their type, operate on the same basic principles of electromagnetic theory, different antenna systems require careful design and a good understanding of the radiation mechanisms involved. Electrical, mechanical, and operating costs usually determine the proper type of antenna, which serves best specific application at hand.

Antennas can be classified into different categories. For our purposes we are interested in antennas in mobile communication systems. Antennas are one of the most important parts in a wireless communication system since they are responsible for the proper transmission and reception of signals. A successful design can relax some of the

complex system requirements involved in a communication link and increase the overall system performance. The choice of an antenna for a specific wireless communication application, either fixed or mobile, depends on (1) the platform to be used, which can be a tower or a building, car, ship, spacecraft, satellite, or other vehicle; (2) the environment in which the communication operates, such as urban, rural area of land, sea, or space; (3) the frequency of operation of the link; and (4) the nature of the application, for example, voice, data transmission, or video.

In the catalog of terrestrial or land mobile systems [2], two of the systems hold a prominent position if we take into account their rapid growth and commercialization: the cellular system and the cordless phones system. The main difference between cellular and cordless is that in a cellular system, the providing service is very similar to those of a normal telephony service with the advantage that the user can be anywhere and in motion, whereas the cordless service is simply a wireless telephone link to your home or your office. There is also a small personal base station that makes it possible to walk around with a direct link to this base station, but there is no provision for taking the cordless phone far away or to another city. Land or terrestrial mobile communications technology has come a long way since the pioneering work of AT&T Bell Laboratories researchers, around 1970, which lead to the first operational cellular system, in 1983. This early analog system, now described as *first-generation*, was relatively simple and developed to suit local requirements, became popular and demonstrated the benefits of communication on the move. The *second-generation* system (1990), designed to use digital transmission, has shown advantages of not only digital over analog processing but also the future of the global standardization. Nowadays third-generation mobile communication systems provide users with advanced communications services having wideband capabilities and using a single worldwide standard.

Mobile satellite communications began in 1976 with satellites in geostationary orbit to provide communications services to ships at sea, and later to aircraft and land-based terminals. The rapid growth of land-mobile communications systems resulted in more efforts for providing global mobile communications services through the use of mobile satellite communication systems in low, medium, and geostationary earth orbits. Second-generation terrestrial and satellite mobile communications systems have existed as two independent environments. However, these two environments are now being combined toward a third-generation global mobile communications system in which the two systems play complementary rather than independent roles, forming a single universal integrated system.

Today, the complexity facing the mobile antenna designer is compounded by the awareness that with clever design the antenna can improve system performance by embodying additional functions, such as diversity reception capability, multipath fading immunity, polarization characteristics selectivity, and matching to different environmental and operational conditions [3]. Modern antenna design for mobile services is no longer confined to small, lightweight, low-profile, or flush-mounted omnidirectional radiators on a well-defined perfect flat ground plane. It is rather the creation of a complicated electromagnetic device that may incorporate active elements and signal processing schemes while operating in a constantly varying time-varying environment.

Antennas now have also to be seen as an integral part of the overall system design, since the equipment onto which an antenna element is mounted can act as a radiator, so that the antenna element and the body of the equipment must be treated together as an antenna system. Another factor is the proximity effects caused by obstacles near to the antenna element, which affect antenna performance and must be considered in the design. The operator of a portable mobile unit can seriously perturb antenna performance, while the human hazard problem is another factor, which always must be of concern.

Mobile communication systems can be divided into two main categories: (1) terrestrial or land-mobile systems and (2) satellite systems. Accordingly, antennas for mobile communication systems can be considered to be part of one of these two main systems, although in many cases there are handsets that were designed to work in both systems.

2. ANTENNAS FOR LAND-MOBILE COMMUNICATION SYSTEMS

2.1. Design Considerations

The two main parts of a land-mobile communication system are the base stations and the mobile stations or units (Fig. 1). Since these stations employ different types of antennas, the key design items are also different. Although *antenna design* usually refers to electrical design, some other aspects must also be considered. Construction costs, easy manufacturing, low installation fees, as well as, key mechanical items such as wind load and seismic load design are important criteria that must be of prime concern. In practice, as a first step, an evaluation that compares electrical and mechanical characteristics and the tradeoff between performance and cost is done, followed in the second step by the determination of the electrical and mechanical design. Also, because these antennas will

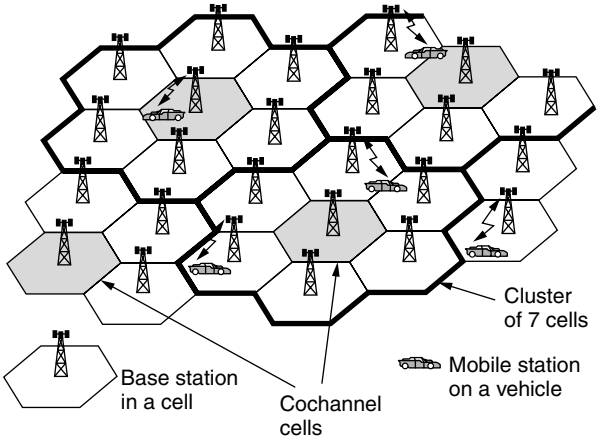


Figure 1. Illustration of a cellular system. Every cell, shown as a hexagon, contains a base station. The cells are organized in clusters of seven cells.

be used in a multipath environment instead of free space, their radiation patterns and gains must be designed for their operating environment. This means that it is useless for an antenna to have superb performance in free space or in an anechoic chamber if it cannot operate well in a real multipath environment. The mobile station antennas, which are classified into two categories, antennas for mobile mounting (on vehicles) and antennas for mounting on portable handsets, are designed for easy handling and subscriber convenience. Besides their electrical characteristics, they must have small volume and weight.

The frequencies used for land-mobile communication systems range from <200 MHz to >60 GHz. The most significant bands are the frequency ranges for analog and digital cellular radio systems from $\sim$800–1000 MHz and 1700–2200 MHz and for the wireless local-area network (WLAN) systems at around 2.4–2.5, 5.1–5.8, and 17 GHz. The bandwidths of cellular systems, which use frequency-domain duplexing (FDD) with two distinct frequency bands for each system, range from $\sim$8–17%, while those of WLAN systems have values of $\sim$5% [4].

2.2. Base-Station Antennas

2.2.1. Basic Requirements. A cellular system services a terrestrial area by dividing it into a number of cells, each containing a base station. The cells are organized in clusters, and every cluster uses all the available channels (Fig. 1). The use of many clusters in an area means an increased number of channels, and this depends on the number of cells per cluster and the area of every cell [5]. Since the same frequencies are used in every cluster, distributed in an ingenious way in its cells in order to reduce cochannel interference, every base station must communicate only with the mobile stations located in its service area, and its radiowave energy must be uniformly radiated inside this area. For this purpose, the radiation pattern and the output power of the base station antenna must be carefully adjusted. For efficient frequency reuse, the uniform illumination of a cell and the suppression of radiation outside of it are accomplished by the use of main beam tilting downward in the vertical plane (see Fig. 2). There are two methods for achieving beam tilting — one is a mechanical beam tilting by leaning the antenna, and the other is an electrical beam tilting by adjusting the relative phases of the antenna elements. The use of a shaped-beam array antenna whose sidelobes, facing the interference cell directions, are suppressed to very low levels assures that interference is kept to minimum possible level. The narrowing of the antenna beam in the vertical plane offers an additional advantage of increased gain [3]. Since the coverage area of each base-station antenna is given, antenna gain cannot be increased by narrowing the beam in the horizontal plane, but only in the vertical plane. So the design parameters for the base-station antenna are the vertical plane pattern shape, which is achieved by an array configuration, and the horizontal plane pattern, which is set by the correct choice of antenna element.

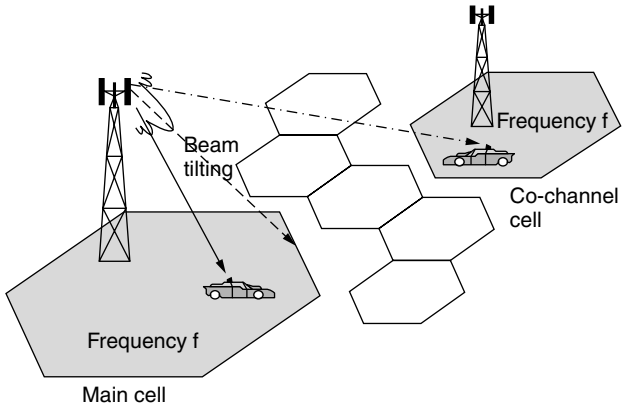


Figure 2. Beam tilt effect for reducing the frequency reuse distance in a cellular system.

Another requirement for the base-station antenna, in order to communicate in all the available channels with the mobile stations, is to be wideband and have a function for branching and combining the channels. Sometimes, because the antenna is shared by several systems, a wider frequency bandwidth is required, and in such cases dual-band or triple-band antennas are used. The multichannel function of a base-station antenna is assured by the use of a broadband antenna element.

As it is well known, communication between base-stations and mobiles antennas rarely occurs within line of sight of each other. A radio transmission link is established in a mobile channel by *multipath propagation*, in which surrounding objects reflect and scatter the transmitted energy causing several waves to arrive at the receiver via different routes. Since these waves have different phases as a result of small pathlength differences between rays coming from scatterers in the near vicinity, or significant time differences if they come from strong scatterers, narrow- or wideband fast fading is produced, respectively. Fast fading is added to slow fading or shadowing, which results from the varying nature of the particular terrain and with the path loss, which is caused by the spreading of waves in space, together constitute the mobile channel fading behavior. As a result, fading occurs constantly at the base and mobile stations and the receiving signal level may fluctuate by $\leq$40–50 dB. In order to keep constant the receiving level and reduce the delay spread, *diversity reception* is used, a technique whose effectiveness has been proved both experimentally and theoretically. It was shown that by placing two antennas $\sim$10 wavelengths apart, a reduction in fading could be achieved, and the value of this reduction depends basically on the correlation between the two antennas [3]. From the three configurations of diversity antennas — space, pattern, and polarization diversity — space diversity is the most widely used [2,6]. In a multipath environment, with Rayleigh distribution, the relationship between the correlation coefficient of the diversity terminals and the carrier-to-noise-level ratio (CNR), with a cumulative probability of 1%, shows that the improvement of CNR does not fall below 8 dB even if the value of correlation coefficient is as high as 0.6 [3]. That means that there is no need to design a diversity antenna with a lower correlation coefficient value. In order to achieve a correlation coefficient of $\leq$0.6 dB, greater distances between antennas for suburban or rural areas are usually required than in urban areas.

Also an increase in antenna height results an increase of the correlation coefficient value.

Finally, lightweight antennas that occupy a small space and have a low wind load are constructed by proper mechanical design. Another subject that must be of concern is the correct choice of materials in order to eliminate *passive intermodulation*, which can arise when the antenna is used for both transmission and reception [7]. According to this phenomenon, as a result of the nonlinear effect of the metal heterojunctions that exist between the antenna elements and the feedline, intermodulation of the transmitting channels can occur, resulting the presence of interference waves having the same frequency of the receiving waves. The use of printed instead of wire elements, the increased contact area between flanges and printed elements, good welding, and the use of measures to prevent oxide generation are some of the common methods used to suppress passive intermodulation.

2.2.2. Types of Base-Station Antennas. The correct choice of a base-station antenna depends on the size of the service area. For cellular systems, a common technique for better frequency reuse is the division of the cell area in zone sectors and the use of sector-beam antennas. It can be shown that in a cellular system, the frequency reuse distance, with a sector zone arrangement, is shorter than that of a circular zone arrangement illuminated by omnidirectional antennas [3]. For a service area limited within a restricted angle in the horizontal plane, a corner reflector antenna is often used either as a single radiator or as an element of a linear array antenna. A corner reflector antenna has the advantage of adjusting its beamwidth by controlling the aperture angle of the reflector. A reflector antenna with its parameters is shown in Fig. 3. Sector beams with beamwidths of 60–180° can be realized by adjusting the aperture angle from 60° to 270°. It should be noted that for the base-station antenna, it is more important to operate with large ratio of desired to undesired signal ratio than to have a high gain value. Fortunately, beam tilting, which is essential for frequency reuse, works in this direction, and if it is combined with the suppression of the sidelobes adjacent to the main beam, by appropriate array antenna pattern synthesis, a very effective antenna is produced. Another point to be noted is that in order to reduce the interference, only several sidelobes near the

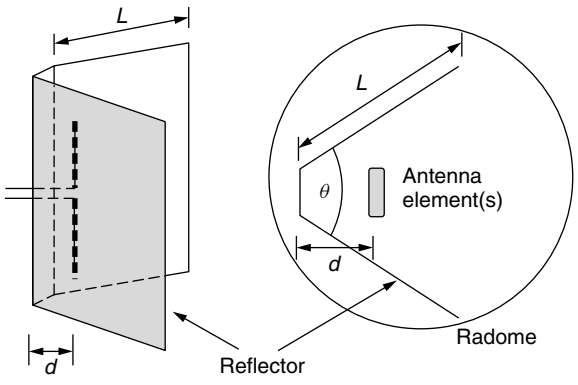


Figure 3. The geometry of a corner reflector antenna.

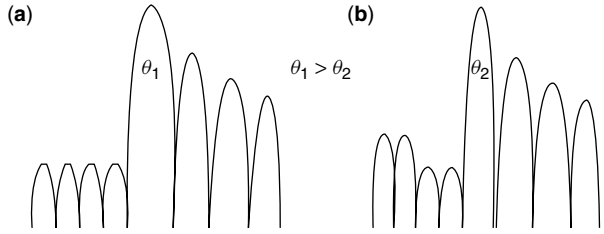


Figure 4. From a uniformly suppressed sidelobe pattern (**a**) a new one is received with the main beam 30% narrower (**b**) by suppressing only several sidelobes near the main beam and setting the others at a higher level.

main beam must be suppressed. In Fig. 4, from a uniformly excited array a new array with a main beam 30% narrower is produced by suppressing only several sidelobes near the main beam and setting the other sidelobes at a higher level [3].

Some basic types of base-station antennas that are used in many systems are the dual-frequency antenna, the dual-beam antenna, and the polarization diversity antenna. In third-generation cellular systems, digital beamforming or smart antennas are used.

Today, several terms are used to refer to the various aspects of smart-antenna system technology, including intelligent antennas, phased array, space-division multiple access (SDMA), spatial processing, digital beamforming, and adaptive antenna systems. A "smart antenna" consists of an antenna array combined with signal processing in both space and time. Spatial processing leads to more degrees of freedom in the system design, which can help improve the overall performance of the system. Smart-antenna systems are usually categorized as either switched-beam or adaptive array systems [8]. Although both systems attempt to increase gain in the direction of the user, only the adaptive array system offers optimal gain while simultaneously identifying, tracking, and minimizing interfering signals. It is the adaptive system's active interference capability that offers substantial performance advantages and flexibility over the more passive switched-beam approach. Smart antennas communicate directionally by forming specific antenna beam patterns. They direct their mainlobe, with increased gain, in the direction of the user, and nulls in directions away from the mainlobe. Different switched beam and adaptive smart antennas control the lobes and the nulls with varying degrees of accuracy and flexibility.

The traditional switched-beam method is considered as an extension of the current cellular sectorization scheme in which a typical sectorized cell site is composed of three 120° macrosectors. The switched-beam approach further subdivides the macrosectors into several microsectors. Each microsector contains a predetermined fixed beam pattern, with the greatest gain placed in the center of the beam. Typically, the switched-beam system establishes certain choices of beam patterns before deployment and selects from one of several choices during operation (Fig. 5). When a mobile user is in the vicinity of a macrosector, the switched-beam system selects the microsector containing the strongest signal. During the call, the system monitors the signal strength and switches to other fixed

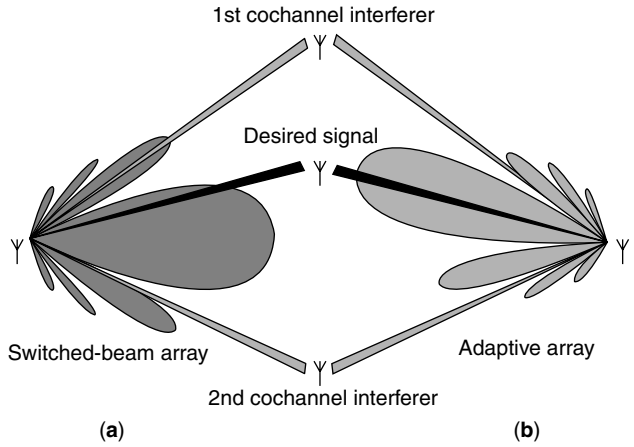


Figure 5. Beamforming lobes and nulls that switched beam (**a**) and adaptive array (**b**) systems might choose for identical user signals and cochannel interferers.

microsectors if required. All switched-beam systems offer similar benefits even though the different systems utilize different hardware and software designs. Compared to conventional sectored cells, switched-beam systems can increase the range of a base station from 20 to 200% depending on the circumstances of operation [8,9]. The additional coverage means that an operator can achieve substantial reduction in infrastructure costs.

There are, however, limitations to switched-beam systems. Since the beams are predetermined, the signal strength varies as the user moves through the sector. As a mobile unit approaches the far azimuth edges of a beam, the signal strength degrades rapidly before the user is switched to another microsector. Moreover, a switched-beam system does not distinguish between a desired signal and interfering ones. Thus, if an interfering signal is around the center of the selected beam and the user is away from the center, degradation in the quality of the signal for the mobile user occurs.

Adaptive antennas take a very different approach. By adjusting to an RF environment as it changes, adaptive antenna technology can dynamically alter the signal patterns to optimize the performance of the wireless system (Fig. 5). The adaptive antenna utilizes sophisticated signal processing algorithms to continuously distinguish between desired signals and multipath and interfering signals, and also, calculate their directions of arrival [8,9]. A block diagram of an adaptive antenna system is shown in Fig. 6. The adaptive approach continuously updates its beam pattern according to changes in both the desired and interfering signal locations. The ability to smoothly track users with mainlobes and interferers with null ensures that the link budget is constantly maximized.

2.3. Mobile-Station Antennas

2.3.1. Design Considerations. Mobile-station antennas or handheld antennas had to follow the dramatic decrease in size and weight of portable phones while maintaining the same antenna performance in terms of radiation pattern, gain, and bandwidth. These changes have necessitated a rapid evolution of antenna structures and techniques for

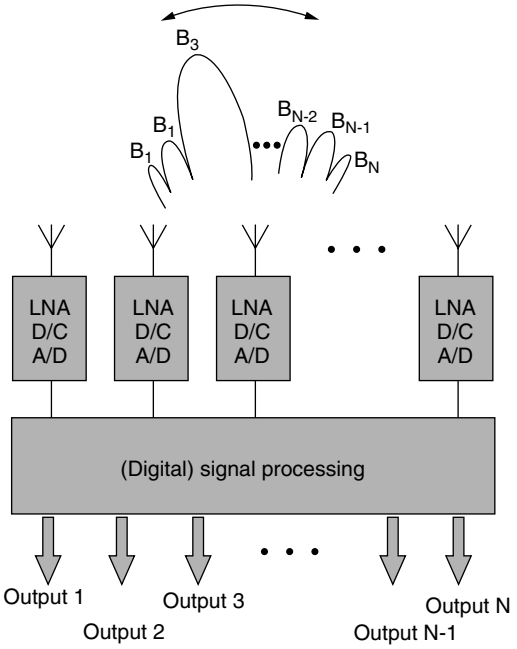


Figure 6. A block diagram of an adaptive antenna system (LNA — low-noise amplifier; D/C — downconverter; A/D — analog–digital converter).

the mobile stations or portable phones. With a volume of <100 cm^3 for a typical mobile phone today, it is difficult to achieve the necessary bandwidth of 10% without inducing currents on the handheld unit, which leads the antenna element to act more as a coupling structure than a radiating element. We can separate the antennas here according to their application: for vehicular and for portable phones. For the first category (vehicular), since the typical mobile user moves randomly in a radio cell, an omnidirectional azimuth pattern is required. Particularly in suburban or rural areas it is common for the mobile station and the base station to be in line of sight, so the omnidirectional pattern assures that the received signal level will not vary considerably. For the second category, although the requirements for operating frequency zones and bandwidth are the same, the transmitting power of portable phones must be less, because of the limited battery capacity and the size limitations. The achievable gain is also generally less than that possible in vehicular antennas, since only small antennas can be used for which adequate bandwidth and high efficiency are very difficult to exist simultaneously [11,12]. Also, their proximity to the human body is another factor of gain degradation. For these reasons, the main antenna requirement for mobile phones is to develop the highest possible gain over the required bandwidth.

It should be added here that the gain performance of a mobile antenna operating in its practical environment is different from that of the same antenna in isolation. For this reason the concept of the mean effective gain (MEG) in a multipath mobile radio environment is introduced. The *mean effective gain* of an antenna is defined as the ratio between the power that the mobile antenna actually receives to the total power available. All power values are

considered averages taken after the mobile station has moved along a path of several wavelengths, and the total average power incident on the mobile is composed of both horizontally and vertically polarized components [3,6].

Among the distinctive characteristics of antennas for portable phones is the fact that their radiation pattern and the polarization direction cannot be considered as fixed, because of the random direction of the phones when used and their proximity to the user's body, which absorbs and scatters the electromagnetic energy. This makes it very impractical for antennas of portable phones to be characterized by omnidirectional radiation patterns and by vertical polarization.

In conclusion, the design considerations to be considered are the relatively large bandwidth ($\sim$10%), the need for a uniform coverage over the azimuthal plane with a high mean effective gain value, and the operating frequency bands if the antenna is mounted on a multiple-band phone, while the small size is more than essential. Under these conditions a wide variety of antenna structures, such as the monopole antenna in different forms, the helical antenna, and the planar inverted-F antenna (PIFA), as well as the microstrip patch, meander line and chip-type built-in antennas, are used in the phone technology. The majority of portable phones use a monopole antenna as the main element and a PIFA as the subelement forming a two-element diversity antenna, while other combinations are a monopole antenna and a helical or a meander line. These antennas are described here.

2.3.2. Types of Practical Antennas for Mobile Stations. Cellular phone antennas were developed initially from communication radio antennas used at lower frequencies. A *monopole antenna* is a quarter-wave whip antenna, which was the original type from which by a distributed inductive loading a significantly shortened antenna was produced (Fig. 7). An antenna of this type, known as a "rubber duck" on communication radios, can be accomplished by using a spiral enclosed in plastic or rubber having a total length of 5–15% of a wavelength [1,3]. Electrically, these antennas are still quarter-wave whips, which are tuned to a shorter length by distributed inductive loading, which is embodied by a helically wound wire. Generally the loading, both inductive and capacitive, can be added to a basic monopole type antenna to obtain better operating characteristics with a reduced physical height by maintaining a more constant current distribution for larger field strength or effective height and improved impedance matching. For all these antennas, which are basically quarter-wave elements, the cavity of the phone and, partially, also the user, are very important parts of overall antenna function. These antennas can be considered electrically as asymmetrically fed half-wave elements, with the feeding point located at the point where the antenna element enters the phone. At the commonly used frequencies, around 900 MHz, the body of the phone is of the order of half wavelength, and a short stubby antenna element is widely used. This is an essential way to extend the electrical size of the antenna when a low frequency is used. The evident disadvantage is that the main antenna current is flowing through the phone case and consequently also through the user, causing losses and probable undesirable medical effects at high powers levels. In order to eliminate these effects, in another approach half-wave whips having high feeding impedance can be used, resulting in low currents along the phone case. Half-wave whip antennas can also be made shorter than half wavelength if an inductive load is used, taking care that sufficient bandwidth remains assured. For easy transportation, the still rather long whip can be retracted into the phone case. It has to be noted that with the widespread use of cellular phones and the continuing improvement of cellular networks, more emphasis is presently being placed on a handy design rather than on maximum performance.

Helical antennas are usually constructed by a wire helically wound around a dielectric core as shown in Fig. 8. This type of antenna has an axial mode of radiation and a normal mode of radiation that is perpendicular to the axis of the helix. Although the axial-mode helix has been widely used as endfire directional elements, resonant normal-mode helical Antennas are useful as short, vertically

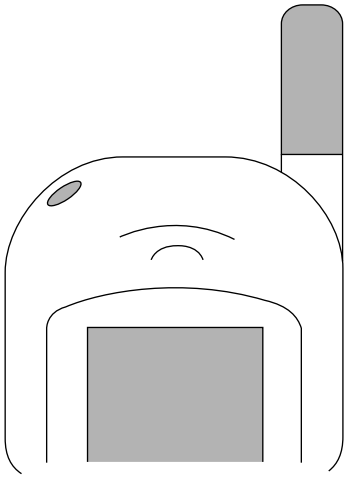


Figure 7. A reduced-length quarter-wave whip antenna by inductive loading.

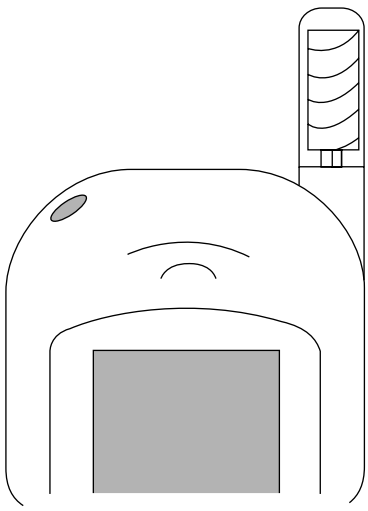


Figure 8. A normal-mode helical antenna.

polarized radiators, similar to the monopole. Using helical antennas, combined with inductive and dielectric loading, a considerable decrease in wave velocity can be achieved, which manifests a behavior similar to that of a quarter-wavelength antenna but with a physical length reduced to only 6–13% of a wavelength [1,3]. Another advantage is that the radiation resistance depends mainly on the exterior physical length and only slightly on the kind of helical winding. Also, the normal-mode helical antenna is more wideband in comparison to a straight monopole of the same length tuned by an internal inductance, because the axial current distribution gives up to 2.5 times higher radiation resistance. For cellular systems operating at 900 MHz, a normal-mode helical antenna has a length equal to 20–40 mm, with the minimum value determined by the necessary bandwidth (8–10%) [3].

A classical helical antenna with good performance, regardless of the grip and orientation of the phone, is shown in Fig. 9, where the helix configuration is on the bottom when the extendable whip is fully taken out. This antenna category is known as *helical Antennas with and without a whip* and can be developed for dual-band applications [3]. Since the whip has a nonmetallic top, when the whip is fully retracted the operation is similar to that of a fixed normal-mode helical antenna. When the whip is fully extended, it is connected in parallel to the helix, via a metallic connection that exists in its bottom. In this case, a part of the whip passes through the helix and detunes it so only the whip itself is fed. During this operation, the higher-frequency bands are slightly less sensitive to the influence of the user, and this can be explained by the fact that the distance from the skin of the user, expressed in wavelengths, is different. Moreover, a slight asymmetry in the azimuthal plane away from the head has been observed, which is desirable since it tends to decrease the losses in the head.

Instead of the classical helical conductor inside a normal-mode helical antenna, a printed meander pattern can be used. This kind of antennas is known as *meander patch antennas* [3,9]. Because of the easy fabrication process, virtually any meander shape can be obtained at low cost, and this advantage can be utilized for the creation of multiband operation and matching networks. The meander pattern can be printed on a flexible board, which then is rolled on a core. Because of the many possibilities for controlling the pattern, better optimization can be obtained than with a helical wire antenna, particularly for multiband functions. The different shapes that can be used as meander patterns include fractal patterns and patterns generated by genetic algorithms. For all these cases the added inductance is the important factor that guarantees the use of the small antenna at the lowest-frequency band, while the special shape of the pattern assures good multiband performance. A simple way to accomplish multiband operation is to connect two or more quarter-wavelength meanders in parallel, each tuned to its own desired frequency. The use of two or three frequency bands has become very common, typically with GSM 900, 1800, and 1900 MHz, allowing dual-band operation in GSM countries as well as in PCS (1900 MHz) areas in the United States. Because radiation resistance increases with the frequency quadratically, the result is that not only the absolute but also the relative bandwidth at higher frequencies will generally be higher. Thus for GSM, both systems at 1800 and 1900 MHz may very well be fitted within the higher band, while at 900 MHz a problem is eliminated if the stubby antenna is short. A dual-band meander antenna corresponding to a fixed normal-mode helical antenna is shown in Fig. 10.

In many new GSM phones there are *built-in antennas*; this means that the antennas are not visible from the exterior of the phone. Since at low-frequency bands (GSM 900 MHz) the antenna element plays the role of a kind of feed structure inducing currents on the phone body, which is actually the main radiator source, there

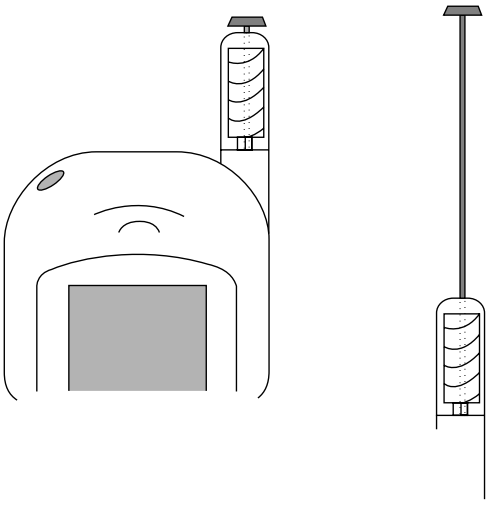


Figure 9. A helical antenna with a whip for dual-band applications.

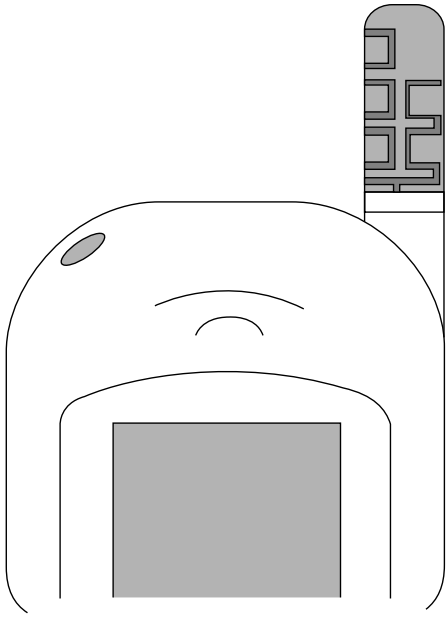


Figure 10. A meander patch antenna. The meander pattern has been printed on a flexible board and rolled on a core.

is really very little difference in whether the antenna elements are actually visible. At higher-frequency bands (GSM 1800/1900 MHz) the antenna element mainly operates independently from the phone body, but generally the same antenna element is used for both lower- and higher-frequency bands.

The radiation properties and the near field around the antenna elements are dependent mainly on the surface of the antenna element, while for the bandwidth and losses the most important characteristic is the volume occupied by the internal field. Usually, for installation, a surface on the upper back of the phone is preferred because the field very near to the antenna element should be kept away from the user to avoid unnecessary losses there. Also the antenna element should not be located too low on the back of the phone, as such a location could also result in increased losses in the hand of the user. The typical *planar inverted-F antenna* (PIFA) has been studied extensively because of its advantages, its low profile, and the easy incorporation into phone units [10]. It can be considered as a quarter-wave stub, which, when viewed from the open end, has a real admittance at quarter-wave resonance (Fig. 11). This L or inverted-L antenna can be considered as a variation of the monopole antenna where the monopole is bent over in an L shape with respect to a ground plane. The basic monopole structure is modified to reduce the height of the antenna while obtaining a lower resonant frequency than that of a comparable electrically short monopole. The short arm of the L antenna radiates in an omnidirectional pattern in the plane perpendicular to the vertical element, and some radiation is also obtained from the long arm of the L antenna.

The disadvantage of PIFA, which is its narrow bandwidth (1–2% in the relative bandwidth in free space) can be corrected when it is mounted on a finite ground plane. In practice, the bandwidth of an antenna system, which uses a built-in PIFA element in a phone, can be designed to have wider bandwidth due to the attribute from the phone body, which acts as a part of the radiator due to the current flow excited by the PIFA element. The physical length, which corresponds to a quarter-wavelength, can become shorter either by a capacitive loading (high ε) or by adding inductance (meandering, etc.). Thus, the characteristic impedance can be increased or decreased by inductive or capacitive loading, respectively, but the radiating conductance will remain the same. As a consequence, the bandwidth will be much wider for the inductively loaded case (meander, etc.) compared to the case with high ε. As already mentioned, volume will again be the critical parameter, and in both cases bandwidth will be smaller in comparison to that of the full-size PIFA. Regarding the implementation with meander patches, there is a limitation in that losses will increase if lines that are too narrow are used. The built-in antenna can also implement an optional whip, which can help in decreasing the influence of the hand of the user, combining the advantages of the built-in antenna with the performance of the extended whip.

It is worth mentioning that in the personal digital cellular system (PDC) used in Japan, mobile phones incorporate a system of a whip and a built-in PIFA antennas, with improved reception quality due to the diversity reception scheme. As can be seen in Fig. 12, the monopole element (whip), which has a length of either $\frac{3}{8}$ or $\frac{5}{8}$ wavelength in order to minimize the current flowing on the handset, has a normal-mode helical antenna on its top, which is encapsulated in a plastic cover. Because the two antenna elements are placed very closely, with a separation distance of only $\sim$0.1–0.2 wavelengths on the ground plane of the phone, the mutual coupling between them may degrade the radiation efficiency and also diversity function. Optimum values have been shown for the length of the whip element and the length of the phone unit in order to achieve low correlation coefficient without causing much degradation in radiation efficiency. The analyses have shown also that the diversity antenna performance depends not only on the correlation coefficient but also on the mean effective gain (MEG) of the antenna system.

Another interesting element is the *chip antenna*, which is a small normal-mode helical element molded in a ceramic chip, where ceramic materials, having a relative permeability of $\geq$20, are used. In order to assure wider bandwidth operation, a matching circuit or a PIN diode

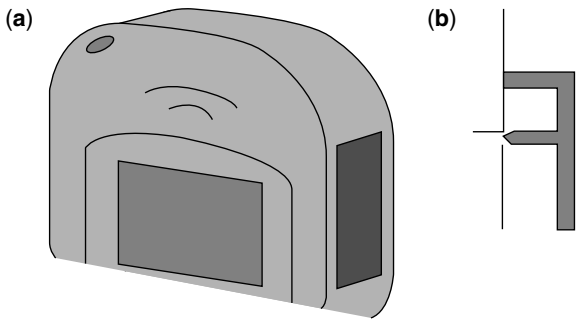


Figure 11. A PIFA antenna mounted on a side of a handset (**a**); side view of a PIFA (**b**).

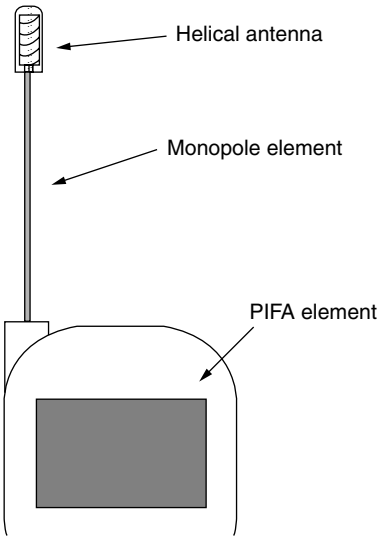


Figure 12. A handset of PDC with a PIFA element on its back, and a whip antenna with a normal-mode helical antenna on top.

circuit, which switches frequency to cover the necessary bandwidth, can be used.

3. ANTENNAS FOR MOBILE SATELLITE COMMUNICATION SYSTEMS

3.1. Introduction

In many cases terrestrial cellular communication systems cannot provide complete coverage over large global rural regions, due to the inability of installing base stations. A satellite-based system can fulfill this need by using either a few fixed geostationary satellites or a large number of low or medium earth-orbiting satellites.

The concept of artificial satellite was introduced in 1945 by A. C. Clarke, and since then more than 1000 satellites are in the geostationary or Clarke orbit, with more added at frequent intervals. Nearly all communication satellites are in geostationary earth orbit (GEO), for which a satellite appears stationary with respect to an observer on the ground because of its velocity match with that of the earth surface. Among the many existing GEO satellite systems, probably the best known is the *International Maritime Satellite System* (INMARSAT), which has provided international maritime satellite communication services since 1982, and is expanding its services to aircraft and land mobiles [9]. Many other systems such as AMSC in the United States, MSAT in many countries, AUSSAT in Australia, and ACES in Asia are using dedicated satellites to provide domestic satellite communication services mainly for land mobiles, such as voice and low-speed data.

Medium- and low-earth-orbiting satellite systems (MEOs and LEOs) have been introduced and use groups of low-altitude orbiting satellites (up to ~10,000 and ~1000 km for MEO and LEO, respectively). Because of the reduced distance between transmitting and receiving sites, as compared to those of GEO satellites systems (Fig. 13), less power and low-gain omnidirectional antennas can be used, and signal delay is also reduced. Typical examples of MEO/LEO satellite systems are Iridium (1.65 GHz, 66 satellites at 780 km), Globalstar (1.6/2.5 GHz, 48 satellites at 1414 km), Teledesic (19/29 GHz, 288 satellites at 1400 km), Ellipso, and ECCO [9].

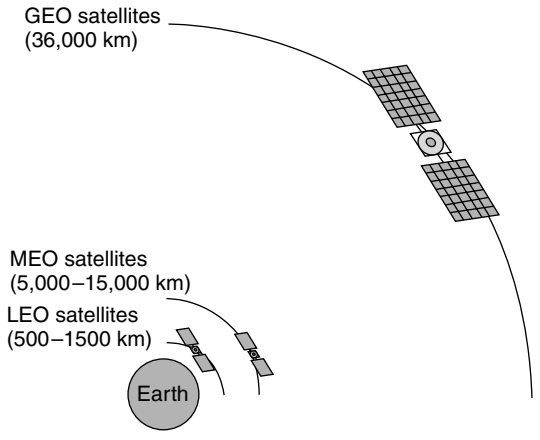


Figure 13. GEO, MEO, and LEO satellite orbits.

3.2. Antenna Design Considerations

Antennas play a significant role in the development and operation of satellite communications. They form the input and output ports to the satellite communication system, which can be divided generally in two segments: the ground and space segments. In every system, the signal is beamed into space by an uplink antenna and after the appropriate processing onboard the satellite is sent back to earth using a downlink antenna to be received by the earth-station antenna.

The types of antennas used depend on a number of factors related mainly to the distance between the satellite and the earth. Thus, for GEO satellite systems, because of the large distance (36,000 km), both satellite and ground-station antennas are characterized by high directive gain in order to overcome space loss, which translates into the demand for large aperture, pencil-beam antennas. Since earth-station antennas look upward at the sky, ground reflections are eliminated but the presence of the ground remains through its effect on the antenna noise temperature via sidelobe and backlobes. In all systems using GEO satellites, the L band is used (1.6/1.5 GHz) and the required frequency bandwidth to cover transmitting and receiving channels is ~8%. As a result, in using a narrowband antenna element such as a patch antenna, efforts have to be made to produce a wider bandwidth. The required gain is calculated by a link budget analysis taking into account the required channel quality and the satellite capability. Although gain is an essential parameter in antennas, the figure of merit G/T, which is the ratio of gain to system noise temperature, is a more commonly used factor in satellite communications. Since the antenna beam must cover $0-90^\circ$ in elevation and $0-360^\circ$ in azimuth directions, a tracking capability generally is needed, while in order to eliminate the need for polarization tracking, circular polarized waves are used.

Radiowave propagation over earth–space links for maritime mobile satellite systems lead to problems substantially different from those arising in the fixed satellite service. Thus, the effects of reflections and scattering by the sea surface become quite severe, especially in case of antennas with wide beamwidths, signal level attenuation due to blocking by the ship superstructure is not negligible, and the effect of ionospheric scintillation for L-band frequencies must also be taken into account. Multipath fading due to sea reflection must also be considered. The reflected waves are composed of a coherent (specular reflection) component and an incoherent (diffuse) component that fluctuates due to the motion of the sea waves [3].

The requirement for high-gain antennas for satellite systems can be relaxed using MEO or LEO satellites. The reduced distance also decreases signal delay, but because the satellites become nonstationary with respect to the ground surface, more complicated communications links are required. Shadowing by buildings is more severe for these systems than the GEO ones and generates polarization-sensitive reflection and diffraction that leads to multipath effects. Also, Doppler shifts (36 kHz–1.6 GHz and 55 kHz–2.5 GHz) have to be corrected with due regard to the relative direction of flight [3].

In implementing mobile satellite communications, of significant importance is the vehicle antenna. For an antenna system to be mounted on a mobile station, for satellite communications, it must be compact and lightweight. Additionally it must be easy to install and ensure mechanical strength. The installation requirement is not so severe for shipborne antennas, because even in small ships there is the space necessary to install an antenna system. However, in the case of automobiles, especially for small, private cars, low-profile and lightweight equipment is an essential requirement. The same demands generally hold for aircraft, plus the more severe required conditions in order to satisfy the standards for avionics, where one of the most important requirements for an aircraft antenna is low drag.

Antennas for mobile satellite communications can be classified as omnidirectional and directional antennas [3]. In the following sections typical examples of these antennas are described, and the most common types of antennas systems for aeronautical mobile communications are also presented.

3.2.1. Omnidirectional Antennas for Mobile Satellite Systems. In the category of omnidirectional antennas for mobile satellite systems, the most common ones are the quadrifiliar helical, the crossed–drooping dipole and printed antennas, and the microstrip patch and cavity-backed cross-slot antennas. These antennas, which are also used as elements in directional array antennas, are attractive because they have small size, are lightweight, and operate with circular polarization. Also, since their gain in the L band is 0–4 dBi, they do not require satellite tracking.

The *quadrifiliar helical antenna* (QHA) (Fig. 14) is composed of four identical helixes wound, equally spaced, on a cylindrical surface. The helixes are fed with signals equal in amplitude and 0°, 90°, 180°, and 270° in relative phase. The QHA has a height equal to 40 cm and presents a gain with a minimum value equal to −4 dBi, and axial ratio of ≤3 dB [3]. The *crossed-drooping dipole antenna* is the best choice for land-mobile satellite systems, when the required angular coverage must be narrow in elevation and almost constant in azimuth. The variation of the separation distance between the dipole elements and the ground plane adjusts the elevation pattern to ensure an optimum pattern for the coverage region of interest

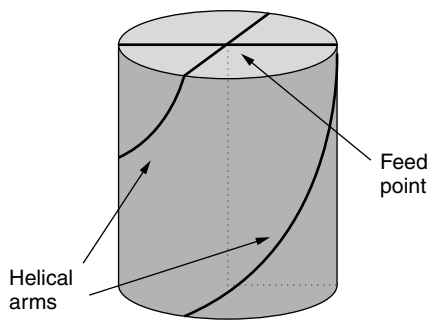


Figure 14. The quadrifiliar helical antenna.

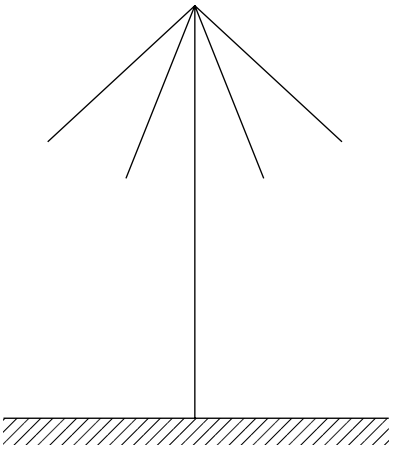


Figure 15. The crossed–drooping dipole antenna.

(Fig. 15). The crossed–drooping dipole antenna is characterized by a minimum gain equal to 4 dBi and a maximum axial ratio value equal to 6 dB for a height of 15 cm [3].

The *microstrip patch antenna* (MSA) [10,13,14] incorporates a circular metallic disk on a dielectric grounded substrate, and in order to produce circularly polarized waves, it is excited at two points orthogonal to each other with signals equal in amplitude and 90° phase difference (Fig. 16). A higher-mode patch antenna can also be designed with a similar radiation pattern to the drooping dipole. To produce conical radiation patterns (null on axis) suitable for land-mobile satellite applications, the antenna is excited at higher-order modes. The circular microstrip patch antenna is characterized by 3.5 dBi minimum gain value and 4 dB maximum axial ratios for a height of 1 cm when RT/Duroid is used as dielectric substrate. Because the available frequency bandwidth of this patch antenna is very narrow, the two-layer patch antenna is used in which the upper and lower parts play a role for transmission and reception, respectively. In a two-layer patch antenna each layer is individually fed at two points with a phase difference of 90° for circular polarization. Another useful form of printed antenna is the *cavity-backed cross-slot antenna* (XSA) [3]. In this case each slot antenna is fed with an equal amplitude and in-phase condition at two points equidistant from the center. One advantage of this antenna is that the input impedance can be matched for a wider frequency band than in the case of the slot antenna. The general

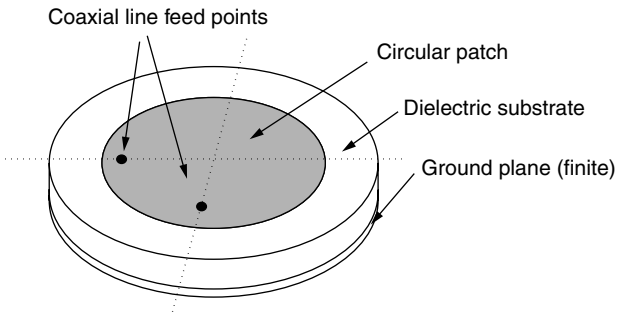


Figure 16. A circular microstrip antenna excited to produce circularly polarized waves.

characteristics of the MSA are antenna gain at boresight $\sim$15.2 dBi and at a scanning angle $\theta = 45^\circ$ the gain is equal to about 13.5 dBi, while for the XSA, the boresight gain is $\sim$15.7 dBi and the gain at $\theta = 45^\circ$ is $\sim$14 dBi [3].

One of the main and most common uses of these antennas is in the Navigation System with Time and Ranging/Global Positioning System (NAVSTAE/GPS), which is the most widely used navigation system at present. Since a GPS satellite transmits two radiofrequencies (1575.42 and 1227.6 MHz), with right-hand circular polarizations and for accurate positioning, both frequencies must be used in order to compensate for the excess delay of radiowaves in the ionosphere, and an antenna operating equally well at both frequencies is needed [3]. From the omnidirectional antennas, QHA and MSAs have been widely used because of their simplicity, small size, and low cost, with slightly modified design parameters to ensure that their radiation patterns are as uniform as possible over the upper hemisphere. Thus, for dual-frequency operation two QHAs are used, one for every frequency, into one structure by coaxially mounting them one into the other or by one on the top of the other. In the case of MSAs, which are extensively used as GPS antennas because of their printed antennas advantages, one effective method of obtaining a broadbeam for GPS reception is the reduction of the size of their ground planes.

3.2.2. Directional Antennas for Mobile Satellite Systems. Directional antennas for mobile satellite communication systems are used in all INMARSAT systems, as well as in PROSAT, ETS-V, MSAT, and MSAT-X research programs for mobile satellite communications. A short description of the basic antennas for the various INMARSAT systems will be given here.

The typical antennas for INMARSAT-A, -B, and -F systems (a description of available INMARSAT systems can be found in Ref. 9) are aperture antennas such as a *parabolic antenna*, which is characterized as a simple structure with high-aperture efficiency. The parabolic antenna can have a gain of 20–23 dBi, a bandwidth of $\sim 10^\circ$, while satellite tracking is required because of ship motions and the small half-power beamwidth. In the case of the INMARSAT-C antenna system, the simplest and most compact configuration is required, that is, without mount systems and tracking/pointing systems. Therefore the antennas used are usually omnidirectional antennas, and the most suitable ones are the antennas described above, namely, the QHA, the cross-drooping dipole, and the MSA. The QHA is the most appropriate for ships because of its good performance of widebeam coverage under the condition of ship motion, while in handheld or briefcase terminals, where a very low profile characteristic is necessary, the obvious choice is the MSA.

Since the INMARSAT-M is used mainly for small ships, such as fishing boats and land vehicles, the kinds of antennas used have to be of small size and low cost. Concerning the efficient utilization of satellite power and the required G/T, antenna gain ranges from 13 to 16 dBi. A *short-backfire antenna* (SBF) is one of the most commonly used M-terminal antennas for maritime applications, especially in the improved form in which the flat-disk

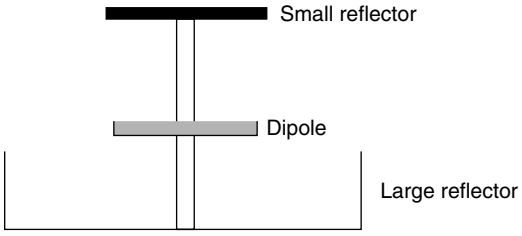


Figure 17. The short-backfire antenna.

main reflector is replaced by a conical or step plate plus a second small reflector in order to achieve better aperture efficiency (80%) and frequency bandwidth (20% instead of 8% of the normal SBF) (Fig. 17). Another antenna suitable for the INMARSAT-M is the *quad helical antenna*, used as a single antenna or as array elements. This antenna, which can be considered as a compromise between the dipole and loop antenna, operates in the so called axial mode, when the helix has a pitch angle of $12-15^\circ$ and the circumference is about one wavelength. The two-turn helical antenna is characterized by very good polarization characteristics for its size and high gain, affected mainly by the size of the reflector. The quad-helical array antenna is composed of four two-turn helical antennas in a square arrangement whose elements are oriented in the manner shown in Fig. 14. This antenna has gain of $\sim$13 dBi (at HPBW = 38°), an axial ratio of $\sim$1.0 dBi, and an aperture efficiency of $\sim$100%. Microstrip and slot antennas can also be used in the INMARSAT-M system in the form of planar arrays. Regarding the INMARSAT-Aero system, phased-array antennas are the best candidates for airborne antennas because of their low profile and mechanical strength. At the present time, two types of phased-array antennas have been used, the conformal type, with the important advantage of the low air drag because of its low profile, and the top-mount type. The first one consists of two sets of phased arrays on both sides of a fuselage, while the second one has a set of phased arrays on the top of the fuselage.

3.2.3. Antenna Systems for Aeronautical Mobile Communications. Today, many different types of antennas are mounted on modern aircraft and serve many individual functions associated with specific avionics systems such as navigation, identification, or radar. These antennas, which are characterized as airborne antennas, must be designed and manufactured in such a way as to satisfy specific electrical requirements under the presence of the stringent environmental and aerodynamic conditions. Thus, airborne antennas must have structures that do not increase the aerodynamic drag and must operate without degrading their basic electrical characteristics under the influence of great scale pressure, temperature, and humidity variations. Moreover, they have to withstand great acceleration differences, static electricity, and lightning. Finally, they must be lightweight, and of small size and low profile. Their individual electrical characteristics are dependent mainly on the function the antenna serves. Generally, these are strongly influenced by the shape and the size of the airframe in relation to the wavelength used. When the wavelengths are significantly

larger than the maximum dimensions of the aircraft, which occurs at low and medium frequencies, the antennas are characterized by low radiation efficiency, which results in a high Q value. In this case, careful matching of the antenna over the necessary frequency band is needed. When the airframe size can be considered larger than the wavelength, for very high and ultrahigh frequencies, the antenna or a part of the airframe can become resonant and in this case more degrees of freedom, regarding the design and the position of installation of the antenna, exist. In any case the influence of the airframe on the radiation pattern of antennas must be considered, since shadowing and reflection on the airframe can result in significant distortions and shape changes of the desired radiation pattern.

Between the various types of antennas installed on an aircraft, those used for satellite systems are the most interesting and complicated. A common antenna element is the microstrip patch antenna, which can be used as a high-gain circularly polarized radiator or as an element to form phased-array and shape-beam antennas [10]. In practical applications, circular and rectangular shapes of the patch radiator are used to achieve circularly polarized patterns. Its main advantages are that it can be made conformal to metallic surfaces; is low-profile, lightweight, and small-size; and can be produced at low cost. Its basic disadvantage of narrow frequency bandwidth, of the order of 2%, can be overcome by either (1) increasing the thickness of the substrate and decreasing its dielectric constant or (2) using stacked patches, electromagnetically coupled together [10,13]. In the first technique, countermeasures must be taken to improve the axial ratio, which degrades with the generation of higher-order modes. The second technique also allows the generation of a dual-frequency antenna system, when transmit and receive frequency bands are well separated, by assigning each frequency band to every patch. In both cases, the degenerate modes for circular polarization can be produced, by giving an appropriate perturbation to the patch dimensions and selecting the suitable feedpoint. This eliminates the necessity of an external circuit.

Other elements that can be used instead of the microstrip patch are the electromagnetically complementary radiators, and crossed-dipole and crossed-slot antennas. Using a pair of orthogonally positioned dipoles or slots fed with equal amplitudes and quadrature phase, circular polarization can be obtained. The two sets differ mainly in terms of the feeding method, which in both cases is quite complicated, while special techniques have been developed to improve the axial ratio off boresight.

Another element is the quadrifiliar helical antenna in the resonant form, which is characterized by small size, no ground-plane requirement, and immunity to effects due to the presence of nearby metal structures.

BIOGRAPHIES

Christos Christodoulou received the B.Sc. degree in physics and math from the American University of Cairo in 1979, and the M.S. and Ph.D. degrees in Electrical Engineering from North Carolina State University, Raleigh, in 1981 and 1985, respectively. He served as a faculty member in the University of Central Florida, Orlando, from 1985 to 1998, where he received numerous teaching and research awards. In 1999, he joined the faculty of the Electrical and Computer Engineering Department of the University of New Mexico, Albuquerque as a Chair. In 1991 he was selected as the AP/MTT Engineer of the year (Orlando Section). He is an IEEE Fellow and a member of URSI (Commission B). He has published over 150 papers in journals and conferences. He also has a book on "neural network applications in electromagnetics." He is, currently, the co-editor for a column on "Wireless Communications" for the IEEE AP Magazine and the associate editor for the IEEE Transactions on Antennas and Propagation. His research interests are in the areas of wireless Communications, modeling of electromagnetic systems, smart antennas, neural network applications in electromagnetics, and reconfigurable/MEMS antennas.

Michael T. Chryssomallis received the Diploma in Electrical Engineering from Democritus University of Thrace, Greece, in 1981. He also received the Ph.D. degree in Electrical Engineering from Democritus University in 1988. In 1982, he joined Democritus University as a Scientific Collaborator up to 1989, then up to 1994 as a Lecturer and up to now as an Assistant Professor.

He worked with the Communications Group (director Prof. P. S. Hall) of the University of Birmingham for the period of Oct. 1997–Jan. 1998, in the areas of Active Antennas, and with the Wireless Group (director Prof. C. G. Christodoulou) of the University of New Mexico for the periods of April to June 2000, and April to July 2002, in the areas of Microstrip Antennas and Arrays, and Smart Antennas.

He is serving as a reviewer for the IEEE Transactions of Antennas and Propagation and has published several journal and conference papers. His current research interests are in the areas of microstrip antennas, RF-Mems, smart antennas and propagation channel characterization. He is member of the IEEE since 1988 and senior member since 2000.

BIBLIOGRAPHY

1. C. A. Balanis, *Antenna Theory, Analysis and Design*, Wiley, New York, 1997, Chap. 14.
2. J. D. Gibson, ed., *The Mobile Communications Handbook*, CRC Press and IEEE Press, 1996, Sec. II.
3. K. Fujimoto and J. R. James, eds., *Mobile Antenna Systems Handbook*, Artech House, Boston, 2001.
4. J. D. Kraus and R. Marhefka, *Antennas*, McGraw-Hill, New York, 2001, Chap. 21.
5. T. S. Rappaport, *Wireless Communications, Principles and Practice*, IEEE Press and Prentice-Hall, New York, 1996, Chap. 2.
6. S. R. Saunders, *Antennas and Propagation for Wireless Communication Systems*, Wiley, Chichester, UK, 1999.
7. R. J. Holbeche, ed., *Aerials and Base Station Design*, IEE Telecommunications Series 14, Peter Peregrinus, 1985, Chap. 4.

8. M. Chryssomallis, Smart Antennas, *IEEE Antennas Propag. Mag.* **42**(3): 129–136 (June 2000).

9. L. C. Godara, ed., *Handbook of Antennas in Wireless Communications*, CRC Press, Boca Raton, FL, 2002.

10. R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, *Microstrip Antenna Design Handbook*, Artech House, Boston, 2001.

11. K. Siwiak, *Radiowave Propagation and Antennas for Personal Communications*, Artech House, Boston, 1998, Chap. 11.

12. K. Fujimoto, A. Henderson, K. Hirasaura, and J. R. James, *Small Antennas*, Research Studies Press, UK, 1987.

13. J. R. James and P. S. Hall, eds., *Handbook of Microstrip Antennas*, Peter Petegrinus, London, 1989, Vols. 1 and 2.

14. D. M. Pozar and D. H. Schaubert, *Microstrip Antennas, The Analysis and Design of Microstrip Antennas and Arrays*, IEEE Press, Piscataway, NJ, 1995.

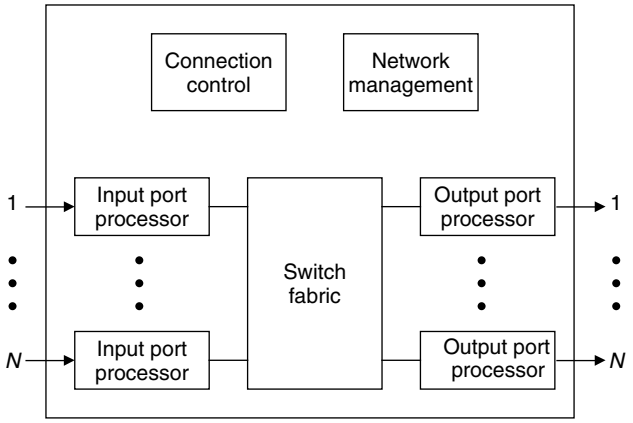


Figure 1. A generic ATM switch architecture.

ATM SWITCHING

THOMAS M. CHEN
Southern Methodist University
Dallas, Texas

STEPHEN S. LIU
Verizon Laboratories
Waltham, Massachusetts

1. INTRODUCTION

ATM (asynchronous transfer mode) is an internationally standardized connection-oriented packet switching protocol designed to support a wide variety of data, voice, and video services in public and private broadband networks [1,2]. ATM networks generally consist of ATM switches interconnected by high-speed transmission links. ATM switches are high-speed packet switches specialized to process and forward ATM cells (packets). Since ATM is a connection-oriented protocol, ATM switches must establish a virtual connection from one of its input ports to an output port before forwarding incoming ATM cells along that virtual connection.

An ATM cell consists of a 5-byte header followed by a 48-byte information field or payload. The main purpose of the ATM cell header is to identify the virtual connection of the cell that occupies most of the header bits. An ATM virtual connection is specified by the combination of a 12-bit virtual path identifier (except the first 4 bits are used for generic flow control at the user-network interface) and a 16-bit virtual channel identifier. Virtual paths are bundles of virtual channels. VP cross-connects are designed to route ATM traffic on the basis of virtual paths only, which is convenient when large amounts of traffic must be routed or rerouted at the same time. The VPI/VCI fields are followed by a 3-bit payload type (PT), 1-bit cell loss priority (CLP), and 8-bit header error control (HEC) field. The PT field is used to distinguish control cells from data cells, and explicit forward congestion indication (EFCI). The CLP flag is used to indicate that lower priority (CLP = 1) cells should be discarded before CLP = 0 cells in the event of congestion. The HEC field allows single bit error correction and multiple bit error detection over the cell header.

A generic ATM switch architecture with N input ports and N output ports is shown in Fig. 1 (note that switches can have any dimensions). The functions of an ATM switching system may be divided broadly into user cell forwarding, connection control, and network management [3]. ATM cells containing user data are received at the input ports, and the input port processors prepare the cells for routing through the switch fabric. The fabric in the center of the switching system provides the interconnections between input port processors and output port processors. The output port processors prepare the outgoing user cells for transmission from the switch. User cell forwarding is characterized by parallelism and high-speed hardware processing. The ATM protocol was intentionally streamlined to allow incoming cells to be processed simultaneously in hardware and routed through the switch fabric in parallel. Thus, ATM switches have been able to realize high-end performance in terms of throughput and cell forwarding delay.

Connection control, sometimes called the *control plane*, refers to the functions related to the establishment and termination of ATM virtual connections. Connection control functions generally encompass: exchange and processing of signaling information, participation in routing protocols, and decisions on admission or rejection of new connection requests.

Network management is currently carried out by SNMP (Simple Network Management Protocol), the standard protocol for managing data networks. ATM switches typically support an SNMP agent and an ATM MIB (management information base). The SNMP agent responds to requests from a network manager to report status and performance data maintained in the MIB. The agent might also send alarms to the network manager when prespecified conditions are detected. Since ATM switches can be viewed as a specific type of network element covered within the SNMP framework, the details of SNMP functions in ATM switches are not discussed in detail here.

Network management should also include standardized ATM-layer OAM (operations and maintenance) functions. ATM switches carry out OAM procedures by generating,

exchanging, processing, and terminating OAM cells. OAM cells are used for fault management, performance management, and possibly other ATM-layer management functions.

2. INPUT AND OUTPUT PORT PROCESSING

The input port processors carry out several important functions. First, the physical layer signal is terminated. For the common case of SONET/SDH (synchronous optical network/synchronous digital hierarchy), the SONET/SDH framing overhead fields are processed, and the payload is extracted from the frame. Individual 53-byte ATM cells are delineated in the payload.

Next, each cell header undergoes a number of processing steps. The cell header is checked for bit errors using the HEC field, and cells with uncorrectable header errors are discarded. The traffic rate of each virtual path or virtual channel is monitored according to an algorithm called the *generic cell rate algorithm* (GCRA), which is essentially a leaky-bucket algorithm. A switch may be configured to allow, discard, or "tag" cells (by setting CLP = 1) exceeding the allowed traffic rate. The VPI/VCI value in each cell header is used to index a routing table to determine the proper output port and outgoing VPI/VCI values. Incoming VPI/VCI values must be translated to outgoing VPI/VCI values by every ATM switch. Cells requiring special handling, such as signaling cells and OAM cells, must be recognized and routed to the appropriate processors in the switch. User cells are prepared for routing through the switch fabric, often by prefixing a routing tag to the cell. The routing tag may consist of the output port, service priority, type of cell, timestamp, or other information for routing and housekeeping purposes. Since the routing tag exists only within the switch, its contents may be chosen entirely by the switch designer. Before entering the switch fabric, cells may be queued in a buffer in the input port processor.

The output port processors have the opposite role of the input port processors, namely, preparing ATM cells for physical transmission from the switch. Cells from the switch fabric may be queued in a buffer in the output port processor, in which case the switch is called an *output-buffered switch*. If routing tags are used, the output port processors remove the routing tag from each user cell. If special cells, such as signaling cells and OAM cells, need to be transmitted, they are inserted into the outgoing cell stream. A new HEC field is calculated and inserted into each cell header. Finally, the ATM cells are transmitted as a physical layer signal. In the case of SONET/SDH, cells are mapped into the payloads of SONET/SDH frames.

3. SWITCH FABRICS

It is often convenient to visualize the basic operation of an $N \times N$ switch synchronized to periodic time intervals equal to the transmission time of one cell, referred to as a "cell time," assuming that the transmission rate on all links are equal. For example, the cell time for a 155-Mbps (megabit-per-second) transmission link would be approximately 53 bytes/155 Mbps = 2.7 μs. In each cell time, a new set of N incoming cells may appear at the input ports and up to N outgoing cells may depart from the output ports. The N incoming cells are processed in parallel by the input port processors and presented simultaneously to the switch fabric. The switch fabric attempts to route the cells in parallel to their appropriate output ports.

There is a chance that more than one cell may attempt to reach the same output port at the same time, called *output contention*, which has four significant consequences:

1. One cell may reach the output port but the other cells would be lost without buffers existing somewhere in the switch to temporarily store these cells. Switch fabric designs differ in their choice of buffer placement.
2. Queues may accumulate in the buffers resulting in random cell delay and cell delay variation, which are usually two performance metrics of interest.
3. Buffers are necessarily finite implying the possibility of buffer overflow and cell loss. Some fabric designs may not be able to handle a full traffic load without a probability of cell loss. A performance metric for switch fabrics is the normalized throughput or utilization defined as the overall fraction of a full traffic load that can be forwarded successfully through the fabric. Ideally, switch fabrics should be capable of 100 percent utilization.
4. Some switch fabrics must operate at a rate faster than the transmission link rate. The ratio of the switch fabric rate to the transmission link rate is sometimes referred to as a "speedup factor." A speedup factor is often related to the *scalability* of a switch fabric design, that is, the difficulty of constructing an arbitrarily large fabric [3].

3.1. Shared Memory and Shared Medium

A speedup factor of N is evident in switch fabric designs based on a shared memory or shared medium, which are shown in Fig. 2. In a shared memory design, incoming cells are first converted from serial to parallel form. They are written sequentially into a dual-port random-access memory [4]. Their cell headers with routing tags are directed to a memory controller that keeps track of the memory location of all cells associated with each output port. The memory controller links the memory location of outgoing cells to maintain virtual output queues. The outgoing cells are read out of the memory, demultiplexed, and converted from parallel to serial form for delivery to the output port processors. Since the cells must be written into and read out from the memory one at a time, the shared memory must operate at the total throughput rate. Hence, it must be capable of writing N cells and reading N cells in one cell time, implying a speedup factor of N. As a consequence, the size of the fabric, N, will be limited by the memory access time. On the other hand, a shared memory design has been popular due to its simplicity and efficient sharing of memory space.

Similarly, cells may be passed from input port processors to output port processors through a high-speed time-division multiplexed (TDM) bus. Incoming

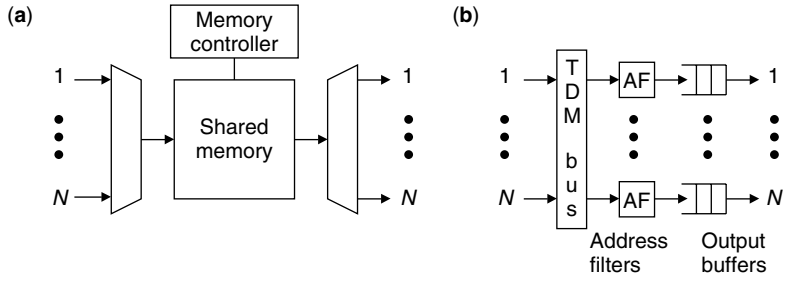


Figure 2. Prototypical switch fabric designs based on (**a**) shared memory and (**b**) shared medium.

cells are sequentially broadcast on the bus. At each output, address filters examine the routing tag on each cell to determine whether the cell is addressed for that output. The address filter passes the appropriate cells through to an output buffer. Shared bus designs have been used in traditional router architectures due to their simplicity and modularity. On the other hand, the buffer space is not shared as efficiently as a shared memory. Also, the bus speed must be fast enough to carry up to N cells in each cell time, corresponding to a speedup factor of N. The address filters and output queues must operate at the bus speed as well. The size of a shared bus fabric will be limited by the expense and complexity of high-speed hardware for the bus, address filters, and output queues.

3.2. Space Division

A simple example of a space division fabric is a crossbar switch shown in Fig. 3, which was originally developed for telephone switching. An $N \times N$ matrix of crosspoints can connect any of the N inputs to any of the N outputs. While a crossbar switch has the advantages of simplicity and no speedup factor, it has two major disadvantages:

1. A crossbar switch will have output blocking, meaning that only one cell may be delivered to an output port at a time. Other cells contending for the same output port may be queued at the input ports, but the normalized throughput for an input buffered fabric is well known to be only $2 - 2^{1/2} = 0.586$ for large N, assuming uniform random traffic; that is, an incoming cell attempts to go to any output port with equal probability independent of all other conditions [5].
2. The N^2 number of crosspoints does not scale well to large fabrics. As an alternative, multistage interconnection networks (MINs) have been studied extensively over many years of development of telephone switches [6]. MINs are constructed by connecting a number of small switching elements, often 2×2 switching elements, in a regular pattern. Banyan networks are a popular class of MINs used for ATM switch fabrics. Figure 4 shows an example of an 8×8 banyan network. The dashed outlines emphasize that the 8×8 banyan network is constructed by adding a third stage to interconnect 4×4 banyan networks, which are in turn constructed by an interconnection of two stages of 2×2 switching elements. An n-level banyan may be constructed by connecting several $(n-1)$-level banyans with an additional stage of switching elements. This recursive and modular construction of larger fabrics is a significant advantage for implementation.

Another advantage is the simplicity of the 2×2 switching elements. Each 2×2 switching element routes a cell according to a control bit. If the control bit is 0, the cell is routed to the upper output (address 0); otherwise, the cell is routed to the lower output (address 1). Delta networks are a subclass of banyan networks with the "self-routing" property: the output address of a cell also controls the route of that cell. For example, the cell shown in Fig. 4 is addressed to output port 010. The nth bit of the address "010" is used as the control bit in the nth stage to route the cell to the proper output port, and this self-routing works regardless of which input port the cell starts from.

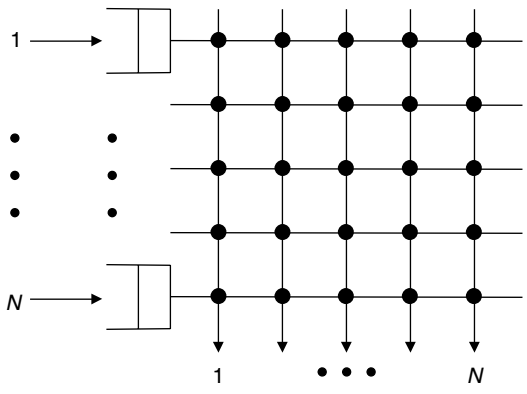


Figure 3. $N \times N$ crossbar switch fabric.

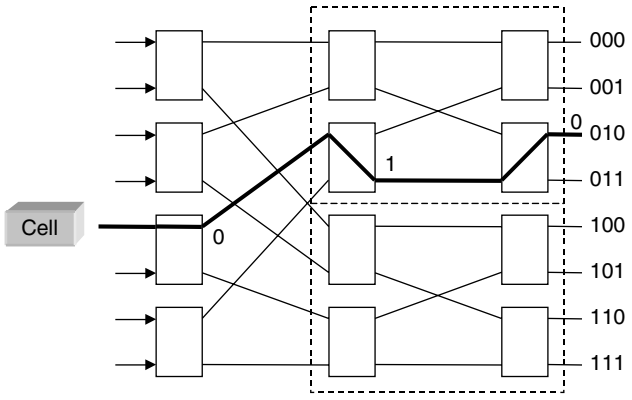


Figure 4. Example of an 8×8 banyan network.

The self-routing property simplifies the control of the delta network switch fabric.

Delta networks can take different forms, depending on their method of construction, including omega, flip, cube, shuffle exchange, and baseline networks [7]. A delta network of size $N \times N$ constructed of $M \times M$ switching elements will have $\log_M N$ stages, each stage consisting of N/M switching elements. Hence, if $M = 2$, the total number of crosspoints will be on the order of $N \log_2 N$, which compares favorably to N^2 crosspoints in a crossbar switch.

Unfortunately, the savings in number of crosspoints comes at the cost of possible internal blocking, meaning that the routes of two cells addressed to different outputs might conflict for the same internal link in the fabric before the last stage. In this situation, only one of the two cells for a link can be passed to the next stage, while the other cell stays behind, queued either in a buffer within each switching element or in an input buffer. Thus, internal blocking will cause a loss of throughput. A well-known solution is to add a Batcher sort network to rearrange the cells according to an increasing or decreasing order of addresses before the banyan network [8]. A combined Batcher–banyan network will be internally nonblocking in that a set of N cells addressed to N different outputs will not cause an internal conflict. However, output blocking can still occur if two cells are addressed to the same output, and it must be resolved by buffering.

An obvious possibility is input buffering before the Batcher–banyan network. If more than one cell is addressed to the same output, one cell is allowed to pass through the Batcher–banyan network while the other cells remain in the input buffers. Naturally, throughput will be lost due to the so-called head-of-line blocking, where a delayed cell prevents the other cells waiting behind it from going through the fabric. Many approaches are possible to overcome the head-of-line blocking problem and increase the throughput of the fabric, such as increasing the speedup factor, distributing the traffic load to multiple banyan networks in parallel, cascading multiple banyan networks in tandem, or virtual output queueing where N separate virtual queues corresponding to the output ports are maintained at each input port. Although these solutions add complexity to the fabric implementation, space-division fabrics are still attractive for their ability to scale to large sizes. Large fabrics may be constructed as MINs composed of small switching modules, where the small switching modules can be any type of fabric design.

3.3. Input and Output Buffering

The placement of buffers in the switch can have a significant effect on the switch performance. Fortunately, this issue has been studied extensively. Figure 5 shows three basic examples: input buffering, output buffering, and internal buffering. Input buffering is known to suffer from head-of-line blocking without special provisions to overcome it. Output buffering is generally agreed to be optimal in terms of throughput and delay [5]. However, output buffering often involves a speedup factor which limits the scalability to large fabrics.

The addition of buffers within the switching elements of a banyan network to resolve internal blocking has not been shown to improve the throughput substantially. An interesting fabric design is a crossbar switch with buffers at each crosspoint [9]. Incoming cells are dropped into the appropriate buffer corresponding to the output. Each output multiplexes the cells queued in N buffers. The buffered crossbar switch (also called *bus matrix switch*) is actually an output buffered fabric as illustrated in Fig. 6. It offers the desirable performance of output buffering with no speedup factor. However, it does not share buffer space efficiently, and the number of output buffers scales exponentially as N^2. The knockout switch shown in Fig. 6 reduces the number of output buffers to NL, where L is a constant by the addition of N : L concentrators at each output [10]. It has been noted that under uniform random traffic conditions, the probability of more than L cells

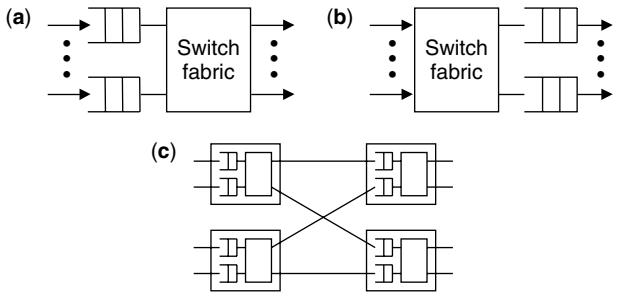


Figure 5. Examples of (**a**) input buffering, (**b**) output buffering, and (**c**) internal buffering.

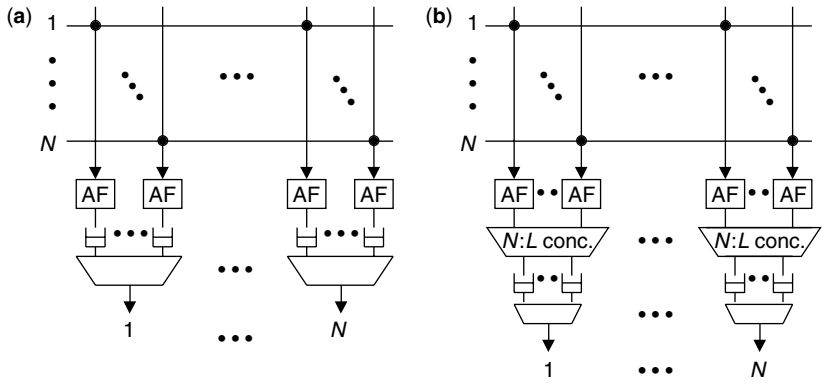


Figure 6. (**a**) Crossbar switch with buffers at each of N^2 crosspoints; (**b**) knockout switch with NL buffers.

addressed to the same output port in the same cell time will be very small if L is chosen appropriately large. The N : L concentrators allow up to only L cells to pass in one cell time to L output buffers at each output; additional cells are lost. However, if L is chosen to be 8 or greater, the cell loss ratio will be 10^{-6} or less. At the cost of a small cell loss, the scalability of the fabric becomes linear with N instead of exponential.

4. CONNECTION CONTROL

Since ATM is a connection-oriented protocol, virtual connections must be established before any user cells can be forwarded. Virtual connections may be permanent, semipermanently controlled through network management, or dynamically established by means of ATM signaling in response to user requests. ATM switches exchange signaling messages along a selected route and make decisions about allocation of switch resources to new user requests. Usually route selection is carried out by a separate process. Routes may be static or dynamically chosen through a routing protocol. The PNNI (private network node interface) routing protocol is a dynamic link-state routing protocol similar to the OSPF (open shortest path first) protocol used in the Internet.

4.1. Signaling

ATM switches must participate in signaling protocols, either access signaling between the user and edge switch or interoffice signaling between two switches. The ATM access signaling protocol is the ITU-T standard Q.2931, which was derived from the ISDN access signaling protocol Q.931. Q.2931 signaling messages are encapsulated in ATM cells using a signaling ATM adaptation layer (SAAL) protocol. Signaling cells are exchanged on a preestablished signaling virtual channel (VCI = 5) or another signaling virtual channel dynamically established through metasignaling (a preestablished metasignaling virtual channel identified by VCI = 1).

The high-layer interoffice signaling protocol is the ITU-T standard BISUP (broadband ISDN user part) derived from the ISDN user part of Signaling System 7 (SS7). BISUP messages may be exchanged directly between ATM switches, where BISUP messages would be encapsulated into ATM cells using SAAL, or sent through the existing SS7 packet-switched network.

Figure 7 shows a typical exchange of signaling messages between ATM switches to successfully establish and release a virtual connection. Basically, a Q.2931 "setup" message is first sent by the user to request a new virtual connection. If each switch decides to accept the request, a BISUP "initial address" message (IAM) is forwarded along a selected route. The IAM message includes all information required to route the connection request to the destination user, such as destination user address, service class, ATM traffic descriptor, connection identifier, quality-of-service (QoS) parameters, and additional optional parameters. A Q.2931 "setup" message notifies the destination user. If the connection is accepted, a series of signaling messages are returned in the reverse direction to alert the calling party that the connection is established. The reverse signaling messages also serve to finalize the resource reservations that were made earlier tentatively in each switch. When the virtual connection is no longer needed, a "release" message will free the reserved resources at each switch to be used for another connection.

The complete ATM signaling protocol, including additional signaling messages, options, and timing requirements, is elaborate to implement. Obviously, signaling cells require special processing within the switch. Incoming signaling cells are recognized and diverted to a signaling protocol engine for processing. Outgoing signaling cells from the signaling protocol engine are multiplexed into the outgoing cell streams.

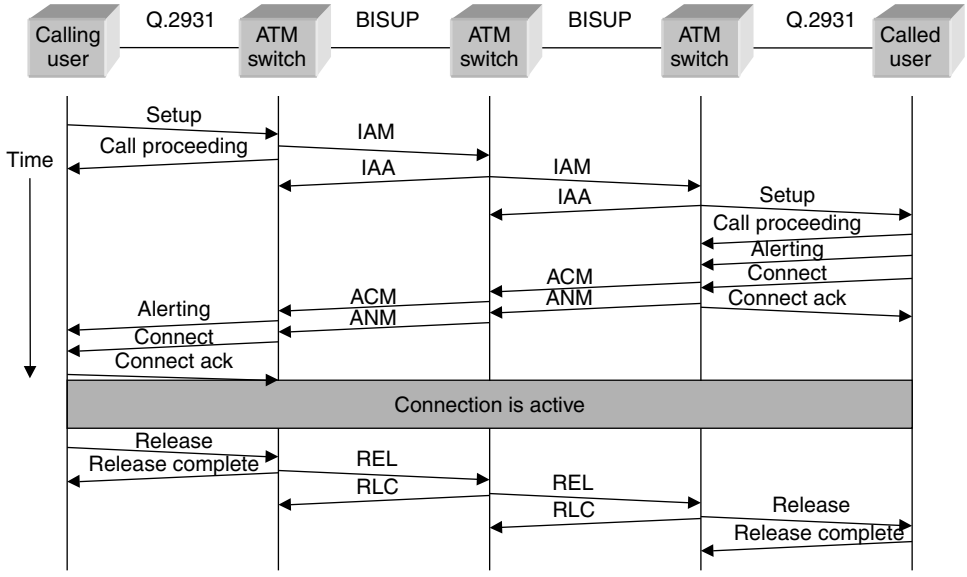


Figure 7. Exchange of signaling messages involved in a successful connection.

4.2. Connection Admission Control

ATM supports the notion that accepted virtual connections will be guaranteed their requested level of QoS — mainly in terms of maximum cell delay, cell delay variation, and cell loss ratio — or otherwise, a new connection request should be rejected. Hence, the acceptance of a virtual connection is an implicit agreement between the user and the network on a mutual understanding of their respective obligations, often called a "traffic contract." The user side of the traffic contract involves conformance to the ATM traffic descriptor or traffic rate parameters. The network side of the traffic contract is a guarantee of the requested QoS for the conforming traffic.

Naturally, not every connection request may be accepted because network resources are shared for higher efficiency. If too much traffic is admitted, the QoS for existing connections may deteriorate below their guaranteed levels. On the other hand, the network should attempt to accept as much traffic as possible to maximize efficiency and revenue. *Connection admission control* (CAC) refers to the general process for deciding acceptance or rejection of new connection requests. The main issue for an ATM switch is whether sufficient resources are available to satisfy the QoS requirements of the new connection and all existing connections. Because ATM traffic is random, the effect of a new connection cannot be known precisely during CAC. The switch follows a CAC algorithm chosen by the network provider to estimate the impact of a new connection.

The numerous CAC algorithms studied over the years can be broadly classified as deterministic or statistical. Deterministic approaches calculate the effect of a new connection on the basis of a deterministic traffic envelope characterizing a bound on the shape of the expected traffic, such as peak cell rate or a leaky-bucket-limited envelope. Statistical methods usually estimate the effect of a new connection by carrying out a stochastic analysis of a queueing model. Statistical methods can be classified as model-based or measurement-based (or a combination of both). Model-based approaches make an assumption about traffic models as inputs to a queueing model. Measurement-based approaches depend on measurements of actual traffic as inputs to an analytical model. In any case, the CAC algorithm is not a matter for standardization and should be chosen by the network provider.

4.3. Routing

The ATM protocol is not tied to a specific routing protocol. Indeed, a dynamic routing protocol is not needed if routes are static. Also, the concept of semipermanent virtual paths was intentionally included in ATM to simplify the routing process. Virtual paths can serve as large "pipes" with allocated bandwidth between pairs of nodes. If a new connection finds a convenient virtual path to its destination, it can make use of an available virtual channel within that virtual path with minimal setup overhead at intermediate switches.

For dynamic routes, PNNI routing is a link-state routing protocol. ATM switches will periodically advertise information about its links and maintain a topological view of the network constructed from link-state advertisements from other switches. These functions are carried out by a routing protocol engine within the connection control function.

5. TRAFFIC CONTROL CONSIDERATIONS

ATM switches are responsible for a comprehensive set of traffic control mechanisms to support QoS guarantees in addition to connection control [11,12]. For the most part, these other mechanisms operate in various parts of the switch independently of connection control.

5.1. Usage Parameter Control

Although the source traffic is expected to conform to the traffic descriptor negotiated during connection establishment, the actual source traffic may be excessive for various reasons. To protect the QoS of other connections, the source traffic rate needs to be monitored and regulated at the user-network interface by ATM edge switches. Usage parameter control (UPC) is the process for traffic regulation or "policing" carried out by a leaky-bucket algorithm called the *generic cell rate algorithm*. The generic cell rate algorithm involves two parameters, an increment I and a limit L, and is therefore denoted as GCRA (I,L). The parameter I is inversely proportional to the average rate allowed by the GCRA, while the parameter L determines its strictness. The GCRA is activated for a virtual connection after it has been accepted.

The operation of the GCRA is illustrated in Fig. 8. A bucket of capacity $I + L$ drains continuously at a rate of 1 per unit time. A cell is deemed to be conforming if the bucket contents can be incremented by I without overflowing; otherwise, the cell is deemed to be nonconforming or excessive. Conforming cells should be allowed to pass the GCRA without any effect. The network administrator can choose non-conforming cells to be allowed, discarded, or tagged by setting CLP = 1 in the cell header.

A virtual scheduling algorithm offers an alternative but equivalent view of the GCRA. The actual arrival time of the nth cell, $t(n)$, is compared with its theoretical arrival time $T(n)$, which is the expected arrival time assuming that all cells are spaced equally in time with separation I. Cells should not arrive much earlier than their theoretical arrival times, with some tolerance dependent on L. A cell is deemed to be conforming if $t(n) > T(n) - L$; otherwise, it is nonconforming (too early). The theoretical arrival time for the next cell, $T(n + 1)$, is calculated as a function of $t(n)$. If the nth cell is conforming and $t(n) < T(n)$, then the next theoretical arrival time is set to $T(n + 1) = T(n) + I$.

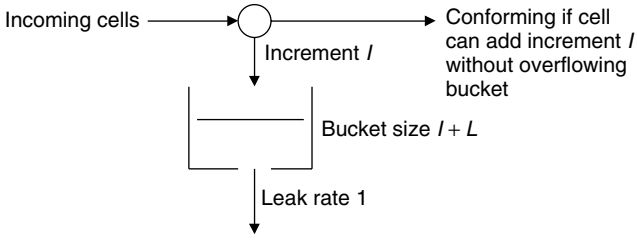


Figure 8. GCRA operation viewed as a leaky-bucket algorithm.

If the n-th cell is conforming and $t(n) \geq T(n)$, then the next theoretical arrival time is $T(n+1) = t(n) + I$. Nonconforming cells are not counted in the update of the theoretical arrival times.

Multiple GCRAs may be used in combination to regulate different sets of parameters. For example, a dual leaky-bucket may consist of a GCRA to regulate the peak cell rate followed by a second GCRA to regulate the sustainable cell rate (an upper bound on the average rate). A conforming cell must be deemed conforming by both GCRAs.

5.2. Packet Scheduling

ATM does not allow indication of service priority on the basis of individual cells. Although service priorities can be associated with virtual connections, it is common to group virtual connections according to their class of service, such as real-time constant bit rate (CBR), real-time variable bit rate (RTVBR), non-real-time VBR (NRTVBR), available bit rate (ABR), and unspecified bit rate (UBR). Real-time services would typically receive the highest service priority; NRTVBR, the second priority; and ABR and UBR, the lowest priority. Packet scheduling is not a matter for standardization and depends on the switch designer.

5.3. Selective Cell Discarding

Selective cell discarding is based on the cell loss priority indicated by the CLP bit in each cell header. CLP =1 cells should be discarded before CLP = 0 in the event of buffer overflows. CLP = 1 cells may be generated by a user who deliberately wants to take a risk with excess traffic or might be tagged cells from the UPC mechanism. In a pushout scheme, a CLP = 0 cell arriving to a full buffer may be allowed to enter the buffer if a queued CLP = 1 cell can be discarded to free space. If more than one CLP = 1 cell is queued, the discarding policy can push out CLP = 1 cells starting from the head or tail of the queue. Pushing out from the tail of the buffer tends to favor more CLP = 1 cells, because the CLP = 1 cells left near the head of the buffer are likely to depart successfully from the buffer. If CLP = 1 cells are pushed from the head of the buffer, the CLP = 1 cells left near the tail of the buffer will take longer to depart and will have a higher risk of being pushed out by the next arriving CLP = 0 cell.

More complicated buffer management strategies are possible. For example, a partial buffer sharing strategy can use a threshold; when the queue exceeds the threshold, only CLP = 0 cells will be admitted into the buffer and arriving CLP = 1 cells will be discarded. This strategy ensures a certain amount of space will always be available for CLP = 0 traffic. Similarly, it is possible to impose an upper limit on the number of CLP = 1 cells queued at any one time, which would ensure some space to be available for only CLP = 0 cells.

5.4. Explicit Forward Congestion Indication

ATM included EFCI as a means for ATM switches to communicate simple congestion information to the user to enable end-to-end control actions. User cells are generated with the second bit in the 3-bit PT field set to 0, signified as EFCI = 0. Any congested ATM switch can set EFCI = 1, which must be forwarded unchanged to the destination user. The algorithm for deciding when to activate EFCI is chosen by the network provider.

EFCI is used for the binary mode of the ABR service [12]. The ABR service is intended to allow rate-adaptable data applications to make use of the unused or "available bandwidth" in the network. An application using an ABR connection is obligated to monitor the receipt of EFCI = 1 cells and change its transmission rate according to a predefined rate adaptation algorithm. The objective is to match the transmission rate to the instantaneous available bandwidth. The ATM switch buffers should be designed to absorb the temporarily excessive traffic caused by mismatch between the actual transmission rate and the available bandwidth. In return for compliance to the rate adaptation algorithm, the ATM network should guarantee a low cell loss ratio on the ABR connection (but no guarantees on cell delay).

5.5. Closed-Loop Rate Control

The binary mode of the ABR rate adaptation algorithm involves gradual decrementing or incrementing of an application's transmission rate. The rate adaptation algorithm for the ABR service also allows an optional explicit mode of operation where the ATM switches along an ABR connection may communicate an exact transmission rate to the application. A resource management cell indicated by a PT = 6 field is periodically inserted into an ABR connection and makes a complete round trip back to the sender. It carries an "explicit rate" field that can be decremented (but not incremented) by any ATM switch along the ABR connection. The sender is obligated to immediately change its transmission rate to the value of the explicit rate field, or the rate dictated according to the binary mode of rate adaptation, whichever is lower.

6. ATM-LAYER OAM

The ATM protocol defines OAM cells to carry out various network management functions in the ATM layer such as fault management and performance management [13]. ATM switches are responsible for the generation, processing, forwarding, and termination of OAM cells according to standardized OAM procedures. OAM cells have the same cell header but their payloads contain predefined fields depending on the function of the OAM cell. F4 OAM cells share a virtual path with user cells. F4 OAM cells have the same VPI value as the user cells in the virtual path but are recognized by the preassigned virtual channels: VCI = 3 for segment OAM cells (relayed along part of a route) or VCI = 4 for end-to-end OAM cells (relayed along an entire route). F5 OAM cells share a virtual channel with user cells. F5 OAM cells have the same VPI/VCI values as the user cells in the virtual channel but have these preassigned PT values: PT = 4 for segment OAM cells and PT = 5 for end-to-end OAM cells.

6.1. Fault Management

OAM cells are used for these fault management functions: alarm surveillance, continuity checks, and loopback testing. If a physical layer failure is detected, a virtual connection failure will be reported in the ATM layer with two types of OAM cells: alarm indication signal (AIS) and

remote defect indicator (RDI). AIS cells are sent periodically "downstream" or in the same direction as user cells effected by the failure to notify downstream switches of the failure and its location. The last downstream ATM switch will generate RDI cells in the upstream direction to notify the sender of the downstream failure.

The purpose of continuity checking is to confirm that an inactive connection is still alive. If a failure has not been detected and no user cells have appeared on a virtual connection for a certain length of time, the switch on the sender's end of a virtual connection should send a continuity check cell downstream. If the switch on the receiver's end of the virtual connection has not received any cell within a certain time in which a continuity check cell was expected, it will assume that connectivity was lost and will send a RDI cell to the sender.

An OAM loopback cell is for testing the connectivity of a virtual connection on demand. Any switch can generate an OAM loopback cell to another switch designated as the loopback point. The switch at the loopback point is obligated to reverse the direction of the loopback cell to the originator. The failure of a loopback cell to return to its originator will be interpreted as a sign that a fault has occurred on the tested virtual connection.

6.2. Performance Management

OAM performance management cells are used to monitor the performance of virtual connections to detect intermittent or gradual error conditions caused by malfunctions. At the sender's end of a virtual connection, OAM performance monitoring cells are inserted between blocks of user cells. Nominal block sizes may range between 2^7, 2^8, 2^9, or 2^{10} cells but do not have to be exact. The OAM performance monitoring cell includes fields for the monitoring cell sequence number, size of the preceding cell block, number of transmitted user cells, error detection code computed over the cell block, and timestamp to measure cell delay. The switch at the receiver's end of the virtual connection will return the OAM cell in the reverse direction with additional fields to report any detected bit errors and any lost or misinserted cells. The timestamp will reveal the roundtrip cell delay. The measurements of cell loss and cell delay reflect the actual level of QoS for the monitored virtual connection.

BIOGRAPHY

Thomas M. Chen received the B.S. and M.S. degrees in electrical engineering in 1984 from Massachusetts Institute of Technology, Cambridge, Massachusetts, and the Ph.D. degree from the University of California, Berkeley, in 1990. He worked at GTE Laboratories on ATM traffic control and network management research until 1997. Since 1997, he has been an Associate Professor in Electrical Engineering at Southern Methodist University, Dallas, Texas, where he has been working on traffic modeling, network management, programmable networks, and network security. He was the recipient of the 1996 IEEE Communications Society's Fred W. Ellersick best paper award. He currently is an associate editor for *ACM Transactions on Internet Technology*, and senior technical editor for *IEEE Communications Magazine* and *IEEE Network*.

Stephen S. Liu received the B.S. degree in electrical engineering from National Cheng-Kung University in Taiwan, and the M. S. and Ph.D. degrees from Georgia Institute of Technology in Atlanta, Georgia. He co-developed the ISO and ANSI standard 32-degree error detection code polynomial used on Ethernet and various high-speed data networks, and co-authored the book entitled *ATM Switching Systems* published by Artech House in 1995. He joined the Verizon Labs (formerly GTE Laboratories) in 1981 and has since been working on packet-switching technology. Dr. Liu's current interest is in unified control plane technology for optical transport networks. He is a senior member of the IEEE.

BIBLIOGRAPHY

1. ITU-T Rec. I.361, *B-ISDN ATM-Layer Specification*, Geneva, July 1995.
2. ATM Forum, *ATM User-Network Interface (UNI) Specification Version 4.0*, April 1996.
3. T. Chen and S. Liu, *ATM Switching Systems*, Artech House, Boston, 1995.
4. N. Endo et al., Shared buffer memory switch for an ATM exchange, *IEEE Trans. Commun.* **41**: 237–245 (Jan. 1993).
5. M. Karol, M. Hluchyj, and S. Morgan, Input versus output queueing on a space-division switch, *IEEE Trans. Commun.* **35**: 1347–1356 (Dec. 1987).
6. T.-Y. Feng, A survey of interconnection networks, *IEEE Commun. Mag.* **14**: 12–27 (Dec. 1981).
7. X. Chen, A survey of multistage interconnection networks in fast packet switches, *Int. J. Digital Analog Cabled Syst.* **4**: 33–59 (1991).
8. J. Hui, Switching integrated broadband services by sort-banyan networks, *Proc. IEEE* **79**: 145–154 (Feb. 1991).
9. S. Nojima, E. Tsutsui, H. Fukuda, and M. Hashimoto, Integrated services packet network using bus matrix switch, *IEEE J. Select. Areas Commun.* **SAC-5**: 1284–1292 (Oct. 1987).
10. Y. Yeh, M. Hluchyj, and A. Acampora, The knockout switch: A simple, modular architecture for high-performance packet switching, *IEEE J. Select. Areas Commun.* **SAC-5**: 1274–1283 (Oct. 1987).
11. ITU-T Rec. I.371, *Traffic Control and Congestion Control in B-ISDN*, Geneva, July 1995.
12. ATM Forum, *Traffic Management Specification Version 4.0*, April 1996.
13. ITU-T Rec. I.610, *B-ISDN Operation and Maintenance Principles and Functions*, Geneva, July 1995.

FURTHER READING

Several good surveys of ATM switch fabric architectures can be found in the literature:

H. Ahmadi and W. Denzel, A survey of modern high-performance switching techniques, *IEEE J. Select. Areas Commun.* **7**: 1091–1103 (Sept. 1989).

A. Pattavina, Nonblocking architectures for ATM switching, *IEEE Commun. Mag.* **31**: 38–48 (Feb. 1993).

E. Rathgeb, T. Theimer, and M. Huber, ATM switches — basic architectures and their performance, *Int. J. Digital Analog Cabled Syst.* **2**: 227–236 (1989).

F. Tobagi, Fast packet switch architectures for broadband integrated services digital networks, *Proc. IEEE* **78**: 133–178 (Jan. 1990).

R. Awdeh and H. Mouftah, Survey of ATM switch architectures, *Comput. Networks ISDN Syst.* **27**: 1567–1613 (1995).

A wealth of papers can be found on performance analysis of switch architectures; examples are

A. Pattavina and G. Bruzzi, Analysis of input and output queueing for nonblocking ATM switches, *IEEE/ACM Trans. Network.* **1**: 314–327 (June 1993).

D. Del Re and R. Fantacci, Performance evaluation of input and output queueing techniques in ATM switching systems, *IEEE Trans. Commun.* **41**: 1565–1575 (Oct. 1993).

The difficulties of constructing large ATM switches are explored in

T. Banwell et al., Physical design issues for very large scale ATM switching systems, *IEEE J. Select. Areas Commun.* **9**: 1227–1238 (Oct. 1991).

T. Lee, A modular architecture for very large packet switches, *IEEE Trans. Commun.* **38**: 1097–1106 (July 1990).

ATMOSPHERIC RADIOWAVE PROPAGATION

HAROLD RAEMER
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Radiowave propagation in the terrestrial environment is a very mature science. Its basic principles were understood in the nineteenth century, after the work of Faraday, Maxwell, Hertz, Sommerfeld, and others had resulted in a formulation of the basic equations of electromagnetic theory and their application to radio wave propagation. However, it was in the early 1950s that these principles were intensively applied to the rapidly developing technology of microwave communications. The impetus for that technology was the radar research that occurred during and immediately after World War II. Much of what was learned about the propagation of radar waves [1] was directly applicable to line-of-sight communication links operated at frequencies from 300 MHz through the microwave bands up to about 30 GHz. With the rapid growth of wireless communication systems in the 1990s, it has become increasingly important to engineers involved in the development and improvement of those systems to understand the propagation environment within that band. Information obtained from propagation studies is directly applicable to decisions on such issues as siting of transmitters and receivers to optimize reception and prediction of fading rates and intersymbol interference due to multipath propagation in digital transmission. Since most wireless systems are mobile, it is particularly important, for a specific swath of terrain over which a system is operating, to know how the received signal varies with time as transmitters and/or receivers travel. That knowledge can be acquired through propagation analysis aided by topological information about the environment and can help reduce the amount of expensive and time-consuming field experimentation required for system design decisions.

Terrestrial radio wave propagation is a vast subject when it covers the entire radio spectrum from the kilohertz region through millimeter waves. The subject could not be adequately covered in an article of this length. This article is confined to the frequency range from 30 MHz to 30 GHz and to "space wave" propagation in the troposphere; the region below 16–18 km above the earth's surface [2, pp. 100–141], that is, ground wave [2, pp. 33–61] ionospheric ("sky wave;" see Ref. 2, pp. 62–99 or Ref. 3, pp. 218–255) and satellite transmission [2, pp. 263–295], are not included. The emphasis is on the propagation phenomena important in mobile wireless communication systems [4,5]. Today's wireless links operate at certain key frequencies [450, 900 MHz (mobile cellular), 2.4, 5.8 GHz (indoor wireless)], but much higher frequencies (e.g., 22 GHz) are being investigated for some applications and will probably be in use within the near future. In summary, the range of frequencies of greatest interest for terrestrial wireless communication is between 30 MHz and 30 GHz, namely, the VHF band (30–300 MHz), the UHF band (300 MHz–3 GHz), and the SHF band (3–30 GHz).

The modes of propagation not covered in this article operate primarily in frequency bands outside this range. Ground-wave propagation predominates at frequencies below 3 MHz, for instance, in the very low-frequency (VLF; 3–30 kHz), low-frequency (LF; 30–300 kHz), and medium-frequency (MF; 300 kHz–3 MHz) bands, while sky-wave propagation predominates in the high-frequency band (HF; 3–30 MHz) and also occurs in VLF, LF, and MF bands. Satellite links operate primarily in the SHF bands, overlapping our spectral region of interest, but satellite communication is a major topic in its own right, and its propagation aspects are not included in this article.

2. THEORY

2.1. Free-Space Propagation Equations

The idealized propagation geometry for a wireless communication link is illustrated in Fig. 1. The transmitter at T and the receiver at R are separated by a distance D_{TR}. If the transmitting and receiving antennas were in infinite free space, the vertically polarized (V) or horizontally polarized (H) components of the vector phasor of the electric field of the wave transmitted from T at the receiving point R would be

$$E_R^{(V,H)} = \frac{e^{-jk_0 D_{TR}}}{D_{TR}} \sqrt{\frac{P_T G_{T0} A_{eR0}}{4\pi}} f_T^{(V,H)}(\Delta\boldsymbol{\Omega}_{TR}) f_R^{(V,H)}(\Delta\boldsymbol{\Omega}_{RT}) \tag{1}$$

where P_T is the total power radiated by the transmitting antenna; G_{T0} and A_{eR0} are the peak values of antenna gain and effective aperture area of transmitting and receiving

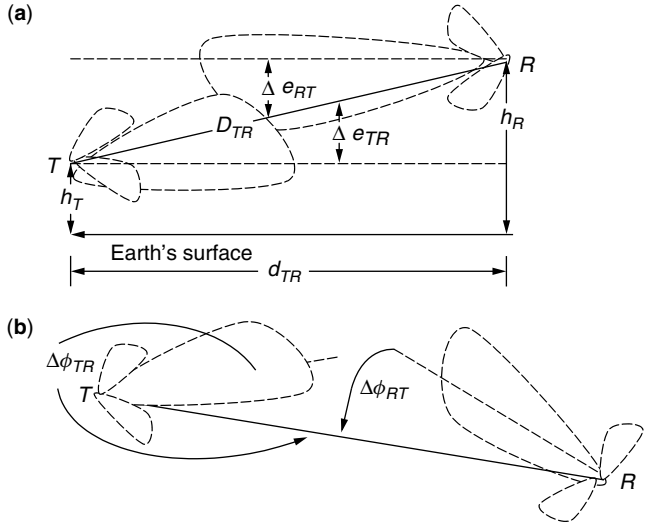


Figure 1. Idealized propagation geometry for wireless communication: (**a**) side view — antennas pointing horizontally in elevation; (**b**) downward view — antennas pointed arbitrarily in azimuth.

antennas, respectively; k_0 is the free-space wave number, equal to $2\pi/\lambda_0$, where λ_0 is the free-space wavelength; and $f_T^{(V,H)}(\Delta\boldsymbol{\Omega}_{TR})$ and $f_R^{(V,H)}(\Delta\boldsymbol{\Omega}_{RT})$ are the angular radiation pattern and receptivity pattern of transmitting and receiving antennas, respectively. The argument of f_T is a two-dimensional vector whose components are $\Delta e_{TR} = e_{TR} - e_{T0}$ and $\Delta\phi_{TR} = \phi_{TR} - \phi_{T0}$, where e_{TP} and ϕ_{TP} denotes the elevation and azimuth angle, respectively, of an arbitrary point P with respect to T and where the subscript 0 denotes the point where the pattern amplitude reaches its maximum. The vector $\Delta\boldsymbol{\Omega}_{RT}$, the argument of f_R, has the same meaning as $\Delta\boldsymbol{\Omega}_{TR}$ except that the reference point is R rather than T. These angles are illustrated in Fig 1.

2.2. Effect of Earth Reflection

The actual field component E_R in the presence of the earth is not given by Eq. (1) alone but includes an additional term due to the reflection of the transmitted wave from a specular point S on the earth's surface, as illustrated in Fig. 2. If the earth's surface is perfectly smooth within the region of interest, then the total field component at R is given by

$$E_R^{(H,V)} = E_{R0}^{(H,V)} F^{(H,V)} \tag{2}$$

where E_{R0} is the free-space wave field given by (1), (the "direct wave") and $F^{H,V}$, known as the *propagation factor* or *path-gain factor*, is given by

$$F^{H,V} = 1 + (R')^{H,V} e^{-jk_0(D_{TS}+D_{SR}-D_{TR})} \tag{3}$$

where, as shown in Fig. 2, D_{TS} and D_{SR} are separation distances between the specular point and transmitter and receiver, respectively, and $(R')^{H,V}$ is a factor of the general form

$$(R')^{H,V} = \frac{f_T^{(H,V)}(\Delta\Omega_{TS}) f_R^{(H,V)}(\Delta\Omega_{RS}) R^{H,V}}{f_T^{(H,V)}(\Delta\Omega_{TR}) f_R^{(H,V)}(\Delta\Omega_{RT})}$$

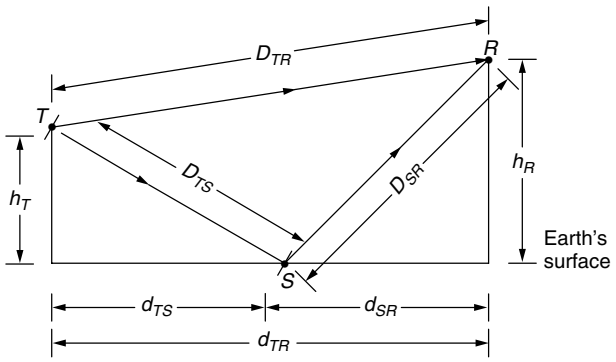


Figure 2. Effect of earth reflection — flat-earth approximation.

where R^H and R^V are the Fresnel reflection coefficients for $H(TE)$ and $V(TM)$ polarizations, respectively, given by

$$R^H = \frac{\mu_R \cos\Theta_i - \sqrt{\nu^2 - \sin^2\Theta_i}}{\mu_R \cos\Theta_i + \sqrt{\nu^2 - \sin^2\Theta_i}} \tag{4a}$$

$$R^V = \frac{\varepsilon_{CR} \cos\Theta_i - \sqrt{\nu^2 - \sin^2\Theta_i}}{\varepsilon_{CR} \cos\Theta_i + \sqrt{\nu^2 - \sin^2\Theta_i}} \tag{4b}$$

where $\varepsilon_{CR} = (\varepsilon_1/\varepsilon_0) - (j\sigma/\omega\varepsilon_0)$, $\mu_R = (\mu_1/\mu_0)$, where ε_0 and μ_0 are respectively the permittivity and magnetic permeability of free-space $\varepsilon_0 = 10^{-9}/36\pi$ farads per meter and $\mu_0 = 4\pi(10^{-7})$ henrys per meter; ε_1, μ_1, and σ, are respectively permittivity, permeability, and conductivity (in siemens per meter) of the earth medium; and ν is the medium's complex refractive index, equal to $\sqrt{\varepsilon_{CR}\mu_R}$.

Invoking the flat-earth approximation, valid for short-range communication links, we can express the pathlength difference appearing in the phase in (3) in the form (see Fig. 2)

$$\Delta L = D_{TS} + D_{SR} - D_{TR} = \sqrt{d_{TR}^2 + (h_R + h_T)^2} - \sqrt{d_{TR}^2 + (h_R - h_T)^2} \tag{5}$$

where h_T and h_R are antenna heights and d_{TR} the horizontal distance between T and R. Given the approximation

$$|h_T \pm h_R| \ll d_{TR} \tag{6}$$

nearly always valid for ground-based transmitting and receiving antennas, we have

$$k_0\Delta L \simeq k_0 d_{TR}\left[1 + \frac{1}{2}\left(\frac{h_R + h_T}{d_{TR}}\right)^2 - 1 - \frac{1}{2}\left(\frac{h_R - h_T}{d_{TR}}\right)^2\right] = \frac{4\pi h_T h_R}{\lambda_0 d_{TR}} \tag{7}$$

Approximation (6) implies that the antenna pattern function for transmitter to ground reflection point (GRP) is nearly the same as that for transmitter to receiver and the same applies to the patterns from receiver to GRP and receiver to transmitter. The assumption that

these patterns are the same and Eq. (7) gives us a simple approximation for (3):

$$F^{H,V} \simeq 1 + R^{H,V} \exp\left(-j\frac{4\pi h_T h_R}{\lambda_0 d_{TR}}\right) \tag{8}$$

The simplified form of the path-gain factor (8) can be used as the basis for "coverage diagrams," examples of which are shown in Ref. 6 (pp. 357–361), which show, for various values of h_T, h_R, d_{TR}, and λ_0, the locations and amplitudes of the peaks and the locations and depths of the troughs of the fields in the propagation plane (defined as the vertical plane containing transmitter, receiver, and GRP) due to interference between the direct and ground-reflected waves. These examples [6] include the effect of earth curvature, to be discussed in Section 2.3 (below). Flat-earth examples are given in Ref. 6 (p. 344).

A further simplification of (8) is achieved by assuming near-grazing incidence, [a further consequence of approximation (6)], namely, $\Theta_i \simeq \pi/2$, in (4), implying that $R^V \simeq R^H \simeq -1$. Under this assumption, the amplitude of (8) reduces to

$$\left|F^{H,V}\right| \simeq 2\left|\sin\left(\frac{2\pi h_T h_R}{\lambda_0 d_{TR}}\right)\right| \tag{9}$$

implying that nulls occur when $h_T h_R/d_{TR}$ is an integral multiple of a half-wavelength and peaks occur when $(h_T h_R/d_{TR}) = (n\lambda_0/2) + (\lambda_0/4)$, where n is an arbitrary integer. The simplified form (9) can be used to roughly estimate the locations of peaks and troughs of the path-gain factor.

2.3. Effect of Earth Curvature: Atmospheric Refraction

As illustrated in Fig. 3, the refractive index of the atmosphere immediately above the earth's surface, although it is very close to unity at all altitudes, decreases approximately exponentially with altitude. The propagation vector, indicating the direction of the radiowave's energy flow, is continuously changing direction if it has a locally vertical component. Governed by Snell's law, in a standard atmosphere, the wave travels in a curved path as shown in Fig. 3. The altitude dependence of the refractivity $N(z)$, defined as $N(z) = [\nu(z) - 1] \times 10^6$, where $\nu(z)$ is the refractive index, is

$$N(z) = N(z_0)e^{-\alpha(z-z_0)} \tag{10}$$

where z_0 is sea level and α is a positive number dependent on atmospheric conditions.

The curvature (reciprocal of the radius of curvature) of the raypath is equal to $-(d\nu(z)/dz)$. The curvature of the earth is $1/r_t$, where r_t is the true earth radius, approximately 6340 km. Using values of α and $N(z_0)$ from the literature, the difference between raypath curvature and earth curvature is about $1.1(10^{-7})\text{m}^{-1}$. The resulting modified earth radius r_e, is between about 1.3 and 1.4 times the true radius (due to variability of atmospheric conditions), but is usually assumed to be $\frac{4}{3}r_t$, giving rise to the concept of the "four-thirds earth radius," commonly used in radio propagation modeling.

Using r_e, the modified earth radius explained above, we then model the propagation as if the rays were straight lines. The maximum distance over which the receiver has a line-of-sight view of the transmitter unobstructed by the earth's curvature (illustrated in Fig. 4a) is given by (where $h_T \ll r_e$, $h_R \ll r_e$)

$$\begin{aligned} D_{\max} = D_{TG} + D_{GR} &= \sqrt{(r_e + h_T)^2 - r_e^2} + \sqrt{(r_e + h_R)^2 - r_e^2} \\ &\simeq \sqrt{2r_e h_T} + \sqrt{2r_e h_R} \end{aligned} \tag{11}$$

If $D_{TR} \le D_{\max}$, then the propagation model is line-of-sight plus a specular earth-reflected ray and the path-gain factor can still be modeled approximately by (8), with the aid of (3) through (7), but with significant modifications. If $D_{TR} > D_{\max}$, then R and T are below the "radio horizon" of T and R, respectively (the radio horizons for T and R are defined as $D_{TG} \simeq \sqrt{2r_e h_T}$ and $D_{GR} \simeq \sqrt{2r_e h_R}$ respectively). This implies that the raypaths between T and R are obscured from each other by the earth's curvature and line-of-sight transmission becomes infeasible. The only feasible modes of propagation in this case are diffraction around the earth [1, pp. 109–112; 6, pp. 369–372], tropospheric scatter [2, pp. 216–237], or satellite transmission [2, pp. 263–295; 3, pp. 100–103].

An important modification is the generalization of (7) to include earth curvature [3, pp. 56–65]. From Fig. 4b, we note that application of the law of sines to the triangles TOS and ROS, the law of cosines to the triangle TSR, and the law of reflection at the point S, together with the observation that $\alpha_1 = d_{TS}/r_e$ and $\alpha_2 = d_{SR}/r_e$, result in the expressions

$$\frac{D_{TS}}{\sin(d_{TS}/r_e)} = \frac{r_e + h_T}{\sin\Theta_i} = \frac{r_e}{\sin(\Theta_i - d_{TS}/r_e)} \tag{12a}$$

$$\frac{D_{SR}}{\sin(d_{SR}/r_e)} = \frac{r_e + h_R}{\sin\Theta_i} = \frac{r_e}{\sin(\Theta_i - d_{SR}/r_e)} \tag{12b}$$

$$D_{TR}^2 = D_{TS}^2 + D_{SR}^2 - 2D_{TS}D_{SR}\cos 2\Theta_i \tag{12c}$$

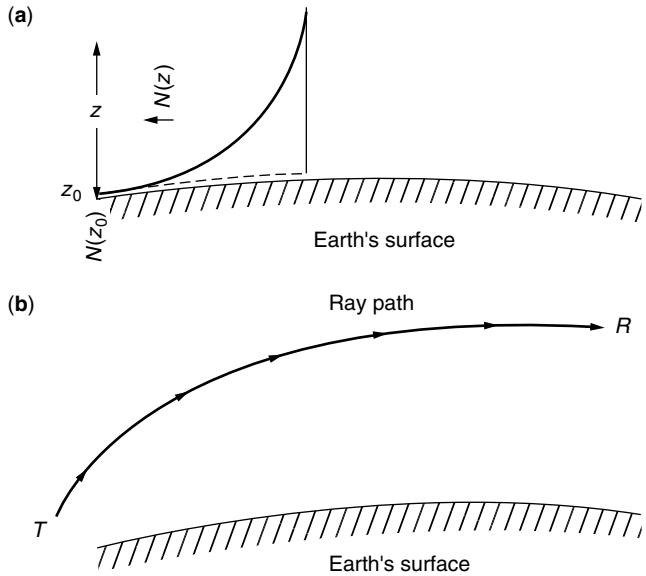


Figure 3. Atmospheric refraction and earth curvature effects: (**a**) variation of refractivity with altitude; (**b**) curvature of ray-path due to variable refractivity.

By adding the assumptions that $d_{TS}/r_e \ll 1$, $d_{SR}/r_e \ll 1$ and the grazing angle $[(\pi/2) - \Theta_i]$ is much smaller than $\pi/2$, applicable to low-altitude propagation paths, manipulations of Eqs. (12a–c) result in the approximate generalization of (5) and (7) to account for earth curvature [6, pp. 349–352]:

$$k_0 \Delta L \simeq \frac{4\pi h_T h_R}{\lambda_0 d_{TR}} \left[1 - \frac{1}{2r_e}\left(\frac{d_{TS}^2}{h_T} + \frac{d_{SR}^2}{h_R}\right)\right] \tag{13}$$

Another effect of earth curvature is the divergence of reflected waves due to the convexity of the reflecting surface near the specular point (Fig. 4c). This can be modeled by multiplication of the reflection coefficient in (4) by a "divergence factor." The general form of this factor is attributed to Vanderpol and Bremmer and its derivation can be found in Ref. 1, (pp. 404–406). Using the usual approximations $h_T \ll r_e$, $h_R \ll r_e$, $(\pi/2) - \Theta_i \ll (\pi/2)$ a simplified form is

$$F_d \simeq \left(1 + \frac{2 d_{TS} d_{SR}}{r_e (d_{TS} + d_{SR}) \cos \Theta_i}\right)^{-1/2} \tag{14}$$

The divergence factor reduces the reflected wave energy relative to the direct wave energy and hence decreases the importance of earth-reflection in the propagation factor.

The location of the ground reflection point and the incidence angle Θ_i at that point are also affected by earth curvature. For the flat-earth case, it is evident from Fig. 2 that

$$d_{TS} = \frac{d_{TR} h_T}{h_T + h_R}; \Theta_i = \cot^{-1}\left(\frac{h_T + h_R}{d_{TR}}\right) \tag{15a}$$

while for the curved earth case [1, p. 113], solution of a cubic equation is required to determine d_{TS}, with the result $d_{TS} = (d_{TR}/2) + p \cos[(\Phi + \pi)/3]$, where

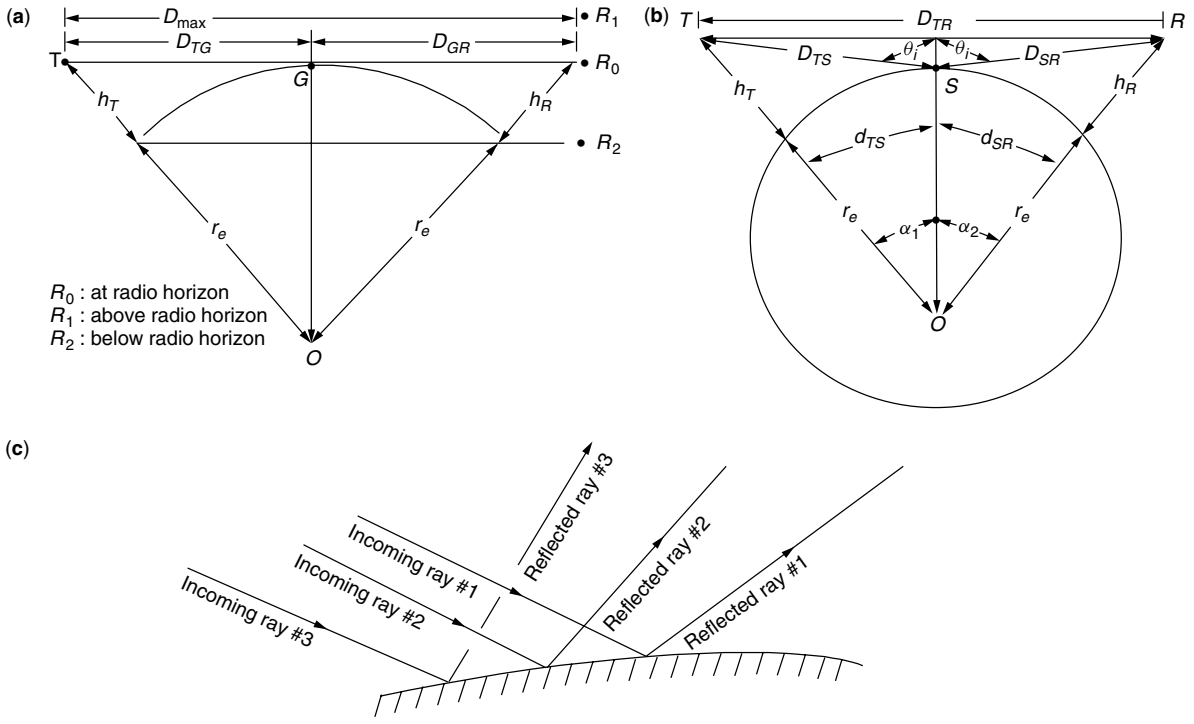
$$p = \frac{2}{\sqrt{3}} \sqrt{r_e (h_T + h_R) + \left(\frac{d_{TR}}{2}\right)^2};$$
$$\Phi = \cos^{-1}\left[\frac{2 r_e |h_R - h_T| d_{TR}}{p^3}\right] \tag{15b}$$

and the angle of incidence Θ_i is approximated by [6, p. 351]

$$\Theta_i \simeq \cot^{-1}\left(\left[\frac{h_T + h_R}{d_{TR}}\right] - \frac{1}{2r_e}\left[\frac{(h_T + h_R)(d_{SR}^2 h_R + d_{TS}^2 h_T)}{d_{TR}(h_T^2 + h_R^2)}\right]\right) \tag{15c}$$

2.4. Effect of Surface Roughness

The effect of surface roughness on the path-gain factor depends on the scale of the roughness. Figure 5 illustrates this in two dimensions by showing three different roughness scales and the corresponding changes in the numbers of specular reflections. It is intuitively evident that there will be a multiplicity of surface points that will result in a specular reflection of a wave from the transmitter into the direction of the receiver. Figure 5c shows that some of the specular reflections from a very rough surface may be mitigated or nullified by shadowing, thus establishing an upper limit on the effect. The theory that leads to (3) and (4a,b) is no longer strictly valid for rough surfaces. However, for a low level of roughness, the effect can



A bundle of nearly parallel incoming rays from transmitter diverges upon reflection due to local earth curvature.

Figure 4. Some effects of earth curvature: (**a**) radio horizon effect; (**b**) effect of pathlength difference; (**c**) divergence of rays reflected from earth's surface.

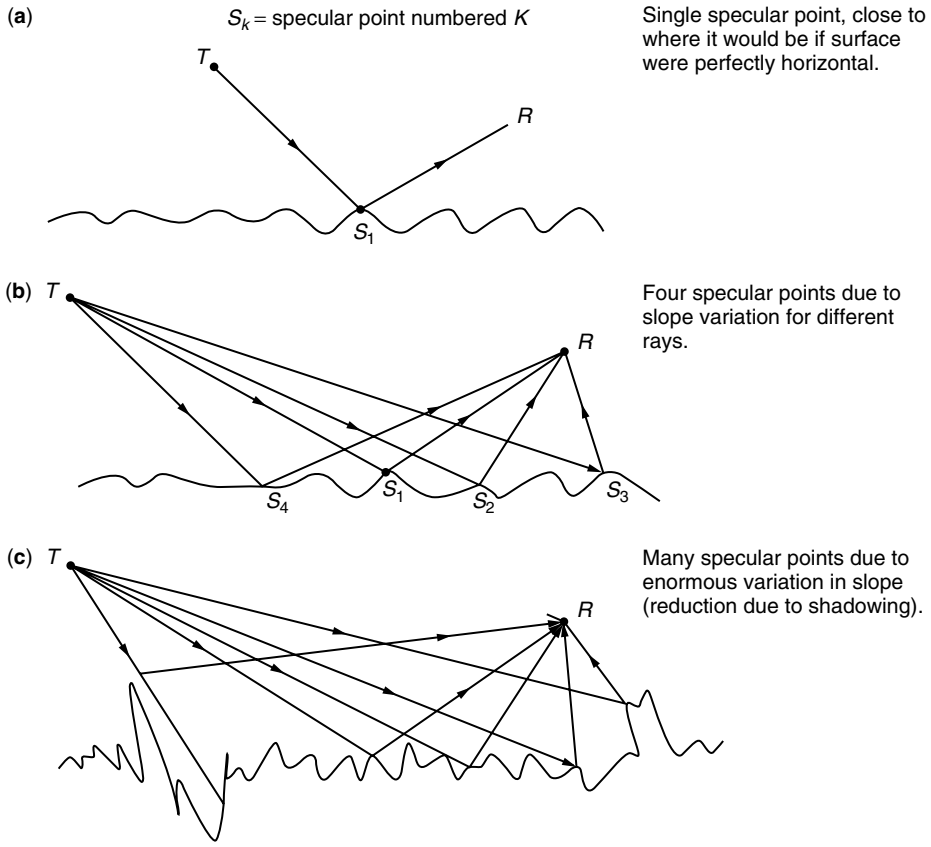


Figure 5. Specular reflections from roughness: (**a**) slightly, (**b**) moderately, and (**c**) very rough surfaces.

be approximated by a "roughness factor" multiplying the reflection coefficient.

Before accounting for roughness, it is desirable to invoke a criterion to determine whether it has a significant effect on the reflected wave, known as the "Rayleigh criterion" where σ is rms surface height, illustrated in Fig. 6a and given by the following rule:

If

$$\sigma \leq \frac{(0.1)\lambda_0}{4\pi \cos \Theta_i}$$

the surface can be approximated as smooth, where σ is the rms height of the rough surface. If σ exceeds

$$\frac{(0.1)\lambda_0}{4\pi \cos \Theta_i}$$

then roughness must be accounted for. The maximum rms height that justifies the smooth surface approximation, $\sigma_{\max}$, is inversely proportional to frequency and increases monotonically with angle of incidence. A surface considered "rough" near normal incidence can be approximated as "smooth" near grazing incidence at a given frequency.

From Fig. 6, it is evident that the difference in pathlength between two ground-reflected parallel rays between T and R, one from the mean surface and the other from a local peak, is $2\Delta h \cos \Theta_i$. The local deviation of the height from its mean value, Δh, is a random variable in a typical ground or sea surface and is often assumed to have a zero-mean Gaussian distribution in the absence of known terrain-specific statistics. The phase difference between two rays is

$$\frac{2\pi}{\lambda_0}(2\Delta h \cos \Theta_i) = \frac{4\pi \Delta h}{\lambda_0} \cos \Theta_i$$

which is also a zero-mean Gaussian random variable. It is shown by more sophisticated methods (e.g., Ref. 7, pp. 80–81 or Ref. 8, pp. 399–401) that this results in 'a "roughness factor" multiplying the reflection coefficient in (8) and given by

$$F_r = \exp\left(-\frac{1}{2}\left(\frac{4\pi}{\lambda_0} \cos \Theta_i\right)^2 \sigma^2\right) \tag{16}$$

where $\sigma^2 = \langle (\Delta h)^2 \rangle$. Like the divergence factor given by (14), this reduces the effect of the reflected wave.

Returning to the basis of the Rayleigh criterion, if $(4\pi \cos \Theta_i/\lambda_0)\sigma$ is less than one-tenth of a radian, or equivalently $\sigma < 0.1\lambda_0/4\pi \cos \Theta_i$, then roughness is considered negligible. Through (16), this implies that $|F_r - 1| < 0.01$, which in turn implies that F_r can be approximated as unity and hence roughness is negligible in the height-gain function. However it should be noted that (16) applies only for small-scale roughness. If σ is far in excess of the Rayleigh limit, then the effect can no longer be modeled in such a simple way and much more complicated theory is required to account for it.

If, as is often the case, large-scale roughness well beyond the Rayleigh limit is distributed throughout the antenna beam coverage region, then specular points with

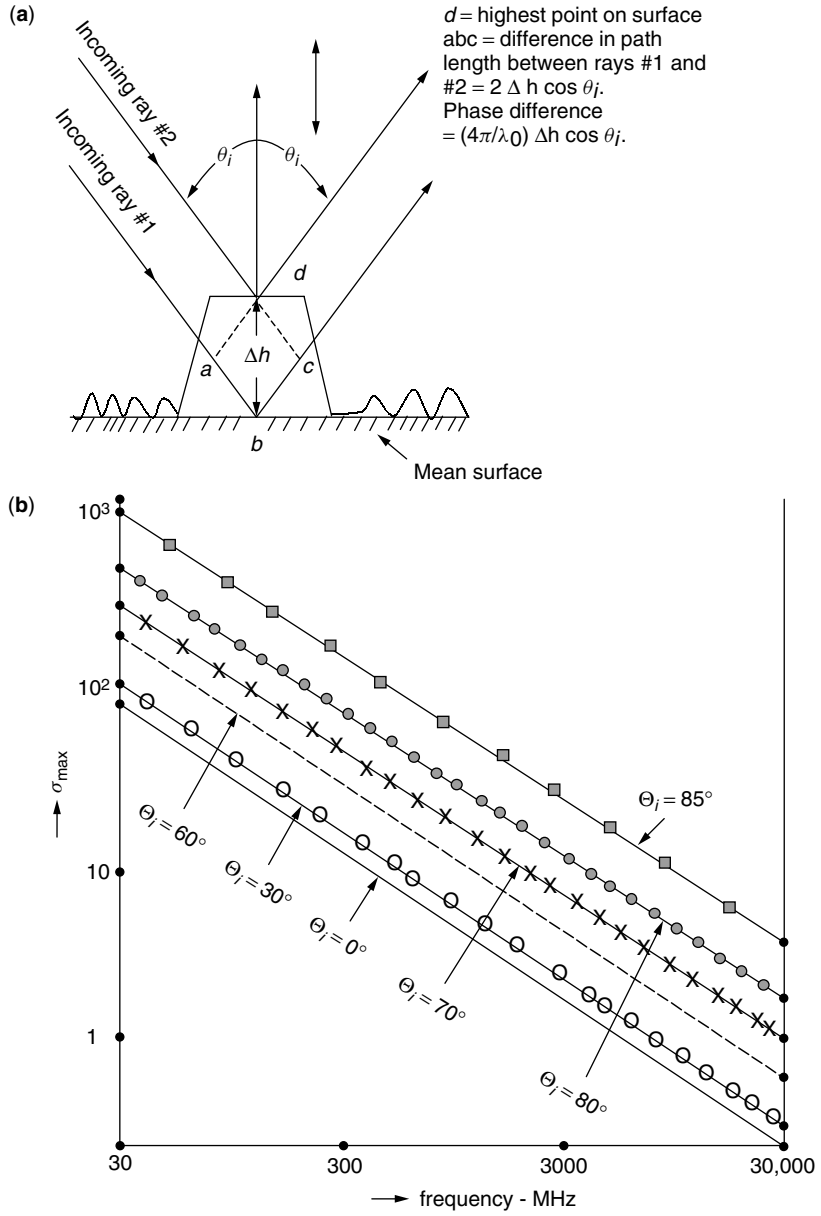


Figure 6. Rayleigh criterion for surface roughness: (**a**) basis of Rayleigh criterion; maximum allowable (**b**) frequency variation of RMS height for smooth surface approximation.

respect to transmitted waves reflected into the receiver's direction exist over a wide swath of terrain. In this case (Fig. 5c), more accurate modeling would require that the height-gain function of (3) be replaced by

$$F^{H,V} \simeq 1 + \sum_{1}^{N} (R')_n^{H,V} e^{-jk_0 \Delta L_n} \tag{17}$$

where $(R')_n^{H,V}$ and the path delay ΔL_n are of the form given in (3) but applicable to the nth specular point. In order to apply (17), one needs either a terrain-specific or a statistically generated surface-height distribution within the coverage area and algorithms to (1) locate the specular points, (2) determine for each such point the angles of incidence of the wave from T and reflection of those waves toward R, and (3) compute the superposition of terms of (17) using stationary phase.

This requires extensive computational power for a swath of very irregular land terrain or a very rough water surface. There are simulation programs available that perform these tasks within reasonable computation times and therefore provide the ability to obtain reality-based coverage diagrams quickly and easily (e.g., SEKE [9]).

2.5. Obstacles Along Propagation Path: Diffraction

Line-of-sight propagation is limited by obstacles along the propagation path, e.g., hills or foliage in irregular open terrain, high waves in a rough sea surface, or buildings in an urban area. The theory of diffraction around obstacles that are opaque to direct waves is required to analyze this situation (Figure 7). Standard line-of-sight theory would predict zero fields behind the obstacle, but diffraction theory shows nonzero fields in the region and can provide

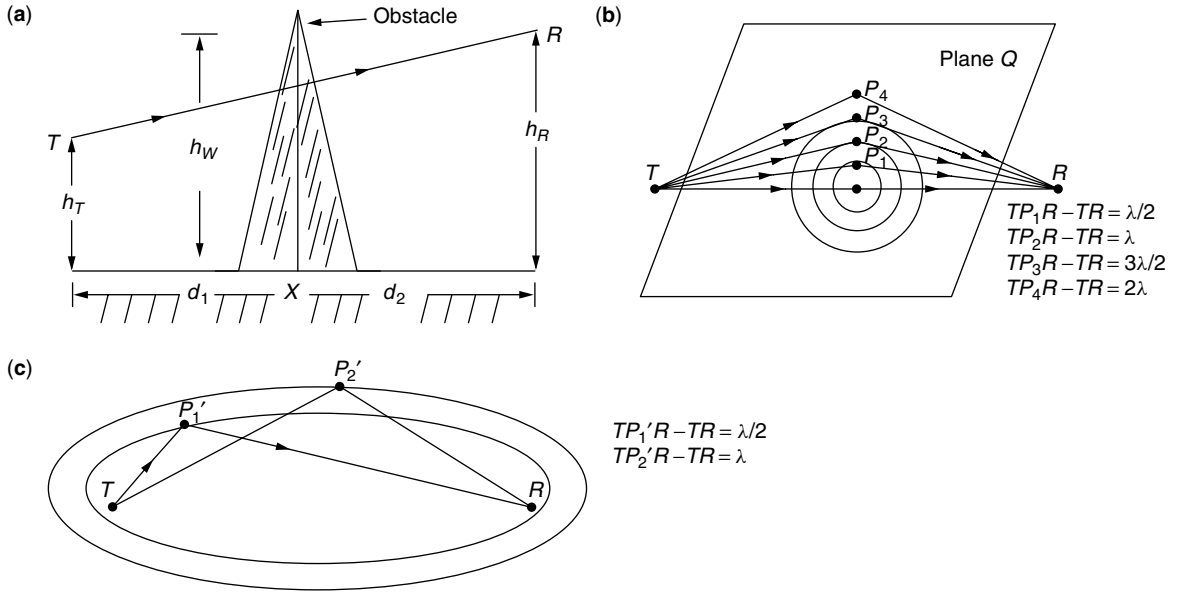


Figure 7. Diffraction around an obstacle: (**a**) obstacle intercepting raypath TR; (**b**) Fresnel zones on plane Q; (**c**) Fresnel ellipsoids.

good predictions of their magnitude [2, pp. 111–112, 122–130; 3, pp. 33–36, 46–50; 5, pp. 90–100].

In Fig. 7, we show a vertical plane Q, normal to the vertical propagation plane P. An opaque obstacle (e.g., hill or building) is placed along the raypath and centered on the plane Q. According to Huyghens' principle, the field at R is a superposition of waves from all points on the plane Q.

To determine whether consideration of diffraction around obstacles is required, we must determine whether an obstacle is within the first Fresnel zone. The Fresnel zone (Fig. 7b) for the transmitted wave is defined as the locus of the points for which the difference in the direct line-of-sight path TR and the indirect path TPR differ by an odd number of half-wavelengths. These 3D (three-dimensional) loci are ellipsoids (Fig. 7c). The circular rings on plane Q illustrated in Fig. 7b correspond to these half-wavelength pathlength differences.

From Fig. 7, with the assumptions that $h \ll d_{TP}$, d_{PR}, the Fresnel zones correspond to the condition

$$\Delta L = \sqrt{d_{TP}^2 + h^2} + \sqrt{d_{PR}^2 + h^2} - d_{TR} \simeq \frac{1}{2}h^2\left(\frac{1}{d_{TP}} + \frac{1}{d_{PR}}\right) = n\frac{\lambda_0}{2} \tag{18}$$

where n is any integer other than zero. The first Fresnel zone is that for which $n = 1$. It follows from (18) that

$$h_{cl} = \sqrt{\frac{\lambda_0 d_{TP} d_{PR}}{d_{TP} + d_{PR}}} \tag{19}$$

where h_{cl} is the "Fresnel zone clearance" height, that is, the minimum height of the transmitter and receiver (assumed to be the same in this simplified analysis) above the top of an obstacle such that line-of-sight conditions are approximated.

To determine the field strength in the shadow zone of an obstacle, diffraction theory must be invoked. Given the wedge shown in Fig. 7a as a simple example of such an obstacle, a field component at R behind the wedge is given approximately in Ref. [8], (pp. 402–406):

$$E_R = \int_{-\infty}^{\infty} dy' \int_{h_w}^{\infty} dz' A(y', z') e^{j\Delta\phi(y', z')} \tag{20}$$

where $A(y', z')$ and $\Delta\phi(y', z')$ are the amplitude and phase (relative to the phase of the straight-line raypath TR) of the field at a point (y', z') on the plane P. If we model the effect as two-dimensional (i.e., uniform in the y direction) and note that $d_1 + d_2 \gg |h_R - h_T|$, after some analysis, we arrive at the expression

$$E_R \simeq \frac{A_0}{\eta} \int_X^{\infty} du\, e^{j\left(\frac{\pi}{2}\right)\mu^2} \tag{21}$$

where $X = \eta[(h_w - h_{av}) + (\Delta h/2)\xi]h_{av} = (h_T + h_R)/2$, $\Delta h = h_R - h_T$

$$\eta = \sqrt{\frac{2}{\lambda_0}\left(\frac{d_1 + d_2}{d_1 d_2}\right)}, \quad \xi = \left(\frac{d_2 - d_1}{d_2 + d_1}\right)$$

Evaluation of the Fresnel integral in (21) leads to the diffraction loss in decibels as a function of the parameter X, shown in Fig 8. That loss is the ratio of field strength at R with the obstacle present to that if the obstacle were not present (i.e., the line-of-sight case). By assigning values to the parameters d_1, d_2, λ_0, h_w, h_{av}, and Δh, calculating X with these values, we can determine the loss in decibels corresponding to that value of X on the curve of Fig 8.

Knife-edge diffraction, as described above, provides a very rough estimate of field amplitude in the shadow zone of an obstacle such as a hill or a building along the propagation path. For realistic terrain, there are usually many vertical protrusions along the path and they are *not* usually simple wedges. There are methods of analysis that

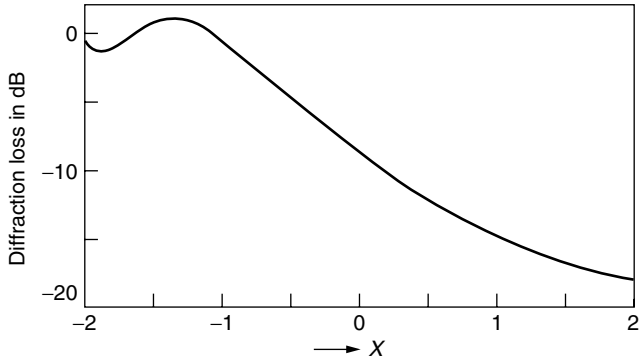


Figure 8. Diffraction loss versus parameter X in Eq. (21).

deal with more complicated diffraction models, summarized by Parsons [4] in 1992 and by Rappaport [5] in 1996.

2.6. Scattering

Scattering from objects, both natural (e.g., trees) and human-made (e.g., distant buildings, particularly in urban environments), occurs in most propagation paths. To differentiate scattering from earth reflection as discussed in Section 2.2, the latter is primarily specular and coherent; that is, the relative phase between direct and reflected waves is mostly deterministic, and hence their superposition contains peaks and troughs due to constructive and destructive interference, respectively. Scattering is from an object sufficiently small and far enough away from both T and R so that it subtends a negligibly small angle with respect to either T or R, appearing as a point when viewed from either site. The electric field at the receiver is determined from the bistatic radar equation [see derivation in, e.g., Ref. 8, (pp. 17–20)] and has the general form

$$E_S^{(V,H)} = \frac{e^{-jk_0(D_{TP}+D_{PR})}}{4\pi D_{TP}D_{PR}}\sqrt{P_T G_{T0} A_{eR0}\sigma_0}\, f_T^{(V,H)}(\Delta\boldsymbol{\Omega}_{TP}) \times f_R^{(V,H)}(\Delta\boldsymbol{\Omega}_{RP}) g^{(V,H)} \frac{\Delta\boldsymbol{\Omega}_{TP}}{\Delta\boldsymbol{\Omega}_{PR}} \tag{22}$$

where the subscript P denotes the location of the scatterer and where the quantities in (22) were all defined following Eq. (1), except σ_0, the peak value of the scattering cross section of the object and $g^{(V,H)}(\Omega_{TP}/\Omega_{PR})$, the angular complex field pattern of scattering from P into the direction of R given a wave originating at T and incident on P. An important feature of the scattered field is that the received power, proportional to $|E_S|^2$, is inversely proportional to the square of $D_{TP}D_{PR}$. This implies that it is usually small compared with the direct field as given by Eq. (1) or the specularly reflected field as obtained from Eq. (3). Hence, in order for this effect to be important in an outdoor communication link, there must be a large number of strong scatterers whose superposed fields generate a signal comparable in magnitude to the received signal. However, in urban or indoor environments, there may be large numbers of scatterers, such as buildings in urban scenes or particularly metallic objects within a room that compete in magnitude with direct signals and whose relative phases generate spatial peaks and troughs, the latter appearing as "dead zones" if the receiver is placed at their locations.

2.7. Tropospheric Scatter

A form of communication mode that was popular in the 1950s and 1960s was "tropospheric scatter" or "troposcatter" [7, pp. 418–453], based on scattering of the transmitted wave from random inhomogeneities in permittivity within a small tropospheric region where transmitter and receiver antenna beams intersect, such that significant power is scattered into the direction of the receiver. The advantage of this propagation mode is that it provides a clear transmission path well above the terrain, circumvents the problem of obstacles along the path, and greatly extends the effective radio horizon. However, satellite links provide those same features with greater reliability and for much longer propagation paths and have largely replaced troposcatter links within recent years.

2.8. Diffraction Around the Earth

In very long-range communication links, where the receiver is beyond the radio horizon of the transmitter, nonzero signal levels can be attained through the mechanism of diffraction around the earth. The theory behind "diffraction zone" propagation is developed in great detail in Ref. 1 (pp. 109–112), and the results are summarized in Ref. 6 (pp. 369–372).

2.9. Attenuation

An attenuation factor of the form $e^{-\alpha D}$ may be present in the electric or magnetic field of a propagating wave, where α is a positive real number measured in nepers per meter and D is the propagation pathlength. It is usually expressed in decibels: $20\log_{10} e^{-\alpha D} = -20(\log_{10} e)\alpha D = 8.686\alpha D$. The important number is $\gamma = 8686\alpha$, the attenuation in decibels per kilometer.

The real part of the refractive index of perfectly dry air is nearly unity, but if there is some degree of moisture content of the air, an imaginary part exists and gives rise to attenuation ranging from the order of 10^{-4} dB/km at 100 MHz to about 10^{-1} dB/km at 30 GHz. Hence within the frequency range of interest attenuation due to atmospheric gases is seldom important over the short pathlengths characteristic of mobile wireless links, even at the resonance peak that occurs at 22.24 GHz. Attenuation due to scattering from and absorption by the tiny water droplets that constitute a very dense fog may be significant at frequencies in the 20–30-GHz region if the density of water droplets is sufficiently high.

Raindrops are another source of attenuation, significant at frequencies above 10 GHz. With sufficiently high rainfall rates, the drops are large enough compared to wavelength to absorb energy in a propagating wave and to scatter it in directions other than that of propagation. Both of these mechanisms result in an exponential decay in wave amplitude, of the order of 0.01 dB/km in a light drizzle to as much as 3–4 dB/km in a very heavy rainfall at 20 GHz and possibly 6–10 dB/km at 30 GHz. Snow particles exhibit similar characteristics. The effects of precipitation on a

wireless link can sometimes be severe at the high end of the microwave region. The details of the theoretical background and key results on atmospheric attenuation in general can be found in Ref. 1 (Chap. 8, pp. 641–692) and on attenuation due to precipitation on in the same work [1, pp. 671–692]. This is still an authoritative source on the subject for engineers requiring data for design purposes.

Another source of attenuation is foliage along the propagation path in a forested environment. The attenuation is due to scattering of wave energy by trees and blockage of the raypath by large aggregates of trees. Since this is a highly specialized situation for wireless links, it will not be discussed further here.

3. PATH LOSS PREDICTION MODELS

A number of models have been used to determine path loss in wireless communication links. Some of these are physics-based and account for the propagation effects discussed above, namely, earth reflection, atmospheric refraction, effects of earth curvature and surface roughness, and diffraction losses due to obstacles in the propagation path. The last of these effects is so important in determination of path loss in real-world environments that it has received enormous emphasis in research on VHF and UHF propagation.

Since propagation environments are often too complex to model physically, some models have been developed from empirical data. Some of these empirical models are widely applied to loss computations for paths along irregular terrain, urban areas, and indoor environments. In what follows, some of the analytical and empirical models will be briefly summarized or at least mentioned with references to the literature for details.

3.1. Analytical Models

The analytical models differ primarily in the degree of complexity in accounting for diffraction. One of the earliest (1947) models, that of Bullington [10], accounted for two knife-edge-type obstacles along the path and set up a diffraction problem for an equivalent single knife edge. This was an attempt to model more than one obstacle. Diffraction from multiple obstacles was later (1953) treated by Epstein and Peterson [11] and much later by Deygout [12] and Longley and Rice [13], Edwards and Durkin [14], and others [15–17]. The Longley–Rice model has been widely used in the United States for irregular terrain modeling and was improved between 1967 and 1985, including a partially empirical extension to urban areas [18]. Covering frequencies from 40 to 100 GHz, the model included most of the effects discussed above — ground reflection, earth curvature effects, diffraction from isolated obstacles, and both tropospheric scatter and earth diffraction for very long propagation paths [4, pp. 57–61; 19].

Another model and associated computer program similar to Longley–Rice, is due to Edwards and Durkin [14] and Dadson et al. [20]. This method was adopted by the Joint Radio Committee (JRC) in the United Kingdom and has been widely used there.

Models like those indicated above are two-dimensional, confined to the vertical propagation plane, and can roughly predict path losses due to ground reflection and diffraction from widely separated obstacles within that plane. But they cannot account adequately for scattering and diffraction due to natural and synthetic terrain features very close to the receiver, and diffraction from obstacles that are close together (and hence interact with each other in a complicated manner). Finally, they do not account for three dimensional effects, such as the multipath arising from constructive and destructive interference between various scatterers distributed horizontally along the terrain both on and off the vertical propagation plane. The rapid fading prevalent in mobile communication links is due largely to these effects.

Another avenue of research to improve the realism of diffraction models is to consider rounded edges, more typical of real hills than the knife edge [e.g., 2, pp. 129,130; 4, pp. 45–47]. Still another is to use geometric theory of diffraction (GTD) or unified theory of diffraction (UTD) in lieu of knife-edge diffraction. There is a body of literature on these latter topics in the context of radio propagation modeling [e.g., 21–23].

3.2. Empirical Models

A set of empirical models primarily designed for urban areas was developed by Okumura et al. [24], using a series of measurements in and near Tokyo, and empirical equations to fit the data, valid from 150 MHz to 1.5 GHz were developed by Hata [25]. Others have since contributed to these kind of models. Physics-based analytical modeling is often difficult for complicated urban scenes. Empirical models, although based on experiments done in specific locations and therefore less flexible, can be useful in development of wireless links.

3.3. Models for Outdoor Urban Areas

Urban areas present a challenging problem in development of path loss models. A model for a scene consisting of a set of buildings arranged deterministically in a horizontal rectangular array is characteristic of an urban area. The major effects contributing to path loss are not necessarily in the vertical plane but usually require three-dimensional modeling. Typically, the direct line of sight between T and R is rarely available. The available paths are usually those involving multiple reflections between buildings and street surfaces and diffraction around buildings in both horizontal and vertical planes. A diffraction model accounting for some of these effects was developed by Walfisch and Bertoni in 1988 [26]. Three-dimensional ray-tracing models have been developed since that time to treat these kinds of urban geometries [e.g., 27–29]. With the increases in computer power that have occurred since the early 1990s, it is possible to compute the path loss in very complicated urban environments within reasonable CPU times.

3.4. Models for Indoor Propagation

With very rapid increases in the use of cell phones since the late 1990s, it has become important to understand propagation within buildings and between sites in different

buildings and to predict path losses for such situations. In this case, the important effects are (1) losses in transmission through walls; (2) multiple reflections between floors walls, and ceilings within rooms; (3) scattering from objects within a room; and (4) diffraction around objects within the direct path between transmitter and receiver. Rappaport [5, pp. 123–132] presents a particularly good summary of indoor propagation models and considerable data, much of it empirical, on losses through various materials found in buildings at various frequencies. Subsequent work has been done by many researchers on indoor propagation modeling and simulation [e.g., 30–32].

BIOGRAPHY

Harold R. Raemer received his Ph.D degree in physics from Northwestern University, Evanston, Illinois in 1959. From 1952 to 1963 he was a research engineer in industrial laboratories, performing analytical studies on problems in radio wave propagation and radar and communication systems. In 1963, he joined the faculty of the Electrical Engineering Department at Northeastern University, Boston, Massachusetts, as an associate professor. He became professor in 1966 and served as chair of the department from 1967 to 1977, and later as acting chair from 1982 to 1984. From 1984 to 1993 he was associate director of radio frequency phenomena and systems at the Center for Electromagnetics Research at Northeastern. He retired from the faculty in 1994 but has remained at the university to the present as a professor emeritus. During his career at Northeastern, he taught undergraduate and graduate courses in a number of subjects within the EE curriculum, conducted sponsored research in plasma dynamics, radio wave propagation, and later in simulation of propagation and scattering in radar systems and wireless communication systems. He is the author of two books and a number of papers in research journals and conference proceedings.

BIBLIOGRAPHY

1. D. E. Kerr, ed., *Propagation of Short Radio Waves*, McGraw-Hill, New York, 1951.
2. J. Griffiths, *Radio Wave Propagation*, McGraw-Hill, New York, 1987.
3. L. Boithias, *Radio Wave Propagation*, McGraw-Hill, New York, 1987.
4. D. Parsons, *The Mobile Radio Propagation Channel*, Wiley, New York, 1992.
5. T. Rappaport, *Wireless Communications*, Prentice-Hall, Upper Saddle River, NJ, 1996.
6. R. E. Collin, *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985, Chap. 6.
7. P. Beckmann and A. Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces*, Artech, Norwood, MA, 1987.
8. H. Raemer, *Radar Systems Principles*, CRC Press, Boca Raton, FL, 1997.
9. S. Ayasli and M. B. Carlson, SEKE: *A Computer Model for Low Altitude Radar Propagation over Irregular Terrain*, Project Report CMT-70, MIT Lincoln Laboratory, Lexington, MA, May 1, 1985.
10. K. Bullington, Radio propagation at frequencies above 30Mc, *Proc. IRE* **35**(10): 1122–1136 (Oct. 1947).
11. J. Epstein and D. W. Peterson, An experimental study of wave propagation at 850Mc, *Proc. IRE* **41**(4): 595–611 (May 1953).
12. J. Deygout, Multiple knife-edge diffraction of microwaves, *IEEE Trans. Antennas Propag.* **AP-14**(4): 480–489 (July 1966).
13. A. G. Longley and P. L. Rice, *Prediction of Tropospheric Radio Transmission Loss over Irregular Terrain, a Computer Method*, ESSA Technical Report, ERL 79-ITS67, 1968.
14. R. Edwards and J. Durkin, Computer prediction of service area for VHF mobile radio networks, *Proc. IEE* **116**(9): 1493–1500 (Sept. 1969).
15. C. L. Giovaneli, An analysis of simplified solutions for multiple knife-edge diffraction, *IEEE Trans. Antennas Propag.* **AP-32**(3): 297–301 (March 1984).
16. L. E. Vogler, An attenuation function for multiple knife-edge diffraction, *Radio Sci.* **17**(6): 1541–1546 (Nov.–Dec. 1982).
17. J. H. Whitteker, A series solution for diffraction over terrain modeled as multiple bridged knife-edges, *Radio Sci.* **28**: 487–500 (July–Aug. 1993).
18. A. G. Longley, *Radio Propagation in Urban Areas*, Office of Telecommunications (OT) Report, April 1978, pp. 78–144.
19. IEEE Vehicular Technology Society Committee on Radio Propagation, Coverage prediction for mobile radio systems operating in the 800/900 MHz frequency range, *IEEE Trans. Vehic. Technol.* **VT-37**(1): 3–72 (Feb. 1988).
20. C. E. Dadson, J. Durkin, and E. Martin, Computer prediction of field strength in the planning of radio systems, *IEEE Trans. Vehic. Technol.* **VT-24**(1): 1–7 (Feb. 1975).
21. R. J. Luebbers, Finite conductivity uniform GTD versus knife-edge diffraction in prediction of propagation path loss, *IEEE Trans. Antennas Propag.* **AP-32**: 70–76 (Jan. 1984).
22. S. Y. Tan and H. S. Tan, UTD propagation model in an urban street scene for micro-cellular communications, *IEEE Trans. Electromagn. Compat.* **EC-37**: 423–428 (Nov. 1993).
23. S. Y. Tan and H. S. Tan, Propagation model for micro-cellular communications in Ottawa city streets, *IEEE Trans. Vehic. Technol.* **VT-44**: 313–317 (May 1995).
24. Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, Field strength and its variability in the VHF and UHF and mobile radio service, *Rev. Electron. Commun. Lab.* **16**(9–10): 825–873 (Sept.–Oct. 1968).
25. M. Hata, Empirical formula for propagation loss in land mobile radio service, *IEEE Trans. Vehic. Technol.* **VT-29**(3): 317–325 (Aug. 1980).
26. S. Walfisch and H. L. Bertoni, A. theoretical model for VHF propagation in urban environments, *IEEE Trans. Antennas Propag.* **AP-36**: 1788–1796 (Oct. 1988).
27. F. Ikegami, S. Yoshida, T. Takeuchi, and M. Umehira, Propagation factors controlling mean field strength on urban streets, *IEEE Trans. Antennas Propag.* **AP-32**(8): 822–829 (Aug. 1984).
28. A. J. Rustako, N. Amitay, G. J. Owens, and R. S. Roman, Radio propagation at microwave frequencies for line-of-sight

microcellular mobile and personal communications, *IEEE Trans. Vehic. Technol.* **VT-40**(1): 203–210 (Feb. 1991).

29. G. Liang and H. L. Bertoni, A new approach to 3-D ray tracing for propagation prediction in cities, *IEEE Trans. Antennas Propag.* **AP-46**(6): 853–863 (June 1998).
30. U. Dersch and E. Zollinger, Propagation mechanisms in micro-cell and indoor environments, *IEEE Trans. Vehic. Technol.* **VT-43**: 1058–1066 (Nov. 1994).
31. C. F. Yang, B. C. Wu, and C. J. Ko, A ray-tracing method for modeling indoor wave propagation and penetration, *IEEE Trans. Antennas Propag.* **AP-46**(6): 907–919 (June 1998).
32. K. W. Chang, J. H. M Sau, and R. D. Murch, A new empirical model for indoor propagation prediction, *IEEE Trans. Vehic. Technol.* **VT-47**(3): 996–1001 (Aug. 1998).

AUTHENTICATION CODES

THOMAS JOHANSSON
Lund University
Lund, Sweden

1. INTRODUCTION

The protection of unauthorized access to sensitive information has been a prime concern throughout the centuries. Still, it was not until Shannon's work in the late 1940s [11] a theoretical model for secrecy was developed. Shannon's work was based on the concept of unconditional security, by which we mean that the enemy faced is assumed to have access to infinite computing power. Under this assumption Shannon developed some rather pessimistic results on the requirements for a cryptosystem to be secure.

More recently, we have understood that one usually needs to protect data not only against unauthorized access but also against unauthorized modifications. In a communication situation, we need to *authenticate* our transmitted messages. We need to check that they are indeed sent by the claimed sender and that they have not been modified during transmission. The threat from the enemy can be viewed as "intelligent noise," the noise taking the worst possible value for the sender and receiver. This means that error correcting codes will not help (because the noise just changes a transmitted codeword to another codeword), but we must introduce secret *keys* that are known to the sender/receiver but unknown to the enemy.

Authentication of transmitted messages can be done in (at least) three fundamentally different ways. We refer to them as *unconditionally secure authentication codes, message authentication codes*, and *digital signatures*.

Unconditionally secure authentication codes is the only solution to the authentication problem for an enemy with unlimited computing power. As this is the topic of the article, we continue the discussion in the next section.

Message authentication codes (MACs) refer to authentication techniques that use symmetric cryptographic primitives (i.e., block ciphers and hash functions) to provide authentication. As for unconditionally secure authentication codes, the sender and receiver are here assumed to share a common secret key. MACs is a very common authentication technique in, for example, banking transactions. MACs appear in many standards, and some common modes of operations for block ciphers provide MACs. We refer to Ref. 9 for more details. Comparing with unconditionally secure authentication codes, MACs are not secure against an unlimited enemy. But they have other practical advantages, such as being able to authenticate many messages without changing the key.

Finally, digital signatures is an asymmetric solution. This means that the sender has a personal secret signing key and the receiver has access to a corresponding public verification key. The sender first hashes the message to be transmitted using a cryptographic hash function. The result is then signed by a signature scheme. Common hash functions are MD5 and SHA-1, and common signature schemes are RSA and DSA. See Ref. 9 for more information. Digital signatures possess several advantages compared to the other two authentication techniques. Since it is an asymmetric technique, there is no need to distribute or establish a common secret key between the sender and the receiver. Basically, the sender generates his/her secret signing key and the corresponding public verification key. The verification key can then be presented in public. This means that anyone can verify the authenticity of a message. This leads to the second important difference, referred to as nonrepudiation. Since the sender is the only person able to generate an authentic message (the receiver cannot), we know that if a message is authentic, it must have been generated by the sender. If the receiver has received an authentic message, the sender cannot deny having sent it. This somewhat resembles a handwritten signature, once you have signed you cannot later deny having signed. There are also drawbacks. Signature schemes rely on the hardness of problems, such as factoring and taking discrete logarithms. This means that we must work with very large numbers, which make the solutions slow compared to the other techniques, especially for short messages.

2. AUTHENTICATION CODES

An unconditionally secure solution to the authentication problem first appeared in 1974 when Gilbert, MacWilliams, and Sloane published their landmark paper "Codes which detect deception" [4]. As mentioned in that paper, Simmons was independently working with the same problems. In the early 1980s Simmons published several papers on the subject that established the authentication model, [12,13,15]. Simmons work on authentication theory has a similar role as Shannons work on secrecy.

This section deals with unconditionally secure authentication codes. We provide some fundamental definition and results. We also include some common constructions. We start by presenting the mathematical model of unconditionally secure authentication due to Simmons.

The communication model for authentication includes three participants, *the transmitter, the receiver*, and *the opponent*. The transmission from the transmitter to the receiver takes place over an insecure channel. The opponent, who is the enemy, has access to the channel in the sense that it can insert a message into the channel, or alternatively, observe a transmitted message and then

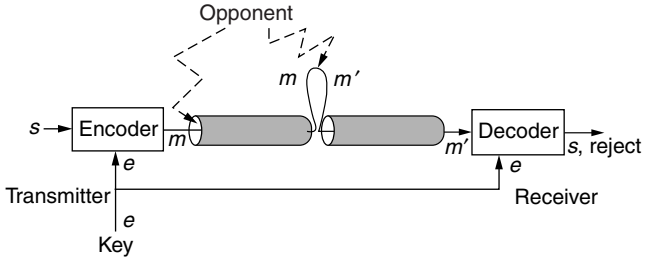


Figure 1. The authentication model.

replace it with another message. The authentication model is illustrated in Fig. 1.

The information that the transmitter wants to send is called a *source message*, denoted by s and taken from the finite set $\mathcal{S}$ of possible source messages. The source message is mapped into a (channel) *message*, denoted by m and taken from the set $\mathcal{M}$ of possible messages. Exactly how this mapping is performed is determined by the secret *key*, which is denoted by e and taken from the set $\mathcal{E}$ of possible encoding rules. The key is secretly shared between the transmitter and the receiver.

Each key determines a mapping from $\mathcal{S}$ to $\mathcal{M}$. Equivalently, the encoding process can be described by the mapping f, where

$$f: \mathcal{S} \times \mathcal{E} \to \mathcal{M}, (s, e) \mapsto m \tag{1}$$

An important property of f is that if $f(s, e) = m$ and $f(s', e) = m$, then $s = s'$ (injective for each $e \in \mathcal{E}$). Two different source messages cannot map to the same message for a given encoding rule, since then the receiver would not be able to determine which source message was transmitted. The mapping f together with the sets $\mathcal{S}, \mathcal{M}$, and $\mathcal{E}$ define an *authentication code* (A-code).

When the receiver receives a message m, it must check whether a source message s exists, such that $f(s, e) = m$. If such an s exists, the message m is accepted as authentic (m is called "valid"). Otherwise, m is not authentic and thus rejected. We can assume that the receiver checks $f(s, e)$ for all $s \in \mathcal{S}$, and if it finds $s \in \mathcal{S}$ such that $f(s, e) = m$, it outputs s; otherwise it outputs a reject signal.

The opponent has two possible attacks at its disposal: the impersonation attack and the substitution attack. The *impersonation attack* simply means inserting a message m and hoping for it to be accepted as authentic. In the *substitution attack*, the opponent observes the message m and replaces this with another message $m', m \neq m'$, hoping for m' to be valid.

We assume that the opponent chooses the message that maximizes its chances of success when performing an attack. The probability of success in each attack is denoted by P_I and P_S, respectively. They are more formally defined by[1]

$$P_I = \max_m P \quad (m \text{ is valid}) \tag{2}$$

[1] We abbreviate expressions such as $\max_{m \in \mathcal{M}}$ as $\max_m$ when no possibility of confusion occurs.

and

$$P_S = \max_{\substack{m, m' \\ m \neq m'}} P \quad (m' \text{ is valid} \mid m \text{ is valid}) \tag{3}$$

Note that this definition considers only transmission of a single message. For transmission of multiple messages, we must introduce a more general definition of the deception probabilities.

Continuing, we define the *probability of deception* P_D as $P_D = \max(P_I, P_S)$. It is convenient to define $\mathcal{E}(m)$ as the set of keys for which a message m is valid:

$$\mathcal{E}(m) = \{e \in \mathcal{E}; \exists s \in \mathcal{S}, f(s, e) = m\} \tag{4}$$

Let us now derive some basic properties for authentication codes. We see that of all the messages in $\mathcal{M}$, at least $|\mathcal{S}|$ must be authentic, since every source message maps to a different message in $\mathcal{M}$. Similarly, for the substitution attack, after the observation of one legal message, at least $|\mathcal{S}| - 1$ of the remaining $|\mathcal{M}| - 1$ messages must be authentic. Thus we have two obvious bounds.

Theorem 1. For any authentication code

$$P_I \geq \frac{|\mathcal{S}|}{|\mathcal{M}|} \tag{5}$$

$$P_S \geq \frac{|\mathcal{S}| - 1}{|\mathcal{M}| - 1} \tag{6}$$

From Theorem 1 we observe two fundamental properties of authentication codes. First, in order to ensure good protection $|\mathcal{M}|$, must be chosen much larger than $|\mathcal{S}|$. This affects the message expansion of our authentication code. For a fixed source message space, an increase in the authentication protection implies an increased message expansion. The second property is that a complete protection (i.e., $P_D = 0$) is not possible. We must be satisfied with a protection where P_D is small.

For example, let an authentication code with $\mathcal{S} = \{H, T\}$, $\mathcal{M} = \{1, 2, 3, 4\}$ and $\mathcal{E} = \{0, 1, 2, 3\}$ be described by the following table

		m			
	s	1	2	3	4
	0	H	T	—	—
e	1	T	—	H	—
	2	—	H	—	T
	3	—	—	T	H

It is easy to verify that $P_I = P_S = \frac{1}{2}$ if the keys are uniformly distributed.

We assume that the reader is familiar with the basic concepts of information theory. As usual, $H(X)$ denotes the entropy of the random variable X, and $I(X; Y)$ denotes the mutual information between X and Y. We are now ready to state the next fundamental result in authentication theory, namely, Simmons' bounds.

Theorem 2: Simmons' Bounds. For any authentication code, we obtain

$$P_I \geq 2^{-I(M;E)}, \tag{7}$$

$$P_S \geq 2^{-H(E|M)}, \quad \text{if } |\mathcal{S}| \geq 2. \tag{8}$$

The bound for the impersonation attack was first proved by Simmons in 1984 with a long and tedious proof [13]. Several new and much shorter proofs have been given since then. The bound for the substitution attack was proved by Simmons and Brickell in [2], but can also be proved in the same way as for the impersonation attack.

Simmons' bounds give a good feeling of how the authentication protection affects the system. For the impersonation attack, we see that P_I is upper-bounded by the mutual information between the message and the key. This means that for a good protection (i.e., P_I small), we must give away a lot of information about the key. On the other hand, in the substitution attack, P_S is lower-bounded by the uncertainty about the key when a message has been observed. Thus we cannot waste all the key entropy for protection against the impersonation attack, but some uncertainty about the key must remain for protection against the substitution attack.

Returning to Theorem 2, we multiply the two bounds together and get

$$P_I P_S \geq 2^{-I(M;E)-H(E|M)} = 2^{-H(E)} \tag{9}$$

From the inequality $H(E) \leq \log|\mathcal{E}|$ we then obtain the *square-root bound*.

Theorem 3: Square-Root Bound. For any authentication code, we have

$$P_D \geq \frac{1}{\sqrt{|\mathcal{E}|}} \tag{10}$$

The bound was originally proved in [4] under other conditions and slightly differently stated. The square root bound gives a direct relation between the key size and the protection that we can expect to obtain. Thus the following definitions are natural.

An authentication code for which equality holds in the square-root bound (10) is called a "perfect" A-code.[2] Furthermore, an A-code for which $P_I = P_S$ is called an "equitable" A-code.

Obviously, a perfect A-code must be equitable. If we can construct A-codes for which equality holds in the square root bound we can be satisfied, since in that case no better authentication codes exist, in the sense that P_D cannot be made smaller. This is a main topic, but also equitable A-codes that are not perfect are of interest. The reason for this is the following.

Theorem 4. The square-root bound (10) can be tight only if

$$|\mathcal{S}| \leq \sqrt{|\mathcal{E}|} + 1$$

The result was discussed in Ref. 4.

The square-root bound motivates a treatment of nonperfect A-codes, since for perfect A-codes a large source size demands a twice as large key size. This is not very practical. On the other hand, if the source size is very modest (i.e., $|\mathcal{S}| \leq \sqrt{|\mathcal{E}|} + 1$), then we will see in the sequel that perfect A-codes can be constructed for any $P_D = 1/q$, where $q = \sqrt{|\mathcal{E}|}$ is a prime power.

The most important kind of authentication code is when the source message s appear as a part of the channel message m. An A-code for which the map $f: \mathcal{S} \times \mathcal{E} \to \mathcal{M}$ can be written in the form

$$f: \mathcal{S} \times \mathcal{E} \to \mathcal{S} \times \mathcal{Z}, \quad (s, e) \mapsto (s, z) \tag{11}$$

where $s \in \mathcal{S}, z \in \mathcal{Z}$ is called a *systematic* (or *Cartesian*) *A-code*. The second part z in the message is called the "tag" (or authenticator) and is taken from the tag alphabet $\mathcal{Z}$. We see that systematic A-codes are codes that have no secrecy at all, the source message is transmitted in the clear, and we add some check symbols to it (the tag). In the sequel we study only systematic authentication codes. Systematic A-codes have the following important property [6].

Theorem 5. For any systematic A-code

$$P_S \geq P_I \tag{12}$$

This means that for systematic A-codes the square-root bound is expressed as

$$P_S \geq \frac{1}{\sqrt{|\mathcal{E}|}}$$

Finally, for systematic A-codes with uniformly distributed keys and $P_I = P_S = 1/|\mathcal{Z}|$, we have the inequality [6]

$$(|\mathcal{Z}| - 1)|\mathcal{S}| \leq |\mathcal{E}| - 1 \tag{13}$$

This bound shows that large source sizes for equitable A-codes require large key sizes, and gives the motivation for the study of nonequitable A-codes.

We next present some ways of constructing equitable A-codes. Equitable A-codes have the lowest possible probability of deception in the sense that $P_D = P_I = P_S$, but they have the disadvantage of having a source size that is quite modest. It is useful to note that the probability of success in a substitution attack can be written as

$$P_S = \max_{\substack{m,m' \\ m \neq m'}} \frac{|\mathcal{E}(m) \cap \mathcal{E}(m')|}{|\mathcal{E}(m)|} \tag{14}$$

provided that the keys are uniformly distributed.

To have some measure of how good a construction is, we introduce two fundamental definitions. An A-code with fixed parameters $|\mathcal{E}|, |\mathcal{M}|, |\mathcal{S}|$, and P_I is said to be weakly optimal if P_S is the lowest possible. A weakly optimal A-code is said to be strongly optimal if, additionally, $|\mathcal{S}|$ has the largest possible value among all the weakly optimal A-codes for fixed parameters $|\mathcal{E}|, |\mathcal{M}|, P_I$.

We start by giving the original *projective plane construction* proposed by Gilbert et al. [4]. Fix a line L in $\mathbf{PG}(2, \mathbb{F}_q)$. The points on L are regarded as source messages, the points not on L are regarded as keys, and the lines distinct from L are regarded as messages. The mapping from $\mathcal{S}$ to $\mathcal{M}$ means joining the source message s and the key e to the unique line m, which is the resulting message.

[2] The definition of the terminology perfect A-code may be different in other literature.

We can easily verify correctness of this construction. The joining of the point e outside L and the point s on L results in a unique line, called m. By running through all pairs (s, e), we find the message space as all lines except L itself. The parameters of the A-code are given by the following theorem.

Theorem 6. The projective plane construction gives parameters

$$|\mathcal{S}| = q + 1, \quad |\mathcal{M}| = q^2 + q, \quad |\mathcal{E}| = q^2$$

and the probabilities of success are $P_I = 1/q$ and $P_S = 1/q$.

The A-codes resulting from this construction are strongly optimal.

Another simple construction is the *vector space construction*. Let $|\mathcal{S}| = q^m, |\mathcal{Z}| = q^m$, and $|\mathcal{E}| = q^{2m}$. Decompose the keys as $e = (e_1, e_2)$, where $s, z, e_1, e_2 \in \mathbb{F}_{q^m}$. For transmission of source message s, generate a message $m = (s, z)$, where

$$z = e_1 + se_2$$

Theorem 7. The above construction described provides $P_I = P_S = 1/q^m$. Moreover, it has parameters $|\mathcal{S}| = q^m, |\mathcal{Z}| = q^m$, and $|\mathcal{E}| = q^{2m}$.

Authentication codes are closely related to combinatorial designs. This relation has been extensively examined in a number of papers. Brickell [2], and later Stinson [16], established a one-to-one correspondence between A-codes with given parameters and certain combinatorial designs. The designs used include transversal designs, orthogonal arrays, balanced incomplete block designs, and perpendicular arrays; see Ref. 17 for an introduction.

3. CONSTRUCTING USEFUL AUTHENTICATION CODES

Our treatment so far has not considered any results of interest for the case when $|\mathcal{S}|$ is large. This case is of great relevance since many practical problems concern very large source sizes. Examples of such problems are authentication of data files or computer programs. We now turn our attention to the problem of solving the authentication problem for sources that have a length of, say, a million bits. This means $\log |\mathcal{S}| = 10^6$.

So, a fundamental problem in authentication theory is to find A-codes such that $|\mathcal{S}|$ is large while keeping $|\mathcal{E}|$ and P_S as small as possible. In many practical situations one has limitations on $|\mathcal{E}|$ and wants P_S to be bounded by some small value. Also, one usually wants the redundancy ($|\mathcal{Z}|$) to be small, since it occupies a part of the bandwidth.

In an earlier study [6] it was shown that authentication codes have a close connection to coding theory. It is, for example, possible to construct authentication codes from error-correcting codes and vice versa. Some of the resulting constructions are illustrated next.

First, we give a construction based on Reed–Solomon codes1,6.

Construction is as follows. Let $\mathcal{S} = \{\mathbf{s} = (s_1, \ldots, s_k); s_i \in \mathbb{F}_q\}$. Define the source message polynomial to be $s(x) = s_1x + s_2x^2 + \cdots + s_kx^k$. Let $\mathcal{E} = \{e = (e_1, e_2); e_1, e_2 \in \mathbb{F}_q\}$ and $\mathcal{Z} = \mathbb{F}_q$. For the transmission of source message $\mathbf{s}$, the transmitter sends $\mathbf{s}$ together with the tag

$$z = e_1 + s(e_2)$$

Theorem 8. The construction gives systematic A-codes with parameters

$$|\mathcal{S}| = q^k, \quad |\mathcal{E}| = q^2, \quad |\mathcal{Z}| = q, \quad P_I = 1/q, \quad P_S = k/q.$$

The construction gives weakly optimal A-codes.

In order to make $|\mathcal{S}|$ really large, here is a final construction from [7] based on dual BCH codes. Let q be a power of a prime p, and let $Tr_{q^m/q}$ denote the trace function from $\mathbb{F}_{q^m}$ to $\mathbb{F}_q$.

Construction is as follows. Let the set $\mathcal{F}_D$ of polynomials of degree $D \le \sqrt{q^m}$, be defined by

$$\mathcal{F}_D = \{f(x): f(x) = f_1x + f_2x^2 + \cdots + f_Dx^D \in \mathbb{F}_{q^m}[x], f_i = 0 \text{ whenever } p \mid i\}$$

Let $\mathcal{S} = \mathcal{F}_D, \mathcal{E} = \{(e_1, e_2): e_2 \in \mathbb{F}_{q^m}, e_1 \in \mathbb{F}_q\}$, and let the tag z be generated as

$$z = e_1 + Tr_{q^m/q}(f(e_2))$$

Theorem 9. The parameters for the systematic A-code are

$$|\mathcal{S}| = q^{m(D-\lfloor D/p \rfloor)}, |\mathcal{E}| = q^{m+1}, |\mathcal{Z}| = q$$
$$P_I = \frac{1}{q}, \; P_S = \frac{1}{q} + \frac{D-1}{\sqrt{q^m}}$$

We can look at the performance of this construction. For example, consider the parameters $P_S = 2^{-19}$ and $m = 4$. By choosing q to be a large prime around 2^{20}, we get $\log |\mathcal{S}| = 20 \cdot 4 \cdot \sqrt{2^{40}} = 80 \cdot 2^{20}$. With a 100-bit key, we can protect a source message of 10 MB (mega bytes)!

Further constructions using, for instance, algebraic geometric codes have appeared. The general topic here has been to optimize the parameters of the authentication code, such as for a fixed source message size and fixed security level (P_S) to find the authentication code using the smallest key size. Several bounds on this problem have also been derived [e.g., 6].

Authentication codes are also very closely related to universal hash functions. Universal classes of hash functions were introduced by Carter and Wegman [3], and it quickly became an established concept in computer science. It found numerous applications, such as cryptography, complexity theory, search algorithms, and associative memories, to mention only a few.

We end our treatment of authentication codes by mentioning a different line of research. Instead of optimizing the code parameters, one could focus on authentication codes with a very fast implementation. An interesting approach, called "bucket hashing," was introduced by

Rogaway [10]. These techniques were eventually refined, resulting in constructions like the UMAC [8].

4. AUTHENTICATION FOR TWO NONTRUSTING PARTIES

We now move to the study of authentication codes where the transmitter and receiver do not necessarily have to trust each other. In this situation we include deceptions from the insiders, like the transmitter sending a message and then later denying having sent it, or conversely, the receiver claiming to have received a message that was never sent by the transmitter. Recalling the short discussion on authentication techniques in the introduction, this is a first step toward digital signatures in the world of unconditional security, since it includes the nonrepudiation aspect.

In order to solve possible disputes we include a fourth participant, called the arbiter. The arbiter has access to all key information and, *by definition*, does not cheat. The arbiter is only present to solve possible disputes and does not take part in any communication activities.

Simmons introduced this extended authentication model, which is called the *authentication with arbitration model* [14], or simply the *A^2-model*. In this model, protection is provided against deceptions both from an outsider (opponent) and from the insiders (transmitter and receiver). Until 2001 or so, it was not known whether it was even possible to construct such schemes in an unconditional setting. Simmons gave a positive answer to this question.

We give a brief description of the A^2-model. For more details, we refer to a paper by Simmons [14], which contains a thorough discussion of the different threats. The A^2-model includes four different participants: the *transmitter*, the *receiver*, the *opponent*, and the *arbiter*. As in conventional authentication, the transmitter wants to send some information, called a *source message*, to the receiver in such a way that the receiver can both recover the transmitted source message and verify that the transmitted message originates from the legitimate transmitter. The source message $s \in \mathcal{S}$, is encoded by the transmitter into a message $m \in \mathcal{M}$. The message m is subsequently transmitted over the channel. The mapping from $\mathcal{S}$ to $\mathcal{M}$ is determined by the transmitter's secret key e_t chosen from the set $\mathcal{E}_T$ of possible keys for the transmitter. We may assume that the transmitter uses a mapping $f: \mathcal{S} \times \mathcal{E}_T \to \mathcal{M}$ such that

$$f(s, e_t) = f(s', e_t) \Rightarrow s = s' \tag{15}$$

In other words, the source message can be recovered uniquely from a transmitted channel message. The opponent has access to the channel in the sense that it can either impersonate the transmitter and send a message, or replace a transmitted message with a different one. The receiver must decide whether a received message is valid or not. For this purpose the receiver uses a mapping, determined by its own key e_r taken from the set $\mathcal{E}_\mathcal{R}$ of possible receiver's keys, which determines whether the message is valid, and if so, also the source message. This mapping is denoted $g: \mathcal{M} \times \mathcal{E}_\mathcal{R} \to \mathcal{S} \cup \{\text{reject}\}$, where for all possible (e_t, e_r), specifically, $P(e_t, e_r) \neq 0$, we have

$$f(s, e_t) = m \Rightarrow g(m, e_r) = s \tag{16}$$

For the receiver to accept all legal messages from the transmitter and to translate them to the correct source message, property (16) must hold for all possible pairs (e_t, e_r). However, in general not all pairs (e_t, e_r) will be possible.

The arbiter solves a possible dispute in the following way. If the channel message m, received by the receiver, could have been generated by the transmitter according to its encoding rule e_t, then the arbiter decides that the message m was sent by the transmitter, and otherwise not. The arbiter is assumed to be honest.

In the A^2-model the following five types of cheating attacks are considered.

Attack I: *impersonation by the opponent* — the opponent sends a message to the receiver and succeeds if this message is accepted by the receiver as authentic.

Attack S: *substitution by the opponent* — the opponent observes a message that is transmitted and replaces this message with another. The opponent is successful if this other message is accepted by the receiver as authentic.

Attack T: *impersonation by the transmitter* — the transmitter sends a message to the receiver and then denies having sent it. The transmitter succeeds if this message is accepted by the receiver as authentic, and if this message is not one of the messages that the transmitter could have generated according to its key.

Attack R_0: *impersonation by the receiver* — the receiver claims to have received a message from the transmitter. The receiver succeeds if this message could have been generated by the transmitter according to its key.

Attack R_1: *substitution by the receiver* — the receiver receives a message from the transmitter, but claims to have received another message. The receiver succeeds if this other message could have been generated by the transmitter according to his key.

In all possible attempts to cheat it is understood that the cheating person chooses the message that maximizes his/her chances of success. For the five possible deceptions, we denote the probability of success in each attack by P_I, P_S, P_T, P_{R_0} and P_{R_1}, respectively. The formal definitions are

$$P_I = \max_m P \quad (m \text{ valid}) \tag{17}$$

$$P_S = \max_{\substack{m, m' \\ m \neq m'}} P \quad (m' \text{ valid} \mid m \text{ valid}) \tag{18}$$

$$P_T = \max_{\substack{m, e_t \\ m \notin \mathcal{M}(e_t)}} P \quad (m \text{ valid} \mid e_t) \tag{19}$$

$$P_{R_0} = \max_{m, e_r} P \quad (m \in \mathcal{M}(e_t) \mid e_r) \tag{20}$$

$$P_{R_1} = \max_{\substack{m, m', e_r \\ m \neq m'}} P \quad (m' \in \mathcal{M}(e_t) \mid m \in \mathcal{M}(e_t), e_r) \tag{21}$$

where $\mathcal{M}(e_t)$ is the set of possible messages for the transmitter's encoding rule e_t, specifically, $\mathcal{M}(e_t) = \{m; f(s, e_t) = m, s \in \mathcal{S}\}$. Furthermore, let us define $P_D = \max(P_I, P_S, P_T, P_{R_0}, P_{R_1})$.

If the source messages are uniformly distributed, we have the following lower bounds on the number of encoding rules and on the number of messages [5].

Theorem 10

$$|\mathcal{E}_\mathcal{R}| \geq (P_I P_S P_T)^{-1}$$
$$|\mathcal{E}_\mathcal{T}| \geq (P_I P_S P_{R_0} P_{R_1})^{-1}$$
$$|\mathcal{M}| \geq (P_I P_{R_0})^{-1} |\mathcal{S}|$$

In particular, if $P_D = 1/q$, then

$$|\mathcal{E}_\mathcal{R}| \geq q^3, |\mathcal{E}_\mathcal{T}| \geq q^4, |\mathcal{M}| \geq q^2 |\mathcal{S}|$$

We end this brief discussion with an example for the simplest possible nontrivial case: $P_D = \frac{1}{2}$. Assume that there are two possible source messages, $\mathcal{S} = \{H, T\}$. The Cartesian product construction due to Simmons [14] gives rise to the matrix shown in Table 1. In the key setup, the receiver chooses (or gets from the arbiter) one of the 16 rows as the key E_R. Assume, for example, that the row e_1 will be the receiver's key. Then the receiver will accept the messages m_1, m_2, m_5, and m_6 as authentic. The messages m_1, m_2 will be interpreted as the source message H, and the messages m_5, m_6 will be interpreted as the source message T. All the other messages are not authentic, and will thus be rejected.

The transmitter's encoding rule E_T tells which message corresponds to the source state H and which message corresponds to the source state T. One of the messages m_1–m_4 corresponds to H, and one of the messages m_5–m_8 corresponds to T. Hence there are 16 possibilities and $|\mathcal{E}_\mathcal{T}| = 16$. However, this choice must be made in such a way that the receiver accepts the messages as authentic and translates them to the correct source state; see (16). In this example, where $E_R = e_1$, the message that corresponds to the source message H must be m_1 or m_2, and the message corresponding to the source message T must be m_5 or m_6. When $E_R = e_1$ is given, there are four possible ways to choose E_T, namely, $\{H \mapsto m_1, T \mapsto m_5\}$, $\{H \mapsto m_1, T \mapsto m_6\}$, $\{H \mapsto m_2, T \mapsto m_5\}$, and $\{H \mapsto m_2, T \mapsto m_6\}$. Note that not all pairs (e_r, e_t) are possible.

Table 1. The Cartesian Product Construction for $|S| = 2$ and $P_D = \frac{1}{2}$

		m							
	$g(m, e_r)$	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
	e_1	H	H	—	—	T	T	—	—
	e_2	H	H	—	—	T	—	T	—
	e_3	H	H	—	—	—	T	—	T
	e_4	H	H	—	—	—	—	T	T
	e_5	H	—	H	—	T	T	—	—
	e_6	H	—	H	—	T	—	T	—
	e_7	H	—	H	—	—	T	—	T
e_r	e_8	H	—	H	—	—	—	T	T
	e_9	—	H	—	H	T	T	—	—
	e_{10}	—	H	—	H	T	—	T	—
	e_{11}	—	H	—	H	—	T	—	T
	e_{12}	—	H	—	H	—	—	T	T
	e_{13}	—	—	H	H	T	T	—	—
	e_{14}	—	—	H	H	T	—	T	—
	e_{15}	—	—	H	H	—	T	—	T
	e_{16}	—	—	H	H	—	—	T	T

We can check the probability of success for any kind of deception. Let us introduce the following notation. Let $\mathcal{E}_\mathcal{R}(m)$ denote the set of receiver's encoding rules for which m is a valid message, i.e., $\mathcal{E}_\mathcal{R}(m) = \{e_r; g(m, e_r) \in \mathcal{S}\}$. Similarly, let $\mathcal{E}_\mathcal{T}(m)$ denote the set of transmitter's encoding rules for which m can be generated, $\mathcal{E}_\mathcal{T}(m) = \{e_t; f(s, e_t) = m, s \in \mathcal{S}\}$. Let $\mathcal{M}(e_r)$ be the set of possible messages for encoding rule e_r, $\mathcal{M}(e_r) = \{m; g(m, e_r) \in \mathcal{S}\}$. Finally, let $\mathcal{E}_\mathcal{R}(e_t)$ be the set of possible e_r values for a given e_t, specifically, $\mathcal{E}_\mathcal{R}(e_t) = \{e_r; g(m, e_r) \in \mathcal{S}, \forall m \in \mathcal{M}(e_t)\}$, and let $\mathcal{E}_\mathcal{T}(e_r)$ be the set of possible e_t values for a given e_r, namely, $\mathcal{E}_\mathcal{T}(e_r) = \{e_t; f(s, e_t) \in \mathcal{M}(e_r), \forall s \in \mathcal{S}\}$.

Each column in Table 1 contains 8 entries out of 16, and thus $|\mathcal{E}_\mathcal{R}(m)| = 8$ for any m, and we have

$$P_I = \max_m \frac{|\mathcal{E}_\mathcal{R}(m)|}{|\mathcal{E}_\mathcal{R}|} = \frac{8}{16} = \frac{1}{2}$$

Any two columns have at most 4 rows for which they both have entries, and thus $|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(m')| \leq 4$. We then have

$$P_S = \max_{\substack{m,m' \\ m \neq m'}} \frac{|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(m')|}{|\mathcal{E}_\mathcal{R}(m)|} = \frac{4}{8} = \frac{1}{2}$$

If the receiver claims to have received a message corresponding to source message H, it must choose either m_1 or m_2. But only one of them is the message that the transmitter would have used. Thus $P_{R_0} = \frac{1}{2}$. Or equivalently, $|\mathcal{E}_\mathcal{T}(e_r)| = 4$ for any e_r, and $|\mathcal{E}_\mathcal{T}(e_r) \cap \mathcal{E}_\mathcal{T}(m)| \leq 2$, so

$$P_{R_0} = \max_{m, e_r} \frac{|\mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(e_r)|}{|\mathcal{E}_\mathcal{T}(e_r)|} = \frac{2}{4} = \frac{1}{2}$$

By similar reasoning, $|\mathcal{E}_\mathcal{T}(e_r) \cap \mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(m')| \leq 1$, and thus

$$P_{R_1} = \max_{\substack{m,m',e_r \\ m \neq m'}} \frac{|\mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(m') \cap \mathcal{E}_\mathcal{T}(e_r)|}{|\mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(e_r)|} = \frac{1}{2} = \frac{1}{2}$$

when $P(e_r, m) \neq 0$. Finally, assume, for example, that the transmitter has the mapping $\{H \mapsto m_1, T \mapsto m_5\}$ as his key. To succeed in his attack it must send a message different from m_1 and m_5, which is accepted by the receiver. From its key it knows that the receiver's key is one of the four keys e_1, e_2, e_5, e_6. For the best choice of message, two keys out of four accept the message as authentic, and thus $P_T = \frac{1}{2}$. Or formally, $|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(e_t)| \leq 2$ when $m \notin \mathcal{M}(e_t)$, and $|\mathcal{E}_\mathcal{R}(e_t)| = 4$, gives

$$P_T = \max_{\substack{m,e_t \\ m \notin \mathcal{M}(e_t)}} \frac{|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(e_t)|}{|\mathcal{E}_\mathcal{R}(e_t)|} = \frac{2}{4} = \frac{1}{2}$$

Thus $P_D = \frac{1}{2}$ when the encoding rules are uniformly distributed. The parameters of the A^2-code are

$$|\mathcal{S}| = 2, \quad |\mathcal{M}| = 8, \quad |\mathcal{E}_\mathcal{R}| = 16, \quad |\mathcal{E}_\mathcal{T}| = 16$$

Expressing the parameters in terms of entropy we get $H(S) = 1, H(M) = 3, H(E_R) = 4$, and $H(E_T) = 4$. We also observe that from the dependence between the keys E_R and E_T we have $I(E_R; E_T) = 2$. This is a typical property of A^2 codes.

BIOGRAPHY

Thomas Johansson was born in Ljungby, Sweden in 1967. He received the M.Sc. degree in computer science in 1990 and the Ph.D. degree in information theory in 1994, both from Lund University, Lund, Sweden. From 1995 he has held various teaching and research positions in the Department of Information Technology at Lund University. Since 2000 he has held a position as Professor of Information Theory in the same department.

His scientific interests include cryptology, error-correcting codes, and information theory. He has served on cryptographic program committees such as EUROCRYPT'98/'00/'01/'02 and FSE'01/'02. He was a recipient of the SSF-JIG (Junior Individual Grant).

BIBLIOGRAPHY

1. B. den Boer, A simple and key-economical unconditionally authentication scheme, *J. Comput. Security* **2**(1): 65–71 (1993).
2. E. F. Brickell, A few results in message authentication, *Congressus Numerantium* **43**: 141–154 (1984).
3. J. L. Carter and M. N. Wegman, Universal classes of hash functions, *Journal of Comput. Syst. Sci.* **18**(2): 143–154 (1979).
4. E. N. Gilbert, F. J. MacWilliams, and N. J. A. Sloane, Codes which detect deception, *Bell Syst. Tech. J.* **53**(3): 405–424 (1974).
5. T. Johansson, Lower bounds on the probability of deception in authentication with arbitration, *IEEE Trans. Inform. Theory* **40**(5): (Sept. 1994).
6. T. Johansson, *Contributions to Unconditionally Secure Authentication*, Ph.D. thesis, Lund Univ., 1994.
7. T. Helleseth and T. Johansson, Universal hash functions from exponential sums over finite fields and Galois rings, *Lecture Notes in Computer Science*, Vol. 1107 Springer-Verlag, Berlin, 1996, (CRYPTO'96), pp. 31–44.
8. J. Black et al., UMAC: Fast and secure message authentication, in *Advances in Cryptology — CRYPTO'99 Lecture Notes in Computer Science*, Springer-Verlag, 1999, pp. 216–233.
9. A. Menezes, P. van Oorschot, S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1997.
10. P. Rogaway, Bucket hashing and its application to fast message authentication, *Proc. CRYPTO'95*, Santa Barbara, CA, LNCS 963, Berlin: Springer-Verlag, Berlin, 1995, pp. 29–42.
11. C. E. Shannon, Communication theory of secrecy systems, *Bell Syst. Tech. J.* **28**: 269–279 (Oct. 1949).
12. G. J. Simmons, A game theory model of digital message authentication, *Congressus Numerantium* **34**: 413–424 (1982).
13. G. J. Simmons, Authentication theory/coding theory, *Proc. CRYPTO'84*, Santa Barbara, CA, 1984, LNCS 196, Springer-Verlag, Berlin, pp. 411–431.
14. G. J. Simmons, A Cartesian product construction for unconditionally secure authentication codes that permit arbitration, *J. Cryptol.* **2**(2): 77–104 (1990).
15. G. J. Simmons, A survey of information authentication, in G. J. Simmons, ed., *Contemporary Cryptology, The Science of Information Integrity*, IEEE Press, New York, 1992. pp. 379–420.
16. D. R. Stinson, The combinatorics of authentication and secrecy codes, *J. Cryptol.* **2**(1): 23–49 (1990).
17. D. R. Stinson, *Cryptography Theory and Practice*, CRC Press, 1995.

AUTOMATIC REPEAT REQUEST

THOMAS E. FUJA
University of Notre Dame
Notre Dame, Indiana

1. INTRODUCTION

Error control is a term that describes methods for protecting the integrity of digitally transmitted signals. Error control techniques are used to provide a desired level of reliability when transmitting digital information over inherently unreliable links.

Error control can be broken down into two broad approaches:

- *Forward error control* (FEC), in which redundancy is added to the digital signal prior to transmission, and the redundancy is used at the receiver to reconstruct the transmitted signal, even if it was corrupted en route.
- *Automatic repeat request* (ARQ), in which a (typically) smaller amount of redundancy is added to the digital signal prior to transmission — redundancy that is used to *detect* corruption that may have occurred during transmission. In an ARQ-based system, received signals that are judged to be corrupt are retransmitted.

Clearly, a major difference between FEC and ARQ is that ARQ requires the existence of a *feedback* (or *return*) channel from the receiver to the transmitter; this feedback channel is used to alert the transmitter when corrupted data have been received and to request retransmission. As a result, ARQ is not feasible for systems in which the return channel does not exist or for some real-time applications where it is impractical to request retransmission in a timely manner. Conversely, ARQ schemes *are* implemented in a wide variety of communications contexts, including satellite-based systems, local area networks, and the ubiquitous Internet.

This article describes the various methods by which an automatic repeat request protocol may be deployed to enhance the reliability of a digital communication system.

First, we consider means by which errors that occur during transmission can be detected. (This can be done via a particular form of *block coding*, and the reader should refer to the article on error control codes for more background.) Then, we consider three different protocols by which the sender and receiver coordinate the retransmission of corrupted information; these three protocols are described in Section 3 and their performances are analyzed in Section 4. Finally, *hybrid* ARQ — incorporating elements of both FEC and conventional ARQ — is described in Section 5.

2. ARQ AND CRC CODES IN THE OSI NETWORK ARCHITECTURE

It is possible to deploy error control methods at many different layers of the seven-layer open system interconnection (OSI) architecture [1,2]. For instance, at the physical layer (layer 1) a form of FEC called *trellis-coded modulation* (TCM) may be used improve the reliability of the virtual bit pipe embodied at the physical layer. At the other extreme, many applications (layer 7) that are used to disseminate multimedia files over the Internet employ error control methods to mitigate the effects of errors that may have "gotten past" the error control techniques in place at the lower layers.

Automatic repeat request schemes are typically deployed at the data-link layer (layer 2). It is at the data-link layer that the data packets formed at the transport layer are augmented with extra symbols at the beginning of the packet ("headers") and extra symbols at the end of the packet ("trailers") to form *frames* (see Fig. 1). What exactly goes into the headers and trailers depends on the particular data-link protocol, but typically the header may include an address and/or a packet sequencing number, while the trailer will include one or more *check bytes* used to detect errors that occur in the frame during transmission.

The contents of these check bytes (also known as *parity bits*) are based on the contents of the rest of the frame, and they are chosen so that the bits in the frame satisfy certain parity constraints — constraints that can be checked for violation at the receiver. The codes most commonly used in this way are the so-called cyclic redundancy check (CRC) codes [3,4].

CRC codes form a class of binary block codes; an (n, k) binary block code is an error control construct in which k data bits are appended with $r = n - k$ redundant bits to form an n-bit codeword. Because $k < n$, not all possible n-tuples are valid codewords; this means that errors can be detected when an invalid codeword is observed at the receiver. Cyclic redundancy check codes take their name from the fact that when n (the "blocklength" of the code) takes on its maximum value, the resulting code is *cyclic* — meaning that cyclically shifting one valid codeword yields another valid codeword.

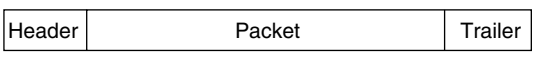


Figure 1. A *frame* consisting of a header, a packet, and a trailer.

A CRC code is defined by a generator polynomial $g(x)$. More specifically, suppose that there are n bits in a frame — call them $[a_{n-1}, a_{n-2}, \ldots, a_1, a_0]$ — and suppose that the r low-order bits (a_0 through a_{r-1}) constitute the parity check bits. [The number of parity check bits is equal to the degree of $g(x)$ — i.e., $r = \deg[g(x)]$.] Then those parity check symbols are chosen so that the polynomial $a(x) = a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ is evenly divisible by $g(x)$. [Here, $a(x)$ and $g(x)$ are both polynomials with binary coefficients, and so all operations are modulo two; moreover, the constraint $g(x)|a(x)$ — read "$g(x)$ divides $a(x)$" — is equivalent to imposing r different parity constraints on the frame.]

As a simple example, consider the generator polynomial $g(x) = x^4 + x + 1$. This is the generator for a [15,11] *Hamming code* — a cyclic code commonly used for error correction and detection. Suppose that we wish to use this code to protect a frame consisting of $n = 15$ bits — 11 "data" (i.e., non-redundant) bits and $r = \deg[g(x)] = 4$ parity bits. Suppose further that the data bits are given by [10011000110]; then the transmitted frame is given by $[10011000110a_3a_2a_1a_0]$ where the parity bits a_0 through a_3 are chosen so that the "codeword polynomial" $c(x) = x^{14} + x^{11} + x^{10} + x^6 + x^5 + a_3x^3 + a_2x^2 + a_1 + a_0$ is divisible by $g(x)$. It can be shown that the appropriate choice is $a_3 = a_1 = 1$ and $a_2 = a_0 = 0$ and so the transmitted frame is [100110001101010]; this is because the corresponding code polynomial $c(x) = x^{14} + x^{11} + x^{10} + x^6 + x^5 + x^3 + x$ is a multiple of $g(x)$ — specifically, $c(x) = g(x) \cdot (x^{10} + x^2 + x)$ — and this choice of the parity bits is the *only* choice that results in a multiple of $g(x)$.

At the receiver, the contents of the frame are checked to see if the "received polynomial" is, indeed, evenly divisible by $g(x)$. If it is, then the frame is declared error-free; if the received polynomial is not evenly divisible by $g(x)$, then the frame is declared corrupt.

CRC codes provide a mechanism through which, at a relatively small cost of overhead, the vast majority of transmission errors can be detected. There are 2^n possible binary n-tuples, but only 2^{n-r} of them satisfy the parity constraints; therefore, if the data are corrupted during transmission into something uniformly random, then the probability the received frame satisfies the constraints, resulting in an *undetected error* — is 2^{-r}. To compute the probability of undetected error for a CRC code on a binary symmetric channel with crossover probability p, we observe that CRC codes are *linear* codes, which means that an undetected error occurs if and only if the error pattern itself is a valid codeword; this means the probability of undetected error is given by $P_u = \sum_{i=1}^{n} A_i p^i (1-p)^{n-i}$, where A_i is the number of valid CRC codewords containing i ones and $n - i$ zeros (i.e., the number of CRC codewords of *Hamming weight* i). Unfortunately, computing the A_i value (known as the code's *weight enumerator*) is difficult for most large codes; however, it has been shown [3] that there exist (n, k) binary codes with an undetected error probability upper bounded by $2^{(n-k)}[1 - (1-p)^n]$.

CRC codes are used in a broad array of communication applications; for instance, a 7-bit CRC is used to detect errors in the compressed speech found in the second-generation TDMA-based digital cellular standard

Table 1. Properties of Three Standard CRC Codes

Code	$g(x)$	$n_{\max}$	$d_{\min}$
CRC-12	$x^{12}+x^{11}+x^3+x^2+x+1$	2,047	4
CRC-ANSI	$x^{16}+x^{15}+x^2+1$	32,767	4
CRC-CCITT	$x^{16}+x^{12}+x^5+1$	32,767	4

[5]. Table 1 shows the generator polynomials for three different CRC codes that have been selected as international standards for use in communication networks. In each case, the degree of $g(x)$ is equal to the number of check bits per frame required to implement the code. Also included in the table is the code's minimum distance ($d_{\min}$) — that is, the smallest number of bits in which two different valid codewords can differ — and the maximum blocklength ($n_{\max}$) that should be used; for instance, CRC-12 should be used with frames no longer than 2047 bits, meaning frames with $2047 - 12 = 2035$ nonparity bits.

3. PROTOCOLS FOR AUTOMATIC REPEAT REQUESTS

Section 2 described how redundancy can be added at the transmitter and used at the receiver to detect frames that have been corrupted during transmission. "Pure" ARQ protocols require that every frame deemed corrupt at the receiver be retransmitted. ("Hybrid" ARQ schemes use a mix of forward error control and retransmission to ensure data integrity; hybrid schemes are described in Section 5.)

This section describes three protocols for alerting the transmitter that retransmission of a frame is necessary. Each protocol is based on *acknowledgments* passed from the receiver to the transmitter: a *positive acknowledgment* ("ACK") to indicate that an error-free frame has been received and a *negative acknowledgment* ("NAK") to indicate that a corrupt frame has been detected. (Obviously, since an acknowledgment is also subject to corruption, the ACK or NAK is also often protected with a CRC code.)

Some ARQ protocols also have a provision for *timeouts*; when a timeout strategy is used, then if a frame has not been acknowledged (either positively or negatively) after a specified amount of time, the transmitter acts as if it had been acknowledged negatively. Timeouts reduce the effects of packets that get "lost" (rather than corrupted) during transmission.

Finally, we observe that the protocol descriptions that follow all assume that the frames transmitted on the forward link and the acknowledgments transmitted on the return link — at least those that are received — are received in the order in which they were sent. (Note this does not preclude the possibility that some frames and/or acknowledgments are lost.) Moreover, we are not taking into account the possibility of undetected errors on the feedback channel; for discussion and analysis of these issues, the reader is referred to Refs. 3 and 4.

3.1. Stop-and-Wait ARQ

sThe simplest ARQ protocol is the *stop-and-wait* strategy (see Fig. 2), which requires that the transmitter, on sending a frame, wait until it has been acknowledged before the next frame is transmitted.

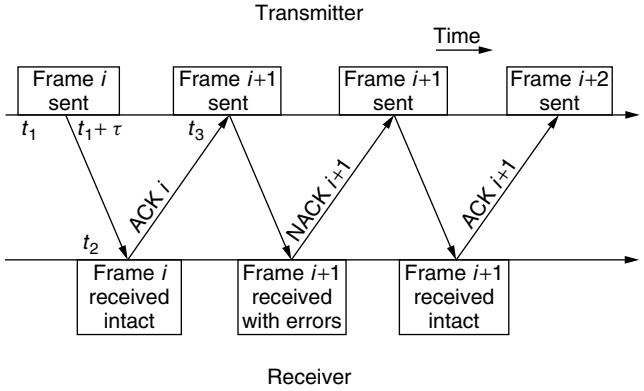


Figure 2. Timeline of stop-and-wait strategy.

The transmitter begins sending frame i at time t_1, and it is received beginning at time $t_2 > t_1$; the receiver then sends an acknowledgment (either positive or negative) through the feedback channel that is received at the transmitter beginning at time t_3. At that point, the transmitter either resends frame i (if a NAK was received) or sends frame $i+1$ (if an ACK was received). This procedure continues in a "pingpong" fashion, with the transmitter and the receiver taking turns sending frames and acknowledgments, respectively. To avoid potential confusion caused by lost frames, a sequence number is often included in the header; in a similar fashion, to avoid confusion caused by lost acknowledgments, the ACK or NAK often will be designed to include the sequence number of the next frame expected by the receiver — and so the ACK of frame i is equivalent to a request for frame $i+1$, while a NAK of frame i is equivalent to a request for frame i. These sequence numbers are incremented modulo some positive integer with each new frame; indeed, simply incrementing the sequence numbers modulo 2 — specifically, identifying the even-numbered and odd-numbered frames — is adequate for stop-and-wait ARQ.

The advantage of the stop-and-wait strategy lies in its simplicity. It does not require a full-duplex channel (i.e., it does not require simultaneous transmission in each direction). Moreover, compared with more sophisticated approaches, stop-and-wait requires minimal storage and processing at both the transmitter and the receiver.

The cost of this simplicity is *throughput* — the efficiency with which information is reliably delivered over the forward channel. Because most communication channels *are* in fact full-duplex and the stop-and-wait strategy does not exploit this capability, the result is a transmitter sitting idle, waiting for acknowledgments, when it could (in principle) be formatting and transmitting additional frames. This is seen in terms of the "idle time" between time $t_1 + \tau$ and time t_3 in Fig. 2.

3.2. Go-Back-*N* ARQ

The go-back-N protocol is designed to address the inherent inefficiency of stop-and-wait. In a communication link implementing the go-back-N strategy, the transmitter

does not have to wait for a request before sending another frame. Rather, the transmitter is permitted to send all of the frames in a "window" maintained at the transmitter. When the first (earliest) frame in the window is ACKed, the transmitter "slides" the window one position, dropping the acknowledged frame and adding a new frame that it then transmits; conversely, if a NAK is received for that first frame in the window, the transmitter "backs up" and resends all the frames in the window, beginning with the corrupt frame.

More specifically, suppose that the transmitter has received an ACK for frame i — equivalently, a request for frame $i+1$. Then, under the go-back-N protocol, the transmitter may send frames $i+1$ through $i+N$ without receiving another ACK; however, the transmitter *must* receive an ACK for frame $i+1$ before transmitting frame $i+N+1$.

Go-back-N is often referred to as a *sliding-window protocol* because the transmitter maintains a sliding window of "active" (or "outstanding") frames — a window of up to N frames that have been transmitted but have not yet been acknowledged. It should be clear that the Stop-and-Wait strategy is simply a go-back-1 strategy.

In execution, the go-back-N protocol is very simple. As long as the receiver judges its frames to be error-free, it sends ACKs to the transmitter, advancing the sliding window. If a corrupt frame is detected, then the resulting NAK tells the transmitter to back up to the corrupt frame and begin retransmitting from that point. Because the transmitter is not permitted to transmit more than N frames beyond what has been ACKed, it will never be required to back up more than N frames. [For example, in a go-back-4 system, the transmitter, having received a positive acknowledgment of (say) frame 10, is free to transmit frames 11–14, but it cannot transmit frame 15 without a positive acknowledgment of frame 11. So, in a worst case, if frame 11 is corrupted, the transmitter will have to back itself up and follow its transmission of frame 14 with a (re)transmission of frame 11.]

Figure 3 illustrates this concept. Frame $i+1$ is corrupted, so as soon as the transmitter receives a NAK for frame $i+1$, it initiates the corresponding backup. Note that, as drawn in Fig. 3, the transmitter had sent only two additional frames (frames $i+2$ and $i+3$) before it received the NAK for frame $i+1$; therefore, this figure describes an action that could have taken place in a go-back-N system for any $N \geq 3$. Moreover, as indicated in Fig. 3, the receiver discards those additional frames sent between the original (corrupt) frame and its retransmitted replica; whether those discarded frames were corrupted during transmission has no effect on the action taken by the receiver.

The appropriate value of N is determined by a number of issues, including the propagation delay of the system and the length of the frames. Obviously, if a "full window" of N frames are transmitted well before the first frame in the window is acknowledged, then the transmitter would sit idle until it gets that acknowledgment — the same kind of inefficiency that makes stop-and-wait (also known as "go-back-1") unattractive. At the opposite extreme, there's no point in making N so large that the first frame in the window is *always* acknowledged before the window reaches its full complement of N frames.

As in the stop-and-wait protocol, frame *sequence numbers* are required in the headers on the forward channel to avoid confusion caused by a lost frame, while frame *request numbers* are designed into the feedback transmission to avoid confusion caused by a lost ACK or NAK. In the stop-and-wait protocol, it was claimed that these counters need only be maintained modulo 2 (i.e., only a single bit needs to be used, indicating whether the transmitted/requested frame has an even or odd frame number). In a go-back-N protocol, the transmitter and receiver must be able to distinguish between all the frames in the window; therefore, the counters must be maintained modulo M for some integer $M > N$. This means that at least $[\log_2(N+1)]$ bits must be dedicated for sequence numbers in each header on the forward link, and the same number of bits must be dedicated for request numbers in each ACK/NAK on the return link.

The go-back-N protocol offers a reasonable balance between efficient use of the channel and reasonably simple implementation.

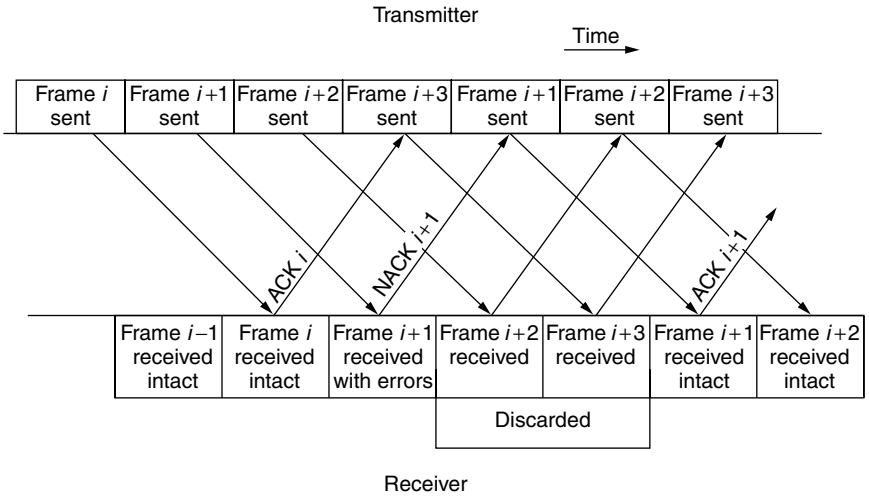


Figure 3. Timeline of go-back-N strategy.

3.3. Selective Repeat ARQ

While the go-back-N protocol clearly offers efficiency advantages relative to stop-and-wait, it is still not as efficient as it could be; when a frame is corrupted under the go-back-N protocol, all the frames that were transmitted *after* the bad frame was sent and *before* it was negatively acknowledged must be retransmitted as well — as many as $N-1$ frames in addition to the corrupt frame. These discarded frames may have, in fact, been received error-free; they are discarded only because they followed too closely on the heels of a corrupt frame.

Selective-Repeat ARQ is designed to overcome this inefficiency. In the Select-Repeat protocol, the transmitter does *not* retransmit the contents of the entire window when a NAK is received; rather, the transmitter re-sends *only* the frame "selected" as corrupt by the receiver.

Selective-repeat is typically implemented as a sliding-window protocol, in the same way as go-back-N. This means that when frame i is ACKed, frames $i+1$ through $i+N$ (but *not* frame $i+N+1$) may be transmitted before frame $i+1$ is positively acknowledged. Once frame $i+1$ is ACKed, all the "oldest" frames in the window that have been positively acknowledged are considered complete and are shifted out of the window; the same number of new frames are shifted into the window and are transmitted.

This process is illustrated in Fig. 4. As in Fig. 3, frame i is received intact and frame $i+1$ is corrupted; however, in the selective-repeat strategy, frames $i+2$ and $i+3$ — the two frames sent *after* frame $i+1$ was sent but *before* it was negatively acknowledged — are *not* discarded at the receiver (and resent by the transmitter) but instead are accepted, assuming they are not corrupt.

As with stop-and-wait and go-back-N strategies, the sequence numbers (on the forward path) and the request numbers (on the return path) used in selective-repeat ARQ are maintained modulo M. For correct operation of selective-repeat, this modulus must satisfy $M \geq 2N$, where N is the window size.

The chief advantage of selective-repeat ARQ is its efficiency; if each frame is corrupted with probability p, then selective-repeat can deliver good frames at a rate approaching $1-p$ good frames per transmitted frame — the highest possible rate. The chief disadvantage of selective-repeat is the memory and logic needed to store the arriving frames and reassemble them in the appropriate order.

4. PERFORMANCE ANALYSIS OF ARQ SYSTEMS

In analyzing the three ARQ protocols described in Section 3, we focus on two criteria — the *reliability* of the information delivered to the receiver and the *efficiency* with which that information is delivered. These two criteria are formalized in the notions of *frame error rate* and *throughput*, respectively.

4.1. Reliability Analysis of an ARQ System

When a frame is transmitted, three things can happen en route:

- It can be delivered error-free to the receiver; henceforth we assume that each packet is delivered error-free with probability P_c [e.g., if each bit in an n-bit frame is "flipped" with independent probability p, then $P_c = (1-p)^n$].
- It can be corrupted by noise in such a way that the receiver can detect the presence of the error via a CRC code parity violation. Let the probability of such an event be denoted P_d. Clearly, P_d depends on the nature of the channel and the particular CRC code under consideration.
- It can be corrupted by noise in such a way that the receiver *cannot* detect the presence of the error via a CRC code parity violation — that is, an *undetected error* occurs. Let the probability of such an event be denoted P_u. Once again, P_u depends on the particular channel and the particular CRC code being used. In Section 2 it was shown that, for a binary symmetric channel with crossover probability p, the undetected error probability is $P_u = \sum_{i=1}^{n} A_i p^i (1-p)^{n-i}$, where A_i is the number of valid codewords containing i ones and $n-i$ zeros.

As noted above, these three events are the only possibilities when a frame is transmitted, and so $P_c + P_d + P_u = 1$.

The *frame error rate* is the probability that a frame is accepted by the receiver as correct when it is, in

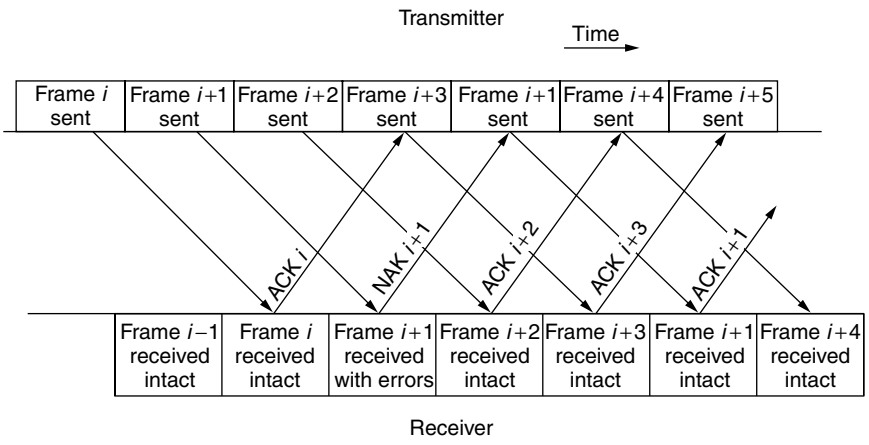


Figure 4. Timeline of selective-repeat strategy.

fact, corrupt. In an ARQ system, it is possible that a frame may be erroneously accepted the first time it is transmitted; alternatively, it could be rejected the first time but erroneously accepted the second time. Or it could be rejected twice before being erroneously accepted the third time. Continuing in this vein, and assuming there is no limit imposed on the number of retransmissions permitted, we arrive at the following expression for frame error rate:

$$\begin{aligned}\text{FER} &= P_u + P_d P_u + P_d^2 P_u + P_d^3 P_u \cdots \\ &= P_u \sum_{i=0}^{\infty} P_d^i \\ &= \frac{P_u}{1 - P_d}\end{aligned}$$

As a simple example, suppose that you were to use the [15,11] Hamming code mentioned in Section 2 to generate $r = 4$ bits of parity for $k = 11$ bits of data, requiring a frame length of $n = 15$. For this code it can be shown [4] that $A_0 = 1, A_1 = 0$, and A_i for $i = 2, 3, \ldots, 15$ can be computed via the recrusion

$$(i+1)A_{i+1} + A_i + (16 - i)A_{i-1} = \binom{15}{i}$$

With this "weight enumeration" it is simple to calculate P_u and P_d [and so $\text{FER} = P_u/(1 - P_d)$] for a binary symmetric channel; this is plotted for crossover probabilities ranging from $p = 10^{-1}$ to $p = 10^{-4}$ as seen in Fig. 5.

4.2. Efficiency Analysis of an ARQ System

The reliability analysis presented above is valid for *any* of the three ARQ protocols described in Section 3. This is because the *reliability* of any ARQ protocol depends only on the ability of the decoder to detect errors that occur during transmission — that is, it depends only on the error-detecting capability of the underlying CRC code and is independent of how any corrupted frames are subsequently retransmitted.

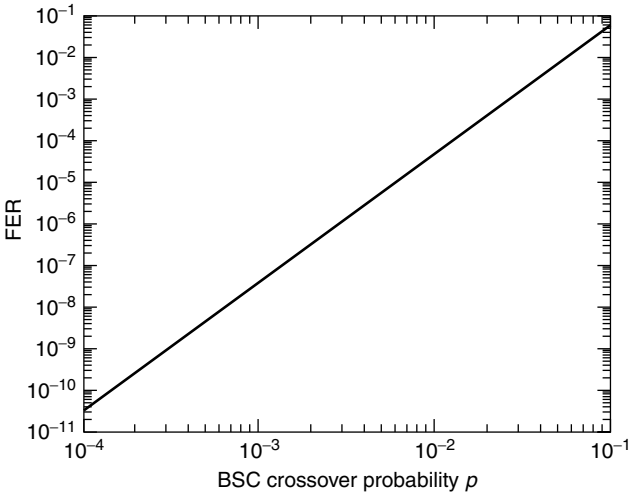


Figure 5. Frame error rate (FER) of an ARQ protocol over a binary symmetric channel assuming that the [15,11] Hamming code is used for error detection.

In contrast, the *efficiency* of an ARQ scheme is strongly tied to the mechanism by which the transmitter and receiver coordinate the retransmission of corrupt frames.

The figure of merit used to assess the efficiency of an ARQ protocol is called its *throughput*, and we denote it by η. The throughput of an ARQ scheme is defined as the ratio of the average number of information bits accepted by the receiver per unit time to the number of bits that can be transmitted over the channel per unit time. Under this definition, the throughput of a scheme employing an (n, k) CRC code (i.e., a code used in an n-bit frame, with k information bits and $n - k$ parity bits) can never be greater than $k/n < 1$.

Conceptually, selective-repeat ARQ is the simplest protocol to analyze in terms of its throughput. Each time a frame is transmitted it has probability $P \triangleq P_c + P_u$ of being accepted by the receiver; therefore, if we let T_{SR} be a random variable equal to the number of frame transmissions required until a particular frame is accepted, then $\Pr(T_{\text{SR}} = 1) = P, \Pr(T_{\text{SR}} = 2) = P(1 - P), \Pr(T_{\text{SR}} = 3) = P(1 - P)^2$, and so on; thus T_{SR} has a geometric distribution with mean $E[T_{\text{SR}}] = 1/P$. If we assume that each n-bit frame contains r bits of redundancy and $k = n - r$ bits of "data" (i.e., nonparity bits), then, under this protocol, the receiver accepts k information bits in the amount of time (on average) it takes to transmit n/P over the channel, so the throughput for selective-repeat ARQ is

$$\eta_{\text{SR}} = \frac{k}{n} P$$

The efficiency of the go-back-N protocol can be analyzed in a similar fashion; the main difference lies in the fact that, when a frame is rejected under the go-back-N protocol, the entire window of outstanding frames must be retransmitted. Once again, let $P = P_c + P_u$ be the probability that a frame is accepted, and let T_{GBN} be a random variable equal to the number of frames that must be transmitted until a specified frame is accepted; then if we assume the worst case (i.e., that a "full window" of N frames must be retransmitted every time a frame is rejected), then

$$\begin{aligned}E[T_{\text{GBN}}] &= 1 \cdot P + (N+1) \cdot (1 - P) \cdot P \\ &\quad + (2N+1) \cdot (1 - P)^2 \cdot P + \cdots \\ &= 1 + \frac{N(1 - P)}{P}\end{aligned}$$

So under the go-back-N protocol the receiver accepts k information bits in the amount of time (on average) it takes to transmit $n(1 + N(1 - P)/P)$ bits over the channel, so the throughput is

$$\eta_{\text{GBN}} = \left(\frac{k}{n}\right)\left(\frac{P}{P + N(1 - P)}\right)$$

Finally, consider stop-and-wait ARQ. In our analysis of the other two protocols, we began by asking a question: "On average, how many frames must be transmitted

before a particular frame is accepted?" In those analyses, we could measure the delay in "frame units" because it was implicitly assumed that transmission was continuous in the forward channel — that is, the frames were transmitted one after the other, with no idle time in between. (By "frame unit" we mean the amount of time required to transmit a single n-bit frame.)

For stop-and-wait, there *is* idle time between subsequent transmissions, and so to use the same approach we shall measure that idle time in terms of the number of frames that *could have been transmitted*. Specifically, let β denote the amount of time the transmitter is idle between frames divided by the amount of time it takes the transmitter to send one frame; in effect, β is the idle time measured in "frame units." [Referring to Fig. 2, $\beta = (t_3 - (t_1 + \tau))/\tau$.] So, if a frame is accepted the first time, the delay is $1 + \beta$ frame units; if a frame is rejected the first time but accepted the second, the delay is $2(1 + \beta)$ frame units. And so, if we let T_{SW} be a random variable representing the delay (in frame units) incurred from the time a frame is first sent until it accepted, we have

$$\begin{aligned} E[T_{\mathrm{SW}}] &= (1+\beta)\cdot P + 2\cdot(1+\beta)\cdot(1-P)\cdot P \\ &\quad + 3\cdot(1+\beta)\cdot(1-P)^2\cdot P + \cdots \\ &= \frac{1+\beta}{P} \end{aligned}$$

As a result, the throughput for the stop-and-wait protocol is given by

$$\eta_{\mathrm{SW}} = \left(\frac{k}{n}\right)\left(\frac{P}{1+\beta}\right)$$

The parameter β is determined by the round-trip propagation delay, the signaling rate of the system (i.e., how many bits per second are transmitted), and the length of the frames and the acknowledgments, as well as the processing delay at the transmitter and receiver.

To compare the three protocols, consider once again a system employing the [15,11] Hamming code for error detection. For the go-back-N protocol we shall set $N = 4$ and for the stop-and-wait protocol we set $\beta = 3$; these are comparable configurations, since each corresponds to a delay between the time the transmission of a frame is completed and the time the frame is acknowledged equal to 3 times the duration of one frame.

Figure 6 shows the throughput of the three protocols as a function of the crossover probability of a binary symmetric channel. We observe that, for low crossover probabilities, the throughput of both selective-repeat and go-back-N approach the rate of the error detection code: $\frac{11}{15} \approx 0.733$. By comparison, because the transmitter in stop-and-wait is sitting idle 75% of the time, its maximum throughput is $0.25 \times \frac{11}{15} \approx 0.183$.

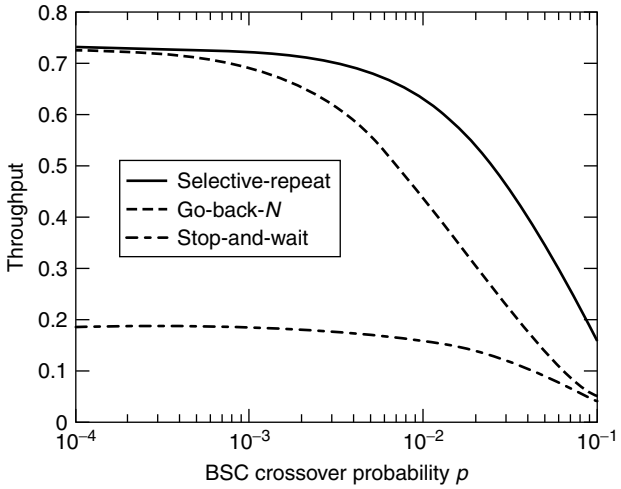


Figure 6. Throughputs of three different ARQ protocols over a binary symmetric channel assuming that the [15,11] Hamming code is used for error detection. The go-back-N protocol uses a window length of $N = 4$, and the stop-and-wait protocol assumes a parameter value of $\beta = 3$.

5. HYBRID ARQ SYSTEMS

It was claimed at the beginning of this article that error control schemes could be broadly classified into two categories: *forward error control* (FEC), in which redundancy added at the transmitter is used to recover the transmitted message even in the face of corruption, and *automatic repeat request* (ARQ), in which redundancy is used only to *detect* corruption and trigger a retransmission.

A hybrid ARQ system [6,7] employs elements of both FEC and ARQ. Hybrid ARQ schemes can be classified as either *type 1 hybrid ARQ* or *type 2 hybrid ARQ*.

The simplest implementation of a type-1 hybrid ARQ system uses two codes: a high-rate error *detection* code and a (typically) lower-rate error *correction* code. Data are first encoded using the error detection code and then encoded again using the error correction code; as a result, the frame trailer contains redundant bits from both codes. At the receiver, the decoder first attempts to reconstruct the frame using the redundancy from the error correction code; it then passes the result to the error detection unit, which checks to see if the error correction decoder was successful. If the error detection unit observes a parity violation, a retransmission is triggered; otherwise, the frame is accepted.

This implementation of a type-1 hybrid ARQ protocol basically uses the "inner code" (i.e., the error correction code) to create a more reliable "virtual" digital channel and then implements a conventional ARQ protocol over that more reliable channel.

Type-1 hybrid-ARQ can also be implemented with a single powerful code. To see how this can be done, observe that an (n, k) block code can be used to simultaneously correct t errors and detect $u > t$ errors provided its minimum distance $d_{\min}$ satisfies $d_{\min} \geq t + u + 1$. (For instance, a code with minimum distance $d_{\min} = 7$ can correct all occurrences of $t = 2$ errors while simultaneously *detecting* — i.e., neither correcting *nor* miscorrecting — three or four errors.) A decoder using such a code within the context of a type-1 hybrid ARQ would correct all error patterns affecting t or fewer bits; if more than t (but no more than u) errors occur during transmission, then a retransmission would be triggered.

Type-2 hybrid ARQ systems operate on the principle of *incremental redundancy*. While a number of type-2 hybrid protocols have been formulated, they all share in common the characteristic that, when a frame is initially rejected, what is retransmitted is *not* the same frame that was originally sent but rather a frame whose "payload" consists of parity bits based on that original frame. In this way, the original corrupted frame — kept in a buffer at the receiver — can be augmented with the newly-received parity bits to form a codeword from a longer, more powerful FEC code.

As one implementation of this approach, let C_0 be a high-rate (n, k) error-detecting code and let C_1 be a more powerful $(2n, n)$ error-correcting code. At the transmitter, k bits of data are encoded using C_0, and the resulting codeword — call it $\mathbf{c}$ — is transmitted; however, before it is sent, $\mathbf{c}$ is itself encoded with the code C_1 to form n parity bits $\mathbf{p}$ [i.e., $\mathbf{c} * \mathbf{p}$ forms a $2n$-bit codeword from C_1; here, "$*$" denotes concatenation]. These n bits of $\mathbf{p}$ are stored at the transmitter while $\mathbf{c}$ is sent. At the receiver, the received version of $\mathbf{c}$ is checked for errors and accepted if none are found; however, if $\mathbf{c}$ has been corrupted, then what is transmitted in response to the NAK is *not* the n bits of $\mathbf{c}$ but rather the n bits of $\mathbf{p}$. As a result, the decoder has a noisy version of $\mathbf{c} * \mathbf{p}$, which can be corrected using a decoder for the powerful code C_1. (If correction is impossible, the process can begin again, with retransmission of $\mathbf{c}$.)

The benefits of type-2 hybrid ARQ systems lie in the fact that they do not "waste" redundancy when the channel is good. In any reasonably designed communication link, frame errors are relatively rare, so to use a powerful error correcting code with every frame (as type-1 hybrid systems do) may be overkill; by only sending error-correcting redundancy when needed, the type-2 hybrid ARQ system makes more efficient use of the channel. As usual, the tradeoff is in buffering and logic complexity.

BIOGRAPHY

Thomas E. Fuja received his undergraduate education at the University of Michigan, obtaining a B.S.E.E. and a B.S.Comp.E. in 1981. He attended graduate school at Cornell University, Ithaca, New York, where he received an M.Eng. in 1983 and a Ph.D. in 1987, both in electrical engineering. Dr. Fuja was on the faculty of the Department of Electrical Engineering at the University of Maryland in College Park, Maryland, from 1987 until 1998; in addition, he served as the program director for communications research at the U.S. National Science Foundation, Arlington, Virginia, in 1997 and 1998. Since 1998, Fuja has been on the faculty of the University of Notre Dame in South Bend, Indiana, where he is a professor of electrical engineering. His research interests lie in channel coding and information theory — most recently focusing on issues related to wireless communications and on the interface between compression and error control. Dr. Fuja has been very active in the IEEE Information Theory Society; in 2002, he served as that organization's president.

BIBLIOGRAPHY

1. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1992.
2. A. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall PTR, Upper Saddle River, NJ, 1996.
3. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
4. S. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
5. T. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall PTR, Upper Saddle River, NJ, 1996.
6. S. Lin, D. Costello, Jr., and M. Miller, Automatic repeat request error control schemes, *IEEE Commun. Mag.* **22**: 5–16 (Dec. 1984).
7. L. Rasmussen and S. Wicker, The performance of type-I trellis coded hybrid-ARQ protocols over AWGN and slowly fading channels, *IEEE Trans. Inform. Theory* **40**(2): 418–428 (March 1994).

B

BANDWIDTH REDUCTION TECHNIQUES FOR VIDEO SERVICES

NELSON L. S. DA FONSECA
State University of Campinas
Institute of Computing
Campinas, Brazil

1. INTRODUCTION

Video-on-demand (VoD) is a client–server application which allows users to watch movies stored in remote servers. It is a critical technology for home entertainment, professional communication, and digital video libraries, to name a few uses. In recent years, it has come to be regarded as the main video application for future broadband multimedia networks.

There are two major kinds of interactiveness in VoD services, varying according to the type of video distribution and scheduling policies implemented by the server. In true video-on-demand, all requests are granted immediately if resources are available, whereas in near-video-on-demand, users may need to wait a certain time before their requests are granted.

The videostream, which constitutes the flow of bytes corresponding to the frames composing a movie, can be delivered only when adequate server resources (I/O bandwidth) as well as network bandwidth (video channels) have been reserved for this purpose. With current technology, there is a mismatch of roughly a whole order of magnitude between I/O data rates and network data rates, so techniques to utilize network bandwidth more efficiently are need.

A single videostream requires a considerable amount of bandwidth, ranging from 6 Mbps (megabits per second) for MPEG-2 streams to 20 Mbps for HDTV streams. Considering a potential population of several million viewing households, the network bandwidth demand would be beyond the present-day network capacity, approaching the order of Tbps. Such demands prevent the deployment of VoD on a large scale. Consequently, numerous techniques to reduce these demands have been developed. One of these techniques consists of the placement of several servers throughout the network so that they will be closer to the user, thus, diminishing the need for allocation of channels along long network paths [1]. In itself, however, this technique is not able to reduce the demand sufficiently, and further reductions are necessary [2]. Some sort of sharing of videostreams, either by a group of viewers (multicast) or by all users (broadcast), must be used. The choice of routing technique depends on the degree of interactiveness required. On one hand, multicasting permits true video-on-demand, but at the expense of high processing overhead costs. On the other hand, broadcasting involves lower overhead costs, but can provide only near-video-on-demand. Various techniques based on multicasting and on broadcasting have been proposed, with greater or lesser success in bandwidth reduction. These techniques will be explained here: piggybacking, patching, and batching, all of which utilize multicasting with different degrees of emphasis in the time that users will have to wait to watch a desired program, and periodic broadcasting, which is more appropriate for high request rates.

2. PIGGYBACKING

Piggybacking is based on the fact that viewers do not perceive an alteration in the display rate when it is no more than 5% of the nominal rate. In a VoD server with piggybacking, a request for viewing a video is immediately granted if a channel is available. However, when a new channel is allocated, and another exhibition of the same video is in progress for another user, the display rate of the original showing is slowed down, while the display rate of the recently admitted request is increased. When the faster streams catch up with the slower one, the two are merged, and, as a consequence, one of the video channels is released (Fig. 1) [3]. The aim of piggybacking policies is to reduce the total number of displayed frames for a set of streams, which is equivalent to reducing the bandwidth required to display this set of movies.

The change in display rate is affected by dynamically compressing or expanding the video being displayed through the insertion of additional frames produced by the interpolation of proceeding and succeeding frames, or through the deletion of frames, with neighboring frames altered to reduce the abruptness of the change. On-line alteration of the display rate has been shown to be a difficult task, however, especially for MPEG-encoded movies. Another solution would be to store copies of movies corresponding to different display rates, although this greatly increases storage demands. For a 100-min-long MPEG-2-encoded movie, for example, the storage demand is about 4.5 GB (gigabytes) per movie.

Merging two streams is possible only if the difference between the frames being displayed by the two streams is such that the faster stream will exhibit the same frame displayed by the slower one at some time prior to the end of the exhibition of the latter. In other words, a merge can occur only if the faster stream is able to catch up with the slower one. The maximum catchup window is the difference in frames for which merging is feasible. This window depends on the discrepancy of the display rate of two streams and is given by $((S_{\max} - S_{\min}) \times L)/S_{\max}$, where $S_{\min}$ and $S_{\max}$ denote the minimum and the maximum display rates, respectively and L, the movie length in frames.

Each piggybacking policy defines its own catchup window, which is not necessarily the maximum possible, because there is a tradeoff between window size and the reduction in number of frames displayed. If the window is

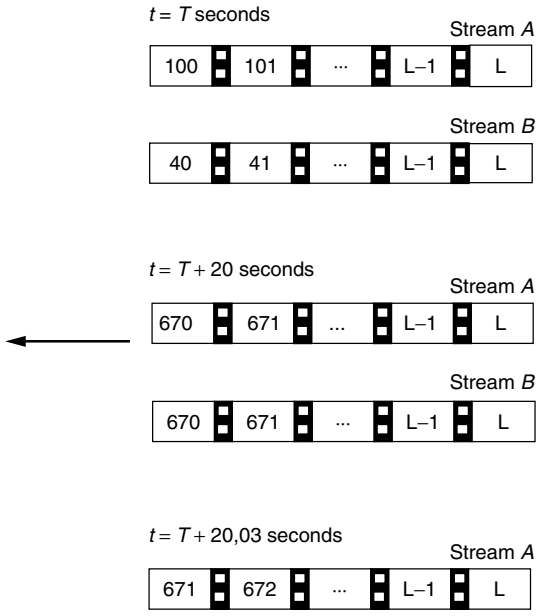


Figure 1. Stream A moves at a rate $S_{\max} = 31.5$ frames per second (FPS) while stream B moves at a rate $S_{\min} = 28.5$ FPS. At time T, the 100th frame of stream A and is the 40th frame of stream B are displayed. Twenty seconds later, at time $T + 20$ s, both streams displays the 670th frame. At $T + 20$ s a channel is released and only one stream is displayed for both viewers at a rate $S_n = 30$ frames.

large, a larger number of streams can be merged, but it is likely that the merging of these streams will occur too near the end of the movie, thus accruing less benefit. When the window is small, merges occur nearer the beginning of the movie, but fewer movies tend to be merged.

Various piggybacking policies are available. One of the simplest policies is the *odd–even* one, which tries to merge each pair of streams that arrived consecutively at the server [3]. On arrival, a stream is moving at the nominal rate, but it is then set to $S_{\min}$ if there is no stream ahead of it, or if there is one already moving at $S_{\max}$ Once a stream crosses the maximum catchup window at a rate of $S_{\min}$ and there is no stream behind it moving at $S_{\max}$, the display rate is reset to the nominal one. If there is a stream behind it moving at $S_{\max}$, however, the display rate remains at $S_{\min}$ to give the latter movie a chance to catch up.

The next policy to be discussed is the "greedy policy," which tries to merge as many videostreams as possible [3]. The greedy policy defines its catch up window based on the current frame. Whenever a merge occurs or the catchup window is crossed, a new window is computed. When a new video request is granted, the video rate is set as in the odd–even policy. Once a merge occurs, the display rate is set to the nominal one if there is no other stream in the new catchup window. Otherwise, the rate of the stream in front is set to $S_{\min}$ and the rate of the stream itself is set to $S_{\max}$. When it crosses the first catchup window, if there is another stream in front of it at the nominal rate, then the rate of this front stream is reduced to $S_{\min}$ and the rate of the following one is set to $S_{\max}$. Otherwise, therefore, if there is no stream ahead of it, the stream moves at the nominal rate.

Simple merging is another policy for merging groups; it guarantees that all streams in a group are eventually merged [3]. One stream is chosen to be the leader of the group, and all streams within this leader's maximum catchup window participate in the merging. As in the odd–even policy, on the arrival of a new stream, the rate of this new stream is set to $S_{\min}$, if there is no other stream within the maximum catchup window moving at $S_{\min}$. Otherwise, it moves at $S_{\max}$. Whenever it leaves the maximum catchup window, the rate is tuned to the nominal rate.

The *generalized simple merging* policy differs from the simple merging policy in that it computes window size to minimize the number of frames displayed by assuming that requests for video exhibition arrive according to a Poisson process with rate λ [4]. This window size is given by

$$W = -\frac{S_{\min}}{\lambda} + \sqrt{\left(\frac{S_{\min}}{\lambda}\right)^2 + 2\frac{LS_{\min}(S_{\max} - S_{\min})}{\lambda S_{\max}}}$$

The "snapshot policy" applies the generalized simple merging policy to a group of streams to form various merges in a given window [4]. At the end of this window, however, some streams may not have been merged. The rate of display of these streams must be adjusted so that further merges are possible. To do this, the stream positions will be represented by a binary tree, constructed in a bottom–up fashion using a dynamic programming solution. In such a tree, the streams are located at the leaf nodes, and merges are represented by interior nodes. The root node represents the final merge of all the streams. In this way, rates can be assigned so that all the remaining streams can merge.

The S2 policy reapplies the snapshot policy a second time to gain further reduction in the number of frames displayed [5]. It also constructs a merging tree, but in a top–down fashion, using a divide-and-conquer strategy, which allows the reduction of the computational complexity of the construction of the binary merging tree. It has been shown, however, that a third application of the snapshot policy does not bring additional benefits.

When comparing these policies, one can see that some of the policies are more efficient than others. The odd–even policy alone presents the least efficient performance, followed by the simple merging policy and by the generalized simple merging policy. The greedy policy outperforms the generalized simple merging policy, but the snapshot policy always produces even higher savings in relation to the number of frames displayed. S2, however, is the most effective of these piggybacking policies, especially since the savings on the number of frames displayed increases with the length of the movie and with the arrival rate of requests. For instance, S2 produces 9% savings over the number of frames displayed by the snapshot policy for a 30-min movie and an interarrival time of 500 s, but 90% savings for a 4 hour movie and an interarrival time of 15 s.

3. PATCHING

Patching policies exploit the client's buffer to reduce the waiting time required to watch a movie. They also

exploit the simultaneous reception of multiple channels. If a request to watch a movie is issued before a certain threshold time after the initiation of another exhibition of the same movie, the client joins the multicast group in the ongoing transmission with the current frames being stored in the client's buffer, namely, in the client's *settop box* (STB) buffer, until that client has seen the initial part recuperated from the video server. Once this has been seen, the client views the frames from the STB buffer (Fig. 2). Various patching policies have been proposed since the late 1990s.

In greedy patching, if the client's buffer is smaller than the initial part of the movie, then a new channel must be allocated for the individual display of the entire movie, with the viewer joining the multicast group of the ongoing transmission only for the storage of the final part to store in the buffer [6].

Grace patching also allocates a new channel for a transmission whenever the client's buffer is smaller than the initial clip, but this will involve multicast routing [6].

The *periodic buffer reuse with thresholding* (PBR) policy exploits the client's buffer to the maximum [7]. In this policy, the sequence of frames shown is drawn alternatively from the server and the ongoing transmission which was initiated the shortest time before the viewer's request. The buffer is initially filled with frames from the ongoing transmission while the client is watching frames from the server. The client then watches frames being drained from the buffer while it is being renewed by the server.

In the *greedy buffer reuse* (GBR) policy, parts of the video can be obtained from multiple ongoing transmissions, not only from that which was initiated the closest in time to the viewer's request [7]. It schedules the receipt of a sequence of frames as late as possible in order to utilize the client's buffer more efficiently. The buffer size and the current frame position of ongoing transmissions determine from where the sequence of frames will be fetched.

As with piggybacking, different patching policies result in different advantages and disadvantages. Greedy patching results in less data sharing than grace patching, since the latter increases the chances that a new request joins a multicast group. By making efficient use of the client's buffer, both PBR and GBR can provide greater bandwidth savings than grace patching, however, especially with large buffers, with GBR providing more savings than PBR. Under high loads, PBR may demand the display of 60% more frames per viewer than GBR. Such savings do not imply the involvement of a large number of channels, as no more than three channels are typically used during the exhibition of a single movie.

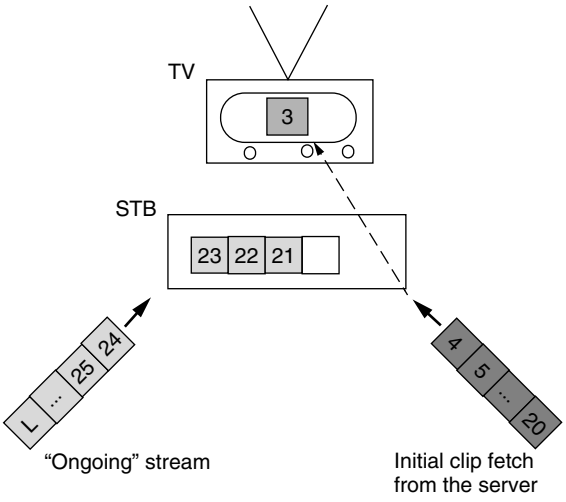


Figure 2. The initial part of the movie is fetched from the server while the frames from an ongoing transmission are stored in the STB for exhibition after the initial part is displayed.

4. BATCHING

In a VoD server with batching, requests are not granted as soon as they arrive. They are delayed so that several requests for the same film within a certain interval can be collected [8]. A single videostream is then allocated to the whole batch of requests (Fig. 3). If, on one hand, batching increases the server throughput (i.e., the rate of granted requests), on the other hand, users may not be willing to wait for long periods of time, and may cancel their requests (reneging).

Given that in batching users share the entire sequence of frames of the whole video, policies are compared by the number of users admitted into the system as well as by the number of users who renege.

Batching policies can be classified according to users' reneging behavior. Policies that do not consider reneging are first-come first-served (FCFS), maximum queue length (MQL), and maximum factor queue length (MFQL).

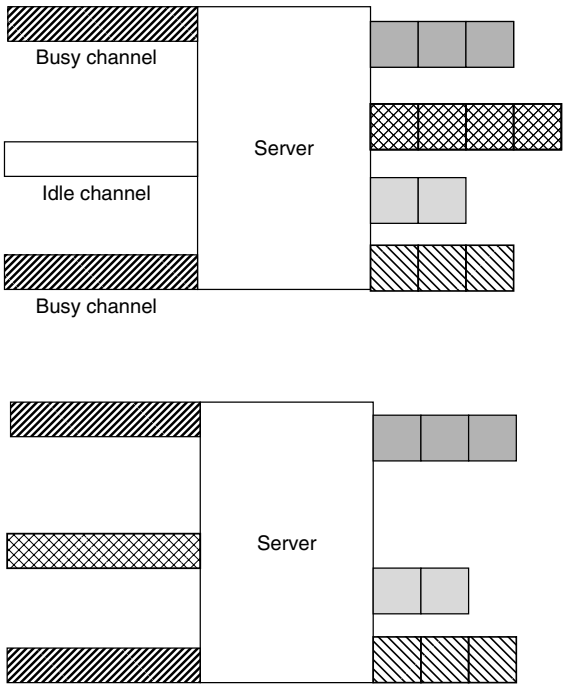


Figure 3. The figure in the top shows a server with 4 batches of requests waiting to be served and all channels busy except one, which was just released. The idle channel is then assigned to the largest batch.

FCFS serves requests according to their arrival order; it gives poor throughput, but treats all video requests equally. MQL allocates a video channel to the longest queue as soon as it becomes available; it produces considerably higher throughput than does FCFS. MFQL, a variation of the MQL policy, assigns a weighting factor to each queue, and allocates an available channel to the queue with the highest weighted length [9].

The second group of policies takes reneging into account. There are two batching schemes in this group: the Max_Batch scheme and the Min_Idle scheme. In the Min_Idle scheme, videos are classified as either "hot" or "cold" according to their popularity. Only hot videos are subject to batching. Moreover, hot videos have higher priority than cold videos for channel allocation. Two sets are defined: H and C. Hot videos, which have at least one request pending that exceeds a certain delay threshold, belong to the set H. Cold videos belong to the set C. Whenever a channel becomes available, a video in H is scheduled, either according to the longest queue criterion (IMQ) or to the highest expected number of losses (IML) criterion. If H is empty, a video in C is scheduled, regardless of how long the requests have been in queue. A cold video may migrate to the set H if any of its pending requests exceeds a certain threshold.

In the Max_Batch scheme, whenever a channel becomes available, a decision is made as to which queue (batch of requests) the channel should be allocated. A channel is allocated to a queue if and only if at least one of the enqueued requests exceeds a certain delay threshold. Two Max_Batch policies have been defined: the Max_Batch maximum queue length (BMQ) and the Max_Batch with minimum loss (BML) policies. BMQ allocates the available channel to the longest queue, whereas BML allocates the channel to the queue with the highest expected number of losses up to the next time a channel will become available.

The *look-ahead-maximize-batch* (LAMB) policy is a variant of the Max_Batch scheme. LAMB considers all videos in a server eligible for batching. Any channel will be allocated on demand, according to the expected number of losses [10].

LAMB considers a queue eligible for channel assignment only if one of its head-of-the-line (HoL) user is about to exceed his/her delay tolerance. In other words, a channel is allocated to a queue if and only if the HoL user is about to leave the system without being served. Moreover, instead of minimizing the number of losses expected by the next scheduling point, as is done in BML and IML, LAMB minimizes the losses in a batching window. This batching window is lower bounded by the current scheduling time and upper bounded by the most distant reneging time of a pending request.

LAMB maximizes the number of users admitted by considering all potential losses in the batching window. Whenever a user is about to leave the system, a decision is made about whether to allocate a channel to his/her queue. If the number of queues is less than the number of available channels, a channel is automatically allocated to the about-to-leave user's queue. Otherwise, an analysis of the implications of such an allocation, at the current scheduling time is made in relation to the admission of an overall larger number of users during the batching window. In other words, it is verified whether allocating a channel at the current scheduling time will cause a shortage of channels, which are associated with longer queues, at future scheduling times. Note that whenever a channel is allocated to a queue, all users in that queue are served at once. Otherwise, only the about-to-leave user is lost.

To maximize the number of users admitted during the batching window, it is necessary to determine when a channel should be allocated to a queue by considering all the information available at the current scheduling time, including the reneging time of all users, and the time when each channel will be released at the end of an exhibition.

Batching gives much better results than piggybacking. Allocating channels on demand to single users, even if temporarily, may result in a future shortage of channels, thus leading to a long-term rejection of a high number of users. Nevertheless, piggybacking produces fair systems, because it does not provide differentiated services for hot movies. If, on one hand, piggybacking makes it unnecessary for users to wait for a channel when it is available, on the other hand, batching significantly increases the server throughput, which is of paramount importance when deploying VoD services on a large scale.

LAMB overperforms all other existing batching policy, taking into consideration the number of admitted users (throughput) and the reneging probability, i.e., percentage of users who give up watching a movie. This trend is more striking for high loads (high arrival rate of requests) and in servers with a large number of video channels. For instance, LAMB admits the greatest number of users, 20% more than those admitted by MBQ, and the lowest reneging probabilities, 0.1 lower than MQL.

Although both batching and piggybacking furnish a single videostream to a group of viewers, they represent a clear trade-off between minimizing the delay to serve a request (piggybacking), and maximizing the server throughput (batching). One approach to enhance the throughput provided by batching is to merge streams in exhibition, which increases the number of channels available for new batches — in other words, to use a combination of batching and piggybacking. A system with both batching and piggybacking admits 20% more users than a system with batching only and produces reneging probabilities 0.05 lower than when only batching is used.

5. PERIODIC BROADCASTING

The top 10 or 20 most popular movies will be responsible for most of the requests for viewing. One possibility for coupling with bandwidth demands generated by these requests is to exhibit them from time to time (Fig. 4), taking a proactive approach rather than a reactive one (on demand), as done in techniques based on multicast [11].

Conventional broadcasting allocates a certain number of channels for showing a given video, staggering the beginning of each session evenly across the channels. The major drawback of such broadcasting is the number of channels needed to provide a low waiting time [12].

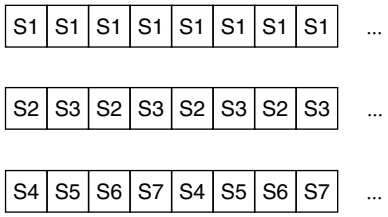


Figure 4. In periodic broadcasting, the video is divided into segments that are periodically broadcast in different channels.

One option is *periodic broadcasting*, which divides the video into a series of segments and broadcasts these periodically on dedicated channels. While the user is watching one segment, the following one is being transmitted so that it arrives just in time to assure continuous playback.

Three basic type of broadcasting protocols have been proposed. The first involves protocols dividing the video into increasing-size segments that are broadcast in channels of the same capacity. In the *pyramid broadcasting* (PB) scheme, for example, the segment size follows a geometric series, where the nth segment of each video is transmitted sequentially in each channel [13]. Although a low waiting time can be assured, high transmission rates are required, and this implies in high I/O bandwidth demand of and large STB buffers.

Permutation-based pyramid broadcasting tries to overcome the major drawbacks of the classical pyramid broadcasting protocol by dividing each channel into a specific number of subchannels. The substreams are staggered with each other to perform the same kind of timing as in PB, although the transmission rate is lower. Since the same geometrical series is used to divide the video, the client buffer cannot be reduced because the last segment is quite large [14].

Skyscraper broadcasting (SB) is similar except that the size of the segments is determined by a recursive function which generates the following sequence [1, 2, 2, 5, 5, 12, 12, 25, 25, 52, 52, ...], where the first segment is the unit size for all following segments. Each segment is broadcast periodically in a specific channel, and the client will need to download from at most two streams at the same time. The client's buffer size must still be maintained to accommodate the final film segment [15].

Fast broadcasting is similar to other pyramid schemes in that equal capacity channels are used, but the segments broadcast are of equal size, where group of segments are transmitted together. Groups of $2i$ contiguous segments are transmitted in the ith channel. One advantage of this protocol is that no buffer is required at the client [16].

In the second family of protocols, videos are divided into equal sized segments transmitted on channels of decreasing capacity. The first of these to be considered here is *harmonic broadcasting* (HB), in which the ith segment is divided into i subsegments. The first segment is repeatedly broadcast in the first channel and contiguous subsegments of the other segments are periodically transmitted on the channel dedicated to these segments; each channel has a broadcasting capacity inversely proportional to the sequential order of the segment. The main idea of HB is that when the client is ready to receive the ith segment he will already have received the $i - 1$ subsegment, and the last subsegment will be received while the client retrieves the first segments from the buffer. The bandwidth demand of HB increases harmonically as a function of the number of segments of the video, and the storage demand is about 40% of the entire video. However, HB does not always deliver data on time [17].

All the variants of HB overcome this timing problem. In *caution harmonic broadcasting* (CHB), for example, the first segment is transmitted as in HB, whereas the second and the third are transmitted in a second channel, while the remaining segments are transmitted in other channels with a capacity inversely proportional to the segment order minus one [18].

Another option for harmonic broadcasting is *quasiharmonic broadcasting* (QHB). Again, the first segment is transmitted as for HB and CHB. The remaining segments are divided in such a way that the ith segment is divided into $im - 1$ subsegments, where m is a positive ith integer, but the subsegments are not transmitted in order. The first $i - 1$ subsegments of the segment are transmitted at the end of the segment slot and the remaining $i(m - 1)$ subsegments are transmitted according to a specific rule, so that the client always has the first $i - 1$ subsegments stored in his buffer. Although QHB demands more bandwidth than do HB demands, the overhead tends to be compensated for an increase in the number of subsegments [18].

The third group consists of protocols which are a hybrid of pyramid-based and harmonic protocols. Like harmonic protocols, *pagoda broadcasting* partitions the video into equal-sized segments, but unlike them, these segments are broadcast at the same rate, although with different periodicity [19]. The effect of channel dedication is achieved by time-division multiplexing. The main advantage of this protocol is that it avoids the problems due to low transmission rates, although the determination of proper segment-to-channel mapping and periodicity is critical. The *new pagoda broadcasting* protocol uses a more sophisticated segment-to-stream mapping than that used by pagoda to further reduce the bandwidth demands [20].

6. CONCLUSIONS

Video-on-demand has been considered "the" video application for the future broadband multimedia networks. Considerable effort has been invested in making it efficient, since the huge amounts of bandwidth needed to serve a large population preclude the deployment of VoD on a large scale. Moreover, the implementation of VoD services has to be competitive with traditional video rental and pay-per-view. In this article, various proactive and reactive bandwidth reduction techniques have been described, techniques that can be used jointly with the distribution of additional servers throughout the network. Their effectiveness depends on viewers' behavior, since at high demand rates schemes based on multicasting tend to transmit videostreams with the same periodicity as schemes based on broadcasting. On the

other hand, periodic broadcasting wastes bandwidth if the rate of request is not high. The most appropriate way for implementing video-on-demand can be determined only when the dimensions of use have been established, and will be understood only when services are available at large. Moreover, providing interactiveness (VCR capability) implies additional costs in terms of careful dimensioning and signaling [21,22].

BIOGRAPHY

Nelson Fonseca received his Electrical Engineer (1984) and M.Sc. in Computer Science (1987) degrees from The Pontifical Catholic University at Rio de Janeiro, Brazil, and the M.Sc. (1993) and Ph.D. (1994) degrees in Computer Engineering from The University of Southern California (Los Angeles). Since 1995 he has been affiliated to the Institute of Computing of the State University of Campinas, Brazil, where is currently an Associate Professor.

He is the recipient of Elsevier Editor of the Year 2000, the 1994 USC International Book Award, and the Brazilian Computing Society First Thesis and Dissertations Award. Mr. Fonseca is listed in Marqui's *Who's Who in the World* and *Who's Who in Science and Engineering*.

He served as Editor-in-Chief for the *IEEE Global Communications Newsletter* from 1999 to 2002. He is an Editor for *Computer Networks*, an Editor for the *IEEE Transactions on Multimedia*, an Associate Technical Editor for the *IEEE Communications Magazine*, and an Editor for the *Brazilian Journal on Telecommunications*.

BIBLIOGRAPHY

1. J. P. Nussbaummer, B. V. Patel, F. Schaffa, and J. P. G. Sterbenz, Networking requirements for interactive video on demand, *IEEE J. Select. Areas Commun.* 779–787 (June 1995).
2. N. L. S. Fonseca, C. M. R. Franco, and F. Schaffa, Network design for the provision of distributed home theatre, *Proc. IEEE Int. Conf. Communications*, 1997, pp. 816–821.
3. L. Golubchik, J. C. S. Lui, and R. Muntz, Adaptive piggybacking: A novel technique for data sharing in video-on-demand storage servers, *Multimedia Syst.* **4**(3): 140–155 (1996).
4. C. C. Aggarwal, J. Wolf, and Philip S. Yu, On optimal piggybacking merging policies for video-on-demand systems, *Proc. ACM Sigmetrics* **24**: 200–209 (1996).
5. R. A. Façanha, N. L. S. Fonseca, and P. J. Rezende, The S2 piggybacking policy, *Multimedia Tools Appl.* **8**(3): 371–383 (May 1999).
6. K. Hua, Y. Cai, and S. Sheu, Patching: A multicast technique for true video-on-demand services, *Proc. 6th ACM Int. Multimedia Conf.*, 1998, pp. 191–200.
7. S. Sen, L. Gao, J. Rexford, and D. Towsley, Optimal patching schemes for efficient multimedia streaming, *Proc. IEEE NOSSDAV*, 1999.
8. H. Shachnai and P. S. Yu, Exploring wait tolerance in effective batching for video-on-demand scheduling, *Multimedia Syst. J.* **6**(6): 382–394 (Dec. 1998).
9. A. Dan, D. Sitaram, and P. Shahabuddin, Dynamic batching policies for an on-demand video server, *Multimedia Syst.* **4**: 112–121 (1996).
10. N. L. S. Fonseca and R. A. Façanha, The look-ahead-maximize-batch batching policy, *IEEE Trans. Multimedia* **4**(1): 1–7 (2002).
11. A. Hu, Video-on-demand broadcasting protocols: A comprehensive study, *Proc. IEEE InfoCOM*, 2001.
12. K. Almeroth and M. Ammar, The use of multicast delivery to provide a scalable and interactive video-on-demand service, *IEEE J. Select. Areas Commun.* **14**(5): 1110–1122 (Aug. 1996).
13. S. Viswanathan and T. Imielinski, Pyramid Broadcasting for video on demand service, *IEEE Multimedia Computing and Networking Conf.*, San Jose, CA, 1995, Vol. 2417, pp. 66–77.
14. C. Aggarwal, J. Wolf, and P. Yu, A permutation-based pyramid broadcasting scheme for video-on-demand systems, *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, 1996.
15. K. Hua and S. Sheu, Skyscraper broadcasting: A new broadcasting scheme for metropolitan video-on-demand systems, *Proc. ACM SIGCOMM'97*, 1997, pp. 89–100.
16. L. Juhn and L. Tseng, Fast data broadcasting and receiving scheme for popular video service, *IEEE Trans. Broadcast.* **44**(1): 100–105 (March 1998).
17. L. Juhn and L. Tseng, Harmonic broadcasting for video-on-demand service, *IEEE Trans. Broadcast.* **43**(3): 268–271 (Sept. 1997).
18. J. Paris, S. Carter, and D. Long, Efficient broadcasting protocols for video on demand, *Proc. 6th Int. Symp. Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, July 1998, pp. 127–132.
19. J. Paris, S. Carter, and D. Long, A hybrid broadcasting protocol for video on demand, *Proc. 1999 Multimedia Computing and Networking Conf.*, 1999, pp. 317–326.
20. J. Paris, A simple low-bandwidth broadcasting protocol for video-on-demand, *Proc. 8th Int. Conf. Computer Communications and Networks* (IC3N'99), 1999, pp. 118–123.
21. N. L. S. Fonseca and H. K. Rubinszjtein, Dimensioning the capacity of interactive video server, *Proc. Int. Teletraffic Congress 17*, 2001, pp. 383–394.
22. J. K. Dey-Sircar, J. D. Salehi, J. F. Kurose, and D. Towsley, Providing VCR capabilities in large-scale video servers, *Proc. 2nd ACM Int. Conf. Multimedia*, 1994, pp. 25–36.

BCH CODES — BINARY*

Arnold M. Michelson
Allen H. Levesque
Marlborough, Massachusetts

1. INTRODUCTION

Bose–Chaudhuri–Hocquenghem (BCH) codes are a broad class of cyclic block codes used for detection and correction

* Preparation of this article supported in part by the Raytheon Corporation.

of transmission errors in digital communications systems. This article describes binary BCH codes while the succeeding article treats *nonbinary BCH codes*. These codes were originally described in two papers, the first by A. Hocquenghem in 1959 [1] and the second by R. C. Bose and D. K. Ray-Chaudhuri in 1960 [2]. It would therefore be more accurate to say "HBR-C codes," but the commonly used abbreviation is BCH.

Error control coding is a field finding wide application in modern digital communications. Described in simple terms, error control coding involves adding redundancy to transmitted data to provide a means for detecting and correcting errors that inevitably occur in any real communication process. Coding can be used to provide a desired level of accuracy in data transmitted over a noisy communication channel and delivered to a user. There are, however, other ways to achieve accurate transmission of data.

For example, an alternative to the use of coding is to provide sufficient signal energy to ensure that uncoded information is delivered with the required accuracy. The energy needed might be achieved by setting signal power to a sufficiently high level or, if power limitations prevail, by using some form of diversity transmission and reception. However, error control coding may provide the required accuracy with less energy than uncoded operation and can be the economically preferred solution in spite of an increase in system complexity. Cost savings through the use of coding can be dramatic when very high accuracy is needed and power is expensive. Furthermore, in some applications, the saving in signal power is accompanied by important reductions in size and weight of the communication equipment.

To describe BCH codes, it is first necessary to provide some background in finite fields, extension fields, and polynomials defined on finite fields. This is done in Section 2. Binary BCH codes are described as cyclic codes in Section 3, and the design of generator polynomials, encoders, and decoders are also covered. In the succeeding article, nonbinary BCH codes are treated. Reed–Solomon (RS) codes are the most widely used class of nonbinary BCH codes, and the design, encoding, and decoding of RS codes are treated there.

2. MATHEMATICAL BACKGROUND

This section describes the essential mathematics needed for understanding the design and implementation of BCH codes. BCH codes are cyclic codes and are conveniently represented as polynomials with coefficients in a *finite field*. We begin with a discussion of finite fields.

2.1. Finite Fields

A *field* is a set of elements with two operations defined, addition (+) and multiplication (·). Two other operations, subtraction and division, are implied by the existence of inverse elements under the defining operations. Stated more completely, the elements in a field F, taken together with the operations + and ·, must satisfy the following conditions:

1. F is closed under the two operations, that is, the sum or product of any two elements in F is also in F.
2. For each operation, the associative and commutative laws of ordinary arithmetic hold, so that for any elements u, v, and w in F:

$$(u+v)+w = u+(v+w)$$

$$u+v = v+u$$

$$(u\cdot v)\cdot w = u\cdot(v\cdot w)$$

$$u\cdot v = v\cdot u$$

3. Connecting the two operations, the distributive law of ordinary arithmetic holds, so that

$$u\cdot(v+w) = u\cdot v + u\cdot w$$

for any u, v, and w in F.

4. F contains a unique additive identity element 0 and a unique multiplicative identity, different from 0 and written as 1, such that

$$u+0 = u$$

$$u\cdot 1 = u$$

for any element u in F. The two identity elements are the minimum elements that any field must contain. We call these two elements zero and unity.

5. Each element u in the field has a unique additive inverse, denoted by $-u$, such that

$$u+(-u) = 0$$

and, for $u \neq 0$, a unique multiplicative inverse, denoted by u^{-1}, such that

$$u\cdot u^{-1} = 1$$

From these observations, the inverse operations subtraction (−) and division (÷) are defined by

$$u - v = u + (-v), \quad \text{any } u, v \text{ in } F$$

$$u \div v = u\cdot(v^{-1}), \quad v \neq 0$$

where $-v$ and v^{-1} are the additive and multiplicative inverses, respectively, of v.

Thus a field provides four elementary operations and the familiar rules of ordinary arithmetic. Common examples of fields are the set of real numbers and the set of rational numbers under ordinary addition and multiplication. The set of real numbers equal to or greater than zero does not constitute a field under the rules of ordinary arithmetic, since the set does not include additive inverses for nonzero numbers. Similarly, the set of integers under ordinary arithmetic is not a field, since integers other than 1 do not have multiplicative inverses in the set.

The number of elements in a *field*, called the *order* of the field, may be finite or infinite, but we consider only

fields having a finite number of elements. A field having a finite number of elements is called a *finite field* and is denoted by GF(q), where q is the number of elements in the field. The notation is related to the designation *Galois field*, which is used interchangeably with "finite field" in the literature. A finite field GF(p^m) exists for any p^m, where p is a prime and m is an integer. The simplest example of a finite field is a *prime field*, GF(p), consisting of the set of all integers modulo p, where p is a prime number greater than 1 and the addition and multiplication operations are addition and multiplication modulo p. The simplest prime field is GF(2), which contains only the zero and unity elements 0 and 1. As another example, the addition and multiplication tables for GF(5) are shown below.

GF(5) Addition

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

GF(5) Multiplication

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

2.2. The Primitive Element

An important property of finite fields is that every finite field GF(q) contains at least one *primitive element*, called α, which has the property that the $q-1$ powers of α are the $q-1$ nonzero elements of the field. This means that the nonzero field elements can be represented as $\alpha, \alpha^2, \ldots, \alpha^{q-1}$.

If we take an arbitrary nonzero element β in the field and raise it to successive powers, we eventually arrive at some exponent e such that $\beta^e = 1$. For an arbitrary β in the field, the smallest positive integer e such that $\beta^e = 1$ is called the *order of the element*. (This is not to be confused with the *order of the field*, defined as the number of elements in the field, which is equal to q in the present discussion.) In the generation of the nonzero elements of GF(q) as powers of a primitive element α, we always find that $\alpha^{q-1} = \alpha^0 = 1$, but no smaller power of α equals 1, so that the order of a primitive element is $q-1$. In general, the various elements of the field can have different orders, but there is a theorem (due to Lagrange), stating that the order of an arbitrary element must be either $q-1$ or a divisor of $q-1$.

For example, consider the prime field GF(5), which consists of the integers modulo 5. Since $q-1=4$, we anticipate that the orders of various elements can be 1, 2, or 4. The element 1 has order 1. Taking successive powers of 2, we find $2^1 = 2, 2^2 = 4, 2^3 = 8 = 3 \bmod 5$, $2^4 = 6 = 1 \bmod 5$, and thus 2 has order 4. One also finds that 3 has order 4, and 4 has order 2. Therefore 2 and 3 are primitive elements of GF(5), while 1 and 4 are *nonprimitive* elements.

Since all the nonzero field elements can be expressed as the first $q-1$ powers of a primitive element α, we note that we can represent the field elements in terms of their exponents. The exponents are in effect *logarithms to the base* α. As in ordinary arithmetic, the logarithm of zero is undefined, although for convenience, the notation $0 = \alpha^{-\infty}$ is often used. Below we show the logarithm tables for GF(5) formed with $\alpha = 2$ and $\alpha = 3$:

β	$\log_2 \beta$	$\log_3 \beta$
0	$-\infty$	$-\infty$
1	0	0
2	1	3
3	3	1
4	2	2

Just as in ordinary arithmetic, multiplication of field elements can be done by adding logarithms. For example, in GF(5), using $\alpha = 2$, we can multiply 2 times 4 by adding the logarithms $(1+2=3)$ and looking up the resulting element (3 mod 5) in an antilogarithm table. In coding implementations, finite-field multiplications are often done with logarithm and antilogarithm tables.

2.3. Vectors of Field Elements and Polynomials Defined on Finite Fields

To encode and decode BCH codes, we need to find an algebraic system for doing calculations with vectors, or m-tuples, of finite-field elements and a representation for field elements that is convenient for implementation in a digital machine. First note that we can enumerate all the m-tuples of elements in a field GF(q), q^m in number, and note that they in fact constitute an m-dimensional vector space over GF(q). Thus we can add and subtract vectors, using vector (element-by-element) addition and subtraction in GF(q), and the result in every case is another vector in the vector space. However, we shall also want to do multiplication and division of vectors. To accomplish this, we associate each vector with a polynomial having coefficients corresponding to the elements in the vector. For example, the set of four 2-tuples on GF(2) can be represented by 0, 1, x, and $x+1$, corresponding to 00, 01, 10, and 11, respectively. Clearly, we can do term-by-term addition of the polynomials just as we would add the vectors. All we have done is replace the set of all 2-tuples defined on GF(2) with the set of all degree-1 polynomials defined on GF(2).

Just as we have closure with addition of vectors, we must also have closure under multiplication. In fact, if we can find a way to multiply the polynomials that conforms to all the properties of multiplication in a finite field, we will have constructed a finite field with q^m elements. First, we want the product of any two polynomials in the set to be another polynomial in the set (closure). This is no problem if the product is a polynomial of degree $(m-1)$ or less. But, what do we do with a polynomial product of degree m or greater? Clearly, we can reduce the product by taking its remainder with respect to a fixed polynomial of degree m. The remainder will always be of degree $m-1$ or less, and closure is achieved. However, we need to know what sort of polynomial to use in this reduction so that the other properties of a finite field are assured.

We can gain some insight into this question by observing that the product of any two nonzero field elements must be nonzero. For example, let two nonzero elements α^i and α^j be represented by $a(x)$ and $b(x)$, respectively, each of degree $m-1$ or less. Then, assuming

a *reduction polynomial* $p(x)$ of degree m, we can write the product $\alpha^i\alpha^j$ as

$$\alpha^i\alpha^j = a(x)b(x) \bmod p(x)$$

Now, let us set this product equal to zero and see what type of reduction polynomial would allow this to happen. That is, we write

$$a(x)b(x) \bmod p(x) = 0$$

or equivalently

$$a(x)b(x) = c(x)p(x) \tag{1}$$

which says that the left-hand side of Eq. (1) must be evenly divisible by $p(x)$. Now, if $p(x)$ is *factorable*, that is, expressible as the product of two or more polynomials of degree $m-1$ or less, there may well be polynomials $a(x)b(x)$ that are evenly divisible by $p(x)$. However, if $p(x)$ is chosen to be a degree-m *irreducible polynomial* (a polynomial that cannot be factored), then $p(x)$ must be a factor of either $a(x)$ or $b(x)$. We can readily see that neither factoring is possible, since the polynomials $a(x)$ and $b(x)$ are each of degree $m-1$ or less and $p(x)$ is of degree m. We, therefore, conclude that if $p(x)$ is chosen to be an irreducible polynomial of degree m, the equality in Eq. (1) cannot be satisfied unless $a(x)$ or $b(x)$ equals zero, in which case $c(x) = 0$. By similar arguments, we could show that the requirements for uniqueness of the products $a(x)b(x)$, and hence the uniqueness of the inverse for each polynomial, again results in choosing the reduction polynomial $p(x)$ to be an irreducible degree-m polynomial in GF(q).

Returning now to the simple example of the four 2-tuples defined on GF(2), we can use $p(x) = x^2 + x + 1$, since $x^2 + x + 1$ cannot be factored into any lower-degree polynomials on GF(2). (The only candidates for factors are x and $x+1$, and it is easily verified that none of the products of these two polynomials equals $x^2 + x + 1$.) With $p(x)$ chosen, we can now write the multiplication table for the degree-1 binary polynomials as follows:

$\cdot$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	$x+1$	1
$x+1$	0	$x+1$	1	x

We see from the multiplication table that each nonzero polynomial has a unique multiplicative inverse, x being the inverse of $x+1$ and vice versa, while 1 is its own inverse, as always. Thus, we have defined a representation of a finite field with four elements, which we denote by GF(4).

We complete this example by describing GF(4) in terms of a primitive element. We can test for a primitive element simply by taking a nonzero element other than 1 and raising it to successive powers until we find its order. For example, testing x, we have $x^1 = x, x^2 = x+1, x^3 = x^2 + x = (x+1) + x = 1$, where calculation of x^2 and x^3 required reduction modulo $x^2 + x + 1$. We see therefore that x has order $e = 3$, and since $q - 1 = 3$, x is a primitive element. It can be seen that the polynomial $x+1$ is primitive as well. Having found two primitive elements in GF(4), we can now use either one to generate a list of the nonzero field elements as powers of α. This is shown here with a table of field elements for each primitive element. The table also shows a representation of the four elements in GF(4) that is convenient for implementation in a digital machine. With each polynomial, we associate a binary 2-tuple, for example, $0 = 00$ and $x + 1 = 11$. Addition of the digital representations of field elements is then conveniently implemented with the exclusive OR operation:

REPRESENTATIONS FOR FIELD ELEMENTS

$\alpha = x$					$\alpha = x+1$				
$\alpha^{-\infty}$	=	0	=	00	$\alpha^{-\infty}$	=	0	=	00
α^0	=	1	=	01	α^0	=	1	=	01
α^1	=	x	=	10	α^1	=	$x+1$	=	11
α^2	=	$x+1$	=	11	α^2	=	x	=	10

Thus, we have a complete representation for elements in GF(4) and a consistent set of operations for addition and multiplication of elements. For multiplication, the appropriate logarithm tables can be used. In the next section, we generalize these results and define somewhat more formally the properties of fields constructed with m-tuples of field elements.

2.4. Extension Fields and Primitive Polynomials

In general, a finite field GF(p^m) exists for any number p^m, where p is a prime and m is a positive integer. For $m = 1$, we have the prime number fields GF(p). The fields GF(p^m) for $m > 1$ are commonly called *prime-power fields*, where p is the *characteristic of the field*. That is, p is the smallest integer such that

$$\sum_{i=1}^{p} \alpha^0 = 0$$

where α^0 is the multiplicative identity element. For fields of characteristic 2, each element is its own additive inverse and a minus sign is unnecessary.

The relationship between GF(p) and GF(p^m) is such that GF(p) is a *subfield* of GF(p^m); that is, the elements of GF(p) are a subset of the elements in GF(p^m), the subset itself having all the properties of a finite field. Equivalently, GF(p^m) is called an *extension field*, or simply an *extension*, of GF(p).

The procedure followed previously for constructing GF(4) from GF(2) serves as an example of how one constructs an extension field from a subfield. The procedure generalizes in a straightforward way to any extension field GF(p^m). That is, we represent elements in GF(p^m) as the p^m polynomials of degree $m-1$ or less with coefficients in GF(p). Polynomials are added by adding coefficients of corresponding powers of x, addition being done in GF(p). To define multiplication, a degree-m irreducible polynomial over GF(p) is selected and a primitive element α for GF(p^m) is found. Then the polynomials corresponding to the $p^m - 1$ distinct powers of

α are constructed. We see that the irreducible polynomial $p(x)$ provides the key link between the addition and multiplication tables and thus fixes the structure that allows us to define the two arithmetic operations and their inverses in a consistent way. Thus, we can say that the set of all polynomials in GF(p) reduced with respect to a degree-m irreducible polynomial over GF(p) forms the field GF(p^m). The role of the irreducible polynomial is seen to be directly analogous to the use of a prime number p to define the finite field GF(p).

Note that a polynomial $p(x)$ of degree m with coefficients in GF(p) is said to be irreducible if it is not divisible by any polynomial with coefficients in GF(p) of degree less than m and greater than zero. For example, consider the polynomial $p(x) = x^3 + x + 1$ having degree 3 and coefficients in GF(2). We can quickly convince ourselves that $x^3 + x + 1$ is not factorable in GF(2), as follows. If it is factorable, it must have at least one factor of degree 1. Of course x is not a factor of $p(x)$, since the lowest-order term in $p(x)$ is $x^0 = 1$. Thus, the only candidate is $x + 1$, but if this were a factor, then $x = 1$ would be a root of $p(x)$. It is easily verified that this is not the case, since $p(x)$ has an odd number of terms, and therefore, $p(x)$ evaluated at $x = 1$ sums to 1 mod 2. Therefore, $p(x) = x^3 + x + 1$ is irreducible in GF(2).

Therefore, we are able to generate the 3-tuples representing elements of GF(2^3) simply by listing all 2^3 polynomials of the form $a(x) = a_2x^2 + a_1x + a_0$ and taking each 3-tuple as the vector of coefficients a_2, a_1, a_0. It is convenient to list the polynomials $a(x)$ in a sequence that automatically provides a consecutive ordering by logarithms. This can be done here by using the polynomial $a(x) = x$ as the primitive element, multiplying repeatedly by x, and reducing the result modulo $p(x)$. This is shown in Table 1. We see from the table that by forming successive powers of x, reduced modulo $x^3 + x + 1$, we obtain all the polynomials defining the nonzero elements of GF(2^3). In order for the procedure to generate the full list of 3-tuples, it is necessary that x be a primitive element, which is clearly the case in this example.

Although $p(x) = x^3 + x + 1$ is an irreducible binary polynomial and consequently has no roots in GF(2), it does have roots defined in an extension field. In fact, it is a simple matter to find one of its roots, since from Table 1, we see that we could as easily have generated the table using powers of α letting $\alpha = x$ and $\alpha^3 = \alpha + 1$, and therefore α is a root of $p(x)$. An irreducible polynomial having a primitive element as a root is called a *primitive irreducible polynomial* or simply a *primitive polynomial*. While an irreducible polynomial with coefficients in GF(p) has no roots in GF(p), it has roots in the extension field GF(p^m). In fact, the degree-m polynomial $p(x)$ must have exactly m roots in the extension field GF(p^m).

Table 1. A Representation of GF(2^3) Generated from $x^3 + x + 1$

Zero and Powers of x		Polynomials Over GF(2)		Vectors Over GF(2)
0	=	0	=	000
x^0	=	1	=	001
x^1	=	x	=	010
x^2	=	x^2	=	100
x^3	=	$x + 1$	=	011
x^4	=	$x^2 + x$	=	110
x^5	=	$x^2 + x + 1$	=	111
x^6	=	$x^2 + 1$	=	101

It is important to note that not all irreducible polynomials are primitive and both can be used to generate a representation for a finite field. However, it is convenient to use a primitive polynomial since the field elements can be generated as powers of x. As a practical matter, tables of irreducible polynomials with primitive polynomials identified are available in the literature [3].

In summary, to construct a representation of GF(p^m), we go to a table of irreducible polynomials on GF(p), and find a polynomial $p(x)$, preferably primitive, of degree m. We then generate the list of p^m polynomials modulo $p(x)$ and take the vectors of polynomial coefficients as m-tuples representing the elements of GF(p^m). Consistent addition and multiplication tables can then be constructed for GF(p^m). The addition table is formed by adding corresponding elements in m-tuples, modulo p. The multiplication table can be formed by adding exponents of α. The addition and multiplication tables for GF(2^3), formed with the use of Table 1, are shown in Table 2. Note that we constructed the addition and multiplication tables using powers of α although we expressed field elements using polynomials in x in Table 1. However, since $x^3 + x + 1$ is a primitive polynomial, it has α as a root, so that $\alpha^3 + \alpha + 1 = 0$. Thus, Table 1 might as easily have been written with α replacing x, as was observed earlier.

It should be noted that while the multiplication table is most easily constructed by addition of exponents of α, it can also be constructed by multiplying the polynomial representations of two elements and reducing the result modulo $p(x)$. For example, we can use Table 1 to calculate

$$\begin{aligned}\alpha^2\alpha^4 &= x^2(x^2 + x) \bmod x^3 + x + 1\\ &= x^4 + x^3 \bmod x^3 + x + 1\\ &= x^2 + x + x + 1\\ &= x^2 + 1\\ &= \alpha^6\end{aligned}$$

This is analogous to generating the multiplication table for a prime field GF(p) by multiplying integers and reducing the product modulo p.

2.5. Key Properties of Irreducible Polynomials

In Section 2.3, we utilized certain properties of irreducible polynomials to provide a consistent set of rules for performing addition and multiplication in a finite field GF(p^m). It is now necessary to present further details on the properties of these polynomials, which form the basis for describing the structure of cyclic codes. In our discussion of binary codes, we confine attention to fields of characteristic 2, GF(2^m).

Our discussion concentrates on polynomials that are irreducible in GF(2), that is, degree-m binary polynomials

Table 2. Addition and Multiplication Tables for GF(2^3)

$+$	0	1	α	α^2	α^3	α^4	α^5	α^6
0	0	1	α	α^2	α^3	α^4	α^5	α^6
1	1	0	α^3	α^6	α	α^5	α^4	α^2
α	α	α^3	0	α^4	1	α^2	α^6	α^5
α^2	α^2	α^6	α^4	0	α^5	α	α^3	1
α^3	α^3	α	1	α^5	0	α^6	α^2	α^4
α^4	α^4	α^5	α^2	α	α^6	0	1	α^3
α^5	α^5	α^4	α^6	α^3	α^2	1	0	α
α^6	α^6	α^2	α^5	1	α^4	α^3	α	0

$\cdot$	0	1	α	α^2	α^3	α^4	α^5	α^6
0	0	0	0	0	0	0	0	0
1	0	1	α	α^2	α^3	α^4	α^5	α^6
α	0	α	α^2	α^3	α^4	α^5	α^6	1
α^2	0	α^2	α^3	α^4	α^5	α^6	1	α
α^3	0	α^3	α^4	α^5	α^6	1	α	α^2
α^4	0	α^4	α^5	α^6	1	α	α^2	α^3
α^5	0	α^5	α^6	1	α	α^2	α^3	α^4
α^6	0	α^6	1	α	α^2	α^3	α^4	α^5

that have no factors of degree less than m and greater than 0. It has already been stated that every degree-m polynomial $f(x)$ on GF(2) has m roots (as in ordinary arithmetic), and if $f(x)$ is irreducible all m roots are in the extension field GF(2^m). The properties of these roots in extension fields are of central importance in the theory of cyclic codes, and thus we summarize the key points required in the subsequent discussion. For convenience of presentation, certain points made earlier are repeated in this summary.

2.5.1. Properties of Polynomials Defined on Finite Fields

1. Given a polynomial $f(x)$ with coefficients in GF(2), we say that β is a root of $f(x)$ if and only if $f(\beta) = 0$, where β is an element of GF(2), or some extension GF(2^m). The multiplications and additions required for the evaluation of the polynomial can be performed in the consistent arithmetic system GF(2^m), since GF(2) is contained in any of its extensions.
2. Every polynomial of degree m has exactly m roots, some of which may be repeated.
3. For any m, there is at least one degree-m polynomial on GF(2) that is irreducible.
4. If $f(x)$ is a degree-m irreducible polynomial ($m \geq 2$) on GF(2), it has no roots in GF(2), but all its roots lie in some extension of GF(2). If $f(x)$ has a root that is a primitive element of GF(2^m), $f(x)$ is called a *primitive irreducible polynomial*, or simply a *primitive polynomial*. Since it can be shown that all the roots of an irreducible polynomial are of the same order, all the roots of a primitive polynomial are primitive. For any m, there is at least one irreducible polynomial on GF(2) that is primitive.
5. For every element β in an extension field GF(2^m), there is a polynomial on GF(2), called the *minimal polynomial* of β, which is the lowest-degree *monic* (the highest-order term has coefficient 1) polynomial having β as a root. Of course, all polynomials defined on GF(2) are monic. Minimal polynomials, sometimes called *minimum functions*, have an important place in the design of cyclic codes, and we shall have more to say about them.
6. If $f(x)$ is an irreducible degree-m polynomial on GF(2) and has a root β, then $\beta, \beta^2, \beta^4, \beta^8, \ldots, \beta^{2^{m-1}}$ are all the roots of $f(x)$. This is an important property relating to the structure of cyclic codes.

Associated with every element β in an extension field GF(2^m) is its minimal polynomial $m_\beta(x)$ with coefficients in GF(2). There is a minimal polynomial for every element in the field, even if the element lies in GF(2) itself. The important properties of minimal polynomials are summarized as follows.

2.5.2. Properties of Minimal Polynomials. The minimal polynomial $m_\beta(x)$ of any field element β must be irreducible. If this were not the case, one of the factors of $m_\beta(x)$ would have β as a root and would be of lower degree than $m_\beta(x)$ and contradict the definition. In addition

1. The minimal polynomial of β is unique, that is, for every β there is one and only one minimal polynomial of β. However, different elements of GF(2^m) can have the same minimal polynomial. (See property 6 of polynomials defined on finite fields.)
2. For every element in GF(2^m), the degree of the minimal polynomial over GF(2) is at most m.
3. The minimal polynomial of a primitive element of GF(2^m) has degree m and is a primitive polynomial.

Consider again the case of the extension field GF(2^2), which we represent as the polynomials of degree 1 or less, modulo the irreducible polynomial $y^2 + y + 1$. We have

$$\beta_0 = 0$$

$$\beta_1 = 1$$

$$\beta_2 = y$$

$$\beta_3 = y + 1$$

The minimal polynomials of β_0 and β_1 are simply

$$m_{\beta_0}(x) = x \quad \text{and} \quad m_{\beta_1} = x + 1$$

To find the minimal polynomial of $\beta_2 = y$, we use property 6 in Section 2.5, which tells us that the irreducible degree-2 polynomial having β_2 as a root has β_2 and as β_2^2 roots, and no others. Therefore, we can write

$$\begin{aligned} m_{\beta_2}(x) &= (x - y)(x - y^2) \\ &= (x + y)(x + y + 1) \\ &= x^2 + xy + x + yx + y^2 + y \\ &= x^2 + x + 1 \end{aligned}$$

where we use $y^2 = y + 1$ to reduce powers of y greater than unity. Similarly

$$\begin{aligned} m_{\beta_3}(x) &= (x + y + 1)(x + y^2 + 1) \\ &= x^2 + xy^2 + x + yx + y^3 + y + x + y^2 + 1 \\ &= x^2 + x + 1 \end{aligned}$$

Thus, we see that β_2 and β_3 have the same minimal polynomial. (We could have shown this directly by noting that $\beta_2^2 = \beta_3$.) Sets of elements having this property are called *conjugates*.

This example is given only to provide a clearer explanation of the concept of a minimal polynomial. We shall see that, fortunately, it is not necessary to derive minimal polynomials in most cases of binary code design, since they are available in published lists.

We conclude our description of minimal polynomials with an important property of polynomials that have minimal polynomials as factors. Let $\beta_1, \beta_2, \ldots, \beta_L$ be elements in some extension field of GF(2), and let the minimal polynomials of these elements be $m_{\beta_1}(x), m_{\beta_2}(x), \ldots, m_{\beta_L}(x)$. Then the smallest degree monic polynomial with coefficients from GF(2) having $\beta_1, \beta_2, \ldots, \beta_L$ as roots, say $g(x)$, is given by

$$g(x) = \text{LCM}[m_{\beta_1}(x), m_{\beta_2}(x), \ldots, m_{\beta_L}(x)]$$

where LCM denotes the *least common multiple*.

We might well refer to $g(x)$ as the minimal polynomial of the set of elements $\beta_1, \beta_2, \ldots, \beta_L$. If the minimal polynomials of these elements are distinct (recall that different field elements can have the same minimal polynomial), then $g(x)$ is simply

$$g(x) = \prod_{i=1}^{L} m_{\beta_i}(x)$$

3. BINARY BLOCK CODES

BCH codes are block codes that form a subclass of a broad class called *cyclic codes*. These codes have a well-defined algebraic structure that has led to the development of efficient encoding and decoding schemes. BCH codes have proven useful in practical applications because, over certain ranges of code parameters, good performance is achieved, and the encoders and decoders have reasonable complexity. We begin with a brief description of binary block codes.

For a binary block code, the information bits to be transmitted are first grouped into k-bit blocks. An encoding rule is applied that associates r redundant check bits to each k bit information set. The resulting group of $n = k + r$ encoded bits forms an n-bit codeword that is transmitted on the channel. Since the r check bits represent overhead in the transmission, we say that the code rate $R = k/n$. The block length of the code is n and the notation (n, k) is used to represent a code with block length n containing k information bits and $r = n - k$ check bits per block. Clearly, an (n, k) code comprises the set of codewords representing all possible k-bit information sets.

There are several ways to represent the codewords in an (n, k) code. For example, codewords may be represented as n-bit vectors or as polynomials with degree up to $n - 1$. For binary codes, the vector components or the polynomial coefficients are the elements of GF(2), namely, 0 and 1. An important property of block codes is linearity. We say that a code is linear if sums of codewords are codewords. Codewords are added by forming bit-by-bit (bitwise) modulo 2 sums of the corresponding vector positions or polynomial coefficients.

An important attribute of a block code is its minimum distance d. The *minimum distance* of a binary block code is the smallest number of codeword bit positions in which an arbitrary pair of codewords differ. A code with minimum distance d provides the capability to correct all error patterns containing $t \leq (d - 1)/2$ errors for d odd, and $t \leq (d/2) - 1$ errors for d even.

Binary block codes may also be systematic or nonsystematic. A systematic code has the feature that the k information bits appear unaltered in each codeword. With nonsystematic codes, they do not. Systematic codes are generally preferred, but we consider both.

3.1. Cyclic Block Codes

A binary block code is said to be cyclic if the following two properties hold:

1. The code is linear.
2. Any cyclic ("end around") shift of a codeword is also a codeword.

The first property means that sums of codewords are codewords, and the second means that if $c = (c_0, c_1, c_2, \ldots, c_{n-1})$ is a codeword, then so are all cyclic shifts, that is, $(c_{n-1}, c_0, c_1, c_2, \ldots, c_{n-2})$, $(c_{n-2}, c_{n-1}, c_0, c_1 \ldots, c_{n-3})$, and so forth.

In describing the structure of cyclic codes, the polynomial representation of codewords is more convenient than the vector representation. We note that for the linearity property, two codeword polynomials are summed by adding in GF(2) coefficients of corresponding terms of each power of x, and the second property means that if $c(x)$ is a codeword polynomial, then

$$x^j c(x) \bmod x^n - 1$$

is also a codeword polynomial for any cyclic shift j. This is true since multiplication of the codeword polynomial by x^j, setting $x^n = 1$, is equivalent to a cyclic shift of the codeword. [We use $x^n - 1$ instead of $x^n + 1$ since this is the general form applicable with polynomials over any finite field.]

To see how cyclic codes are constructed, consider the codeword polynomial that has the smallest degree, $g(x)$. The degree of $g(x)$ is r and it is straightforward to see that all codeword polynomials can be represented as linear combinations of $g(x)$ and cyclic shifts of $g(x)$. Consequently, all codeword polynomials are divisible evenly by $g(x)$, and

we call $g(x)$ the *generator polynomial* of the code. Thus, a cyclic code with block length n can be represented as all the polynomials of the form

$$c(x) = a(x)g(x) \bmod x^n - 1$$

We next show that a cyclic code of block length n is formed from any polynomial $g(x)$ that divides $x^n - 1$; that is, the generator polynomial must be such that

$$x^n - 1 = g(x)h(x)$$

We can verify that a code generated in this manner is cyclic, as follows. We wish to prove that a cylic shift of a codeword

$$x^j c(x) \bmod x^n - 1 = x^j a(x)g(x) \bmod x^n - 1$$

is also a codeword. If it is, the polynomial $x^j c(x) \bmod x^n - 1$ must be divisible by $g(x)$, that is, we must have

$$[x^j a(x)g(x) \bmod x^n - 1] \bmod g(x) = 0$$

Now, if $g(x)$ is a factor of $x^n - 1$, then $[b(x) \bmod x^n - 1] \bmod g(x)$ is simply $b(x) \bmod g(x)$, so that we can write

$$[x^j a(x)g(x) \bmod x^n - 1] \bmod g(x) = x^j a(x)g(x) \bmod g(x) = 0$$

showing that the cyclically shifted codeword is divisible by $g(x)$ and is therefore itself a codeword.

Since r is the degree of the generator polynomial and the generator divides $x^n - 1$, it is a simple matter to show that the resulting cyclic code has 2^k codewords, where $k = n - r$. This follows from the fact that all polynomials $a(x)$ of degree less than k produce distinct codewords, since the products $g(x)a(x)$ must have degree less than n, which in turn means that each product modulo $x^n - 1$ will simply be the polynomial $a(x)g(x)$ itself. Since there are 2^k distinct polynomials of degree $k - 1$ or less, there must be 2^k distinct codewords in the code. Therefore, using the notation adopted previously, we say that the vectors of coefficients of the codeword polynomials generated by $g(x)$, where $g(x)$ divides $x^n - 1$, form an (n, k) cyclic block code. For convenience, the term *codeword* is used interchangeably with *codeword polynomial*.

Cyclic codes may be encoded using the property that $g(x)$ divides all codewords evenly. Let $i(x)$ be the degree $k - 1$ polynomial representing a set of k information bits; then the corresponding codeword polynomial for a nonsystematic code is

$$c(x) = i(x)g(x)$$

and the systematic code may be encoded using

$$c(x) = x^r i(x) + [x^r i(x)] \bmod g(x)$$

A linear feedforward shift register, a polynomial multiplication circuit, may be used to encode the nonsystematic code, and a linear feedback shift register, a polynomial division circuit, may be used to encode the systematic code.

Since each codeword in a cyclic code contains $g(x)$ as a factor, each code polynomial will have roots [from solution of $c(x) = 0$] that must include the roots of $g(x)$. It then follows that since the cyclic code is completely described by $g(x)$, we may define the code by specifying the roots of $g(x)$.

The foregoing discussion of cyclic codes constructed from generator polynomials illustrates the usefulness of polynomial algebra in representing block codes. The factorizations of $x^n - 1$ provide a number of generator polynomials for cyclic codes with block length n, the degree r of the generator determining the number of parity check bits, and $k = n - r$ the number of information bits in the code. The description of codewords as multiples of the generator polynomial provides a characterization of the codewords as the set of polynomials whose roots are the roots of the generator polynomial. The great value of this approach in describing cyclic codes has been to enable coding theorists to draw upon the extensive body of mathematical theory on the algebra of polynomials and their roots in finite fields.

3.2. Binary BCH Codes

The *Bose–Chaudhuri–Hocquenghem codes*, usually referred to as *BCH codes*, are an infinite class of cyclic block codes that have capabilities for *multiple-error detection and correction* [1,2]. For any positive integers m and $t < n/2$, there exists a binary BCH code with block length $n = 2^m - 1$, and minimum distance $d \geq 2t + 1$ having no more than mt parity check bits. Each such code can correct up to t random errors per codeword, and thus is a *t-error-correcting code*.

A BCH code being cyclic can be defined in terms of its generator polynomial $g(x)$. Let α be a primitive element of the extension field GF(2^m). The generator polynomial for a *t-error-correcting* BCH code is chosen so that $2t$ consecutive powers of α, such as $\alpha, \alpha^2, \alpha^3, \ldots, \alpha^{2t}$, are roots of the generator polynomial and consequently are also roots of each codeword.

This defines the subclass of *primitive BCH codes* because the roots are specified to be consecutive powers of a primitive element of GF(2^m). The block length of a BCH code is the order of the element used in defining the consecutive roots. Since α is a primitive element in GF(2^m), the block length of a primitive BCH code is $2^m - 1$. To generalize the definition, if $2^m - 1$ is factorable, $2t$ consecutive powers of some nonprimitive element β of GF(2^m) may instead be specified as roots of the codewords. The resulting code is a *nonprimitive BCH code* and will have a block length that divides $2^m - 1$.

In general, the definition of a BCH code allows the powers of the roots to range over any interval of consecutive values, say, $m_0, m_0 + 1, \ldots, m_0 + 2t - 1$. The parameter m_0 is usually chosen to be zero or one. For the present discussion, $m_0 = 1$.

Since the BCH codes are cyclic, codewords are assured of having the desired set of roots by choosing the generator polynomial so that it has $\alpha, \alpha^2, \ldots, \alpha^{2t}$ as roots. This is done by letting $g(x)$ be the least common multiple of the minimal polynomials of $\alpha, \alpha^2, \ldots, \alpha^{2t}$; that is, we write

$$g(x) = \text{LCM } [m_{\alpha^1}(x), m_{\alpha^2}(x), \ldots, m_{\alpha^{2t}}(x)] \qquad (2)$$

We note from property 6 in the earlier discussion of irreducible polynomials that if a binary irreducible polynomial has β as a root, where β is an element of an extension field of GF(2), then it also has β^2 as a root. Therefore, in Eq. (2), each even-power element α^{2i} and the corresponding element α^i are roots of the same minimal polynomial, $m_{\alpha^i}(x)$, and we can condense the sequence of minimal polynomials and write instead

$$g(x) = \text{LCM}[m_{\alpha^1}(x), m_{\alpha^3}(x), \ldots, m_{\alpha^{2t-1}}(x)]$$

Since we know from property 3 of minimal polynomials that the minimal polynomial of an element in GF(2^m) will have degree no greater than m, we know that $g(x)$ will have degree no greater than mt, and thus the number of parity bits $r = n - k$ will be $\leq mt$. For high-rate codes, $n - k$ is exactly equal to mt, and as t is increased $n - k$ can be smaller than mt. The single-error-correcting primitive BCH codes, $n = 2^m - 1$, $n - k = m$, are the Hamming codes. The generator polynomial for a Hamming code is the minimal polynomial of the primitive element, $m_\alpha(x)$.

The quantity $2t + 1$ used in specifying the generator polynomial of a BCH code is called the *design distance* of the code, but the true minimum distance will in some cases be greater, that is, $d \geq 2t + 1$. The true minimum distance for an arbitrary BCH code cannot be readily given, as this general problem is as yet unsolved. For a great many cases of practical interest, the true minimum distance is equal to the design distance, and the number of check bits is mt.

Tables of generator polynomials for many codes are available in the literature. A simple example is included here to show how a generator is obtained.

Consider a primitive three-error-correcting BCH code with block length $n = 31$ and $m_0 = 1$. The generator polynomial has α, α^3, and α^5 as roots, where α is a primitive element of GF(32). Therefore, $g(x)$ is obtained by forming the product of the minimal polynomials of α, α^3, and α^5 in GF(32). A table of polynomials set up for our purpose is given in Ref. 3. (This table was originally published by Peterson [4] in the first book devoted to the subject of error-correcting codes. It also appears in Peterson and Weldon [5], which is an expanded and updated edition of the original text. For convenience, we refer to the table as the *Peterson table*). From the table, we find

$$m_\alpha(x) = 45_8 = x^5 + x^2 + 1$$

$$m_{\alpha^3}(x) = 75_8 = x^5 + x^4 + x^3 + x^2 + 1$$

$$m_{\alpha^5}(x) = 67_8 = x^5 + x^4 + x^2 + x + 1$$

We now enumerate all the roots of each of these minimal polynomials to determine whether any one polynomial has more than one of the required three roots. This is done with the aid of property 6 of minimal polynomials. The roots of the three minimal polynomials are as follows:

$$\text{Roots of } m_\alpha(x): \quad \alpha, \alpha^2, \alpha^4, \alpha^8, \alpha^{16}$$

$$\text{Roots of } m_{\alpha^3}(x): \quad \alpha^3, \alpha^6, \alpha^{12}, \alpha^{24}, \alpha^{48} = \alpha^{17}$$

$$\text{Roots of } m_{\alpha^5}(x): \quad \alpha^5, \alpha^{10}, \alpha^{20}, \alpha^{40} = \alpha^9, \alpha^{18}$$

From this enumeration, we see that α, α^3, and α^5 are roots of three distinct polynomials, and therefore, the required generator polynomial is simply the product of the three minimal polynomials just found:

$$\begin{aligned} g(x) &= (x^5 + x^2 + 1)(x^5 + x^4 + x^3 + x^2 + 1) \\ &\quad \times (x^5 + x^4 + x^2 + x + 1) \\ &= x^{15} + x^{11} + x^{10} + x^9 + x^8 + x^7 + x^5 + x^3 + x^2 + x + 1 \end{aligned}$$

From the enumeration of all the roots of $m_{\alpha^1}(x)$, $m_{\alpha^3}(x)$, and $m_{\alpha^5}(x)$, it can be verified that each of the three minimal polynomials must be of degree 5, as we found directly by use of the table of polynomials. This distance-7 code, therefore, has 15 check bits and 16 information bits in each codeword.

A table of generator polynomials for primitive BCH codes of block length up to 255 has been published by Stenbit [6], and a table of generator polynomials for BCH codes with lengths up to 1023 is given in a text by Lin and Costello [7].

A number of codes widely used for error detection with long data packets and files in communication network and computer applications are called the *cyclic redundancy check* (CRC) codes [8]. Some of the standardized CRC codes are distance-4 binary BCH codes in which the generator polynomial is formed by multiplying the generator polynomial of a Hamming code by $x + 1$. In these cases, the generator polynomial for the CRC code is of the form $(x + 1)m_\alpha(x)$ and has consecutive roots α^0, α, and α^2. In some applications, CRC codes are modified so that the all-zeros information set is not associated with the all-zeros check set in order to detect certain types of hardware failures.

A generalization of the BCH codes mentioned previously permits specifying a consecutive sequence of roots that can be powers of any element of GF(2^m). That is, with $m_0 = 1$, the sequence of roots can be selected as

$$\beta^1, \beta^2, \beta^3, \ldots, \beta^{2t}$$

where β need not be a primitive element. If $2^m - 1$ is not prime, some of the elements of GF(2^m) will be nonprimitive, and the order of each such element divides $2^m - 1$. For example, GF(2^4) contains elements of order 3, 5, and 15; GF(2^6) contains elements of order 3, 7, 9, 21, and 63; and so on. BCH codes generated from nonprimitive field elements are called *nonprimitive BCH codes*. Each such code, defined as having roots that are $d - 1$ consecutive powers of an element β, will have design distance d, and block length equal to the order of β.

An important example of a nonprimitive code is the (23,12) Golay code. This three-error-correcting code may be constructed as a nonprimitive BCH code with roots in GF(2^{11}). Since $2^{11} - 1 = 23 \times 89$, GF($2^{11}$) has elements of order 23 and 89 as well as 2047. By selecting $\beta = \alpha^{89}$, a code of length 23 can be constructed. Consider the design of a single-error-correcting code, which we specify as having roots β and β^2. Using property 6 of minimal polynomials once again, the roots of the minimal polynomial of β are enumerated as

$$\beta, \beta^2, \beta^4, \beta^8, \beta^{16}, \beta^9, \beta^{18}, \beta^{13}, \beta^3, \beta^6, \beta^{12}$$

where $\beta^{23} = \alpha^{2047} = 1$ is used to reduce powers of β greater than 22. Notice that the sequence is found to include four consecutive powers of β, namely, β, β^2, β^3, and β^4. Therefore, we see that by using the minimal polynomial $m_\beta(x)$ as a BCH code generator polynomial, with the intention of designing a distance-3 code, we have "discovered" two additional consecutive roots and thus have actually constructed a code with design distance 5. The code has 11 check bits, and its generator polynomial can be found in [3], listed as 89 5343B, which yields

$$g(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$$

Although we do not show it here, the true minimum distance is actually 7 rather than 5, and the code is the three-error-correcting (23,12) code originally described by Golay [9].

In applications of binary block codes, it is often necessary to provide a code with a block length that does not correspond exactly to one of the strict-sense BCH codes. This can usually be accomplished by choosing a code with block length greater than the required length and *shortening* the code by an appropriate amount. The shortening is most readily done by setting a number of the information bits equal to zero. The number of codewords that can be generated is reduced accordingly, and since the reduced set of codewords is a subset of the codewords in the unshortened code, the minimum distance of the shortened code must be at least as great as that of the unshortened code. Depending on the amount of the shortening and which particular bits are omitted, the minimum distance may be unchanged or it may increase.

In general, a shortened BCH code may or may not be cyclic, depending on which particular information bits are omitted. There is no general theory available to give guidance about which bits are best omitted for a required amount of shortening. Typically, the shortening is done in the most convenient manner, which is to set a string of consecutive information bits equal to zero, usually the high-order bits in the codeword.

Another commonly used code modification is the *extension* of a code of odd minimum distance by addition of a single overall parity check. Since this modification causes the weight of any odd-weight codeword in the original code to be increased by 1, the minimum distance of the original code is also increased by 1. It should be noted that this modification is not the same as inserting the factor $x + 1$ into the generator polynomial, although even-valued minimum distance $2t + 2$ results in both cases. In the earlier case, while the parity set was increased by one bit, the information set was simultaneously decreased by one bit, so that the block length was unchanged. With the extension being described here, the information set remains unchanged and the block length is increased by one bit. Furthermore, the code obtained by this one-bit extension is not a cyclic code.

A frequently used extended code is the (24,12) distance-8 code, called the *extended Golay code*, which is obtained by appending an overall parity check bit to the (23,12) distance-7 Golay code. The extended code is attractive partly because the rate k/n is exactly equal to 0.5.

As we have pointed out, a BCH code or any cyclic code can be encoded by using the generator polynomial $g(x)$ in the manner indicated by the basic definition of a cyclic code:

$$c(x) = i(x)g(x)$$

Thus, we associate a polynomial $i(x)$ of degree $k - 1$ with the set of k information bits to be transmitted and multiply by the degree-r polynomial $g(x)$ forming the degree-$(n - 1)$ code polynomial $c(x)$. However, this results in a nonsystematic code structure, and it is preferred instead to form codewords using

$$c(x) = [x^r i(x) \bmod g(x)] + x^r i(x)$$

It is seen that this encoding operation places the k information bits in the k highest-order terms of the code polynomial, while the parity check bits, represented by $x^r i(x) \bmod g(x)$, are confined to the r lowest-order terms.

The encoding of any binary cyclic code can be done in a straightforward manner using a linear feedback shift register. An encoding circuit using a shift register with $r = n - k$ stages is shown in Fig. 1. Each box in the circuit is a binary storage device. The additions indicated are done modulo 2, and the tap connections are specified by the coefficients of the generator polynomial. The operation of the encoder is as follows:

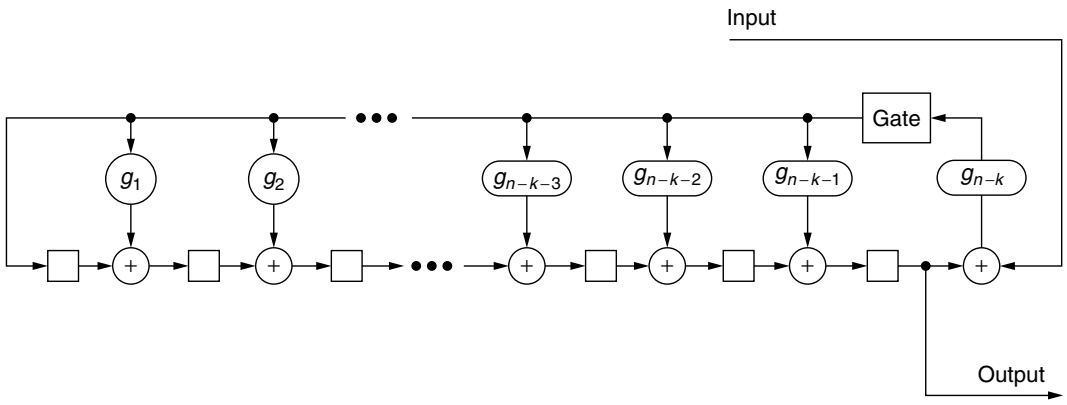


Figure 1. Encoder for a binary BCH code generated by $g(x)$.

1. Shift the k information bits into the encoder and simultaneously into the channel. As soon as the k information bits have entered the shift register, the $r = n - k$ bits in the register are the check bits.
2. Disable the feedback circuit.
3. Shift the contents of the register into the channel.

As an example, consider an encoder for the length-7 Hamming code generated by $g(x) = x^3 + x + 1$. Using a shift register of the form shown in Fig. 1, the feedback tap connections are $g_1 = g_3 = 1$, and $g_2 = 0$. The feedback circuit accomplishes division by $x^3 + x + 1$ in that it sets $x^3 = x + 1$ at each shift of the circuit in step 1.

The encoding circuit that uses an r-stage feedback shift register whose connections are given by the generator polynomial is most convenient for high-rate codes where $k > r$. For low-rate codes, a more convenient encoder realization employs a k-stage feedback shift register whose tap connections are given by $h(x) = (x^n - 1)/g(x)$. For the Hamming code considered, we have $h(x) = x^4 + x^2 + x + 1$, and the circuit shown in Fig. 2 may be used. The operation of the encoder is as follows:

1. With the feedback circuit disabled, shift the $k = 4$ information bits into the k-stage register and simultaneously into the channel.
2. When the k information bits have entered the encoding register, cycle the register $r = 3$ times with the input disabled. The $r = 3$ bits obtained at the output are the encoded parity bits. The parity bits are shifted into the channel.

3.3. Decoding Binary BCH Codes

The problem of decoding a binary BCH code can be described succinctly as follows. For each received word, first determine whether any errors have occurred during transmission. If errors have occurred, determine the most likely locations of the errors in the received word, and make the appropriate corrections. A brute-force approach would be to change one bit at a time, then 2 bits, in all combinations, and so on, until a valid codeword is found. This, of course, is impractical for any but extremely simple codes. Therefore, much work has been devoted to finding efficient algorithms for implementing error correction.

Here we describe the more important decoding techniques developed to date. First, we describe an approach based on solving a set of nonlinear algebraic equations over a finite field, a discussion that leads to the Berlekamp iterative algorithm. Error-trapping decoders are mentioned, and the important example of the Kasami decoder for the Golay code is referenced.

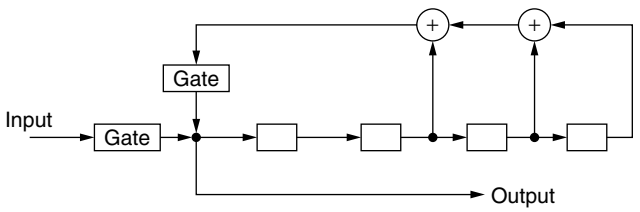


Figure 2. Encoder for the (7,4) Hamming code generated by $h(x) = x^4 + x^2 + x + 1$.

Each of these decoders operates on inputs that consist of hard-bit decisions. In systems where quality measures can be obtained for the received bits, soft-decision decoding techniques can be utilized to increase the power of a code beyond that achievable with algebraic hard-decision decoding. Several such techniques exist, ranging from the simplest forms of *erasure filling* to a set of algorithms termed *channel measurement decoding*.

3.3.1. The Syndrome Equations. Decoding a BCH code begins by calculating a quantity called the *syndrome*. We represent a codeword as the polynomial $c(x)$, the received word as the polynomial $r(x)$, and the corresponding error pattern as $e(x)$, and we can write

$$r(x) = c(x) + e(x)$$

To compute the syndrome values S_k for the received word, we simply substitute the roots of the code generator polynomial into $r(x)$:

$$S_k = r(\alpha^k) = r_0(\alpha^k)^0 + r_1(\alpha^k)^1 + r_2(\alpha^k)^2 + \cdots + r_{n-1}(\alpha^k)^{n-1}$$

which can be evaluated using

$$r(\alpha^k) = \{\cdots[(r_{n-1}\alpha^k + r_{n-2})\alpha^k + r_{n-3}]\alpha^k + \cdots\}\alpha^k + r_0$$

Note that

$$\begin{aligned} S_k &= c(\alpha^k) + e(\alpha^k) \\ &= e(\alpha^k), \quad k = 1, 3, \ldots, 2t-1 \end{aligned} \tag{3}$$

since $c(\alpha^k) = 0$, $k = 1, 3, \ldots, 2t - 1$. That is, each element S_k of the syndrome is simply the error pattern polynomial $e(x)$ evaluated at $x = \alpha^k$ and thus is some element in the extension field GF(2^m). Let us now assume that there are t errors in the received word, so that $e(x)$ has t nonzero coefficients. (To avoid unduly complicated notation, we are letting the actual number of errors be equal to the maximum number correctable by the code, i.e., the value t such that the minimum distance of the code is $d = 2t + 1$. For error patterns having fewer than t errors, one can think of the appropriate subset of the assumed t errors having values equal to 0 rather than 1.) If the ith error $(1 \le i \le t)$ occurs in received symbol $r_j (0 \le j \le n - 1)$, then we define its *error locator* to be $X_i = \alpha^j$, which is an element of GF(2^m). We thus refer to GF(2^m) as the *locator field*. Since we are considering a binary code, all *error values* are 0 or 1, and we can write for any k

$$e(\alpha^k) = \sum_{i=1}^{t} X_i^k \tag{4}$$

where t is the number of errors in the received word.

To make these points clearer, let us say, for example, that there are three errors in the received word, in the first, second, and last bit positions. Then the error polynomial evaluated at each root is simply

$$\begin{aligned} e(\alpha^k) &= c_0(\alpha^k)^0 + e_1(\alpha^k)^1 + e_{n-1}(\alpha^k)^{n-1} \\ &= (\alpha^0)^k + (\alpha^1)^k + (\alpha^{n-1})^k \\ &= X_1^k + X_2^k + X_3^k \end{aligned}$$

where X_1, X_2, and X_3, are the three error locators and the $e_j = \alpha^0 = 1$ are the error values.

From Eqs. (3) and (4) we see that

$$S_k = \sum_{i=1}^{t} X_i^k, \quad k = 1, 3, \ldots, 2t-1 \tag{5}$$

The decoding problem then is simply to find the error locators X_i from the syndrome values $S_1, \ldots, S_{2t-1}$. Note, however, that Eq. (5) represents t nonlinear coupled algebraic equations over the finite field GF(2^m). Direct solution of such equations is generally avoided, and an indirect approach is used instead. To this end, we introduce the polynomial

$$\sigma(x) = \prod_{i=1}^{t}(x + X_i) = x^t + \sigma_1 x^{t-1} + \cdots + \sigma_t \tag{6}$$

having the error locators as roots, and which we therefore call the *error locator polynomial*. The coefficients σ_i are seen to be given by the *elementary symmetric functions* of the error locators [5]:

$$\begin{aligned} \sigma_1 &= \sum_i X_i \\ \sigma_2 &= \sum_{i<j} X_i X_j \\ \sigma_3 &= \sum_{i<j<k} X_i X_j X_k \\ &\vdots \\ \sigma_t &= X_1 X_2 X_3 \cdots X_t \end{aligned}$$

[*Note*: Some authors define the error locator polynomial $\sigma(x)$ as a polynomial with factors of the form $(1 + X_i x)$, so that the roots of $\sigma(x)$ are the reciprocals of the error locators. We find the notation used here to be more convenient for purposes of exposition. However, the reciprocal-root formulation of $\sigma(x)$ will be used in later discussions.]

Several approaches can be taken to decoding a BCH code, each having relative advantages and disadvantages that depend largely on the number of errors the code is designed to correct. Several of the important techniques that are used can be broadly summarized for binary codes as follows:

Step 1. Calculate the syndrome values $S_k = r(\alpha^k), k = 1, 3, \ldots, 2t-1$.

Step 2. Determine the elementary symmetric functions, that is, the coefficients of the error locator polynomial $\sigma(x)$, from the syndrome values.

Step 3. Solve for the roots of $\sigma(x)$, which are the error locators.

Step 4. Correct the errors in the positions indicated by the error locators.

In general, the most difficult part of this procedure is step 2, determination of the coefficients of $\sigma(x)$ from the syndrome values, and it is in this step that the most prominent algorithms differ.

3.3.2. Peterson's Direct Solution Method. We saw in the previous discussion that the syndrome values $S_1, S_3, \ldots, S_{2t-1}$ are the constants in a set of simultaneous nonlinear equations in which the unknowns are the error locators $X_1, X_2, \ldots, X_t$. We now describe a method, due to Peterson [10], for direct solution of these nonlinear equations. In order to describe this method, we write the full set of syndrome values $S_1, S_2, \ldots, S_{2t}$ as

$$\begin{aligned} S_k &= r(\alpha^k) \\ &= c(\alpha^k)) + e(\alpha^k) \\ &= \sum_{i=1}^{t} X_i^k, \quad k = 1, 2, \ldots, 2t \end{aligned}$$

which gives the equations

$$\begin{aligned} X_1 + X_2 + \cdots + X_t &= S_1 \\ X_1^2 + X_2^2 + \cdots + X_t^2 &= S_2 \\ &\vdots \qquad\qquad \vdots \quad \vdots \\ X_1^{2t} + X_2^{2t} + \cdots + X_t^{2t} &= S_{2t} \end{aligned} \tag{7}$$

We call these the *syndrome equations*. The syndrome values $\{S_k\}$ are computed from the received word, and Eq. (7) is to be used to obtain the error locators $\{X_i\}$.

Rather than solving this set of nonlinear equations directly, we convert the equations into linear equations that can be solved in conjunction with the error locator polynomial $\sigma(x)$. This is accomplished by first noting that $\sigma(x)$ evaluated at each error locator value equals zero:

$$\sigma(X_i) = X_i^t + \sigma_1 X_i^{t-1} + \cdots + \sigma_t = 0, \quad i = 1, 2, \ldots, t \tag{8}$$

Clearly, we can multiply Eq. (8) through by any power of X, and the equality is preserved. In particular, let us multiply by, X_i^j, so that we have

$$X_i^{t+j} + \sigma_1 X_i^{t+j-1} + \cdots + \sigma_t X_i^j = 0, \quad i = 1, 2, \ldots, t \tag{9}$$

Now, letting j remain general, we sum Eq. (9) over $i = 1, 2, \ldots, t$, and using the syndrome equations, Eq. (7), we can write

$$S_{t+j} + \sigma_1 S_{t+j-1} + \cdots + \sigma_t S_j = 0 \tag{10}$$

The equations defined by Eq. (10), with t general, are called *Newton's identities*, which for a binary code can be shown to be equivalent to

$$\begin{aligned} S_1 + \sigma_1 &= 0 \\ S_3 + S_2\sigma_1 + S_1\sigma_2 + \sigma_3 &= 0 \\ S_5 + S_4\sigma_1 + S_3\sigma_2 + S_2\sigma_3 + S_1\sigma_4 + \sigma_5 &= 0 \\ &\vdots \end{aligned} \tag{11}$$

In principle, to decode a code of any given minimum distance, we need only truncate Eq. (11) in an appropriate manner and solve a set of linear equations for the $\{\sigma_i\}$ in terms of the given syndrome values.

For example, in decoding a single-error-correcting code, there is only one syndrome value, S_1, and the first line of Eq. (11) gives

$$S_1 + \sigma_1 = 0$$

so that we have

$$\sigma_1 = S_1$$

For $t = 1$, the error locator polynomial, Eq. (6), is simply $x + \sigma_1$, having the trivial root $x = \sigma_1$, which we have just found to be equal to S_1. Thus for a single-error-correcting BCH code, we have the very simple result that the error locator is equal to the syndrome S_1.

For a two-error-correcting code, two syndrome values are computed, S_1 and S_3, and the first two lines of Eq. (11) (with $\sigma_3 = 0$) can be written in matrix form as

$$\begin{bmatrix} 1 & 0 \\ S_2 & S_1 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_3 \end{bmatrix}$$

These simultaneous linear equations are solved using the methods of ordinary algebra except that multiplications, divisions, and additions are done using the rules for GF(2^m). We note here that while the S_k values for k even need not be calculated for a binary BCH code, these even-indexed syndrome values appear in the given formulation of the decoding problem. They are readily obtained since it is easy to show that for binary codes, $S_{2k} = S_k^2$ for any k. That is, for elements A and B_i in a field of characteristic 2, if

$$A = \sum_i B_i$$

then the square of A is simply

$$A^2 = \sum_i \sum_j B_i B_j = \sum_i B_i^2$$

so that we have

$$S_1^2 = \sum_{i=1}^{t} X_1^2 = S_2 \tag{12}$$

Similarly, $S_4 = S_2^2 = S_1^4, S_6 = S_3^2$, and so forth. Thus, in solving the simultaneous equations, the solutions can be expressed in terms of only the odd-indexed syndrome values. For the case of two-error correction, we have

$$\sigma_1 = S_1 \text{ and } \sigma_2 = \frac{S_3 + S_1^3}{S_1} \tag{13}$$

Using standard techniques to solve sets of simultaneous linear equations, direct solutions for the coefficients of the error locator polynomial can be found for any error-correction limit t. The results of such solutions for $t = 3$ and 4 are as follows:

Three-Error Correction	**Four-Error Correction**
$\sigma_1 = S_1$	$\sigma_1 = S_1$
$\sigma_2 = \dfrac{S_1^2 S_3 + S_5}{S_1^3 + S_3}$	$\sigma_2 = \dfrac{S_1(S_7 + S_1^7) + S_3(S_1^5 + S_5)}{S_3(S_1^3 + S_3) + S_1(S_1^5 + S_5)}$
$\sigma_3 = (S_1^3 + S_3) + S_1\sigma_2$	$\sigma_3 = (S_1^3 + S_3) + S_1\sigma_2$
	$\sigma_4 = \dfrac{(S_5 + S_1^2 S_3) + (S_1^3 + S_3)\sigma_2}{S_1}$

In general, however, with use of a t-error-correcting code, any error pattern with fewer than t errors is also correctable, and we do not know at the outset of decoding how many errors there actually are. To use these formulas knowing that the actual number of errors in a received word may be less than t, we start by determining whether the first t lines in Eq. (11) can be solved for $\sigma_1, \sigma_2, \ldots, \sigma_t$. This is done by using the determinant test

$$\det \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ S_2 & S_1 & 1 & 0 & 0 & \cdots & 0 \\ S_4 & S_3 & S_2 & S_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ S_{2t-4} & S_{2t-5} & S_{2t-6} & S_{2t-7} & S_{2t-8} & \cdots & S_{t-3} \\ S_{2t-2} & S_{2t-3} & S_{2t-4} & S_{2t-5} & S_{2t-6} & \cdots & S_{t-1} \end{bmatrix} \overset{?}{\neq} 0$$

It can be shown [10] that if there are t or $t-1$ errors in the received word, the determinant will be nonzero. Given this outcome, we proceed with the formulas for t-error correction. If there are actually t errors, the solutions found for $\sigma_1, \sigma_2, \ldots, \sigma_t$ define a degree-t error locator polynomial. If there are only $t-1$ errors, $\sigma_t = 0$ and thus, $\sigma(x)$ has degree $t-1$.

If the determinant shown above is found to be zero, two rows and columns of the matrix are removed, and the determinant of the resulting $(t-2) \times (t-2)$ matrix is tested in the same manner. This procedure is repeated until a nonzero determinant is found and the error locator polynomial coefficients are determined.

The final steps in decoding a binary BCH code are to find the roots of the error locator polynomial $\sigma(x)$, which are the error locators, and to correct the errors. A procedure called the *Chien search* [11] accomplishes these two processes without explicitly solving $\sigma(x)$. This can be done with the circuit shown in Fig. 3. The circuit steps sequentially through all possible error locator values and corrects the corresponding bits as the locators are found. To see how the circuit operates, consider the error locator

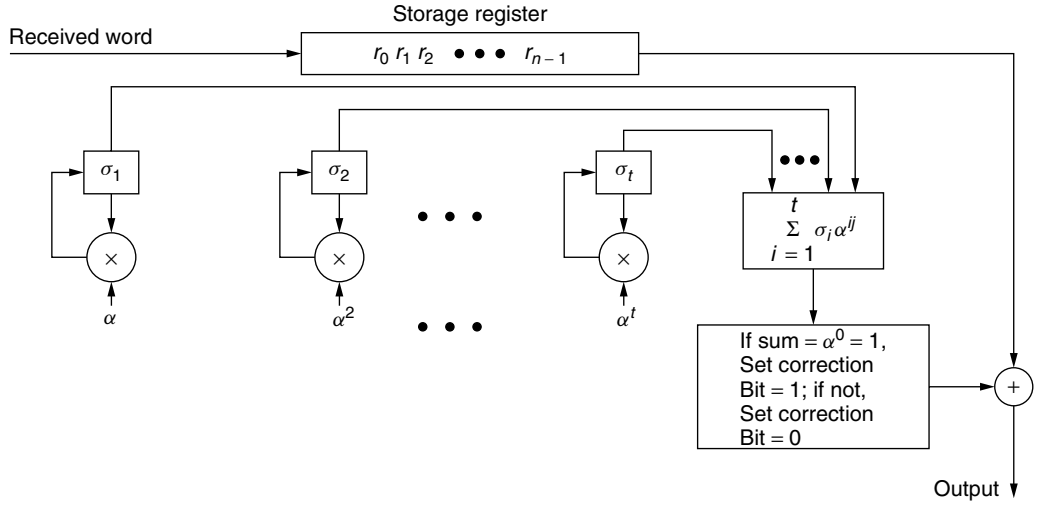


Figure 3. Circuit for implementing Chien search.

polynomial as given by Eq. (6) and divide through by x^t, which gives

$$\frac{\sigma(x)}{x^t} = 1 + \sigma_1 x^{-1} + \sigma_2 x^{-2} + \cdots + \sigma_t x^{-t}$$

The values of x that satisfy $\sigma(x) = 0$ consequently satisfy the equation

$$\sigma_1 x^{-1} + \sigma_2 x^{-2} + \cdots + \sigma_t x^{-t} = 1$$

Assuming the convention of transmitting codewords high-order bits first, it is convenient to apply the root test to locator α^{n-1} first. Note that evaluation of a term x^{-i} at α^{n-1} yields α^{-in+i}, which equals α^i if we are using a full-length BCH code, since we then have $\alpha^n = 1$ and thus $\alpha^{-in} = 1$. Therefore, we see that testing α^{n-1} as a possible root of $\sigma(x)$ is the same as testing for

$$\sigma_1 \alpha + \sigma_2 \alpha^2 + \cdots + \sigma_t \alpha^t \stackrel{?}{=} 1$$

and, in general, testing for α^{n-j} as an error locator is equivalent to finding whether or not α^j satisfies

$$\sum_{i=1}^{t} \sigma_i a^{ij} = a^0 = 1, \quad j = 0, 1, 2, \ldots, n-1$$

3.3.3. The Berlekamp Algorithm. For correction of more than about six errors in a binary BCH codeword, Peterson's direct method of solving for the coefficients of $\sigma(x)$ from the syndrome values becomes cumbersome and inefficient, since the number of finite-field multiplications required increases approximately with the square of the number of errors to be corrected. Instead, it is preferable to use an iterative algorithm developed by Berlekamp [12] for solution of Newton's identities. In contrast with the direct solution method, the *Berlekamp algorithm* has a computational complexity that grows only linearly with the number of errors to be corrected. Another version of this algorithm was given by Massey [13]. The Massey formulation is presented in the succeeding article on nonbinary BCH codes.

In the use of the Berlekamp algorithm, the sequence of calculated syndrome values, $S_1, S_2, \ldots, S_{2t}$, is represented by the polynomial

$$S(z) = S_1 z + S_2 z^2 + \cdots + S_{2t} z^{2t}$$

As a convenience in the algorithm, the error locator polynomial is replaced with an equivalent polynomial $C(z)$ whose *reciprocal roots* are the error locators $X_i, i = 1, 2, \ldots, t$; that is, $C(z)$ is defined by

$$C(z) = \prod_{i=1}^{t} (1 + X_i z)$$

so that $C(z)$ has roots at $z = Z_i$, where $Z_i = 1/X_i$, $i = 1, 2, \ldots, t$. We call the polynomial $C(z)$ the *reciprocal error locator polynomial* to distinguish it from the error locator polynomial $\sigma(x)$. Now writing $C(z)$ in its expanded form, we have

$$C(z) = 1 + \sigma_1 z + \sigma_2 z^2 + \cdots + \sigma_t z^t$$

where the coefficients $\{\sigma_i\}$ are again seen to be the elementary symmetric functions of the error locators $X_1, X_2, \ldots, X_t$.

The Berlekamp algorithm is an efficient iterative procedure for finding the minimum-degree reciprocal error locator polynomial $C(z)$ whose coefficients, taken together with the syndrome values, satisfy all t equations in Newton's identities, Eq. (11). The algorithm begins by constructing the polynomial $C^{(1)}(z)$ of least degree satisfying the first line in Eq. (11), and then setting $C^{(2)}(z) = C^{(1)}(z)$ and determining whether $C^{(2)}(z)$ satisfies the second line in Eq. (11). It is easy to see that these first steps consist in simply letting $C^{(1)}(z) = 1 + S_1 z$, so that $\sigma_1 = S_1$, and then testing (second line) the relationship $S_3 = S_2 \sigma_1$, where $S_2 = S_2^1$ from Eq. (12). This is equivalent to testing $S_3 = S_1^3$, the relationship that must hold, by Eq. (7), if there is only a single error in the received

word. If the test of the second line succeeds, then we set $C^{(3)}(z) = C^{(2)}(z)$ and test the third line of Eq. (11). If the test of the second line of Eq. (11) fails, $C^{(2)}(z)$ is modified by adding a correction term that changes $C^{(2)}(z)$ to a minimum-degree polynomial satisfying the first two lines of Eq. (10). With the corrected form of $C^{(2)}(z)$, we let $C^{(3)}(z) = C^{(2)}(z)$, and then test the third line of Eq. (11), and so forth. The iteration continues until a reciprocal error locator polynomial $C^{(l)}(z), l \leq t$, is found that satisfies all t lines in Eq. (11).

The efficiency of the Berlekamp algorithm is due largely to its provision for constructing the correction term at the ith iteration, if needed, so that the $i-1$ previous lines in Eq. (11) do not have to be retested. It can be shown that if the number of errors in the received word is t or less, the Berlekamp algorithm will end with the correct reciprocal error locator polynomial. We shall simply provide a brief description of the algorithm here. A detailed discussion of the algorithm and a rigorous proof of its error-correction properties can be found in the literature [5,12]. The algorithm as described below is actually a simplification for use with binary BCH codes. There is a more general version of the algorithm applicable to nonbinary codes as well.

The steps in the Berlekamp algorithm are described below. The initialized polynomial $C^{(0)}(z)$ fixes 1 as the leading term in $C(z)$, while $T^{(0)}(z)$ is the initialized correction polynomial. The quantity $\Delta^{(2k)}$ is the discrepancy found when an interim version of $C(z)$ constructed at one line in Eq. (11) fails to satisfy the next line. Superscripts are used to index the steps in the iteration.

The Berlekamp Algorithm for Decoding Binary BCH Codes

1. Initialize: $k = 0, C^{(0)}(z) = 1, T^{(0)}(z) = 1$.
2. If S_{2k+1} is not given, stop. Otherwise, define $\Delta^{(2k)}$ as the coefficient of z^{2k+1} in the product $[1 + S(z)]C^{(2k)}(z)$. Let

$$C^{(2k+2)}(z) = C^{(2k)}(z) + \Delta^{(2k)} z T^{(2k)}(z)$$

where

$$T^{(2k+2)} = \begin{cases} z^2 T^{(2k)}(z) & \text{if } \Delta^{(2k)} = 0, \text{ or if } \deg C^{(2k)}(z) > k \\ \dfrac{z C^{(2k)}(z)}{\Delta^{(2k)}} & \text{if } \Delta^{(2k)} \neq 0 \text{ and } \deg C^{(2k)}(z) \leq k \end{cases}$$

3. Set $k = k + 1$ and return to step 2.

Note that the multiplications and additions indicated are all in the locator field GF(2^m).

3.3.4. Other Decoding Algorithms. At times, other decoding algorithms for binary BCH codes are useful. For example, there are decoders that use "error trapping" to find and correct channel errors. These decoders evaluate the syndrome by calculating $s(x)$ where

$$s(x) = r(x) \bmod g(x)$$

The polynomial $s(x)$ has degree up to $r-1$ and is zero when no channel errors occur. Note that if t or fewer channel errors occur, and all the errors are confined to the check bit position, $s(x)$ contains at most t terms, the error pattern unmodified. It can be shown that, if there is at least one error in an information bit position, the number of terms in $s(x)$ is greater than t. Therefore, if t or fewer terms are found in $s(x)$, the channel error pattern is determined and can be corrected.

This property may be used to decode cyclic codes since if $s(x)$ has more than t terms, the received word may be shifted cyclically by one bit, and a second syndrome $s'(x)$ computed. If the error pattern has been shifted into the check bit positions, $s'(x)$ now contains t or fewer terms, and the error pattern has been successfully trapped in the check bit positions. In total, $n-1$ cyclic shifts of the received word may be tried in this way.

This decoding procedure will succeed when the channel error pattern spans at most r bit positions in the received word. This is not always the case, but for some codes, it is possible to specify additional tests that may be used to detect and correct the other correctable error patterns. The best example of this type of error trapping decoder is the Kasami decoder for the Golay code [14]. The Kasami decoder can be implemented with a simple linear feedback shift register and logic circuits.

3.3.5. Soft-Decision Decoding Techniques. Up to this point, we have discussed the decoding problem as one of finding the number and locations of errors in a received word. It has been assumed that at the receiving end of the communication circuit, a definite binary decision is made on each received digit after demodulation and prior to decoding, that is, a hard binary decision. However, it is sometimes possible to provide for quality or confidence estimates for demodulated data. In the simplest example of such schemes, we might test the demodulator output against a preselected magnitude threshold and erase each digit that falls below the threshold. The decoder is then presented with a sequence consisting of definite zeros and ones as well as erasures, and, given that sequence, the decoder has the task of deciding which of the valid codewords is most likely to have been transmitted. We call this decoding task one of *errors-and-erasures decoding*.

It has been noted that a block code having minimum distance d is capable of correcting any pattern of t or fewer errors, where $d = 2t+1$ or $2t+2$, for d odd or even, respectively. We now state that a distance-d code is capable of correcting any pattern of l errors and s erasures such that $2l + s < d$, and we show a very simple procedure that demonstrates that this is true. Assuming that we have at our disposal a decoder for correcting up to t errors, where $2t+1 = d$, we decode for s erasures and an unknown number of errors as follows:

1. Set all s erased bits in the received word equal to 0, and perform error correction of up to t errors. Note the number of errors corrected if decoding can be completed.
2. Next, set all s erased bits equal to 1, and decode the received word again, noting the number of errors corrected if decoding can be completed.

3. If only one decoding succeeds, accept that output. If both decoding attempts succeed but produce different codewords, accept the decoding result that required correction of the smaller number of errors.

The errors-and-erasures decoding procedure just described is an example of a general class of algorithms that are usually referred to as *soft-decision decoding* techniques. The simplest of such techniques is Wagner coding. In this scheme, encoding is done by appending a single overall parity check to a block of k information bits. The decoding procedure can be described as follows. On reception of each received digit r_i, the a posteriori probabilities $p(0|r_i)$ and $p(1|r_i)$ are calculated and saved, and a hard-bit decision is also made on each of the $k+1$ digits. Overall parity is checked, and if it is satisfied, the k information bits are accepted as first decoded. If parity fails, the received digit having the smallest difference between its two a posteriori probabilities is inverted before the k information bits are accepted. It is seen that this technique is in fact the simplest application of the errors-and-erasures decoding procedure described in the previous section, where here only a single erasure may be filled but no errors corrected, since the minimum distance of the single-parity-check code is only $d = 2$.

A generalization of Wagner coding applicable to any multiple-error-correcting (n, k) code is a scheme called *forced-erasure decoding*. Here we assume that the demodulator, in addition to making a hard binary decision on each received digit, also measures relative reliability; we denote the set of reliability measures by $p_1, p_2, \ldots, p_n$. For many communication channels, the probability of correct bit demodulation is monotonically related to the magnitude of the detected signal, and, in such cases, the detected signal strength can be taken as a measure of reliability for each bit.

Several decoding strategies come under the heading of forced-erasure decoding. They share the feature that decoder performance is improved by use of multiple hard-decision decoding attempts where the order of the decode attempts is based on the bit reliability information. The schemes that permit the largest number of decoding trials provide the best performance but they are complex and cumbersome. Finding efficient soft-decision decoding techniques for BCH codes is still an open problem.

BIOGRAPHIES

Arnold M. Michelson received his BSEE degree from the Johns Hopkins University, his MSEE from the University of Rochester, New York, and he did further graduate work at the Polytechnic Institute of Brooklyn, New York. In 1968, Mr. Michelson joined Sylvania Electric Products, which later became GTE Government Systems. At Sylvania and GTE he worked on the development and implementation of advanced communication techniques, including error-control coding for military applications. That work focused primarily on long-wave communications. Since 2000, he has been with the Raytheon Company where he is involved in the development of high-performance coding techniques for military and commercial satellite applications. In 1997, Mr. Michelson received GTE's Leslie H. Warner Technical Achievement Award, and, in 2002, Raytheon's Excellence in Technology Award, Distinguished Level.

Allen H. Levesque received his BSEE degree from Worcester Polytechnic Institute in 1959 and his MSEE and PhDEE degrees from Yale University in 1960 and 1965, respectively. Following completion of his graduate studies, he joined the GTE Corporation, where, over a 36-year career, he worked on and led a variety of digital communications research and development projects, with application to both defense and commercial systems. Much of his early work concerned applications of error-control coding techniques in radio networks. For the past decade, his work has concentrated on mobile and wireless communications networks. In early 1999, he retired from GTE Laboratories to begin and independent consulting practice and to take a part-time teaching and research position at WPI. He currently teaches graduate courses in modulation and coding and is a member of WPI's Center for Wireless Information Network Studies. His areas of research interest include communication theory and techniques, communication networks, wireless communications, spread-spectrum, secure communications, and digital signal processing. He has published numerous journal and conference papers, and coauthored two books, as well as chapters in several communications handbooks. He is a life fellow of the IEEE and a Registered Professional Engineer in the Commonwealth of Massachusetts.

BIBLIOGRAPHY

1. A. Hocquenghem, Codes correcteurs d'erreurs, *Chiffres* **2**: 147–156 (1959).
2. R. C. Bose and D. K. Ray-Chaudhuri, On a class of error-correcting binary group codes, *Inform. Control* **3**: 68–79 (1960).
3. A. M. Michelson and A. H. Levesque, *Error-Control Techniques for Digital Communications*, Wiley, New York, 1985.
4. W. W. Peterson, *Error Correcting Codes*, MIT Press, Cambridge, MA, 1961.
5. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2nd ed., MIT Press, Cambridge, MA, 1972.
6 J. P. Stenbit, Table of generators for Bose-Chaudhuri codes, *IEEE Trans. Inform. Theory* **IT-10**: 390–391 (1964).
7. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
8. A. Leon-Garcia and I. Widjaja, *Communication Networks*, McGraw-Hill, New York, 2000.
9. M. J. E. Golay, Notes on digital coding, *Proc. IRE* **37**: 657 (1949).
10. W. W. Peterson, Encoding and error-correction procedures for the Bose-Chaudhuri codes, *IRE Trans. Inform. Theory* **IT-6**: 459–470 (1960).

11. R. T. Chien, Cyclic decoding procedures for Bose-Chaudhuri-Hocquenghem codes, *IEEE Trans. Inform. Theory* **IT-10**: 357–363 (1964).

12. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

13. J. L. Massey, Shift-register synthesis and BCH decoding, *IEEE Trans. Inform. Theory* **IT-15**: 122–127 (1969).

14. T. Kasami, A decoding procedure for multiple-error correcting cyclic codes, *IEEE Trans. Inform. Theory* **IT-10**: 134–138 (1964).

BCH CODES — NONBINARY AND REED-SOLOMON*

ARNOLD M. MICHELSON
ALLEN H. LEVESQUE
Marlborough, Massachusetts

1. INTRODUCTION

Bose–Chaudhuri–Hocquenghem (BCH) cyclic block codes include both binary and nonbinary codes. The preceding article describes the fundamental properties and structure of *binary BCH codes*. This article treats nonbinary BCH codes and a closely related class of nonbinary codes called *Reed–Solomon codes*. Certain important and useful modifications of BCH and Reed–Solomon codes are also discussed. The approach to describing the structure of nonbinary codes closely parallels that used for binary codes in the preceding article, and reference to that discussion is made where appropriate.

The description of a nonbinary cyclic code follows directly from the binary case; that is, an (n, k) cyclic code defined on the Galois Field GF(q) can be generated as the set of all polynomials of the form $a(x)\ g(x)$, where $a(x)$ is any polynomial of degree $k-1$ or less with coefficients in GF(q) and the generator polynomial $g(x)$ divides $x^n - 1$ and has coefficients in GF(q). As in the binary case, we shall see that the design of a nonbinary code rests upon selection of a generator polynomial having prescribed roots in a field that is an extension of GF(q), say, GF(q^m).

1.1. Nonbinary BCH Codes

The binary BCH codes are a special case of a class of cyclic codes that can be constructed for any symbol alphabet defined on a finite field, say, GF(q), which can be a prime field or some extension of a prime field. As a generalization of the binary case, a t-error-correcting BCH code on GF(q) is a cyclic code, and all the codewords have roots that include $2t$ consecutive powers of some element β contained in GF(q^m), an extension field of GF(q). It will be convenient to distinguish between the two fields by calling GF(q) the *symbol field* and GF(q^m), the *locator field*. As with the binary codes, BCH codes on GF(q) can be primitive or nonprimitive, depending on whether a primitive or nonprimitive element of GF(q^m) is used to specify the consecutive roots of the codewords. For the present discussion, attention is restricted to the case of primitive codes, so that the code is specified to be a set of code polynomials whose roots include the elements $\alpha, \alpha^2, \ldots, \alpha^{2t}$, where α is a primitive element of GF(q^m). The design distance is one greater than the number of consecutive roots, and the true minimum distance can be equal to or greater than the design distance. The generator polynomial of a BCH code on GF(q) is defined as the least common multiple of the minimal polynomials of $\alpha, \alpha^2, \ldots, \alpha^{2t}$:

$$g(x) = \mathrm{LCM}[m_{\alpha^1}(x), m_{\alpha^2}(x), \ldots, m_{\alpha^{2t}}(x)]$$

The block length of the code is the order of the element chosen to prescribe the consecutive roots, and therefore, for the primitive codes, where we choose a primitive element of GF(q^m), the block length is $n = q^m - 1$.

In general, a t-error-correcting code may have either odd or even minimum design distance, given by $d = 2t + 1$, or $d = 2t + 2$, respectively. Furthermore, the sequence of powers of α can begin with an arbitrary power, say, m_0, so that we can specify the roots as $\alpha^{m_0}, \alpha^{m_0+1}, \ldots, \alpha^{m_0+d-2}$, that is, $d-1$ consecutive powers of α. Similarly, we can define the generator polynomial as

$$g(x) = \mathrm{LCM}[m_{\alpha^{m_0}}(x), m_{\alpha^{m_0+1}}(x), \ldots, m_{\alpha^{m_0+d-2}}(x)]$$

As mentioned previously, nonprimitive nonbinary BCH codes can be defined on GFq) as well. If $q^m - 1$ is factorable, a nonprimitive code with design distance d can be formed by specifying its roots to be $d-1$ consecutive powers of β, some nonprimitive element of GF(q^m). The block length n of the code is the order of β, that is, n divides $q^m - 1$. However, the most widely used nonbinary BCH codes are the Reed–Solomon codes, which we discuss next.

1.2. Reed–Solomon Codes

An important subclass of nonbinary BCH codes is obtained by choosing the locator field to be the same as the symbol field. These codes are called *Reed–Solomon codes* [1], often abbreviated as *RS codes*.

Specifically, an RS code on GF(q) with minimum distance d has as roots $d-1$ consecutive powers of α, a primitive element of GF(q). The minimal polynomial over GF(q) of any element γ in GF(q) is just $x - \gamma$. This means that the generator polynomial $g(x)$ for a design-distance-d RS code is

$$g(x) = (x - \alpha^{m_0})(x - \alpha^{m_0+1}) \cdots (x - \alpha^{m_0+d-2}) \qquad (1)$$

where m_0 is an arbitrary integer, usually chosen as 0 or 1. Since the order of α is $q-1$, the block length of an RS code is $q-1$. For any BCH code, the design distance is one greater than the number of consecutive roots in the locator field, and since from Eq. (1) the number of check symbols is always equal to the number of prescribed roots, we have for any RS code

$$d = n - k + 1$$

* Preparation of this article supported in part by the Raytheon Corporation.

where n is the block length and k is the number of information symbols in each block. An important property of any RS code is that the true minimum distance is equal to the design distance. No (n, k) linear block code can have minimum distance greater than $n - k + 1$, and a code for which the minimum distance equals $n - k + 1$ is called a *maximum-distance-separable* (MDS) code, or simply a *maximum code* [2]. Therefore, every RS code is an MDS code. Furthermore, shortening the block length of an RS code by omitting information symbols cannot reduce its minimum distance, and, therefore, we can state that any shortened RS code is also an MDS code.

2. ENCODING NONBINARY BCH CODES AND RS CODES

The formation of codewords in an RS code on GF(q) from its generator polynomial $g(x)$ extends directly from the binary case. Thus, the words in an (n, k) code correspond to the set of all polynomials over GF(q) of degree $n - 1$ or less that are divisible by $g(x)$, where the degree of $g(x)$ is $r = n - k$.

The codewords can be generated by multiplying all polynomials over GF(q) having degree $k - 1$ or less by $g(x)$. As was seen in the binary case, this will not produce a systematic code and is generally avoided. As in the case of the binary codes, systematic structure can be provided by forming codewords as

$$c(x) = x^r i(x) \bmod g(x) + x^r i(x)$$

where $i(x)$ denotes the k information symbols on GF(q) to be encoded represented as a polynomial of degree $k - 1$ or less.

Encoding can be implemented with a polynomial division circuit of the form described previously for binary BCH codes (see Fig. 1 in the article on binary BCH codes). However, multiplications and additions are now done in GF(q). As an example, we consider the (63,57) $d = 7$ RS code defined on GF(64). Assuming $m_0 = 1$, and letting α be the primitive element of GF(64), we have

$$g(x) = \prod_{i=1}^{6} (x + \alpha^i)$$

If we use the primitive binary polynomial $p(x) = x^6 + x + 1$ to provide a representation of GF(2^6), it is straightforward to show that

$$g(x) = x^6 + \alpha^{59}x^5 + \alpha^{48}x^4 + \alpha^{43}x^3 + \alpha^{55}x^2 + \alpha^{10}x + \alpha^{21}$$

An encoder for the (63,57) $d = 7$ RS code defined GF(64) is shown in Fig. 1. Each stage of the register is a 64-ary storage device, and the feedback lines require multiplication in GF(64). The feedback weights are the coefficients of the generator polynomial. The circuit shown in Fig. 1 operates as follows:

1. Enable gates G_1, G_2, and G_3. Disable gate G_4. Clock the information symbols to be encoded into the feedback shift register and simultaneously into the codeword buffer.
2. Disable gates G_1, G_2, and G_3, and enable gate G_4. The six parity symbols are now contained in the six storage elements of the feedback shift register. Clock these six symbols into the buffer to complete formation of the codeword.
3. All stages of the feedback shift register are now reset to zero, and the encoded word is shifted out while the next information set to be encoded is shifted in. Return to step 1.

3. DECODING RS CODES

We now describe algorithms for decoding RS codes that are generalizations of the bounded-distance decoding algorithms presented for binary BCH codes. Given the set of syndrome values calculated for the received word, the decoding task is to find the most likely error pattern, within the error-correction limit of the code, which produces the observed syndrome values. Therefore, as in the binary case, decoding is viewed as a problem of solving a set of simultaneous syndrome equations, but one where the set of unknowns now includes the error values or *error magnitudes* in addition to the error locators.

We use the notation adopted for the binary case, letting a transmitted codeword be represented by a polynomial

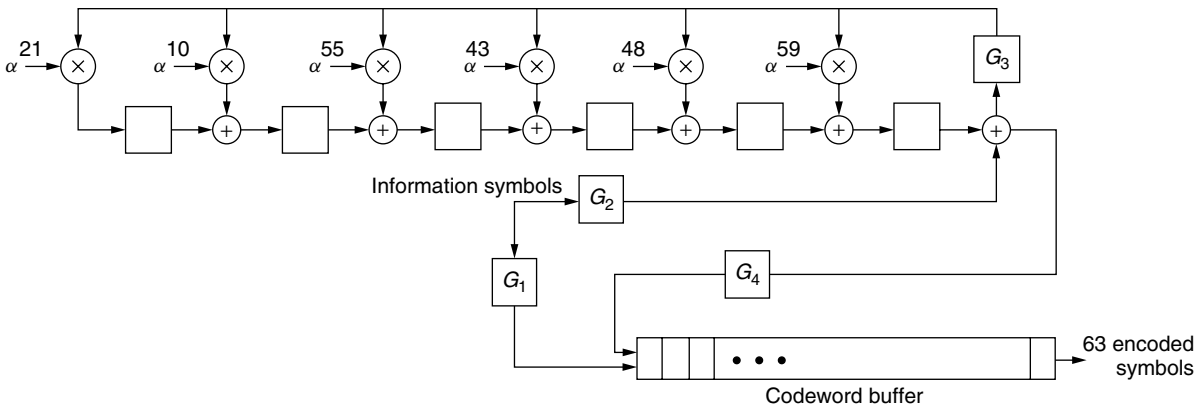


Figure 1. Encoder for the (63,57) RS code on GF(64).

$c(x)$, where here the coefficients of $c(x)$ are elements in GF(q). Similarly, a received error pattern is represented by a polynomial $e(x)$, again with coefficients in GF(q), and the received word is represented by $r(x)$, where

$$r(x) = c(x) + e(x)$$

The syndrome values are obtained by evaluating $r(x)$ at the prescribed roots of the generator polynomial:

$$\begin{aligned} S_k &= r(\alpha^k) \\ &= c(\alpha^k) + e(\alpha^k) \\ &= e(\alpha^k), \quad k = m_0, m_0+1, \ldots, m_0+d-2 \end{aligned} \tag{2}$$

where we let the roots of the code be any arbitrary sequence of consecutive powers of α, although m_0 is usually chosen to be 0 or 1.

The error polynomial $e(x)$ has nonzero terms only in those positions where errors have occurred, so that if there are t errors in the received word, we can write the syndrome values as

$$S_k = \sum_{i=1}^{t} Y_i X_i^k, \quad k = m_0, m_0+1, \ldots, m_0+d-2 \tag{3}$$

where X_i is the error locator for the ith error and Y_i is its value. Therefore, the decoding task is, given the S's, find the X and Y values. In a generalization of the procedure outlined for binary codes, syndrome decoding of an RS code proceeds as follows:

1. Calculate the syndrome values S_k, $k = m_0, m_0+1, \ldots, m_0+d-2$.
2. Determine the error locator polynomial $\sigma(x)$ from the syndrome values.
3. Solve for the roots of $\sigma(x)$, which are the error locators.
4. Given the error locators, calculate the error values.
5. Correct the indicated errors.

The fundamental difference between this sequence of steps and the procedure outlined for the binary case is step 4, calculation of the error values. However, once the error locations have been determined, finding the error values is straightforward, since, given the S and X values, Eq. (3) is simply a set of simultaneous linear equations having the error values as unknowns. As in the binary case, the most difficult part of the procedure is usually step 2, determination of the error locator polynomial $\sigma(x)$ from the syndrome values.

3.1. Peterson's Direct Solution Method

Peterson's direct solution method for finding the coefficients of the error locator polynomial $\sigma(x)$, generalizes in a straightforward way to the case of nonbinary codes, although there are a few important differences. The set of simultaneous nonlinear (in the X values) syndrome equations, Eq. (3), can be converted into a set of linear equations to be solved in conjunction with $\sigma(x)$. To begin, exactly as we did in Eqs. (6–9) in the article on binary BCH coding, we can operate repeatedly on $\sigma(x)$ and invoke the syndrome equations, Eq. (3), to establish the relationship

$$S_{t+j} + \sigma_1 S_{t+j-1} + \cdots + \sigma_t S_j = 0, \quad \text{for all } j \tag{4}$$

where the σ terms are coefficients of the error locator polynomial, $\sigma(x)$:

$$\sigma(x) = x^t + \sigma_1 x^{t-1} + \cdots + \sigma_t \tag{5}$$

The equations defined by Eq. (4) are *Newton's identities*.

Let us consider a t-error-correcting nonbinary BCH or RS code, for which we have computed $2t$ syndrome values $S_1, S_2, \ldots, S_{2t}$. From Eq. (4), we can construct t simultaneous equations, linear in coefficients of $\sigma(x)$, by letting j range from 1 through t. To illustrate this with an example, we consider the case of a three-error-correcting code so that we have

$$\begin{aligned} S_1\sigma_3 + S_2\sigma_2 + S_3\sigma_1 &= -S_4 \\ S_2\sigma_3 + S_3\sigma_2 + S_4\sigma_1 &= -S_5 \\ S_3\sigma_3 + S_4\sigma_2 + S_5\sigma_1 &= -S_6 \end{aligned} \tag{6}$$

The three equations have been written in a form suggesting their use, that is, as a set of simultaneous linear equations, with coefficients and constants that are the syndrome values. These equations are then solved for the three coefficients of $\sigma(x)$ when three errors are assumed to have occurred.

The reader should compare Eq. (6) with Eq. (10) in the article on binary BCH codes and note certain differences. First, unlike the binary case, Eq. (6) includes equations beginning with the even-indexed syndrome values. This is because the relationships $S_j^2 = S_{2j}$ are specific to the binary case and do not hold for nonbinary codes. Second, the simpler forms of the uppermost lines in Eq. (10) for the binary case are also specific to the binary case and do not apply here. Finally, negative signs are retained when the constants S_4, S_5, and S_6 are moved from the left side in Eq. (4) to the right side in Eq. (6), since addition and subtraction are identical only when the field is of characteristic 2.

As with binary codes, determining the locations of a given number of errors is done by constructing an appropriate set of simultaneous equations of the form given by Eq. (6) and solving the equations for the σ's in terms of the syndrome values $\{S_k\}$. The number of equations to be used is equal to the actual number of errors in the received code block, which must be determined as part of the decoding operation. This is done by testing determinants of various sizes corresponding to the possible numbers of errors. The equations for the σ terms in the three-error case are given in Eq. (6). We now write the sets of equations for the one-error and two-error cases, in the more compact matrix form, as follows:

$$[S_1][\sigma_1] = [-S_2] \tag{7}$$

$$\begin{bmatrix} S_1 S_2 \\ S_2 S_3 \end{bmatrix} \begin{bmatrix} \sigma_2 \\ \sigma_1 \end{bmatrix} = \begin{bmatrix} -S_3 \\ -S_4 \end{bmatrix} \tag{8}$$

Define D_2 as the determinant of the coefficient matrix in Eq. (8), that is $S_1S_3 - S_2^2$, and D_3 as the determinant of the 3×3 coefficient matrix in Eq. (6), which is

$$S_1S_3S_5 + S_2S_3S_4 + S_2S_3S_4 - S_3^3 - S_1S_4^2 - S_2^2S_5$$

Now, tests of D_2 and D_3 can be used to determine how many errors have occurred, and therefore, which set of equations should be used to solve for the σ terms. For example, if only one error has occurred, D_2 and D_3 will equal zero, and therefore, the Eqs. (6) and (8) will be indeterminate, and Eq. (7) is to be used.

Once the σ values have been determined, the error locator polynomial $\sigma(x)$ is formed and its roots obtained. The Chien search, already described for binary codes, can be used. The roots of $\sigma(x)$ are the error locator values, the X values. Once the X values have been determined, they are inserted into the syndrome equations, Eq. (3), which are then solved as linear equations for the error values, the Y terms. The steps in a direct solution decoding algorithm are described next using an example.

We describe the use of Peterson's direct solution method for decoding the (63,57) RS code defined on a 64-ary alphabet in more detail. Since the code has distance 7, it can be used to correct up to three errors in a received word or to correct combinations of l errors and s erasures such that $2l + s < 7$. In this discussion, however, we confine our attention to error-correction decoding. Combined errors-and-erasures decoding is treated later.

Since for an RS code, the symbol field and the locator field are the same, all computations for decoding are done in the field GF(64). Furthermore, since GF(64) is a field of characteristic 2, addition and subtraction are identical operations, which means, for example, that in determining the coefficients of $\sigma(x)$ and calculating error values, the minus signs can be replaced with plus signs. The 64 elements of the field may be represented conveniently as binary 6-tuples. Addition is then implemented with modulo-2 addition, applied bit by bit. For implementation in a processor, finite field multiplication and division are conveniently done with logarithm and antilogarithm tables and table lookup routines.

An error-correction decoder for the (63,57) RS code can be implemented as follows. Let the polynomial $r(x)$ represent the received word, where the high-order terms correspond to the information symbols and the low-order terms to the check symbols. The steps in the decoding process are

1. Compute the syndrome values S_k, $1 \le k \le 6$, where

$$S_k = r(\alpha^k) = \{\ldots[(r_{62}\alpha^k + r_{61})\alpha^k + r_{60}]\alpha^k + \cdots\}\alpha^k + r_0, \quad 1 \le k \le 6$$

2. Determine the number of errors in the received word:
   a. If $S_k = 0$, $1 \le k \le 6$, the received word is a codeword, and no further processing is necessary.
   b. If $D_3 = S_1S_3S_5 + S_1S_4^2 + S_2^2S_5 + S_3^3 \neq 0$, assume that three errors are present.
   c. If $D_3 = 0$ and $D_2 = S_1S_3 + S_2^2 \neq 0$, assume that two errors are present.
   d. If $D_2 = D_3 = 0$ and $S_1 \neq 0$, assume that one error is present.
3. Compute the coefficients of the error-locator polynomial:
   a. If three errors are present, compute

$$\sigma_1 = \frac{1}{D_3}[S_1S_3S_6 + S_1S_4S_5 + S_2^2S_6 + S_2S_3S_5 + S_2S_4^2 + S_3^2S_4]$$

$$\sigma_2 = \frac{1}{D_3}[S_1S_4S_6 + S_1S_5^2 + S_2S_3S_6 + S_2S_4S_5 + S_3^2S_5 + S_3S_4^2]$$

$$\sigma_3 = \frac{1}{D_3}[S_2S_4S_6 + S_2S_5^2 + S_3^2S_6 + S_4^3]$$

   b. If two errors are present, compute

$$\sigma_1 = \frac{1}{D_2}[S_1S_4 + S_2S_3]$$

$$\sigma_2 = \frac{1}{D_2}[S_2S_4 + S_3^2]$$

   c. If one error is present, compute

$$\sigma_1 = X_1 = \frac{S_2}{S_1}$$

4. If three errors are indicated in step 3, find (using the Chien search) the roots of the polynomial $\sigma(x)$, where

$$\sigma(x) = x^3 + \sigma_1x^2 + \sigma_2x + \sigma_3$$

   If two errors are indicated, find the roots of

$$\sigma(x) = x^2 + \sigma_1x + \sigma_2$$

   Of course, in the case of three errors, three distinct roots of $\sigma(x)$ must be found, and for the two-error case, $\sigma(x)$ must have two distinct roots. If the correct number of roots is not found, error detection is announced.
5. After the error locators are determined, the error values are obtained by solving the syndrome equations.
   a. *One-error case*:

$$Y_1 = \frac{S_1^2}{S_2}$$

   b. *Two-error case*:

$$Y_1 = \frac{S_1X_2 + S_2}{X_1X_2 + X_1^2}$$

$$Y_2 = \frac{S_1X_1 + S_2}{X_1X_2 + X_2^2}$$

c. *Three-error case*:
Let

$$C = X_1X_2^2X_3^3 + X_1^3X_2X_3^2 + X_1^2X_2^3X_3 + X_1^3X_2^2X_3 + X_1X_2^3X_3^2 + X_1^2X_2X_3^3$$

Then

$$Y_1 = \frac{1}{C}\left[S_1X_2^2X_3^3 + S_2X_2^3X_3 + S_3X_2X_3^2 + S_1X_2^3X_3^2 + S_2X_2X_3^3 + S_3X_2^2X_3\right]$$

$$Y_2 = \frac{1}{C}\left[S_1X_3^2X_1^3 + S_2X_1X_3^3 + S_3X_1^2X_3 + S_1X_1^2X_3^3 + S_2X_1^3X_3^3 + S_3X_1X_3^2\right]$$

$$Y_3 = \frac{1}{C}\left[S_1X_1^2X_2^3 + S_2X_1^3X_2 + S_3X_1X_2^2 + S_1X_1^3X_2^2 + S_2X_1X_2^3 + S_3X_1^2X_2\right]$$

It should be noted that when the denominators in the expressions for Y_1, Y_2, and Y_3 are written out, the expressions can be simplified.

6. Correct the received word by adding the computed error values to the corresponding symbols received in positions identified as error locations.
7. Compute the syndrome of the corrected word, and if it is not zero, announce error detection.

The correction of both errors and erasures is discussed in Section 3.3. We first present an efficient iterative decoding algorithm for correction of errors in nonbinary BCH and RS codes.

3.2. The Massey–Berlekamp Algorithm

For correction of moderate to large numbers of errors with a nonbinary BCH or RS code, Peterson's direct method of solving for the coefficients of $\sigma(x)$ from the syndrome values becomes cumbersome and inefficient due to the large number of multiplications and divisions that must be performed. Instead, it is preferable to use either of two algorithms developed by Berlekamp [3] and by Massey [4] for solution of Newton's identities. The two algorithms are closely related and are often referred to as one procedure, the *Massey–Berlekamp algorithm*. The approach used by Massey in presenting the technique is particularly instructive, and thus we shall follow Massey closely here. Berlekamp's formulation with simplifications applicable to decoding binary codes was described for binary codes. Both Massey's and Berlekamp's versions of the algorithm can be used for binary and nonbinary codes.

We let $m_0 = 1$ and return to the error-locator polynomial $\sigma(x)$ in Eq. (5). Then, substituting an error locator X_j for x, we obtain

$$X_j^t + \sigma_1 X_j^{t-1} + \cdots + \sigma_t = 0, \quad j = 1, 2, \ldots, t \tag{9}$$

Multiplying Eq. (9) by X_j^k and summing for $j = 1, 2, \ldots, t$, we obtain

$$S_{k+t} + \sigma_1 S_{k+t-1} + \cdots + \sigma_t S_k = 0, \quad k = 1, 2, \ldots \tag{10}$$

which are again Newton's identities. Letting $j = k + t$, we obtain

$$S_j + \sigma_1 S_{j-1} + \cdots + \sigma_t S_{j-t} = 0, \quad j = t+1, t+2, \ldots \tag{11}$$

With Newton's identities written in this form, one can recognize that they describe the operation of a linear feedback shift register (FSR) with initial states $S_1, S_2, \ldots, S_t$ and tap connections given by $C_i = \sigma_i$. A diagram of a linear FSR is shown in Fig. 2. From the figure, it is seen that the FSR implements the equations.

$$S_j = -C_1S_{j-1} - C_2S_{j-2} - \cdots - C_tS_{j-t}, \quad j = t+1, t+2, \ldots \tag{12}$$

or

$$S_j + C_1S_{j-1} + \cdots + C_tS_{j-t} = 0, \quad j = t+1, t+2, \ldots \tag{13}$$

With $C_i = \sigma_i$, the correspondence with Eq. (11) is immediate.

Recognizing the relationship just obtained, Massey established the equivalence between the problem of determining the coefficients of the error locator polynomial from the syndrome values and that of synthesizing an FSR with minimum length that generates the given sequence of syndromes. We shall provide a rationale for this in the following.

We define the *connection polynomial* as a convenient representation for the coefficients of the syndrome values in Eq. (13):

$$C(x) = 1 + C_1x + C_2x^2 + \cdots + C_tx^t \tag{14}$$

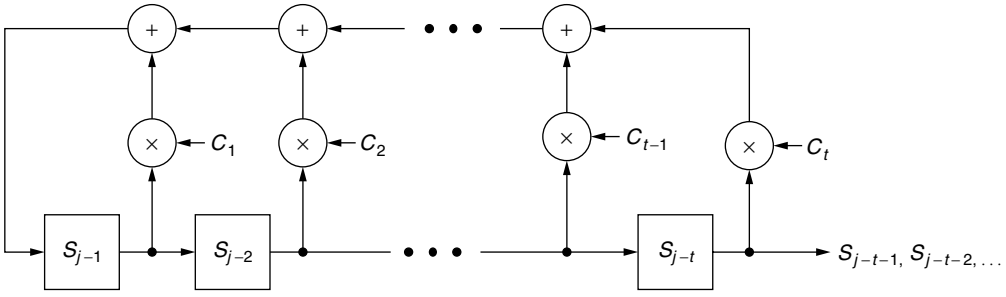


Figure 2. Linear feedback shift register for generating a sequence of syndrome values.

We now state that the problem of determining the error locator polynomial $\sigma(x)$ is equivalent to that of determining a connection polynomial $C(x)$ for a linear FSR that generates the syndrome values $S_{t+1}, S_{t+2}, \ldots$, given that the FSR is initialized with $S_1, S_2, \ldots, S_t$.

Note that in Eqs. (11)–(14), as well as in Fig. 2, the assumed length of the FSR is t stages, where t is the error correction limit of the code. However, the iterative algorithm is designed to correct l errors, where $l \leq t$. The number of errors l is not known at the start of decoding and is determined as part of the decoding procedure.

Without delving into the details of the properties of FSRs and the sequences that they generate, we simply point out that for a given sequence of syndrome values, a determinable number of connection polynomials of various lengths will generate the syndromes. This corresponds directly to the fact that, in general, a number of error patterns can account for a given set of syndrome values. However, the task of bounded-distance decoding is to find the lowest-weight error pattern corresponding to the given syndrome. Therefore, in the FSR synthesis problem, we seek the lowest degree connection polynomial $C(x)$ that generates the syndrome. In his 1969 paper [4], Massey described an algorithm that finds the minimal-length FSR. He further showed that given an error pattern of weight $l \leq t$, the algorithm yields the connection polynomial that uniquely corresponds to the correct error-locator polynomial. Massey's algorithm is often called the *FSR synthesis algorithm*. Shift register sequences are also described in [5].

Before describing the FSR synthesis algorithm in detail, we outline the procedure as follows. The FSR algorithm synthesizes the minimal-length shift register with an iterative routine that begins by postulating the shortest possible shift register and then attempts to generate the entire sequence of given syndrome values in order. The actual syndrome sequence is repeatedly compared with the output of the postulated FSR until either the entire sequence of given syndrome values is reproduced or a discrepancy is encountered. At the first discrepancy, the postulated FSR is modified with a specified rule, and the sequence generation is restarted and continued until all the remaining syndromes are reproduced or another discrepancy is encountered, and so forth. The modification rule is designed to ensure that for a correctable error pattern, the FSR eventually settles into the correct configuration. The FSR synthesis algorithm is described in detail below.

The Massey FSR Synthesis Algorithm

0. *Compute syndrome values*: S_n, $1 \leq n \leq d-1$.
1. *Initialize algorithm variables*:

   Let $C(x) = 1 \quad D(x) = x$

   $L = 0 \qquad n = 1$

2. *Take in new syndrome value and compute discrepancy*:

$$\delta = S_n + \sum_{i=1}^{L} C_i S_{n-i}$$

3. *Test discrepancy*. If $\delta = 0$, go to step 8; otherwise, go to step 4.
4. *Modify connection polynomial*. Let $C^*(x) = C(x) - \delta D(x)$.
5. *Test register length*. If $2L \geq n$, go to step 7 (i.e., do not extend register); otherwise, go to step 6.
6. *Change register length and update correction term*. Let $L = n - L$ and $D(x) = C(x)/\delta$.
7. *Update connection polynomial*. Let $C(x) = C^*(x)$.
8. *Update correction term*. Let $D(x) = xD(x)$
9. *Update syndrome counter*. Let $n = n + 1$.
10. *Test syndrome count*. If $n < d$, go to step 2; otherwise, stop.

In the algorithm, $C(x)$ is the FSR connection polynomial. The algorithm is designed to expediently build up the polynomial $C(x)$ of lowest degree that generates the given sequence of syndromes, $S_1, S_2, \ldots, S_{d-1}$. The connection polynomial is first initialized to its simplest possible form, $C(x) = 1$, and is subsequently modified as needed to correctly reproduce the syndrome values in sequence. The other polynomial formed in the algorithm, $D(x)$, is a correction term that is used to modify $C(x)$ at each iteration in which a discrepancy is encountered between a generated value and the corresponding syndrome value. The syndromes are examined by the algorithm in sequence, one in each iteration. At each iteration, the discrepancy δ, the difference between the newly entered syndrome value and the value generated by the FSR in the corresponding sequence position, is computed, using the connection polynomial $C(x)$ as it was structured at the end of the previous iteration. Note that δ is defined in such a way that at the first entry into step 2, it is given the value of the first syndrome S_1, even though there are no previous syndrome values from which S_1 could be generated. At each appearance of a nonzero value for δ, the connection polynomial is modified using the computed value of δ and the correction term (step 4). The formation and use of the correction term is the most important part of the algorithm. One reason is that in addition to zeroing out the encountered discrepancy, the modification of $C(x)$ is such that the new $C(x)$ also correctly generates all the previous syndrome values. This obviates the necessity of having to reexamine previous syndromes each time $C(x)$ is modified, and provides an algorithm in which the number of computations required per decoding is a linear function of l rather than some geometric function. Another important characteristic of the polynomial modification procedure is that it accomplishes the needed sequence modification with the smallest possible increase in the degree of the connection polynomial. The other variable used in the algorithm is L, which is the current length of the FSR. If the algorithm terminates with an FSR connection polynomial of degree greater than t, that is, $2L > d-1$, then we are not assured that the corresponding error locator polynomial is correct, and error detection is announced.

As an example, consider the case of the (31,25) RS code on GF(32), with $m_0 = 1$, for which codewords all

have the six consecutive roots $\alpha, \alpha^2, \ldots, \alpha^6$. Let the all-zeros codeword be transmitted, and assume the received word

$$000\alpha^7 00000000\alpha^3 0000000\alpha^{22} 0000000000$$

Thus, we have

$$r(x) = \alpha^7 x^3 + \alpha^3 x^{12} + \alpha^{22} x^{20}$$

To represent elements in GF(32), we use the primitive polynomial $p(x) = x^5 + x^2 + 1$. Then the syndrome values are as follows:

$$\begin{aligned} S_1 &= r(\alpha) = \alpha^{29} \\ S_2 &= r(\alpha^2) = \alpha^{28} \\ S_3 &= r(\alpha^3) = \alpha^9 \\ S_4 &= r(\alpha^4) = \alpha^4 \\ S_5 &= r(\alpha^5) = \alpha^{24} \\ S_6 &= r(\alpha^6) = \alpha^{19} \end{aligned}$$

We next use the FSR synthesis algorithm to find the shortest connection polynomial $C(x)$ that generates the six syndrome values in order. The iterative solution is summarized as follows:

n	S_n	$C(x)$	δ	L
1	α^{29}	1	α^{29}	0
2	α^{28}	$1 + \alpha^{29}x$	α^{14}	1
3	α^9	$1 + \alpha^{30}x$	α^{10}	1
4	α^4	$1 + \alpha^{30}x + \alpha^{12}x^2$	α^{11}	2
5	α^{24}	$1 + \alpha^4 x + \alpha^{23}x^2$	α^{10}	2
6	α^{19}	$1 + \alpha^4 x + \alpha^{12}x^2 + \alpha^{30}x^3$	α^{19}	3
7		$1 + \alpha^6 x + \alpha^{30}x^2 + \alpha^4 x^3$	(STOP)	

Thus, the minimal-length connection polynomial is found to be

$$C(x) = 1 + \alpha^6 x + \alpha^{30}x^2 + \alpha^4 x^3$$

and with $\sigma_i = C_i$, we can write the error locator polynomial as

$$\sigma(x) = x^3 + \alpha^6 x^2 + \alpha^{30} x + \alpha^4$$

The three roots of $\sigma(x)$, the error locator numbers, are found to be

$$X_1 = \alpha^3, \quad X_2 = \alpha^{12}, \quad X_3 = \alpha^{20}$$

which point to errors in the 4th, 13th, and 21st symbol positions.

The error magnitudes, Y_1, Y_2, and Y_3 are now computed using the equations shown previously for the three-error case. The reader may verify that the computations yield

$$Y_1 = \alpha^7, \quad Y_2 = \alpha^3, \quad Y_3 = \alpha^{22}$$

Finally, error correction is completed by subtracting [or adding, in GF(32)] the error values from the corresponding received symbols, which yields the all-zeros word as the corrected codeword.

3.3. Errors-and-Erasures Decoding

If some procedure is being used to erase unreliable symbols in a received word, then the function of the decoder is to fill in the proper values of the erasures and at the same time locate and correct any unknown errors. We recall from earlier discussions that a code of minimum distance d is capable of correcting any pattern of l errors and s erasures as long as $2l + s < d$. We now outline an efficient procedure for simultaneous errors-and-erasures decoding, which was suggested by Forney [6] for nonbinary BCH codes. First, we define the *erasure locator polynomial* $\sigma'(z)$, which is the polynomial of degree s whose roots are the erasure locators:

$$\begin{aligned} \sigma'(z) &= \prod_{i=1}^{s} (z + Z_i) \\ &= \sigma_0' z^s + \sigma_1' z^{s-1} + \cdots + \sigma_s' \end{aligned} \tag{15}$$

where Z_i gives the location of the ith erasure.

It should be noted that $\sigma'(z)$ is written in much the same form as the error locator polynomial, Eq. (5), except that for notational convenience we have given the term of highest degree the coefficient σ_0' even though it always has value 1. Since, by definition, the erasure location values are known, the coefficients of $\sigma'(z)$ may be computed directly. We also assume use of a primitive code with $m_0 = 1$.

Combined errors-and-erasures decoding begins, as does error correction decoding, with calculation of the syndrome values, which are

$$S_k = \sum_{i=1}^{n} r_i \alpha^{ik}, \quad 1 \le k \le d - 1$$

where we denote $d - 1$ rather than $2t$ syndrome values, to allow for both odd and even values of d. The reader may well ask what values should be assigned to the erasures for the syndrome calculation, but it will be seen shortly that these values are immaterial to the decoding procedure. As a practical matter, it is usually advantageous to assign zeros for all the erasure values.

To take account of the known erasure-location information in forming the syndromes, Forney introduced a linear transformation on the syndromes:

$$T_i = \sum_{j=0}^{s} \sigma_j' S_{i+s+1-j}, \quad 0 \le i \le d - s - 2 \tag{16}$$

The T values are called the *modified syndromes*. Notice that there are s fewer T than S symbols. Thus, if one symbol is erased, the $d - 1$ original syndromes are transformed by Eq. (16) into $d - 2$ modified syndromes, and so forth. We shall see how this transformation lets us establish a useful recursion among the T symbols.

Let us assume the presence of l errors and s erasures. Let the errors be at locations $X_1, X_2, \ldots, X_l$ and have values

$Y_1, Y_2, \ldots, Y_l$. Let the known erasure locations be denoted by $Z_1, Z_2, \ldots, Z_s$, and let $D_1, D_2, \ldots, D_s$ designate the erasure-discrepancy values, that is, the difference between the correct symbol values and the values arbitrarily assigned before the syndromes are computed. We can now express the syndromes as

$$S_k = \sum_{m=1}^{l} Y_m X_m^k + \sum_{n=1}^{s} D_n Z_n^k, \; 1 \le k \le d-1$$

From Eq. (16), we write the modified syndromes as

$$T_i = \sum_{j=0}^{s} \sigma_i' \left[\sum_{m=1}^{l} Y_m X_m^{i+s+1-j} + \sum_{n=1}^{s} D_n Z_n^{i+s+1-j} \right], \quad 0 \le i \le d-s-2$$

or

$$T_i = \sum_{m=1}^{l} Y_m X_m^{i+1} \sum_{j=0}^{s} \sigma_j' X_m^{s-j} + \sum_{n=1}^{s} D_n Z_n^{i+1} \sum_{j=0}^{s} \sigma_j' Z_n^{s-j} \quad (17)$$

However, from Eq. (15), we see that the second summation in the last term of Eq. (17) is the erasure locator polynomial evaluated at a root Z_n, which equals zero. Further, we recognize from Eq. (15) that the second summation in the first term of the right-hand side of Eq. (17) is simply the erasure locator polynomial evaluated at the error location X_m, which we write as $\sigma'(X_m)$. Therefore, if we define a new quantity E_m as

$$E_m = Y_m X_m \sigma'(X_m) \quad (18)$$

we can rewrite Eq. (17) as

$$T_i = \sum_{m=1}^{l} E_m X_m^i, \quad 0 \le i \le d-s-2 \quad (19)$$

What is important to note here is that Eq. (19) defines the modified syndrome values in a manner essentially the same as that in which the ordinary syndrome values are defined for l-error correction, for example, by Eq. (3). Thus, we see that for the simultaneous decoding of l errors and s erasures, the transformation in Eq. (16) has the effect of folding the known erasure locators into the original syndromes in such a way that we preserve the form of the syndrome equations in terms of the error locators. Now, by starting with Eq. (19) as the formulation of a new decoding problem, where l error locators X_m are to be determined, we can perform decoding with much the same overall procedure as is used for the case of ordinary error correction. That is, we first find the l error locators from the T values, and then compute the values of code symbols in the $l+$ s error and erasure locations.

If Peterson's direct solution method is used, error locator polynomial coefficients are computed from the T values in the same way as they are computed from the S values in the earlier discussion of errors-only decoding. After solving for the roots of $\sigma(x)$, any $l+s$ of the syndrome equations can be used as a set of simultaneous linear equations to solve for the Y and D values.

Alternatively, the Massey FSR synthesis technique may be applied in almost the same way as for ordinary error correction. That is, using Eq. (19) instead of Eq. (3), we treat the relationship of the T values to the σ values in a manner that exactly parallels the discussion in Eqs. (9)–(14), developing along the way a recursion relationship for the T values equivalent to Eq. (11):

$$T_j + \sigma_1 T_{j-1} + \cdots + \sigma_l T_{j-l} = 0, \quad j = l, l+1, \ldots$$

Thus, the problem of finding the coefficients of the error locator polynomial can be formulated again as an FSR synthesis problem, where the FSR must now be synthesized to generate a given sequence of modified syndrome values $\{T_j\}$ rather than original syndrome values $\{S_k\}$.

Once the error-locator polynomial is obtained, the roots are found efficiently using the Chien search. The l error locators, taken together with the s known erasure locators, in effect constitute $l+s$ erasures whose values are to be computed from the original syndromes. This can be done by solving $l+s$ syndrome equations, as in the direct method. However, a more efficient method of determining the erasure values has also been given by Forney [6]. Although we mention this method as part of the errors-and-erasures decoding procedure, it is also applicable in the case of errors-only decoding, since it is applied at the point in decoding where all of the unknown errors have been located. The suggested erasure-filling procedure is now described.

Let us denote the given erasure locators and computed error locators together by $Z_1, Z_2, \ldots, Z_{l+s}$. Now consider deleting Z_1 from the set of $l+s$ erasure locators, and forming the erasure locator polynomial ${}_1\sigma(z)$, which has as roots the remaining $l+s-1$ locators. Next, we calculate the coefficients of

$${}_1\sigma(z) = {}_1\sigma_0 z^{l+s-1} + {}_1\sigma_1 z^{l+s-2} + \cdots + {}_1\sigma_{l+s-1}$$

Then the erasure correction value, to be subtracted from the received or assigned value in the location Z_1, is given by

$$D_1 = \frac{\sum_{k=0}^{l+s-1} {}_1\sigma_k S_{l+s-k}}{\sum_{j=0}^{l+s-1} {}_1\sigma_j Z_1^{l+s-j}}$$

or in general by

$$D_i = \frac{\sum_{k=0}^{l+s-1} {}_i\sigma_k S_{l+s-k}}{\sum_{j=0}^{l+s-1} {}_i\sigma_j Z_i^{l+s-j}}, \quad 1 \le i \le l+s \quad (20)$$

By deleting one erasure at a time, all erasure values are calculated in turn by Eq. (20). Another procedure requiring even fewer computations can also be used. If,

after computing D_1, the syndrome values are modified using

$$S'_k = S_k + D_1 Z_1^k$$

it is only necessary to form an $(l+s-2)$-order erasure locator polynomial, with coefficients ${}_2\sigma_1, {}_2\sigma_2, \ldots, {}_2\sigma_{l+s-2}$, in order to find D_2 from Eq. (20), and so forth.

We now summarize the procedure for errors-and-erasures decoding, assuming use of the FSR synthesis algorithm, as follows:

1. Inspect the received word for erasures, assign erasure values (e. g., all 0s) and compute the syndrome.
   a. If $s > d-1$, declare the word undecodable.
   b. Otherwise, compute the syndrome values $S_1, S_2, \ldots, S_{d-1}$. If all syndromes are zero, the received word is a valid codeword, and no further processing is to be done.
2. If no symbols have been erased ($s = 0$), follow the procedure for errors-only decoding.
3. Compute the modified syndrome (if necessary).
   a. If $s = d-1$, go to step 6.
   b. If $0 < s < d-1$, compute the modified syndrome values using Eq. (16).
4. Determine the number of errors in the received word.
   a. If all $T_i = 0$, $0 \le i \le d-s-2$, assume that no errors are present and go to step 6.
   b. If some $T_i \neq 0$, use the FSR synthesis algorithm to find $\sigma_1, \sigma_2, \ldots, \sigma_l$, the coefficients of the error locator polynomial $\sigma(x)$.
5. Determine the error locators, the roots of $\sigma(x)$, using the Chien search. Put the l computed error locators together with the given s erasure locators to make up the new set of erasure locators $Z_1, Z_2, \ldots, Z_{l+s}$.
6. Compute the $l+s$ erasure magnitudes, using Eq. (20) or the more efficient procedure discussed immediately following Eq. (20).

4. FINAL COMMENTS

Errors and erasures decoding is the simplest form of *soft-decision decoding*. More complex techniques have been proposed for decoding Reed Solomon codes, for example, generalized minimum distance (GMD) decoding [7]. GMD is a technique that uses a sequence of decode attempts to find the most likely transmitted codeword. Each time a word is received, a trial decode list is executed. First, errors-only decoding is attempted and the outcome is recorded. Then errors and erasures decoding attempts are executed where we first erase the least reliable received symbol, the one with the smallest matched-filter output, and execute a decode attempt. Next, we erase the three least reliable symbols, then five, and so on up to a decode attempt in which the $d-1$ least reliable received symbols are erased where d is the minimum distance of the code. Clearly, this procedure can result in more than one decoded output, which we resolve by choosing the candidate codeword with highest likelihood.

Other soft-decision decoding schemes for RS codes are described by Cooper [8]. However, efficient soft-decision decoding of Reed–Solomon codes is still considered an open issue. As with the binary codes, what is needed is a way to decode beyond the code's minimum distance with an efficient soft-decision algorithm.

In this and the article on binary BCH coding, binary BCH codes and Reed–Solomon codes have been considered and the commonly used encoding and decoding algorithms have been described. The presentation closely follows the early developments as they were originally published. More recent treatments have proved useful as well, providing new insights into the underlying fundamentals. In some cases, more efficient encoding and decoding algorithms have resulted. A good example is Blahut's description of binary and nonbinary BCH codes in terms of Fourier transforms [9].

To conclude this article, we note again that binary BCH codes and RS codes have been incorporated into many communication systems. The wide ranges of block lengths and error correction power afforded by these codes have enabled designers to tailor solutions for particular applications, and provide significant performance gains relative to uncoded transmission. Given the continuing growth and advancement of the digital communications field, the number of applications for BCH and RS codes will certainly increase.

BIOGRAPHIES

Arnold M. Michelson received his BSEE degree from the Johns Hopkins University, his MSEE from the University of Rochester, New York, and he did further graduate work at the Polytechnic Institute of Brooklyn, New York. In 1968, Mr. Michelson joined Sylvania Electric Products, which later became GTE Government Systems. At Sylvania and GTE he worked on the development and implementation of advanced communication techniques, including error-control coding for military applications. That work focused primarily on long-wave communications. Since 2000, he has been with the Raytheon Company where he is involved in the development of high-performance coding techniques for military and commercial satellite applications. In 1997, Mr. Michelson received GTE's Leslie H. Warner Technical Achievement Award, and, in 2002, Raytheon's Excellence in Technology Award, Distinguished Level.

Allen H. Levesque received his BSEE degree from Worcester Polytechnic Institute in 1959 and his MSEE and PhDEE degrees from Yale University in 1960 and 1965, respectively. Following completion of his graduate studies, he joined the GTE Corporation, where, over a 36-year career, he worked on and led a variety of digital communications research and development projects, with application to both defense and commercial systems. Much of his early work concerned applications of error-control coding techniques in radio networks. For the past decade, his work has concentrated on mobile and wireless communications networks. In early 1999, he retired from GTE Laboratories to begin and independent consulting

practice and to take a part-time teaching and research position at WPI. He currently teaches graduate courses in modulation and coding and is a member of WPI's Center for Wireless Information Network Studies. His areas of research interest include communication theory and techniques, communication networks, wireless communications, spread-spectrum, secure communications, and digital signal processing. He has published numerous journal and conference papers, and coauthored two books, as well as chapters in several communications handbooks. He is a life fellow of the IEEE and a Registered Professional Engineer in the Commonwealth of Massachusetts.

BIBLIOGRAPHY

1. I. S. Reed and G. Solomon, Polynomial codes over certain finite fields, *J. SIAM* **8**: 300–304 (1960).
2. R. C. Singleton, Maximum distance q-nary codes, *IEEE Trans. Inform. Theory* **IT-10**: 116–118 (1964).
3. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
4. J. L. Massey, Shift register synthesis and BCH decoding, *IEEE Trans. Inform. Theory* **IT-15**: 122–127 (1969).
5. S. W. Golomb, *Shift Register Sequences*, Holden Day, San Francisco, 1967 (revised ed., Aegean Park Press, Laguna Hills, CA, 1982).
6. G. D. Forney, Jr., On decoding BCH codes, *IEEEE Trans. Inform. Theory* **IT-9**: 64–74 (1963).
7. G. D. Forney, Jr., Generalized minimum distance decoding, *IEEE Trans. Inform. Theory* **12**: 125–131 (1966).
8. S. B. Wicker and V. K. Bhargava, eds., *Reed–Solomon Codes and Their Applications*, IEEE Press, New York, 1994.
9. R. E. Blahut, *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, MA, 1983.

BISDN (BROADBAND INTEGRATED SERVICES DIGITAL NETWORK)

ENDER AYANOGLU
University of California
Irivne, California

NAIL AKAR
Bilkent University
Ankara, Turkey

1. BROADBAND ISDN

We first outline the history of the BISDN vision and then move on to the ATM technology that is envisioned to fulfill this vision.

1.1. History of BISDN

Shortly after the invention of the telephone by A. G. Bell in 1876, means to interconnect or network telephones at different locations were devised. Within only 2 years, the first switching center was built [4]. In the United States during the twentieth century, a public company, the Bell System and its parent, AT&T, emerged as the national provider of telephony services. The fundamental principle, formulated by AT&T president T. Vail in 1907, was that the telephone would operate most efficiently as a monopoly providing universal service, by nature of its technology. The U.S. government accepted this principle in 1913. The Bell System made steady progress toward its goal of universal service, which came in the 1920s and 1930s to mean that everyone should have a telephone. Percentage of American households with telephone service reached 50% in 1945, 70% in 1955, and 90% in 1969. This network was based on analog technology for transmission, signaling, and switching.

The Bell System studied digital telephony, first starting from its theoretical principles during the late 1940s. Most of the principles of digital telephony, such as theory of sampling, theory of quantization, and fundamental limits in information transfer, were invented or perfected by Bell System scientists such as H. Nyquist, J. R. Price, S. P. Lloyd, and C. E. Shannon in the late 1940s. Parallel with this progress in theory was a fundamental breakthrough in device technology known as the *transistor*, introduced, again by the Bell System, in 1948. The transistor would make the digital telephony revolution possible, while many years later, powerful integrated circuits would spark the dream of BISDN.

Digitization of the telephony network was useful since it provided a number of advantages:

- Ease of multiplexing
- Ease of signaling
- Integration of transmission and switching
- Increased noise tolerance
- Signal regeneration
- Accommodation of other services
- Performance monitoring
- Ease of encryption

The first deployment of digital transmission was in 1962 by the Bell System, while the first digital commercial microwave system was deployed in Japan in 1968. Research on digital switching was initiated in 1959 by Bell Labs. The first deployment of a digital switch in the public network was in 1970 in France while in the United States, the Bell System deployed an electronic switch known as 4ESS in 1976 [4].

CCITT (Comité Consultatif International de Télégraphique et Téléphonique, or Consultative Committee for International Telegraph and Telephone), is a committee of the International Telecommunications Union (ITU), which is a specialized agency of the United Nations. ITU was originally established after the invention of telegraphy in 1865 and became a specialized agency of the United Nations in 1947, shortly after the formation of the United Nations. Similar to ITU, CCITT was originally established as a standardization organization in the field of telegraphy, in 1925. In 1993, standardization aspects of CCITT and those of the sister radio standardization committee, CCIRR, were unified

under the name ITU-T (International Telecommunications Union — Telecommunication Standardization Sector). Members of ITU-T are governments. ITU-T is currently organized into 13 study groups that prepare standards, called Recommendations. There are 25 Series of Recommendations (A–V, X–Z). Work within ITU-T is conducted in 4-year cycles.

In 1968, CCITT established Special Study Group D to study the use of digital technology in the telephone network. This study group established 4-year study periods beginning with 1969. The first title of the group was "Planning of Digital Systems." By 1977, the emphasis of the study group was on overall aspects of integrated digital networks and integration of services. As of 1989, the title of the study shifted to "General Aspects of Integrated Services Digital Networks." The concept of an integrated services digital network was formulated in 1972 as one in which "the same digital services and digital paths are used to establish for different services such as telephony and data" [29]. The first ISDN standard was published in 1970, under the title "G.705 Integrated Services Digital Network (ISDN)." Although this first document of an ISDN standard is in the Series G Recommendations, most of the ISDN standards are in the Series I Recommendations, with some also in G, O, Q, and X Series Recommendations.

Three types of ISDN services are defined within the ISDN Recommendation I.200:

- Bearer services
- Teleservices
- Supplementary services

Bearer services (I.140) provide the means to convey information in the form of speech, data, video, and other forms of communication between users. There is a common transport rate for bearer services: it is the 64 kbps (kilobits per second) rate of digital telephony. Various bearer services are defined as multiples of this basic 64-kbps service, for example, 64, 2×64, 384, 1536, and 1920 kbps [29]. Teleservices cover user applications and are specified in I.241 as telephony, teletex, telefax, mixed mode, videotex, and telex. Supplementary services are defined in I.250. These services are related to number identification (such as calling line identification), call offering (such as call transfer, call forwarding, and call deflection), call completion (such as call waiting and call hold), multiparty (such as conference or three-party calling), community of interest (such as a closed user group), charging (such as credit card charging), and additional information transfer (such as the use of the ISDN signaling channel for user-to-user data transfer).

Toward the end of 1980s and almost two decades after the first study group on ISDN was formed at the CCITT, ISDN was still not being deployed by service providers at a commercial scale, especially in the United States. It is important to review the reasons for this absence of activity. ISDN required digitization of both the telephony network and the subscriber loop (connection between a residence or a business and the central office of the telecommunications service provider). While the network was becoming digital, and doing so involved economies of scale (and thus was relatively inexpensive), making the subscriber loop digital required replacement of the subscriber front end equipment at the central office. This was a labor-intensive, expensive process. In addition, there was no compelling push from consumers demanding ISDN. With the network becoming digital, the quality and reliability advantages of voice transmission were achieved. In addition, it was possible to offer supplementary services (such as caller ID and call waiting) as defined by ISDN Recommendations without making the subscriber loop digital. At the time, modem technology enabled data transmission over the subscriber loop at rates up to ~30 kbps, and that was sufficient for most of the available residential data services available (which were text-based). Business data communications needs were restricted to large businesses. These needs were being served with dedicated digital lines (T1 lines) at speeds of 1.5 Mbps in the United States. Although these lines were very expensive, the market for them was relatively small. In addition, it was becoming clear that in order to serve any future ISDN service needs, ISDN transmission speeds would not suffice and packet switching was going to become necessary. At the time, some overestimates were made as to the needed transmission speeds. For example, it was considered that entertainment video was one of the services that service providers would offer on such an integrated network and that the required transmission speeds for these services were in excess of 100 Mbps. ISDN was certainly insufficient to provide these speeds, and its packet switching recommendations were not yet developed.

In 1988, CCITT issued a set of Recommendations for ISDN, under the general name of "broadband aspects of ISDN" (I.113: *Vocabulary of Terms for Broadband Aspects of ISDN*, and I.121: *Broadband Aspects of ISDN*). This was a time when packet switching was proved in the Internet (although the Internet was still a research network), there was increased activity in video coding within the contexts of HDTV (high-definition television) and MPEG (video coding specification by the Moving Picture Experts Group), voice compression was beginning to achieve acceptable voice quality at rates around 8 kb/s, and first residential data access applications were appearing in the context of accessing the office computer and electronic bulletin boards. Consequently, telecommunications industry representatives came to the conclusion that a need for broadband services in the telecommunication network was imminent. Since ISDN was not capable of answering high-speed and packet-based service needs of such services, the concept of BISDN was deemed necessary. Aiding in this process was the availability of high-speed transmission, switching, and signal processing technologies. It became clear that even higher processing speeds would become available in the near future (e.g., the fact that the speed of processing doubles every 1.5 years, also known as *Moore's law*). CCITT considered these signs so important that the usual 4-year cycle of a study group to issue Recommendations was considered too long and an interim set of broadband ISDN (BISDN) Recommendations were

first issued in 1990. It should be emphasized at this point that for the telecommunications industry, and specifically for the service providers, the vision of B-ISDN involves the integration of voice, video, and data services *end-to-end* and with quality-of-service (QoS) guarantees.

1.2. ATM Fundamentals

The concept of asynchronous transfer mode (ATM) was first unveiled in an international meeting in 1987 by J. P. Coudreuse of CNET, France [9]. The basic goal of ATM was to define a networking technology around the basic idea of fast packet switching. In doing so, it was recognized that integration of services is desirable, but requires true packet switching in order to be effective and economical. Since new services were expected to operate at multimegabit rates, a fast packet-switching technology was desired. This implied a number of choices (made for simplification purposes):

- Fixed packet size (known as *cells*)
- Short packet size
- Highly simplified headers
- No explicit error protection
- No link flow control

Since ATM was an effort to define BISDN by telephone equipment vendors and service providers, voice was a major part of the BISDN effort from the onset. In fact, the decision on short cell size (53 bytes total, with a 48-byte payload) was made with considerations of echo cancellation for voice. For 64-kbps voice, the use of echo cancellation equipment becomes necessary if packetization delay is more than 32 bytes (4 ms). Although the public telephone network in the United States has echo cancellers installed, smaller European countries do not. To avoid echo cancellation equipment, European countries proposed that the payload for ATM be 32 bytes. The U.S. proposal was 64 bytes and 48 bytes were chosen as a compromise. The maximum tolerable overhead due to the header was considered 10%, and thus the 5-byte header was chosen.

1.3. ATM Protocol Reference Model

The protocol reference model for ATM is shown in Fig. 1. This model is different from that of ISDN. In this reference model, the ATM layer is common to all services. Its function is to provide packet (cell) transfer capabilities. The ATM adaptation layer (AAL) is service-dependent. The AAL maps higher-layer information into ATM cells. The protocol reference model makes reference to three separate planes:

- *User plane* — information transfer and related controls (flow and error control)
- *Control plane* — call control and connection control
- *Management plane* — management functions as a whole, coordination among all planes, and layer management

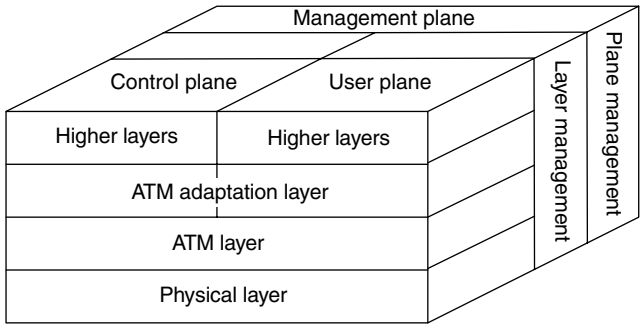


Figure 1. BISDN protocol reference model.

1.4. ATM Layer

We will first describe the ATM layer. ATM headers are very simple by design. The cell header has a different structure at the user–network interface (UNI) and at the network–network interface (NNI) (Figs. 2 and 3). Routing in ATM is achieved by an identifier field. It is the contents of this field that drives the fast hardware switching of an ATM cell. This field consists of two parts: the virtual circuit identifier (VCI) and the virtual path identifier (VPI). VCI is simply an index to a *connection* [14]. This "connection" is known as a *virtual circuit* (VC). A number of VCs are treated as a single entity known as a *virtual path* (VP). Thus, inside the network, cell switching can be performed on the basis of VPI alone. The VPI field is 8 bits at the UNI and 12 bits at the NNI. The VCI field is 16 bits long at both interfaces. It should be noted that VCIs and VPIs are not addresses. They are explicitly assigned at each segment (link between ATM nodes) of a connection when a connection is established, and they remain so for the duration of the connection. Using the VCI/VPI, the ATM layer can asynchronously interleave (multiplex) cells from multiple connections. As a historical remark, we would like to note that origins of the VPI/VCI concept can be traced back to the Datakit virtual circuit switch, developed by A. Fraser of Bell Labs during the 1970s [13,14]. Datakit was a product manufactured by AT&T for the data transmission needs of local exchange carriers.

HEC is an error check field, based on an 8-bit cyclic redundancy code (CRC), restricted to the cell header only. Three bits in the header (PT) are used to define the payload type. One bit (CLP) is reserved to indicate cell loss priority. This allows an ATM network to drop packets in case of congestion with the recovery mechanism provided by higher layers; such dropped cells will be detected by the higher layers of networking protocols (such as TCP/IP) and will be retransmitted. In passing, we would like to note that some earlier design decisions for ATM were later criticized when ATM was used to carry data belonging to the TCP/IP protocol. The most common type of IP packets carried in an IP network are TCP acknowledgment packets. Those packets are 44 bytes long and constitute about half of the packets carried in an IP network. Therefore, about half of IP packets are carried in an ATM network at an efficiency loss of about 10%. This inefficiency was later criticized by service providers in deploying IP-over-ATM networks and was termed *cell tax*.

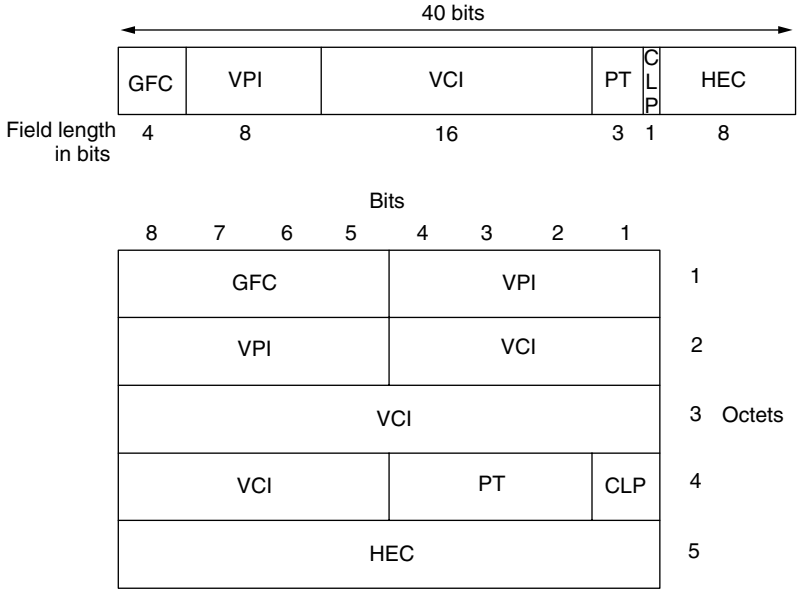


Figure 2. UNI cell header.

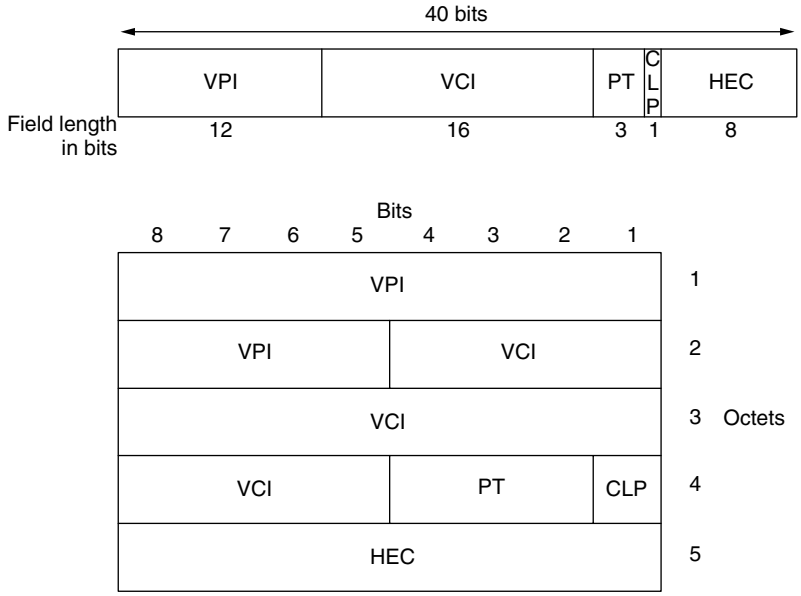


Figure 3. NNI cell header.

We stated above that VCI is an index to a connection. Thus, by this concept, ATM networks emulate connections between two points in a network and therefore are termed as *connection-oriented*. VP identifiers do not have to be universal in a network, they can be mapped from a set of values to another at the subnetwork boundary, albeit at a hardware cost. Virtual connections (consisting of VPs and VCs together) can be established permanently or on a per need basis. Permanent VCs (PVCs) are established once and all and are simple to work with. For bursty applications, switched VCs (SVCs) are designed. At a network node, SVCs can be established (added to the VC list) and removed from the VC list on a per need basis. Although this is a desirable property since not all connections in a network can be known in advance and the goal of developing the technology is indeed provided for bursty, unpredictable traffic, the hardware cost of incorporating this capability is high. In particular, this flexibility of being able to support bursty connections was later criticized since it limits the scalability of the ATM concept because of the difficulty of its implementation for high-speed backbone networks.

A PVC is not signaled by the endpoints. Both of the endpoint VC values are manually provisioned. The link-by-link route through the network is also manually provisioned. If any equipment fails, the PVC is down, unless the underlying physical network can reroute below ATM. A soft PVC also has manually provisioned endpoint VC values, but the route through the network can be automatically revised if there is a failure. Failure of a link causes a soft PVC to route around the outage and remain available. A switched VC (SVC) is established

by UNI signaling methods. So an SVC is a connection initiated by user demand. If a switch in the path fails, the SVC is broken and would have to be reconnected. The difference between an SVC and a soft PVC is that an SVC is established on an "as needed" basis through user signaling. With a soft PVC, the called party cannot drop the connection.

1.5. ATM Adaptation Layer

In order for ATM to support many kinds of services with different traffic characteristics and system requirements, it is necessary to adapt the different classes of applications to the ATM layer. This function is performed by AAL, which is service-dependent. Four types of AAL were originally recommended by CCITT. Two of these have now been merged into one, and a new one is added, making the total four once again. The four AALs are now described briefly:

- *AAL1* — supports connection-oriented services that require constant bit rates and have specific timing and delay requirements. Examples are constant bit rate services such as DS1 (1.5 Mbps) or DS3 (45 Mbps) transport.
- *AAL2* — a method for carrying voice-over ATM. It consists of variable size packets with a maximum of 64 bytes encapsulated within the ATM payload. AAL2 was previously known as "composite ATM" or "AAL-CU." The ITU specification, which describes AAL2 is called "ITU-T I.363.2."
- *AAL3/4* — a method intended for both connectionless and connection-oriented variable-bit-rate services. Two original distinct adaptation layers AAL3 and 4 have been merged into AAL3/4.
- *AAL5* — supports connection-oriented variable-bit-rate data services. Compared with AAL3/4, AAL5 is substantially lean at the expense of error recovery and built-in retransmission. This tradeoff provides a smaller bandwidth overhead, simpler processing requirements, and reduced implementation complexity. AAL5 has been proposed for use with both connection-oriented and connectionless services.

There is an additional AAL, AAL0, which normally refers to the case where the payload is directly inserted into a cell. This typically requires that the payload can always be fitted into a single cell so that the AAL is not needed for upper-layer PDU delineation when the upper-layer PDU bridges several cells.

1.6. ATM Traffic Management

During the early 1990s, the computer networking community was looking for a replacement of the 10-Mbps Ethernet standard at higher speeds of 100 Mbps and beyond. ATM, as specified by CCITT, was considered a viable alternative. Various companies active in this field formed an industry consortium, known as the *ATM Forum*. The ATM Forum later made quick progress in specifying and modifying the ATM specifications. ATM Forum defined the following traffic parameters for describing traffic that is injected into the ATM network at the UNI [2]:

- *Peak cell rate* (PCR) — maximum bit rate that may be transmitted from the source
- *Cell delay variation tolerance* (CDVT) — tolerance controlled by the network provider on how the actual peak rate deviates from the PCR
- *Sustainable cell rate* (SCR) — upper limit for the average cell rate that may be transmitted from the source
- *Maximum burst size* (MBS) — maximum number of cells for which the source may transmit at the PCR
- *Minimum cell rate* (MCR) — minimum cell rate guaranteed by the network

The ATM Forum defined different service classes to be supported by ATM. The classes are differentiated by specifying different values for the following QoS parameters defined on a per-connection basis:

- *Maximum Cell Transfer Delay* (*maxCTD*). CTD is the delay experienced by a cell between the transmission of the first bit by the source and the reception of the last bit of the cell by the destination. The CTD of a cell is smaller than the maxCTD QoS parameter of the connection with which it is carried within with a large probability, or equivalently, maxCTD is the $(1 - \alpha)$ quantile of CTD for a small α.
- *Peak-to-peak Cell Delay Variation* (*ppCDV*). The ppCDV is the difference between the $(1 - \alpha)$ quantile of the CTD and the fixed CTD that could be experienced by any delivered cell on a connection during the entire connection holding time.
- *Cell Loss Ratio* (*CLR*). This ratio indicates the percentage of cells that are lost in the network due to error or congestion and are not received by the destination.

The QoS parameters are defined for all conforming cells of a connection, where conformance is defined with respect to a generic cell rate algorithm (GCRA) described in the ATM Forum *User-Network Interface Specification 3.1* [2]. The input to this algorithm is the traffic parameters described above.

The proposed service categories by the ATM Forum are then described as follows [1]:

- *CBR* (*Constant Bit Rate*). The CBR service class is intended for real-time applications, namely, those requiring tightly constrained delay and delay variation, as would be appropriate for voice and video applications. The consistent availability of a fixed quantity of bandwidth is considered appropriate for CBR service. Cells that are delayed beyond the value specified by maxCTD are assumed to be of significantly less value to the application. For the service class CBR, the attributes PCR, CDVT, maxCTD, ppCDV, and CLR are specified.

- *VBR-rt* (*Variable Bit Rate — Real-Time*). The real-time VBR service class is intended for real-time applications, that is, those requiring minimal loss and tightly constrained delay and delay variation, as would be appropriate for voice and video applications. Sources are expected to transmit at a rate that varies with time. Equivalently, the source can be described as "bursty." VRB-rt expects a bound on the cell loss rate for cells conforming to the associated GCRA. Cells delayed beyond the value specified by maxCTD are assumed to be of significantly less value to the application. Real-time VBR service may support statistical multiplexing of real-time sources, or may provide a consistently guaranteed QoS. For VBR-rt, the ATM attributes PCR, CDVT, SCR, MBS, maxCTD, ppCDV, and CLR are specified.
- *VBR-nrt* (*Variable Bit Rate — Non-Real-Time*). The non-real-time VBR service class is intended for non-real-time applications that have "bursty" traffic characteristics and can be characterized in terms of a generic cell rate algorithm (GCRA). VRB-nrt expects a bound on the cell loss rate for cells conforming to the associated GCRA. Bounds for cell transfer delay and cell delay variation are not provided for VBR-nrt. Similar to VBR-rt, non-real-time VBR service also supports statistical multiplexing of connections. For non-real-time VBR, ATM attributes PCR, CDVT, SCR, MBS, and CLR are supported.
- *UBR* (*Unspecified Bit Rate*). The UBR service class is intended for delay-tolerant or non-real-time applications — those that do not require tightly constrained delay and delay variation, such as traditional computer communications applications. Sources are expected to transmit noncontinuous bursts of cells. UBR service supports a high degree of statistical multiplexing among sources. UBR service includes no notion of a per VC allocated bandwidth resource. Transport of cells in UBR service is not necessarily guaranteed by mechanisms operating at the cell level. However, it is expected that resources will be provisioned for UBR service in such a way as to make it usable for some set of applications. UBR service may be considered as interpretation of the common term "best-effort service." For UBR, only PCR and CDVT are specified as a traffic attribute.
- *ABR* (*Available Bit Rate*). Many applications have the ability to reduce their information transfer rate if the network requires them to do so. Likewise, they may wish to increase their information transfer rate if there is extra bandwidth available within the network. There may not be deterministic parameters because the users are willing to live with unreserved bandwidth. To support traffic from such sources in an ATM network will require facilities different from those for peak cell rate of sustainable cell rate traffic. The ABR service is designed to fill this need. The traffic attributes PCR, CDVT, MCR, and CLR are specified for the ABR service class.

There are other service categories proposed by the ITU: ABT (ATM block transfer) and CCT (controlled cell transfer). However, these categories have not gained much acceptance.

2. IP NETWORKS

In the 1990s, while ATM technology was being developed to integrate voice, data, and video, pure data services embraced the TCP/IP protocol, or the IP technology. What made the IP technology attractive is its universal adoption, due mainly to the popularity of the global Internet and the unprecedented growth rates the Internet has reached. Initially, IP was not designed for the integration of voice, data, and video to the end user. Developed under the U.S. Department of Defense (DoD) funding, IP was built around reliability and redundancy so as to allow communication to continue between nodes in case of a failure.

2.1. History of IP Networks

There was a perceived need for survivable command and control systems in the United States during the 1960s. To fulfill this need, early contributors were drawn from the ranks of defense contractors, federally funded think tanks, and universities: the RAND Corporation, Lincoln Laboratories, MIT, UCLA, and Bolt Beranek and Newman (BBN), under DoD funding.

P. Baran of RAND Corporation postulated many of the key concepts of packet-switching networks that were implemented in the ARPANET, the research network Advanced Research Projects Agency (ARPA) of DOD funded in 1967. Baran's motivation was to use novel approaches to build survivable communications systems. The traditional telephone system is based on a centralized switching architecture and the concept of connection or a "circuit" that must be established between the parties of a communications session using the centralized switches. If a link or a switch is broken (or destroyed) during a connection in this architecture, the communications session will fail, which is unacceptable for survivability purposes. Baran's work was built around the replacement of centralized switches with a larger number of distributed routers, each with multiple (potentially redundant) connections to adjacent routers. Messages then would be divided into parts (*blocks* or *packets*), and the packets would then be routed independently. This packet-switching concept allows bursty data traffic to be statistically multiplexed over available communications paths and makes it possible to adapt to changing traffic demands and to use existing resources more efficiently without a need for a priori reservation.

ARPANET was proposed by ARPA as an ambitious program to connect many host computers at key research sites across the country, using point-to-point telephone lines and the packet-switching concept. The idea of using separate switching computers, rather than the hosts, was proposed to serve as the routing elements of network, thereby offloading this function from the timesharing hosts. BBN received the contract to build the interface message processors (IMPs) in this newly

proposed architecture. The engineers at BBN developed the necessary host-to-IMP and IMP-to-IMP protocols, the original flow control algorithms, and the congestion control algorithms. In the BBN model, hosts communicate with each other via messages. When a host sends a message, it is broken down into packets by the source IMP (which is the IMP directly attached to the host). The IMP then routes each packet, individually through the network of IMPs, to the destination IMP. Each packet will be sent along the path that is estimated to be the shortest, and the path taken by each packet may be different. The destination IMP, on receiving all packets for a message, will reassemble an exact replica of the original message and forward the message on to the destination host. On the basis of the implementation of BBN, the ARPANET started to emerge with its first four nodes at UCLA, UCSB, Stanford Research Institute (SRI), and the University of Utah in 1969. ARPANET's purpose was to provide a fast and reliable communication between heterogeneous host machines. The goal of the computer network was for each computer to make every local resource available to any computer in the network in such a way that any program available to local users can be used remotely without much degradation.

In 1969, N. Abramson, motivated by the poor telephone lines in the Hawaiian Islands, launched the Aloha Project at the University of Hawaii, a project funded by ARPA. In this project, the principles underlying a packet-switched network based on fixed-site radio links were investigated. The ALOHA project developed a new technology for contention-based media access, the "ALOHA protocols," and applied these techniques to satellites as well as radio systems. R. Metcalfe at Xerox PARC built on this work, leading to the development of the Ethernet protocols for access to a shared wirelined medium as a local-area networking technology. In 1972, L. G. Roberts and R. Kahn launched the ARPA Packet Radio Program: packet switching techniques on the mobile battlefield. ARPA also created a packet-switched experimental satellite network (SATNet), with work done by Comsat, Linkabit, and BBN. All this work motivated the need for a technology to link these independent networks together in a true "network of networks," the so-called Internet.

In 1973, R. Kahn and V. Cerf developed the concept of a network gateway (or a software packet switch), as well as the initial specifications for the Transmission Control Protocol (TCP). With this breakthrough concept, transmission reliability is shifted from the network to end hosts, thus allowing the protocol to operate no matter how unreliable the underlying link is. This paradigm shift was based on the "end-to-end argument," which states that the underlying network is only as strong as its weakest link and therefore improving the reliability of a single link or even an entire subnetwork may have only a marginal effect on the end-to-end reliability. With this paradigm change, the architecture internal to the network was significantly simplified. V. Cerf then joined ARPA to complete the design of the Internet Protocol (IP) Suite, overseeing the separation of the routing portions of the protocols (IP) from the transport-layer issues (TCP), and the transition of the new protocols into ARPANET.

The global Internet began around 1980, when ARPA started converting machines attached to its research networks to the new TCP/IP protocols. ARPANET, already in place, quickly became the backbone of the new Internet and was used for many of the early experiments with TCP/IP. In 1983, the Defense Communications Agency (DCA) split ARPANET into two separate networks: one for future research and one for military communications, with the research part retaining the name ARPANET. At around the same time, most university computer science departments were running a version of the UNIX operating system available from the University of California's Berkeley software distribution, commonly known as Berkeley UNIX or BSD UNIX. By funding BBN to implement its TCP/IP protocols for use with BSD UNIX, and funding University of California Berkeley to integrate the protocols with its software distribution, ARPA was able to reach over 90% of university computer science departments in the United States. A large number of hosts subsequently connected their networks to ARPANET, thus creating the "ARPA Internet."

By 1985, ARPANET was heavily used and congested. The National Science Foundation (NSF), which needed a faster network, initiated the development of NSFNET in the mid-1980s. NSF selected the TCP/IP protocol suite used in ARPANET. However, as opposed to a single core backbone used in ARPANET, the earliest form of NSFNET in 1986 used a three-tiered architecture that consisted of universities and other research organizations that are connected to regional networks, which are then interconnected to a major backbone network using 56-kbps links. The link speeds were then upgraded to T1 (1.5 Mbps) in 1988 and later in 1991 to T3 (45 Mbps). In the early 1990s, the NSFNET was still reserved for research and educational applications. At this time, government agencies, commercial users, and the general public began demanding access to NSFNET. With the success of private networks using IP technology, NSF decided to decommission the NSFNET backbone in 1995. Commercial Internet providers then took over the role of providing Internet access. These providers have connection points called *point of presence* (PoP). Customers of these service providers are connected to the Internet via these PoPs. The collection of PoPs and the way they are interconnected form the provider's network. Providers may be regional, national, or global, depending on the scope of their networks. Today's Internet architecture is based on a distributed architecture operated by multiple commercial providers rather than a single core network (NSFNET) that are interconnected via major network exchange points. Historically, commercial Internet providers exchange traffic at network access points (NAP) and the metropolitan-area exchanges (MAEs) through a free exchange relationship called *bilateral public peering*. Two connectivity models have emerged as a result of increasing congestion in the major exchange points: (1) private peering among the largest backbone providers and (2) more recently, private transit

connections to multiple backbone providers, which are favored by specialized ISPs.

2.2. IP Fundamentals

The Internet provides three sets of services [8]. At the lowest level, one has a connectionless delivery service. The other two services (transport services and application services) lie on top of this connectionless delivery service. The protocol that defines the unreliable, connectionless delivery mechanism is called the Internet Protocol and is commonly referred to by its initials, IP. IP defines the basic data unit of data transfer and it also performs the routing function. Therefore, IP is also referred to as the *layer 3 protocol* in the Internet suite as it corresponds to the layer 3 (network layer) of the OSI model. Layer 4 protocols such as TCP and UDP run on IP and provide an appropriate higher level platform on which the applications depend.

In addition to internetwork routing, IP provides error reporting and fragmentation and reassembly of information units called *datagrams*. Datagrams of different size are used by IP for transmission over networks with different maximum data unit sizes. IP addresses are globally unique, 32-bit numbers assigned by the network information center. Globally unique addresses permit IP networks anywhere in the world to communicate with each other.

An IP address is divided into three parts. The first part designates the network address, the second part designates the subnet address, and the third part designates the host address. Originally IP addressing supported three different network classes. Class A networks were intended mainly for use with a few very large networks, because they provided only 8 bits for the network address field. Class B networks allocated 16 bits, and class C networks allocated 24 bits for the network address field. Because Internet addresses were generally assigned only in these three sizes, there were many wasted addresses. In the early 1990s only 3% of the assigned addresses were actually being used and the Internet was running out of unassigned addresses. A related problem was the size of the Internet global routing tables. As the number of networks on the Internet increased, so did the number of entries in the routing tables. By this time, Internet standards were being specified by an organization known as the Internet Engineering Task Force (IETF). IETF selected classless interdomain routing (CIDR) [15,24] to be a much more efficient method of assigning addresses and address aggregation to address these two critical issues. CIDR is a replacement for the old process of assigning class A, B, and C addresses with a generalized network prefix. Instead of being limited to network identifiers (or "prefixes") of 8, 16, or 24 bits, CIDR currently uses prefixes anywhere from 13 to 27 bits. This allows for address assignments that much more closely fit an organization's specific needs and therefore avoids address waste. The CIDR addressing scheme also enables route aggregation in which a single high-level route entry can represent many lower-level routes in the global routing tables.

In the 1990s, there were also significant developments in IP routing. There are two main routing infrastructures: flat routing and hierarchical routing. In a flat routing infrastructure, each network ID is represented individually in the routing table. The network IDs have no network/subnet structure and cannot be summarized. In a hierarchical routing infrastructure, groups of network IDs can be represented as a single routing table entry through route summarization. The network IDs in a hierarchical internetwork have a network/subnet/subsubnet structure. A routing table entry for the highest level (the network) is also the route used for the subnets and sub-subnets of the network. Hierarchical routing infrastructures simplify routing tables and lower the amount of routing information that is exchanged, but they require more planning. IP implements hierarchical network addressing, and IP internetworks can have a hierarchical routing structure.

In very large internetworks, it is necessary to divide the internetwork into separate entities known as *autonomous systems*. An autonomous system (AS) is a portion of the internetwork under the same administrative authority. The AS may be further divided into regions, domains, or areas that define a hierarchy within the AS. The protocols used to distribute routing information within an AS are known as *interior gateway protocols* (IGPs). The protocols used to distribute routing information between ASs are known as *exterior gateway protocols* (EGPs). In today's Internet, link-state protocols such as OSPF version 2 [21] and IS-IS [23] are used as IGPs, whereas the path vector protocol BGP-4 [25] is used as a exterior gateway protocol.

With the changes to IP address structure and address summarization with CIDR, and the development of efficient hierarchical routing infrastructures, IP networks have scaled up to the level of universal connectivity today. This has made the Internet a global medium in such a way that any two hosts can communicate with each other as long as they are attached to the Internet. However, currently a packet is transported in the Internet without any guarantees to its delay or loss. Because of this "best effort" forwarding paradigm, the Internet cannot provide integrated services over this infrastructure. As we described previously, the BISDN vision requires end-to-end QoS guarantees for different services. The IETF is working on several QoS models that may potentially realize the BISDN vision using IP. Using IP as opposed to ATM to realize the BISDN vision is a new approach made popular by the widespread use of IP.

2.3. QoS Models in IP Networks

Several QoS architectures are proposed by the IETF for IP networks to enable the support of integrated services over IP networks. We will briefly overview these models below.

2.3.1. Integrated Services (Intserv) Model. The integrated services architecture [6] defines a set of extensions to the traditional best-effort model of the Internet so as to provide end-to-end (E2E) QoS commitments to certain applications with quantitative performance requirements. Two services are defined: guaranteed service [28] and controlled load [31] services. Guaranteed service provides an assured level of bandwidth, a firm end-to-end delay bound,

and no loss due to queueing if the packets conform to an a priori negotiated contract. It is intended for applications with stringent real-time delivery requirements such as audio and video applications with playback buffers. A packet arriving after its playback time is simply discarded by the receiver. In the case of controlled load service, the network will commit to a flow a service equivalent to that seen by a best-effort flow on a lightly loaded network. This service is intended for adaptive real-time applications that can tolerate a certain amount of loss and delay provided it is kept to a reasonable level. The integrated services architecture assumes some explicit setup mechanism such as RSVP (Resource Reservation Protocol) [7]. This setup or signaling mechanism will be used to convey QoS requirements to IP routers so that they can provide requested services to flows that request them. On receiving per flow resource requirements through RSVP, the routers apply Intserv admission control to signaled requests. The routers also enable traffic control mechanisms to ensure that each admitted flow receives the requested service independent of other flows. These mechanisms include the maintenance of per flow classification and scheduling states. One impediment to the deployment of integrated services with RSVP is the use of per flow state and per flow processing, which typically exceeds the flow-handling capability of today's core routers. This is known as the *scalability problem* in RSVP or in Intserv.

The integrated services architecture is similar to the ATM SVC architecture in which ATM signaling is used to route a single call over an SVC that provides the QoS commitments of the associated call. The fundamental difference between the two architectures is that the former typically uses the traditional hop-by-hop IP routing paradigm, whereas the latter uses the more sophisticated QoS source routing paradigm.

2.3.2. Aggregate RSVP Reservations Model. This QoS model attempts to address some of the scalability issues arising in the traditional Intserv model. In the traditional intserv model, each E2E reservation requires a significant amount of message exchange, computation, and memory resources in each router along the way. Reducing this burden to a more manageable level via the aggregation of E2E reservations into one single aggregate reservation is addressed by the IETF [3]. Although aggregation reduces the level of isolation between individual flows belonging to the aggregate, there is evidence that it may potentially have a positive impact on delay distributions if used properly and aggregation is required for scalability purposes.

In the aggregation of E2E reservations, we have an aggregator router, an aggregation region, and a deaggregator. Aggregation is based on hiding the E2E RSVP messages from RSVP-capable routers inside the aggregation region. To achieve this, the IP protocol number in the E2E reservation's Path, PathTear, and ResvConf messages is changed by the aggregator router from RSVP to RSVP-E2E-IGNORE on entering the aggregation region, and restored to RSVP at the deaggregator point. Such messages are treated as normal IP datagrams inside the aggregation region, and no state is stored. Aggregate Path messages are sent from the aggregator to the deaggregator using RSVP's normal IP protocol number. Aggregate Resv messages are then sent back from the deaggregator to the aggregator, via which an aggregate reservation with some suitable capacity will be established between the aggregator and the deaggregator to carry the E2E flows that share the reservation. Such establishment of a smaller number of aggregate reservations on behalf of a larger number of E2E flows leads to a significant reduction in the amount of state to be stored and the amount of signaling messages exchanged in the aggregation region.

Aggregation of RSVP reservations in IP networks is very similar in concept to the virtual path in ATM networks. In this framework, several ATM virtual circuits can be tunneled into one single ATM VP for manageability and scalability purposes. A virtual path identifier (VPI) in the ATM cell header is used to classify the aggregate in the aggregation region (VP switches), and the virtual channel identifier (VCI) is used for aggregation/deaggregation purposes. A VP can be resized through signaling or management.

2.3.3. Differentiated Services (Diffserv). In contrast to the per flow nature of integrated services, differentiated services (Diffserv) networks classify packets into one of a small number of aggregated flows or "classes" based on the Diffserv Codepoint (DSCP) written in the differentiated services field of the packet's IP header [22]. This is known as *behavior aggregate* (BA) classification. At each Diffserv router in a Diffserv domain (DS domain), packets receive a per hop behavior (PHB), which is invoked by the DSCP. Differentiated services are extended across a DS domain boundary by establishing a service-level agreement (SLA) between an upstream network and a downstream DS domain. Traffic classification and conditioning functions (metering, shaping, policing, remarking) are performed at this boundary to ensure that traffic entering the DS domain conforms to the rules specified in the *traffic conditioning agreement* (TCA), which is derived from the SLA. A PHB then refers to the packet scheduling, queueing, policing, or shaping behavior of a node on any given packet belonging to a BA, as configured by a SLA or a policy decision. Four standard PHBs are defined:

- *Default PHB* [22] — provides a best-effort service in a Diffserv-compliant node.
- *Class-selector PHB* [22] — to preserve backward compatibility with any IP precedence scheme currently in use on the network, Diffserv defines a certain DSCP for class selector code points. The PHB associated with a class selector code point is a class selector PHB. Eight class selector code points are defined.
- *Assured forwarding* (*AF*) *PHB* [16] — the AF PHB group defines four AF classes: AF1, AF2, AF3, and AF4. Each class is assigned a specific amount of buffer space and interface bandwidth, according to the SLA with the service provider or a policy decision. Within each AF class, three drop precedence values are assigned. In the case of a congestion indication or

equivalently if the queue occupancy for the AF class exceeds a certain threshold, packets in that class with lower drop precedence values will be dropped. With this description, assured forwarding PHB is similar to the controlled load service available in the integrated services model.

- *Expedited forwarding* (*EF*) *PHB* [10] — EF PHB provides a guaranteed bandwidth service with low loss, delay, and delay jitter. EF PHB can be implemented with priority queuing and rate limiting on the behavior aggregate. EF PHB can be used to provide virtual leased line or premium services in Diffserv networks similar to the guaranteed service in Intserv networks and the CBR service in ATM networks.

Since Diffserv eliminates the need for per flow state and per flow processing, it scales well to large-core networks.

2.3.4. Hybrid Intserv–Diffserv [5]. In this QoS model, intserv and diffserv are employed together in a way that end-to-end, quantitative QoS is provided by applying the Intserv model end- to- end across a network containing one or more Diffserv regions. The Diffserv regions of the network appear to the Intserv-capable routers or hosts as virtual links. Within the Diffserv regions of the network, routers implement specific PHBs (aggregate traffic control) on the basis of policy decisions. For example, one of the AF PHBs can be used to carry all traffic using E2E reservations once an appropriate amount of bandwidth and buffer space is allocated for that AF class at each node. The total amount of traffic that is admitted into the Diffserv region that will receive a certain PHB may be limited by policing at the edge. The primary benefit of Diffserv aggregate traffic control is its scalability. The hybrid Intserv–Diffserv model is closely related to the RSVP reservation aggregation model.

2.3.5. Multiprotocol Label Switching (MPLS). MPLS introduces a new forwarding paradigm for IP networks in that a path is first established using a signaling protocol. A label in the IP header rather than the destination IP address is used for making forwarding decisions throughout the MPLS domain [26]. Such paths are called *label-switched paths* (LSPs), and routers that support MPLS are called *label-switched routers* (LSRs). In this architecture, edge ingress LSRs place IP packets belonging to a certain forwarding equivalence class (FEC) in an appropriate LSP. The core LSRs forward packets only according to the label in the header, and the egress edge LSRs remove the labels and forward these packets as regular IP packets. The benefits of this architecture include but are not limited to

- *Hierarchical Forwarding*. MPLS provides a forwarding hierarchy with arbitrary levels as opposed to the two-level hierarchy in ATM networks. Using this flexibility and the notion of nested labels, several level 1 LSPs can be aggregated into one level 2 LSP, and several level 2 LSPs can be aggregated into one level 3 LSP, and so on. One immediate benefit of this is that the transit provider need not know about the global routes, which makes it very scalable [11] for transit providers.
- *Traffic Engineering*. The mapping of traffic trunks (an aggregation of traffic belonging to the same class) onto a given network topology for optimal use of network resources is called the *traffic engineering problem*. In MPLS networks, traffic trunks are mapped to the network topology through the selection of routes and by establishing LSPs with certain attributes using these routes. A combination of a traffic trunk and the LSP is called an *LSP tunnel*. In its simplest application, in the case of congestion arising from suboptimal routing, LSP tunnels may be rerouted for better performance.
- *Virtual Circuit Emulation*. Another benefit is that other connection-oriented networks may be emulated by MPLS. The advantage is that a single integrated datagram network can provide legacy services such as frame relay and ATM to end customers while maintaining a single infrastructure.

2.3.6. Summary of QoS Models for IP Networks. For elastic applications that can adapt their rates to changing network conditions (e.g., data applications using TCP), a simple QoS model such as "Diffserv" will be suitable. For inelastic applications such as real-time voice and video with stringent delay and loss requirements, end-to-end Intserv is a better fit. The need for per flow maintenance in RSVP capable routers is known to lead to a scalability problem especially in core networks. Therefore, several novel QoS models have been introduced to attack this scalability problem. From the perspective of a network, both models rely on eliminating the per flow maintenance requirement by either aggregating E2E reservations into one single reservation at the border nodes of this network or carrying all E2E reservations in one preprovisioned Diffserv class. However, these architectures pose a burden on the border routers, and their success remains to be seen in the commercial marketplace. MPLS, on the other hand, is promising traffic-engineered backbones with routing scalability for all these QoS models.

3. BISDN AND THE WORLD WIDE WEB

In this section we describe the development of IP versus ATM as the underlying networking technology of BISDN.

The development of ATM realized full progress at ITU-T during 1989/90. This effort was led by telecommunications service providers as well as telecommunications equipment manufacturers. The main goal was to develop the switching and networking technology for BISDN. As cooperation and contributions from telecommunications industry leaders were at a very substantial level, the vision of an integrated wide-area network (WAN) using ATM seemed very likely to happen. This development in the WAN sparked interest in other networking platforms. The first affected was the computer communications industry, specifically the local-area networking (LAN) community. At the time, available LANs (mainly the Ethernet) had a

top speed of 10 Mbps. The technology had improved from coax to twisted pair and from shared media to switched (1991). However, as user needs increased, the top speed of 10 Mbps became insufficient and the industry began to search for a replacement at significantly higher speeds of 100–150 Mbps. At this time, the ATM effort at ITU-T defined a basic transport rate of 155 Mbps. This speed was very convenient for the LAN community. In addition, adopting the same switching and networking technology with the WAN was attractive from the viewpoint of simplifying the WAN gateway. This led to an industry standardization organization known as the *ATM Forum*. The goal was to define a set of specifications common to the member companies, primarily for the LAN. An additional goal was to speed up the standardization process, which, at ITU-T, required long study periods and consensus from national representatives.

Another development related to ATM was the emergence of the ADSL (Asymmetric Digital Subscriber Line) technology in the 1990s [19]. At the time, invoking Shannon's capacity formula, the highest transmission rate for a voiceband modem over a subscriber loop, without changing any equipment at the central office, was considered to be about 30 kbps. The ADSL technology replaces central office channel banks to exploit frequencies above 4 kHz. In addition, it employs sophisticated methods that limit near-end crosstalk and therefore substantially expand the transmission potential of the subscriber loop. As a result, it can operate at rates that are orders of magnitude higher than those of voiceband modems. The ADSL access network includes terminations both within the home and the public network (ATU-R and ATU-C, respectively). The ATU-R is commonly called a "DSL modem," and the ATU-C is commonly called a "DSLAM" (DSL access multiplexer). ATM is used as layer 2 in this "residential broadband" architecture. ADSL provides up to 1.5 Mbps (downlink) rate. It may be used to extend the ATM network, and therefore QoS properties of ATM, all the way to the residential or corporate desktop. In this model [19], the ATM user-to-network interface (UNI) is tunneled through an ADSL link. By having desktop applications talk directly to the ATM network, bandwidth can be allocated end-to- end across the network that was thought to facilitate the deployment of isochronous, delay-sensitive applications such as voice and videoconferencing [17]. In fact, this was the intent of BISDN from the onset. The effort to employ ADSL to provide integrated services to the home was led by potential application service providers [20]. At this time, PC (personal computer) operating systems did not yet include a networking stack as part of the kernel, and beyond computer terminal emulation, there were not yet any major residential networking applications available. With the arrival of the World Wide Web (WWW) and the concept of a Web browser, the need for an IP stack in PCs became apparent. At the time the most popular PC operating system was Windows version 3.1 from Microsoft. As this operating system did not have an IP stack, it was added to the operating system manually by the user. Later, Windows 95 became the first PC operating system to include an IP stack. With this development, the IP stack became an inevitable option in residential broadband networking. Consequently, the original concept of residential ATM was later modified as IP over ATM over DSL [20]. This could have been a cosmetic change, however, and by this time, the vision of BISDN using ATM still seemed likely to happen, with a form of IP over ATM being used mainly for best effort data transfer.

A number of developments that took place in the second part of the 1990s have changed the outlook for ATM as the underlying networking technology of BISDN:

1. IEEE 802.3 Working Group made rapid progress to define a newer version of the Ethernet standard to operate at 100 Mbps over twisted-pair and switched media. This LAN standard did not have any QoS guarantees, but the solution satisfied a much sought-after need for a LAN operating around 100 Mbps. This solution was quickly adopted by the marketplace, and the 100-Mbps Ethernet quickly became a commodity product. The absence of a compelling need for QoS in LANs virtually stopped the local ATM activity. With this development, the ATM Forum lost a major thrust.
2. The development of the WWW and the Web browser, as well as the commercialization of the Internet quickly made Web browsing using a PC a household activity. This development stalled or perhaps even stopped the concept of residential ATM.
3. A possible application of ATM was in digital cable access systems. By making extensions to the coaxial or hybrid fiber/coaxial cable TV plant so that duplex transmission becomes possible (providing amplification in both directions), and using digital technology so that compression can be used to transmit hundreds of TV channels, ATM was under consideration as a potential service offering. Adding data services to this potential offering was attractive. A multiaccess control algorithm was needed to share the uplink channel. A standardization activity was initiated under the IEEE 802.14 Working Group. While this group was working on an access system based on ATM and provide QoS guarantees for delivering a multitude of services, and while some progress was made, cable service providers decided to pursue their own standardization effort. They named this activity the *multimedia cable network system* (MCNS). The main reason for this secession was to make the process of standardization faster. As PC operating systems were beginning to offer IP stacks, MCNS chose IP technology as the basis of their own access system. The resulting system specification is known as the *Data over Cable Service Interface Specifications* (DOCSIS). Although version 1.0 of these specifications was for best-effort data service only, in its version 1.1, DOCSIS supports some QoS guarantees, specifically designed for voice over IP (VoIP). DOCSIS is currently the de facto worldwide standard for digital cable access, while IEEE 802.14 has stopped its activities. With this development, IP, rather than ATM became the underlying technology for digital cable access systems.
4. A significant advantage of ATM was its fast switching property. ATM was designed to be a simple switching technology so that scalable switches at total

throughput values approaching hundreds of gigabits per second could be built. This vision is by and large correct (although segmentation and reassembly at edge routers can become difficult at higher speeds). However, there was a surprising development in this period — throughput values of routers increased substantially. Today the maximum throughput values of core IP routers compete with those of core ATM switches. Implementation of algorithms for IP address lookup and memory manipulation for variable-length packet switching in ASICs is largely responsible for this development.

5. The ATM Forum was founded as an industry organization with the premise of fast standardization. As we noted above, ITU-T requires long study periods and consensus among national representatives. It was thought that the ATM Forum would move faster in reaching a standard. Although that was partially achieved, the industry perception is that signaling became too complex in the ATM Forum.

6. In the 1990s, a number of developments took place in optical transport systems that altered network switching in a major way: (a) invention of Erbium-doped fiber amplification made long-distance optical transmission without intermediate electrical conversions possible; (b) development of wavelength-division multiplexing (WDM) or dense WDM (DWDM) made transport of a large number of wavelengths in a single fiber possible — the number of wavelengths approached >100, while transmission speeds on individual wavelengths approached 10 Gbps; and (3) wavelength routing or wavelength cross-connects made it possible to demultiplex individual wavelengths from a single fiber and multiplex wavelengths from different fibers into a single fiber. The result is a wavelength switch with total throughput in the range of tens of terabits per second. As a result, wavelength routing provided an alternative to electronic switching at the network core, thus making the scalability argument of ATM switching less attractive.

7. A major advantage of ATM was its QoS capabilities. However, as described in the previous section, IP community developed a set of QoS capabilities. Although there are questions and uncertainties about the realization of these capabilities, there is some established confidence in IP QoS. We would like to note that ATM actually was never deployed for the end-to-end QoS vision. The reason for this is the complexity in signaling and the needed per flow queuing. The multiclass and aggregate IP QoS model may indeed be more scalable.

8. IP embraced ATM's VP concept. MPLS essentially implements VPs. Various tunneling mechanisms introduced into IP make switching aggregated traffic in IP possible. Furthermore, the endpoints of a VP implemented by MPLS do not need to be routing peers, which significantly reduces the number of peerings in the network, and therefore routing scalability.

9. Because of its scalable fast switching nature, ATM switches were used to carry and switch IP traffic. However, over time, other solutions were developed that avoid the ATM layer in between. For example, at one point, service providers deployed IP over ATM over SONET over WDM. IP was employed since applications required it, ATM was employed for high-speed packet switching, SONET was employed because of its fast restoration capability via SONET rings, and WDM was employed for higher transmission speeds in a single fiber. The industry sought for ways to simplify this complicated hierarchy. As a result, IP extended to assume many of the functionalities of ATM and even some of those of SONET (e.g., resilient packet rings).

10. We described the 10% inefficiency that results in carrying TCP/IP traffic over ATM, known as the *cell tax*, above. Several service providers claimed this inefficiency was too high. In reality, with IP extending to assume many of ATM's functionalities, the need for IP over ATM was alleviated and the cell tax became irrelevant.

11. In the 1980s there were several attempts made to build private networks for multiple-location enterprises. These typically employed nailed-up leased lines, used voice compression to reduce voice rates, and combined voice and data. Such networks, called *private networks*, were the precursors of integration of services, albeit on a small scale. As discussed above, first ISDN and then BISDN had the vision of integrated services. In an integrated public packet network, security, by means of proper authentication and encryption, enables construction of a virtual private network (VPN). A VP is very useful in the construction of a VPN since it simplifies processing of data belonging to a particular VPN in the network. Thus ATM is a natural way to implement VPNs. However, as described above, solutions were developed to embrace the same concept in the IP world. Examples of such protocols are the Layer 2 Tunneling Protocol (L2TP) [30], IPSec [18], and GRE [12]. MPLS, on the other hand, makes it possible to build provider-provisioned scalable VPNs also making use of BGP4 for routing and label information distribution [27]. Thus ATM is no longer a unique method to implement VPNs.

12. Another aspect of the aggregation property of VPs is the traffic engineering potential it provides. For example, one possibility in integrated networks is to use different routes based on QoS properties of different flows, such as those belonging to the same source and destination pair. There are tools, such as the concept of equivalent bandwidth, that enable traffic engineering for integrated networks. Then, VPs become very useful tools to implement the desired property. Obviously, with the development of VP-like concepts in IP networks, the superiority of ATM in this regard is no longer valid.

To summarize, from the discussion above it appears that two related events stalled the development of BISDN:

1. The appearance of the WWW made IP protocol instantly ubiquitous. Common PC operating systems quickly adopted an IP stack. A similar ATM stack was not needed because there was no immediate application tied to ATM in the way that the WWW was tied to IP.
2. IP quickly extended to assume the advantageous properties of ATM, at least in theory. As a result, ATM lost its role as the underlying technology that glues BISDN all together.

Therefore, it is safe to say that BISDN is not likely to occur as it was originally designed at the ITU-T, frequently described by the acronym BISDN/ATM.

Having said that, we must reiterate that integration of services is certainly useful for the consumer. Furthermore, there appears to be an increasing (albeit at a smaller rate than expected) demand for broadband services. Thus, in the near future some form of a service offering that unifies voice, broadband data, and video can be expected (in fact, it currently exists in digital cable). Whether this offering will eventually become a ubiquitous service such as expected of BISDN/ATM depends on many factors and it is difficult to predict today (in mid-2002). It is clear, however, that voice-over IP (VoIP) will be used to carry some voice traffic, especially in traffic-engineered enterprise networks. The degree of voice compression available for VoIP ($\sim$8 vs. 64 kbps, although with VoIP overhead, this ratio of $\frac{1}{8}$ becomes bigger), statistical multiplexing advantages, and the capability to combine with data in VPNs is an attractive value proposition. Adding the Public Switched Telephone Network and video services to this value proposition successfully in the marketplace in the short term, however, is a taller order.

BIOGRAPHIES

Ender Ayanoglu received his B.S. degree in 1980 from the Middle East Technical University and M.S. and Ph.D degrees in 1982 and 1986, respectively, from Stanford University, all in electrical engineering. He was with the Communications Sciences Research Laboratory of Bell Laboratories (AT&T and Lucent Technologies) during 1986–1999. From 1999 to summer 2002 he was with Cisco Systems. Currently he is with University of California, Irvine, he was the Chairman of the IEEE Communications Society Communication Theory Technical Committee during 1999–2001. He was the Chairman of the IEEE-ISTO Broadband Wireless Internet Forum during 2000/01. Currently he serves as an Editor for the *IEEE Transactions on Communications*. He is the recipient of two best paper awards from the IEEE Communications Society and is an IEEE Fellow.

Nail Akar received the B.S degree in 1987 from Middle East Technical University, Ankara, Turkey, and the M.S. and Ph.D degrees from Bilkent University, Ankara, Turkey, in 1989 and 1994, respectively, all in electrical and electronics engineering. He joined the Computer Science Telecommunications Program in 1994 at the University of Missouri — Kansas City as a Visiting Scholar and was a Visiting Assistant Professor in the same program in 1996. At UMKC, he worked on the development of computational algorithms for the performance analysis of communication networks. Dr. Akar joined the Technology Planning and Integration group at Sprint Long Distance Division in 1996 and was a senior member of technical staff in the same group in 1998–2000. While at Sprint, he worked on ATM traffic management and routing, IP Qos, virtual private networking architectures, and pricing. Since 2000, he has been with the Electrical and Electronics Engineering Department, Bilkent University, Turkey as an assistant professor. His areas of interest include quality of service in IP networks, network design and engineering, and queueing systems.

BIBLIOGRAPHY

1. ATM Forum, *ATM Forum Traffic Management Specification Version 4.0*, 1996.
2. ATM Forum, *ATM User-Network Interface Specification Version 3.1*, 1994.
3. F. Baker, C. Iturralde, F. L. Faucheur, and B. Davie, *Aggregation of RSVP for IPv4 and IPv6 Reservations*, RFC 3175, 2001.
4. J. C. Bellamy, *Digital Telephony*, Wiley, New York, 1991.
5. Y. Bernet et al., *A Framework for Integrated Services Operation over Diffserv Networks*, RFC 2998, 2000.
6. R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, 1994.
7. R. Braden et al., *Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification*, RFC 2205, 1997.
8. D. E. Comer, *Internetworking with TCP/IP: Principles, Protocols, and Architectures*, Prentice-Hall, 2000.
9. J. P. Coudreuse and M. Servel, Prelude: An asynchronous time-division switched network, *ICC Proc.*, Seattle, 1987.
10. B. Davie et al., *An Expedited Forwarding PHB (Per-Hop Behavior)*, RFC 3246, 2002.
11. B. Davie and Y. Rekhter, *MPLS Technology and Applications*, Academic Press, 2000.
12. D. Farinacci et al., *Generic Routing Encapsulation (GRE)*, RFC 2784, 2000.
13. A. G. Fraser, Early experiments with asynchronous time division networks, *IEEE Network* **7**: 12–26 (1993).
14. A. G. Fraser, Towards a Universal Data Transport System, *IEEE J. Select. Areas Commun.* **1**: 803–815 (1983).
15. V. Fuller, T. Li, J. Yu, and K. Varadhan, *Classless Inter-Domain Routing (CIDR): An Address Assignment and Aggregation Strategy*, RFC 1518, 1993.
16. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, *Assured Forwarding PHB Group*, RFC 2597, 1999.
17. M. Humphrey and J. Freeman, How xDSL supports broadband services to the home, *IEEE Network* **11** 14–23 (1997).
18. S. Kent and R. Atkinson, *Security Architecture for the Internet Protocol*, RFC 2401, 1998.
19. T. Kwok, *ATM: The New Paradigm for Internet, Intranet & Residential Broadband Services and Applications*, Prentice-Hall, 1998.
20. T. Kwok, A vision for residential broadband services: ATM-to-the-home, *IEEE Network* 14–28 (1995).
21. J. Moy, *OSPF Version 2*, RFC 2178, 1997.
22. K. Nichols, S. Blake, F. Baker, and D. Black, *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers*, RFC 2474, 1998.
23. D. Oran, *OSI IS-IS Intra-domain Routing Protocol*, RFC 1142, 1990.
24. Y. Rekhter and T. Li, *An Architecture for IP Address Allocation with CIDR*, RFC 1518, 1993.

25. Y. Rekhter and T. Li, *A Border Gateway Protocol 4* (*BGP-4*), RFC 1771, 1995.

26. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, 2001.

27. E. C. Rosen, BGP/MPLS VPNs, *<draft-ietf-ppvpn-rfc2547bis-01.txt>*, 2002.

28. S. Shenker, C. Partridge, and R. Guerin, *Specification of Guaranteed Quality of Service*, RFC 2212, 1997.

29. W. Stallings, *ISDN and Broadband ISDN*, Macmillan, New York, 1992.

30. W. Townsley et al., *Layer Two Tunneling Protocol "L2TP,"* RFC 2661, 1999.

31. J. Wroclawski, *Specification of the Controlled-Load Network Element Service*, RFC 2212, 1997.

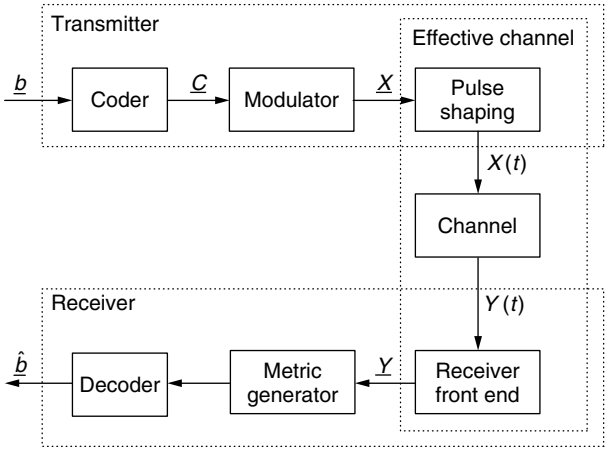


Figure 1. Block diagram of a communication system for conveying the information sequence $\underline{b}$ across a communication channel.

BIT-INTERLEAVED CODED MODULATION

DENNIS L. GOECKEL
University of Massachusetts
Amherst, Massachusetts

1. INTRODUCTION

Bit-interleaved coded modulation (BICM) has emerged as a promising method for transmitting information robustly over many types of communication channels; in particular, BICM has proven to be particularly attractive for the types of channels often found in wireless communication systems. The modern version of BICM is attributed to a 1992 paper of Zehavi [1], but it was the significant investigation of Caire and colleagues published in 1998 [2] that led to its widespread popularity. BICM marks a significant departure from the trend set in coded modulation roughly from 1978 to 1998 [2] and represents a return, at least structurally if not philosophically, to the types of coded modulation employed prior to 1980, which could be decidably suboptimal for the communication system applications to which they were applied at that time.

The goal of the error control coding and modulation, which are often referred to as one unit with the term "coded modulation," are to efficiently convey a sequence of information bits $\underline{b} = (b_0, b_1, b_2, \ldots)$ reliably across a channel, where the channel is defined as the physical entity connecting the transmitter to the receiver as shown in Fig. 1. Although the channel generally accepts the waveform $X(t)$ and produces the waveform $Y(t)$, the coded modulation is generally designed for an effective channel, which includes the transmitter pulseshaping, channel, and the sampled receiver front end. The input to this effective channel is the sequence of complex values $\underline{X}$, and its output is the sequence of complex values $\underline{Y}$. In 1948, Claude Shannon published his seminal work on information theory [3], which introduced the notion of capacity — the maximum rate (in bits per symbol) at which information can be reliably transmitted across this effective channel. Coded modulation strives to obtain this limit in a practical manner.

The ability of a practical coded modulation scheme to operate over the effective channel is often measured by the bit error rate (BER), which is a long-term average of the fraction of bits in the decoder output $\hat{\underline{b}}$ that do not agree with the transmitted sequence $\underline{b}$. Thus, a good coded modulation scheme will assign sequences of transmitted symbols $\underline{X}$ to information bit sequences in such a way that the decoder is able to ascertain which information bit sequence was transmitted with very high probability from the received sequence $\underline{Y}$. If the coding and modulation is performed separately, as is portrayed in Fig. 1, the information bit stream $\underline{b}$ is encoded to produce a coded bit stream $\underline{C}$ that contains carefully introduced redundancy in order to protect against errors that may be introduced by the channel. These coded bits are then taken individually or in groups by the modulator to choose from a number of possible complex values (termed the "constellation of possible signal points") to determine each entry of the sequence $\underline{X}$.

The field of coded modulation has progressed rapidly since the early 1970s. Prior to the late 1970s, error control coding and modulation were considered separable, as has been represented in Fig. 1. At that time, the channel type most often considered in the communication community was the additive white Gaussian noise (AWGN) channel, for which the received signal is that which was transmitted plus Gaussian noise attributed to background radiation and thermal noise in the receiver. For such a channel, the structure shown in Fig. 1 is suitable if the system is employing relatively simple modulation schemes, generally with no more than four signal points in the constellation. However, by the late 1970s, there had been significant work to increase the bandwidth efficiency of communication systems, which, assuming that the pulseshaping is left unchanged, is done by increasing the rate in information bits per symbol sent across the effective channel. To achieve this gain in rate, the rate of the convolutional code or the number of signal points used in the constellation is increased. When the structure of Fig. 1 was employed for constellations with larger numbers of

signal points, it was no longer desirable in many cases. This was demonstrated in the late 1970s and early 1980s by Ungerboeck's seminal work [4], which showed that separating the error control coding and modulation as depicted in Fig. 1 can be decidably suboptimal for the AWGN channel. Ungerboeck's construction that led to this conclusion was termed *trellis-coded modulation* (TCM), the structure of which is shown in Fig. 2. Note that certain information bits go through the encoder while others do not, thus making it impossible to separate coding and modulation. In addition, the method for choosing a signal point is jointly designed with the coder. Trellis-coded modulation with Ungerboeck's basic structure and his rules for building schemes on that structure, often termed "Ungerboeck set partitioning" after one of the key aspects of the rules, soon became an indispensable tool in the communication engineer's toolbox, and, hence, it was readily apparent that coding and modulation were thereafter inseparable.

By the late 1980s, many TCM schemes had been developed for the AWGN channel. However, wireless communication systems, which had been studied with mild intensity over the previous decades, became of increasing importance as the potential of a huge commercial cellular telephony market loomed. The AWGN channel model does not generally represent the wireless communications channel well, because, in wireless systems, the signal reflects off of many objects (automobiles, buildings, mountains, etc.), which leads to the superposition of many replicas of the transmitted signal in the environment. At a given point in the environment, these many waves can add constructively or destructively, depending on the relative phase of the replicas at that point. Thus, depending on the location of the receiver, the received signal power can be significantly more or less than would be expected without the presence of such multiple copies; because of its cause, this fluctuation of the signal power is termed *multipath fading*. From the physics of the problem, it is important to note that the received power of the signal can vary greatly with time because of the movement of the receiver to a different place in the environment or in response to changes in the reflections when objects in the environment move.

Multipath fading can have a significant impact on the performance of a communication system. When the received power drops too low, a burst of bit errors can occur, and such bursts tend to dominate the error probability — even if the occurrence of such system power drops is relatively unlikely. To combat this effect, an effective method is to make it such that possible symbol sequences for $\underline{X}$ that correspond to different information bit sequences are distinguishable even when the received power drops significantly a portion of the time. However, most coded modulation schemes produce many pairs of sequences whose distinguishing characteristics are limited to sequence elements very near each other, which are transmitted across the channel at nearly the same time under the architecture of Fig. 1; thus, a single drop in the received power, which can last many symbol periods, will make such sequences virtually indistinguishable at the receiver. A method of combating such an effect and spreading out the distinguishing characteristics between sequences in time is to reorder the symbols out of the coded modulator before they are transmitted as shown in Fig. 3. This reordering is generally done through a process termed *interleaving*, which essentially permutes the ordering of the symbols. Since symbols in $\underline{X}$ that were close to each other at the input of the interleaver are now separated significantly at its output and thus transmitted across the channel at times relatively far apart, the distinguishing characteristics between transmitted sequences associated with different information bit sequences are unlikely to be lost as a result of a single drop in the received power. This process, which is referred to as achieving "time diversity," greatly improves system performance.

There was significant work in the late 1980s and early 1990s to develop TCM schemes for wireless communication systems based on the architecture given in Fig. 3 with the coded modulation paradigm of Fig. 2. It became readily apparent that different criteria [5,6] were required for multipath fading channels relative to their AWGN counterparts; in particular, because of the importance of achieving time diversity on the multipath fading channel, the ability to distinguish between two possible sequences output from the coded modulation is not determined by the standard Euclidean distance between those sequences, which is appropriate for the AWGN channel, but rather by the number of elements in the two sequences that are different. Thus, an entire new line of coded modulation schemes employing Ungerboeck's construction was developed under this new criterion, and, although this new line of coded modulation schemes mandated the

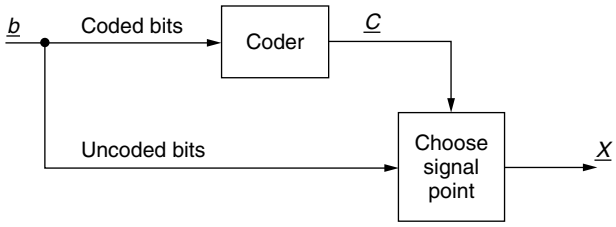


Figure 2. Ungerboeck's structure for coded modulation. Note that the method for choosing the signal point is a critical portion of the architecture, which Ungerboeck specified by a particular method of signal set partitioning.

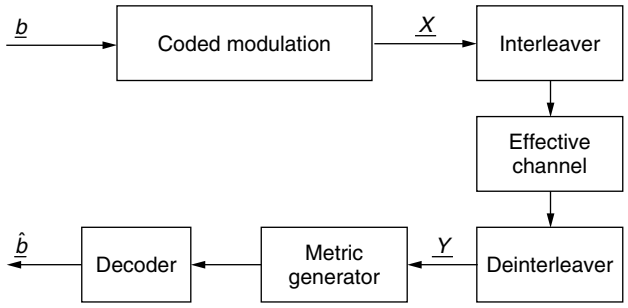


Figure 3. Block diagram of a communication system with interleaving, which is employed to obtain time diversity when conveying the information sequence $\underline{b}$ across a wireless communication channel.

removal of the uncoded bits in Fig. 2, the coding and modulation were still designed jointly, generally following Ungerboeck's rules.

Zehavi's 1992 paper [1] exploited the fact that the metric for sequence distinguishability for multipath fading channels is different from that for AWGN channels; thus, perhaps the structure of the coded modulation should change as well. Thus, bit-interleaved coded modulation was introduced, which employs the coded modulation of Fig. 1 in conjunction with the interleaver of Fig. 3, resulting in the structure shown in Fig. 4. As will be shown below, such a construction immediately leads to a larger number of differences between two possible coded modulation output sequences for a given code complexity. Analytical results and simulation results [2] have confirmed the desirability of such a structure, not only for wireless communication systems but also potentially for a variety of other channels, and it has even been suggested that BICM may be a strong contender for implementation over AWGN channels when powerful codes are employed. In addition, the ability of BICM to protect each information bit with coding has made it a robust choice for related applications [7].

The remainder of this article is organized as follows. In Section 2, the mathematical models for the AWGN channel and the interleaved multipath fading channel, as introduced above, are presented, along with the metrics for determining the quality of a given coded modulation scheme operating on that channel. In Section 3, the various coded modulation constructions discussed above are briefly reviewed. These constructions will motivate the structure of BICM, which is discussed in detail in Section 4. In particular, Section 4 presents encoding and decoding methods for BICM, and discusses its performance versus state-of-the-art TCM schemes for both the multipath fading channel and the AWGN channel. Finally, Section 5 summarizes this article and suggests further reading.

2. CHANNEL MODELS

2.1. The AWGN Channel

The additive white Gaussian noise (AWGN) channel is the most classical of communication channels and the channel studied most often before the explosion in wireless communications research in the late 1980s. The AWGN channel is a channel with additive distortion, and thus the ith element of the output $\underline{Y}$ is given by

$$Y_i = X_i + N_i$$

where X_i is the ith element of the input sequence $\underline{X}$. The sequence of random variables N_i, which correspond to additive noise encountered in the channel and the front end of the receiver, are modeled as independent random variables, each of which is Gaussian with mean zero and variance $N_0/2$. Since the N_i are independent, the statistical characterization of the channel output sequence $\underline{Y}$ given the input sequence $\underline{X} = \underline{x}$ factors into the conditional probability density functions of the individual components:

$$p_{\underline{Y}|\underline{X}}(\underline{y} \mid \underline{x}) = \prod_i p_{Y_i|X_i}(y_i \mid x_i) = \prod_i \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{(y_i - x_i)^2}{N_0}\right)$$

Under this channel model, the maximum-likelihood (ML) detector, which chooses the most likely information sequence given the received sequence $\underline{Y} = \underline{y}$, will choose from among the possible transmitted sequences the one that is closest in squared Euclidean distance to $\underline{y}$, where the squared Euclidean distance between the two sequences $\underline{y}$ and $\underline{u}$ is defined as

$$d_E^2(y, u) = |y - u|^2 = \sum_i (y_i - u_i)^2$$

The performance of the system can be characterized by the probability that the ML detector chooses the wrong information sequence (i.e., one other than the one input to the transmitter). In particular, through the use of familiar union bounding techniques, measures for the performance of the communication system (e.g., bit error rate or frame error rate) are generally a linear combination of elements from the set (over all distinct $\underline{x}$ and $\tilde{\underline{x}}$) of probabilities that the received sequence $\underline{Y}$ is closer to the possible transmitted sequence $\tilde{\underline{x}}$ than to the actual transmitted sequence $\underline{x}$. For the AWGN channel, this probability is given by

$$P(\underline{x} \to \tilde{\underline{x}}) = Q\left(\frac{d_E(\underline{x}, \tilde{\underline{x}})}{(2N_0)^{1/2}}\right) \tag{1}$$

where $Q(x) \triangleq (1/\sqrt{2\pi}) \int_x^\infty e^{-(u^2/2)}\, du$. In particular, of these pairwise error events, the one that is most likely, which is the one corresponding to the possible sequences $\underline{x}$ and $\tilde{\underline{x}}$ that are closest in Euclidean distance, dominates the system performance at high signal-to-noise ratios (SNR), where systems are generally designed. Thus, the goal of coded modulation over the AWGN channel is to map information sequences $\underline{b}$ to transmitted sequences $\underline{X}$ in such a way that the minimum Euclidean distance between any two possible transmitted sequences corresponding to different information sequences is maximized.

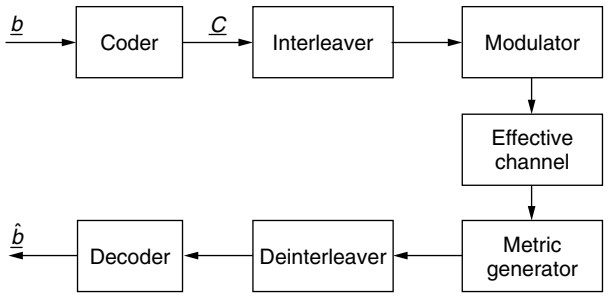


Figure 4. Block diagram of a bit-interleaved coded modulation system.

2.2. Multipath Fading Channels

Per Section 1, interest in the wireless communications channel motivated the study of a channel model very different from the classical AWGN channel. In particular, the following non-frequency-selective slowly fading channel model, which is appropriate for the traditional narrowband communication system or a subcarrier of a modern orthogonal frequency-division multiplexing (OFDM) system, is generally assumed:

$$Y_i = \alpha_i X_i + N_i \tag{2}$$

where N_i is defined in the same manner as for the AWGN model, and $\underline{\alpha}$ is a sequence of attenuations of the signal strength, where each element comes from a common distribution, which depends on the channel. This common distribution is generally assumed to be a Rayleigh, Ricean, or Nakagami-m distribution. The Rayleigh distribution will be employed throughout this paper, which implies that the probability density function of each α_i is given by

$$p_{\alpha_i}(x) = 2xe^{-x^2}, \qquad x \geq 0$$

where the average power due to this multipath fading has been normalized to unity. If the system architecture in Fig. 1 is assumed, then the elements of $\underline{\alpha}$ are correlated with each other, since the mobile moves across the interference pattern set up in the environment slowly relative to the rate at which symbols are sent across the channel. However, if the system architecture of Fig. 3 is assumed and a reasonable amount of system latency is allowed, then the entries of the sequence $\underline{\alpha}$ can be modeled as independent, since the places where $\underline{x}$ and $\underline{\tilde{x}}$ differ following the impact of a given information bit b_i can be assumed to be separated by a time sufficient to render the multipath fading affecting these places independent. Thus, the model of (2) will be adopted throughout this work with the elements of $\underline{\alpha}$ assumed to be independent. Similarly to the AWGN channel, the system performance measures of interest are linear combinations of elements from the set (over all distinct $\underline{x}$ and $\underline{\tilde{x}}$) of probabilities that the sequence $\underline{\tilde{x}}$ is chosen when sequence $\underline{x}$ was sent, which is given by

$$P(x \to \tilde{x}) = E_\alpha \left[Q\left(\frac{\alpha_i d_E(\underline{x}, \underline{\tilde{x}}))}{(2N_0)^{1/2}} \right) \right] \tag{3}$$

$$\leq \prod_{i \in D(\underline{\tilde{x}}, \underline{x})} \frac{1}{1 + \dfrac{|\tilde{x}_i - x_i|^2}{4N_0}} \tag{4}$$

where $D(\underline{\tilde{x}}, \underline{x})$ is the set of indices corresponding to locations where $\underline{\tilde{x}}$ and $\underline{x}$ differ. For high SNRs, the first term in the denominator can be ignored, which leads to

$$P(\underline{x} \to \underline{\tilde{x}}) \approx \frac{1}{\displaystyle\prod_{i \in D(\underline{\tilde{x}}, \underline{x})} \frac{|\tilde{x}_i - x_i|^2}{4N_0}} \tag{5}$$

It can be readily observed that a plot of (5) on a logarithmic scale versus the SNR in decibels will exhibit a slope at high SNRs that is the negative of the size of the set $D(\underline{\tilde{x}}, \underline{x})$. Furthermore, the minimum size of $D(\underline{\tilde{x}}, \underline{x})$ over all possible transmitted sequences $\underline{\tilde{x}}$ and $\underline{x}$ will characterize this same slope for the bit error rate for the code, and thus this is a critical parameter for good performance at high SNRs. It determines the diversity of the system, which is indeed defined as the negative of the slope of the information bit error rate versus SNR at high SNRs. Whereas the size of $D(\underline{\tilde{x}}, \underline{x})$ sets the slope of the curve, the magnitudes of the values $|\tilde{x}_i - x_i|$ for $i \in D(\underline{\tilde{x}}, \underline{x})$ set the horizontal positioning of the curve. This establishes the two criteria for designing coded modulation on the perfectly interleaved Rayleigh fading channel:

1. *Primary* — maximize the minimum number of elements in $D(\underline{\tilde{x}}, \underline{x})$.
2. *Secondary* — maximize the minimum product distance $\displaystyle\prod_{i \in D(\underline{\tilde{x}}, \underline{x})} |\tilde{x}_i - x_i|$.

3. CODED MODULATION

In this section, the evolution of coded modulation for AWGN and Rayleigh fading channels that led to the introduction of BICM is briefly reviewed.

3.1. Coding and Modulation

3.1.1. Encoder. The traditional approach to coding and modulation is shown in Fig. 1. If the overall coded modulation is trellis-based, as will be assumed throughout this article, a convolutional code is employed. In a convolutional code, the information bits are coded with a shift register circuit, an example of which is given in Fig. 5a. For this example, at each timestep, an information bit is input to the left side of the shift register; this input bit, along with the last 2 information bits, each of which is contained in one of the 1-bit memory elements, is used to determine the current pair of outputs by the modulo-2 summations. The memory elements are then clocked, which causes each information bit to move one memory location to the right, thus causing the encoder to reside in a new memory state, and a new information bit to be input to the system from the left. In general, a trellis code is denoted an (n, k) code with memory m, where n is the number of output coded bits for each clock cycle, k is the number of input information bits for each clock cycle, and m is the number of 1-bit memory elements in the circuit. Thus, the convolutional code of Fig. 5a would be denoted a $(2, 1)$ code with memory 2.

Recall that good coded modulation schemes produce, for distinct input sequences, outputs that are as distinct as possible. The number of places where two binary sequences differ is termed the *Hamming distance* between those two sequences. For a convolutional code, the minimum possible Hamming distance between output sequences corresponding to distinct information sequence inputs is

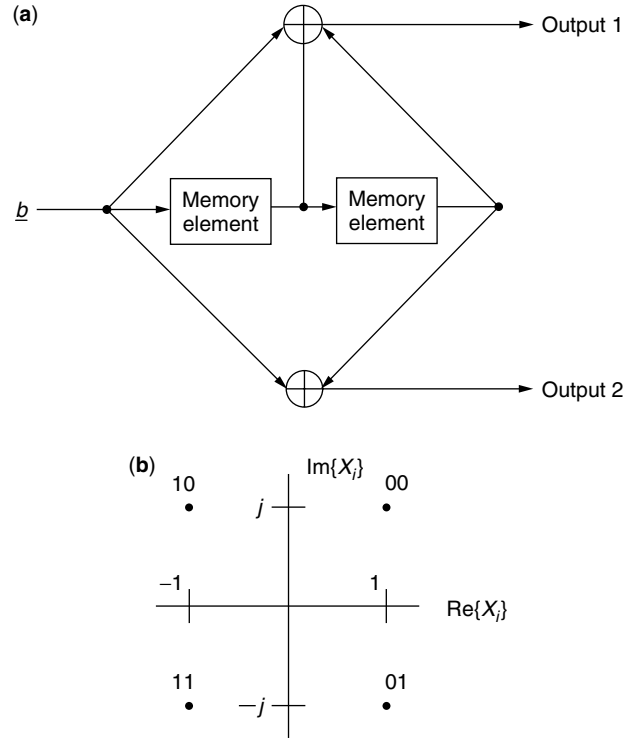


Figure 5. A simple example of coded modulation: (**a**) encoder for a $(2, 1)$ convolutional code with memory 2, (**b**) a possible constellation labeling for QPSK.

termed the *free distance*, which will be denoted by d_f. This free distance, along with other distance properties of the code, depend on which memory elements contribute to which output (i.e., the connections to the modulo-2 summers in Fig. 5a). Because of the importance of convolutional codes, an enormous amount of research has gone into finding convolutional codes of maximal free distance for a given number of memory elements. The connections for binary convolutional codes of maximal free distance can be found from standard references [8, pp. 330–331]; for example, the code shown in Fig. 5a corresponds to the recipe given in the first line of Table 11.1(c) of Ref. 8.

The sequence of bits output from the convolutional coder must then be assigned to a transmitted sequence $\underline{X}$, whose elements are drawn from the space of complex numbers. Assuming that there are M possible signal points in the constellation from which the value for each entry of $\underline{X}$ is drawn, the modulator takes $\log_2 M$ output bits from the output of the convolutional coder and uses these to select one of the M constellation points. The mapping of the $\log_2 M$ bit sequence to a constellation point is termed the "constellation labeling," because it can be specified by labeling each signal point with a $\log_2 M$ bit sequence. A simple example of this is shown in Fig. 5b for a quadrature phase-shift-keyed (QPSK) constellation.

3.1.2. Decoder. A convenient representation of a coded modulation scheme for characterizing the performance and for the decoding of convolutional codes is a trellis diagram, as shown in Fig. 6a for the coded modulation scheme of Fig. 5. Here, each branch of the trellis has been labeled with the output of the coded modulation scheme when that branch is traversed in the convolutional coder. Assuming that the encoder of Fig. 5 starts with a zero in each of its memory elements, any possible path through the trellis enumerates a possible sequence output from the coded modulator, and, any possible sequence output from the coded modulator corresponds to a path through the trellis.

Per the observations discussed above, for performance analyses, interest lies in the distinction between any two possible coded modulation sequences output from the coded modulator, or, equivalently, the distinction between any two different paths through the trellis. From Fig. 6a, we see that there are two coded modulation sequences, $\underline{x}_1 = (+1+j, +1+j, +1+j, \ldots)$ and $\underline{x}_2 = (-1-j, -1+j, -1-j, \ldots)$, which split at time $t = 0$, rejoin at time $t = 3$, and are identical after $t = 3$. The probability for choosing $\underline{X} = \underline{x}_2$ when $\underline{X} = \underline{x}_1$ was sent is then easily found from (1) or (4) for the AWGN or perfectly interleaved Rayleigh fading channel, respectively. Likewise, the pairwise error probabilities for all possible sequence differences can be found using the trellis of Fig. 6a.

The trellis is also employed for decoding of the coded modulation scheme. Recall that the goal of decoding is generally to find the most likely transmitted sequence $\underline{x}$ given the received sequence $\underline{Y} = \underline{y}$. Since each transmitted sequence is represented by a path in the trellis, this reduces to finding the sequence through the trellis that is the most likely given $\underline{Y} = \underline{y}$. At any time step t, a path in

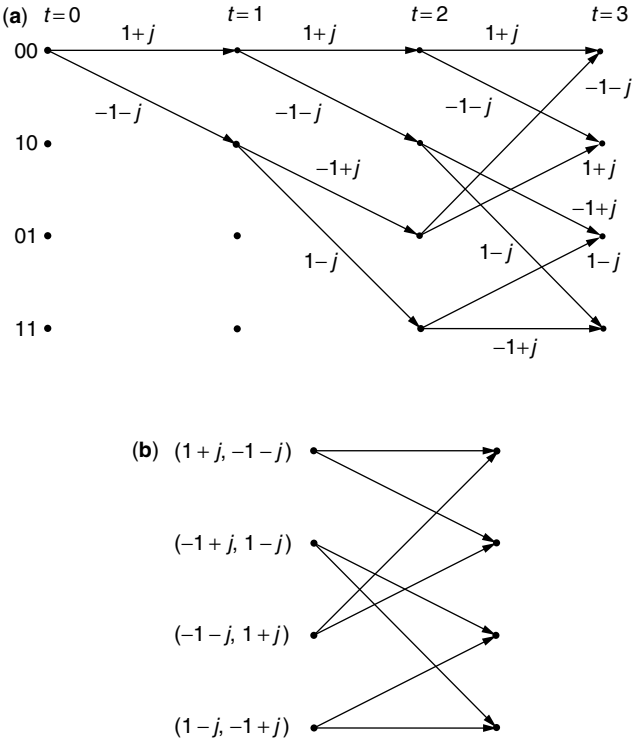


Figure 6. Trellis representation for coded modulation schemes: (**a**) full trellis for decoding and performance analyses, (**b**) compact trellis for specifying the coded modulation scheme.

the trellis is associated with its likelihood

$$p_{Y|X}(y \mid x) = \prod_{i=1}^{t} p_{Y_i|X_i}(y_i \mid x_i)$$

$$= \begin{cases} \displaystyle\prod_{i=1}^{t} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{(y_i - x_i)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{(y_i - \alpha_i x_i)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

where the fact that both the AWGN channel and the perfectly interleaved Rayleigh fading channel considered here are memoryless, has been exploited. In addition, coherent reception, where the carrier phase is available at the receiver, has been assumed, as has perfect estimation of the channel state information (CSI) α_i for the Rayleigh fading channel. Since the natural logarithm is a monotonic function, taking the natural logarithm of the likelihood of each path implies that it is equivalent to search for the path $\underline{x}$ that exhibits the minimum value of

$$\mu(\underline{y}, \underline{x}) = \begin{cases} \displaystyle\sum_{i=1}^{t} (y_i - x_i)^2, & \text{AWGN channel} \\ \displaystyle\sum_{i=1}^{t} (y_i - \alpha_i x_i)^2, & \text{Rayleigh channel} \end{cases} \tag{6}$$

which is easy to evaluate and can often be simplified further as discussed in standard digital communication texts.

Thus, at each time stage t in the trellis, there is a likelihood associated with each possible path, and it is clear from (6) that the metric for each of the paths at time $t + 1$ can be calculated from the metrics of the paths at time t. However, for the example of Fig. 5, there are 2^t possible paths at time t, implying that some sort of simplification is required if the receiver is to be implementable. The solution to this problem is provided by the celebrated Viterbi algorithm. Note from Fig. 6a that two paths merge into each channel state at time $t = 3$. Now, recalling that the receiver's goal is only to find the single path $\underline{x}$ with the *highest* likelihood given the received sequence $\underline{Y} = \underline{y}$, the likelihood for the two paths entering the same state can be compared with one another and only the best path retained without loss of optimality. The reasoning for this is as follows. Consider the supposition that the path with the smaller likelihood entering a given state is the prefix for the path through the entire trellis with the eventual highest likelihood. Then, if one takes the suffix of this "best" path and concatenates it to the path with the higher likelihood entering the given state, a path through the entire trellis with higher likelihood results, thus proving the supposition false. Hence, one can always disregard all but one of the paths entering a given state at a given time. Note that this trims the number of paths retained at time t significantly, from 2^t to the steady-state number 2^m, where $m = 2$ for the example of Fig. 5.

Notationally, it is inconvenient to specify the trellis shown in Fig. 6a for coded modulation schemes, particularly for systems with large numbers of branches. Thus, two notation simplifications are generally employed: (1) only the steady-state trellis is generally manifested, as shown in Fig. 6b, rather than the startup stages; and (2) rather than placing the label actually on each branch, which leads to significant confusion, particularly in high-rate systems, the labels are listed to the left of a given state, with the understanding that they correspond to the branches leaving that state at any given time in order from top to bottom. Figure 6b illustrates this compact notation for the coded modulation scheme of Fig. 5.

3.2. Trellis-Coded Modulation for AWGN Channels

Per the observations mentioned above, the goal of coded modulation design for AWGN channels is to separate distinct possible transmitted sequences (or paths through the trellis) $\underline{x}$ and $\tilde{\underline{x}}$ by a Euclidean distance $|\underline{x} - \tilde{\underline{x}}|^2$ that is as large as possible. In particular, the minimum of all such differences for distinct paths is critical. If the constellation employed by the modulator is the binary antipodal set $\{-1, +1\}$, often implemented as binary phase-shift keying (BPSK), or quadrature phase shift keying $\{-1 - j, -1 + j, +1 + j, +1 - j\}$, the coded modulation structure shown in Fig. 1 with a convolutional code of maximal free distance is a good solution.

If the size of the constellation employed by the modulator is larger than BPSK and QPSK, the best method of coded modulation is not so clear. For example, suppose that one would like to send 2 information bits per symbol. One possibility would be to employ QPSK with no coder, but this is often a poor choice. Before the late 1970s, the structure of Fig. 1 would have been retained, and a convolutional code of maximal free distance would have been employed. Using the rule of thumb for the constellation size of coded modulation systems operating on AWGN channels, which states that the constellation size M should be such that $\log_2 M - 1$ is the number of bits per symbol, results in the choice of a rate-$\frac{2}{3}$ convolutional code followed by modulation with a 8-ary phase shift keyed (8-PSK) signal set. A reasonable choice for the labeling of coded bits to constellation points is to employ Gray labeling, whereby the label for each signal point differs by exactly 1 bit from each of its nearest neighbors. With this labeling, 3-bit sequences separated by large Hamming distances correspond to signal points separated by large Euclidean distances.

Per Section 1, Ungerboeck [4] revolutionalized the art of coded modulation by introducing the concept of trellis-coded modulation, as shown in Fig. 2. Two key philosophical concepts became readily apparent: (1) coding and modulation could no longer be considered separately in high-rate systems, and (2) the constellation labeling plays a key role. Note that some bits in the Ungerboeck scheme do not go through the convolutional encoder; these "uncoded bits" show up in the trellis representations in Fig. 6 as parallel branches, which originate and end at the same state with a single transition. Although two sequences that differ only on these parallel branches are allowable codewords, the

large Euclidean distance generally obtained between parallel branches keeps these codewords from greatly restricting the minimum distance of the coded modulation scheme.

3.3. Trellis-Coded Modulation for Fading Channels

The concept of trellis-coded modulation produced a large number of good coded modulation schemes for the AWGN channel during the 1980s. As the Rayleigh fading channel rose in importance late in the 1980s, it was natural to extend the idea of coded modulation to this environment. However, as noted in Section 2.2, the criteria for the distance between two paths is quite different for a Rayleigh fading channel. In particular, the parallel branches of a scheme employing uncoded bits as in Fig. 2, which result in parallel paths in the trellis of Fig. 6 and, hence, sets $D(\underline{\tilde{x}}, \underline{x})$ with only a single entry, result in poor performance on the perfectly interleaved Rayleigh fading channel. Essentially, the uncoded information bit is not protected from deep signal fades since its impact is concentrated on only a single modulated symbol, and thus it is decoded in error with very high probability.

Thus, it was soon recognized [5,6] that trellis-coded modulation schemes for Rayleigh fading channels should avoid the uncoded bits often employed on the AWGN channel. One of the later instantiations of trellis-coded modulation schemes for fading channels was the I-Q TCM scheme of Al-Semari and Fuja [9], which cleverly performed trellis-coded modulation on two 4-ary amplitude shift-keyed (4-ASK) $\{-3, -1, +1, +3\}$ streams and then combined these together by using one 4-ASK stream as the in-phase component and one 4-ASK stream as the quadrature component to get a stream of 16-ary quadrature amplitude modulation (QAM) symbols. When rate-$\frac{1}{2}$ encoders are used, the resulting scheme conveys 2 information bits per symbol. Such an I-Q TCM scheme employing a four-state code is shown in Fig. 7. Note the diversity increase from 1 for a TCM scheme employing uncoded bits per Fig. 2 to 3 for the scheme of Fig. 7, which, per Section 2.2, will greatly improve performance on the perfectly interleaved Rayleigh fading channel.

4. BIT-INTERLEAVED CODED MODULATION

4.1. Motivation

The schemes discussed in Section 3 are all built on the structure shown in Fig. 2, often without the uncoded bits in the case of the multipath fading channel. In particular, all the schemes discussed in Section 3 have assumed that each set of n bits along a branch in the convolutional coder impacts a single output symbol. Under such a paradigm, the maximum diversity for a coded modulation scheme is upper-bounded by the symbolwise Hamming distance of the coded modulation, which is defined for the schemes of Section 3 as the number of branches that differ between two possible paths through the trellis. Thus, for the convolutional encoder shown in Fig. 5, the I-Q TCM of Fig. 7 achieves the upper bound of three on the diversity under such a construction paradigm.

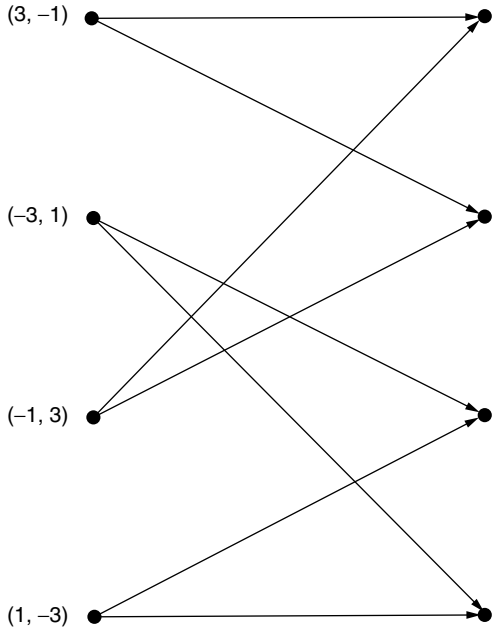


Figure 7. Specification for an I-Q TCM scheme using the encoder of Fig. 5 mapped to a 4-ASK signal set via Gray labeling.

However, the bitwise Hamming distance between codewords for the encoder of Fig. 5 is 5 instead of 3, thus motivating the method pictured in Fig. 4, which is termed *bit-interleaved coded modulation*. The goal of such a construction is to increase the diversity of the coded modulation scheme from 3 to 5 by interleaving at the bit level so that the five symbols that carry the difference between the two possible sequences out of the encoder are temporally separated. These five symbols will then see roughly independent fading, and thus a diversity of 5 should be achieved.

The BICM construction breaks from the standard coded modulation in a number of ways. Prominent among these is that the coded bits that choose a constellation point for a given modulated symbol no longer come sequentially from the encoder; in fact, to achieve the maximum diversity, it is desirable that they are drawn from places in the encoder sequence that are as distant as possible. Thus, in both encoding and decoding, the bits joining the bit on the trellis branch of interest will generally be unknown, which is why a given bit in BICM is often said to be randomly modulated. Because of this essential loss of control of the way that a given code bit affects a transmitted symbol, new rules for the design of such a system and new analysis techniques must be developed relative to those considered in Section 3.

4.2. Encoder

The generic BICM system is shown in Fig. 4. In general, assuming a constellation of M signal points, $\log_2 M$ signal points are taken from the output of the interleaver and used to choose a constellation point. However, to make this discussion more concrete, consider the system shown in Fig. 8a, which carries 2 information bits per 16-QAM

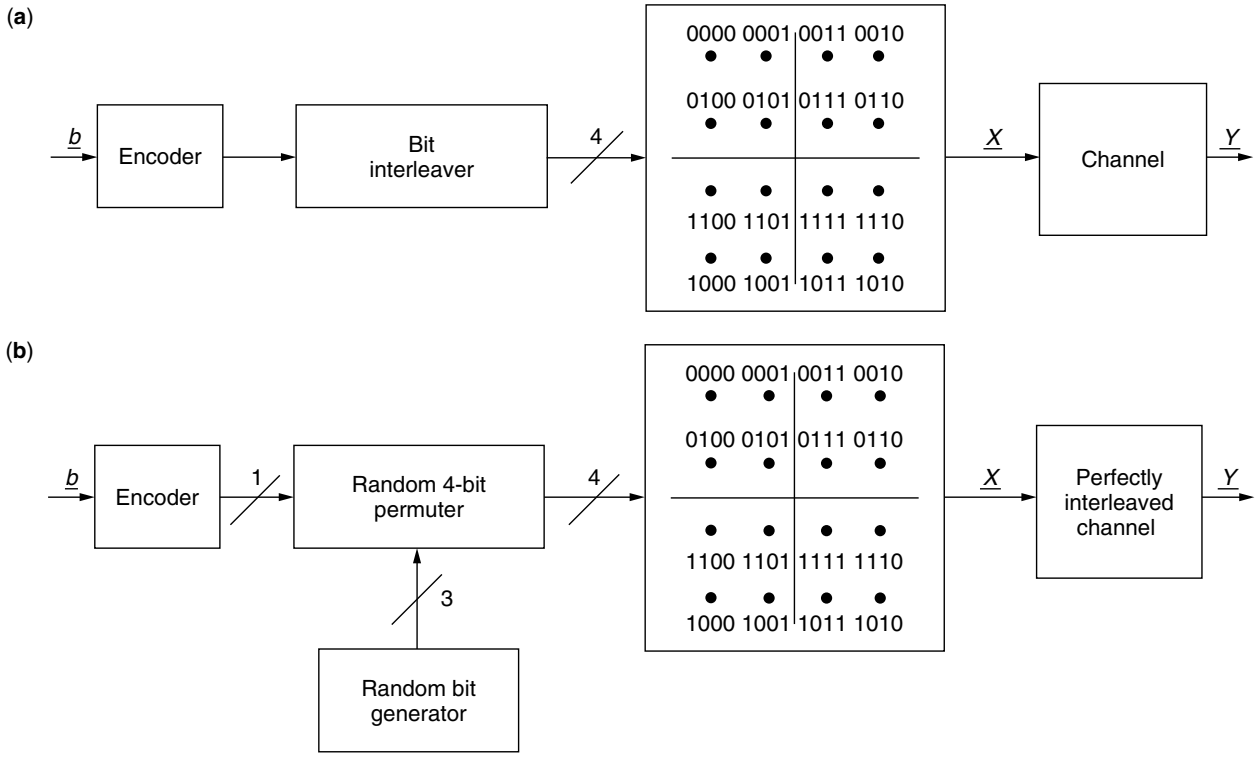


Figure 8. The BICM block diagram: (**a**) a BICM scheme that transmits 2 information bits per channel symbol; (**b**) the equivalent representation for such a scheme assuming a perfect bit interleaver.

channel symbol and can thus be compared to the sample systems of Section 3. Assuming a perfect interleaver, which separates the bits of the input sequence essentially infinitely far apart in time at the output (and thus suppressing interleaver design questions), two questions arise for the design of the encoder:

1. What are desirable properties of the convolutional code?
2. How should the bits taken at the output of the bit interleaver 4 at a time be mapped to the 16-QAM constellation points, or, equivalently, what should be the 4-bit labels of the 16-QAM constellation points?

To help answer these questions, consider the equivalent system shown in Fig. 8b for a BICM system under the assumption of perfect interleaving.

First, consider the question of how to choose the convolutional encoder. The minimum Euclidean distance squared between the two sequences caused by a given information bit is proportional to the free distance of the convolutional code (note that this result does not depend on the assumption of independent fading). More importantly, the diversity of the BICM scheme is equal to the minimum bitwise Hamming distance of the convolutional code. Thus, it is straightforward to conclude that a convolutional code of maximal free distance is a good choice.

The best mapping from each set of $\log_2 M$ bits taken at the output of the bit interleaver to one of the M-ary signal points is not as obvious. It seems reasonable to assume that sets of $\log_2 M$ bits that are separated at large Hamming distances at the output of the bit interleaver should be separated at large Euclidean distances at the output of the constellation mapper. Thus, the technique of Gray labeling, whereby the label for each signal point differs by exactly one bit from each of its nearest neighbors, is a logical choice. In fact, Gray labeling has been shown to have some optimality properties [2], although it has not been shown to be optimal for practical systems in general. Indeed, as discussed in Section 4.4, very different labelings can be desirable for certain applications and certain types of decoders.

Thus, the standard convention for BICM is to employ a convolutional code of maximal free distance in conjunction with Gray labeling of the constellation to achieve good performance. Note that this is an advantage of BICM — its design is relatively straightforward and thus less of an art than in previous TCM schemes.

4.3. Decoder

4.3.1. The Optimal Decoder. Unlike the TCM schemes discussed earlier, it is far too complex to decode BICM optimally, because the bit interleaver intertwines the bits on a given branch in the trellis with those from other branches far removed in the trellis, thus essentially requiring the complicated soft-decision decoding of a block code with length greater than the depth of the interleaver. Since a large interleaver is generally employed to achieve independent fading on the coded bits for a given branch,

this complexity is not feasible for current receivers. Since iterative methods can approach the performance of maximum-likelihood (ML) decoding as described in Section 4.4, it is unlikely that ML decoding will ever be considered for BICM.

4.3.2. The Optimal Metric Generator. Since decoding over the entire interleaving depth is not viable, it is common to treat the bits that join a bit from the branch of interest in the convolutional encoder as random as shown in Fig. 8b. Note that not only are these randomly generated bits joined with the bit of interest but also that the bit of interest is randomly assigned to one of the $\log_2 M$ label positions. The decoder knows the bit position where the bit of interest is located on a given symbol, but it does not know the value of the other bits that combine with that bit for that symbol.

As with the TCM schemes discussed in Section 3.1.2, the decoder for the BICM scheme attempts to calculate the most likely bit sequence in the trellis given the received sequence $\underline{Y} = \underline{y}$. Assuming each of the transmitted sequences is equally likely, this is equivalent to finding the bit sequence for which the likelihood of the received vector given that bit sequence is maximized. Assuming an AWGN channel or a perfectly interleaved Rayleigh fading channel, the likelihood that $\underline{Y} = \underline{y}$ given the bit sequence $\underline{C} = \underline{c}$ was transmitted is given by

$$p_{Y|C}(y \mid c) = \prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i)$$

where the received symbol Y_i corresponds to the channel symbol on which the ith bit was placed. Because of the random modulation caused by the other random bits that join C_i on a given symbol, calculating this likelihood in BICM is a bit more involved than it is for TCM. Let $j_i \in \{0, 1, \ldots, \log_2 M - 1\}$ be the position in the signal set label where C_i was mapped by the bit permuter, and, following [2], define χ_b^j to be the set of signal points in S such that the jth label position is equal to b. Then, using the law of total probability, we obtain

$$\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) = \prod_{i=1}^{t} \frac{1}{|\chi_{c_i}^{j_i}|} \sum_{x \in \chi_{c_i}^{j_i}} p_{Y_i|X_i}(y_i \mid x)$$

$$= \begin{cases} \displaystyle\prod_{i=1}^{t} \frac{1}{|\chi_{c_i}^{j_i}|} \sum_{x \in \chi_{c_i}^{j_i}} \frac{1}{\sqrt{\pi N_0}} \\ \qquad \times \exp\left(-\dfrac{(y_i - x)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \frac{1}{|\chi_{c_i}^{j_i}|} \sum_{x \in \chi_{c_i}^{j_i}} \frac{1}{\sqrt{\pi N_0}} \\ \qquad \times \exp\left(-\dfrac{(y_i - \alpha_i x)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

which can be simplified by removing common terms to

$$\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) \sim \begin{cases} \displaystyle\prod_{i=1}^{t} \sum_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - x)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \sum_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - \alpha_i x)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases} \tag{7}$$

and thus it is readily apparent that the metric generation process can be quite a bit more complicated than that for the standard TCM schemes in Section 3.1.2, particularly for large signal sets. It requires a summation over half of the signal set of a nonlinear function of the distance to form the metric contribution for a single bit, whereas the metric contribution for an entire branch can be calculated for the TCM scheme with only simple addition and multiplication involving one signal point.

4.3.3. A Suboptimal Metric Generator. As noted above, the metric given in (7) is much more complicated than that for the standard TCM schemes. Thus, in this section, a suboptimal bit metric suggested by Caire et al. [2] is reviewed. The suboptimal bit metric relies on the fact that the sum of a number of quantities that are quite disparate in magnitude is well approximated by the maximum of those quantities. Thus, the summations in (7) are replaced by maximums to yield

$$\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) \sim \begin{cases} \displaystyle\prod_{i=1}^{t} \max_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - x)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \max_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - \alpha_i x)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

and, taking the natural logarithm of the quantities on the right and recognizing the monotonicity of the logarithm, yields that it is equivalent to minimize

$$-\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) \sim \begin{cases} \displaystyle\sum_{i=1}^{t} \min_{x \in \chi_{c_i}^{j_i}} (y_i - x)^2, \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\sum_{i=1}^{t} \min_{x \in \chi_{c_i}^{j_i}} (y_i - \alpha_i x)^2, \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

which requires only that the point in $\chi_0^{j_i}$ and $\chi_1^{j_i}$ that is closest to the received symbol Y_i be found, which is relatively simple for small constellations. For larger or multidimensional constellations, this can be accomplished though a technique known as *sphere decoding* [10].

4.4. Iterative Decoding

As discussed in Section 4.3.1, the optimal decoder for BICM is not practically implementable, due to the interlacing of different bits from different portions in the trellis onto the channel at the same time, since soft-decision decoding would essentially have to be done across the entire interleaver depth. Thus, as shown in Section 4.3.2, the typical decoder assumption is that the other bits that were used with the bit of interest to choose a constellation point are randomly generated. However, the sets $\chi_0^{j_i}$ and $\chi_1^{j_i}$ associated with the coded bit i could be reduced to a single point if the values of these other bits were known, which should substantially improve decoder performance. This notion was exploited by Li and Ritcey in a series of papers (see Ref. 11 and references cited therein) that introduced bit-interleaved coded modulation with iterative decoding. In this technique, the BICM scheme is decoded in a manner similar to that described in Section 4.3.2, but now the decoder generates "soft" information, which tells not if the bit is a 0 or a 1 but instead the probability of such. The decoding is then repeated many times, but this time with the soft information from the previous decoding as an additional input. If the bit error rate of the original BICM scheme is reasonably low, one expects that the estimates of the soft information for the bits joining the coded bit of interest on a given symbol will be very accurate much of the time, and thus the sets $\chi_0^{j_i}$ and $\chi_1^{j_i}$ will be effectively reduced to single points, as desired.

The use of iterative decoding also motivates a change in the encoding. As illustrated in Fig. 8b, bit i is often said to be randomly modulated in BICM in the sense that its effective channel varies depending on the values of the other $\log_2 M - 1$ bits that join with a given coded bit to choose a signal. Note from Fig. 8b that, regardless of the position to which the coded bit i gets mapped (i.e., $j_i = 1, 2, 3, 4$), there always exists 3 bit values such that the minimum distance between the point corresponding to a bit value 0 and that point corresponding to a bit value 1 that are separated by the minimum distance of the signal set. For example, if the coded bit of interest is mapped to the first label location, the fact that the values of the other 3 bits are 1,0, and 0, respectively, implies that the effective signaling points for the binary channel for this bit are labeled 0100 and 1100, which is at the minimum distance of the signal set. Although such a phenomenon occurs only some fraction of the time, it dominates the error probability, particularly at high SNRs and on the AWGN channel. Thus, iterative decoding suggests various relabelings of the constellation points, as described by Li et al. [11].

Finally, note that iterative decoding of BICM is reminiscent of one of the greatest advances in coding theory: concatenated codes with iterative decoding. In fact, the block diagram in Fig. 4 is very similar to that of a serial concatenated Turbo code [12], which would suggest that, if the modulator is viewed as a rate $\log_2 M$ inner code, iterative decoding may provide significant gains. As will be discussed below, iterative detection does provide significant gains, although the analogy to serial concatenated codes must be done very carefully, because there are significant differences.

4.5. Performance

In this section, the performance of BICM is compared to that of other popular coded modulation schemes. From Refs. 2 and 13, it can be inferred that the performance of BICM employing the suboptimal metric of Section 4.3.3 is generally only slightly inferior to the performance when the optimal metric generator of Section 4.3.2 is employed; thus, the suboptimal metric generation of Section 4.3.3 is most often used. In Fig. 9, the BER performance of BICM operating over a perfectly interleaved Rayleigh fading channel is shown for the case where the metric of Section 4.3.3 is employed at the receiver. If the curves in Fig. 9 are compared to a 2 bits/symbol I-Q TCM (see Fig. 6 of Ref. 9), which is a good method of coded modulation based on the paradigm of Fig. 2 in the architecture of Fig. 3, it can be observed that BICM shows a significant performance gain. Thus, when employing low-complexity noniterative decoders, BICM is an effective method of communication on the perfectly interleaved Rayleigh fading channel that motivated its development .

A number of papers have characterized the performance of BICM in a more theoretical fashion, which will allow the consideration of its potential performance with future coding techniques. In particular, Caire et al. [2] showed that the BICM structure incurs only a small loss in Shannon capacity versus coded modulation on the perfectly interleaved Rayleigh fading channel, and, somewhat surprisingly, on the AWGN channel. To attempt to capture the effects of codes of finite complexity, those authors [2] then investigated the system cutoff rate, which indicated that the BICM structure, while incurring a slight loss in the cutoff rate versus coded modulation on AWGN channels, shows significant gains versus coded modulation for the perfectly interleaved Rayleigh fading channel. This supports the gains that BICM shows over coded modulation schemes for practical coded schemes on the perfectly interleaved Rayleigh fading channel. Through a coding exponent analysis, Wachsmann, et al. [14] also investigated the performance of BICM and concluded that, on the AWGN channel, it was inferior to multilevel coding with multistage decoding, particularly for small blocklengths (including trellis codes). This supported the results of Schramm [15], which indicated that BICM is slightly inferior to multilevel techniques when employing trellis codes on the AWGN channel or the Rayleigh fading channel. Wachsmann et al. [16] then investigated the performance of BICM versus multilevel codes when Turbo codes [17] are employed, which are long block codes. Simulation results indicate that BICM is a very effective structure in such a situation, particularly for the Rayleigh fading channel, and that it possesses some universality properties in the sense that it performs very well on a variety of channels.

When standard convolutional codes are employed, numerical results from the iterative decoding of BICM employing a signal set that is not Gray-labeled as described in Section 4.4 indicate a significant performance improvement through such iteration over Gray-labeled BICM with noniterative decoding. In fact, when employing such a labeling and iterative decoding, BICM significantly outperforms standard trellis-coded modulation on the AWGN channel. This seems to conflict with the results

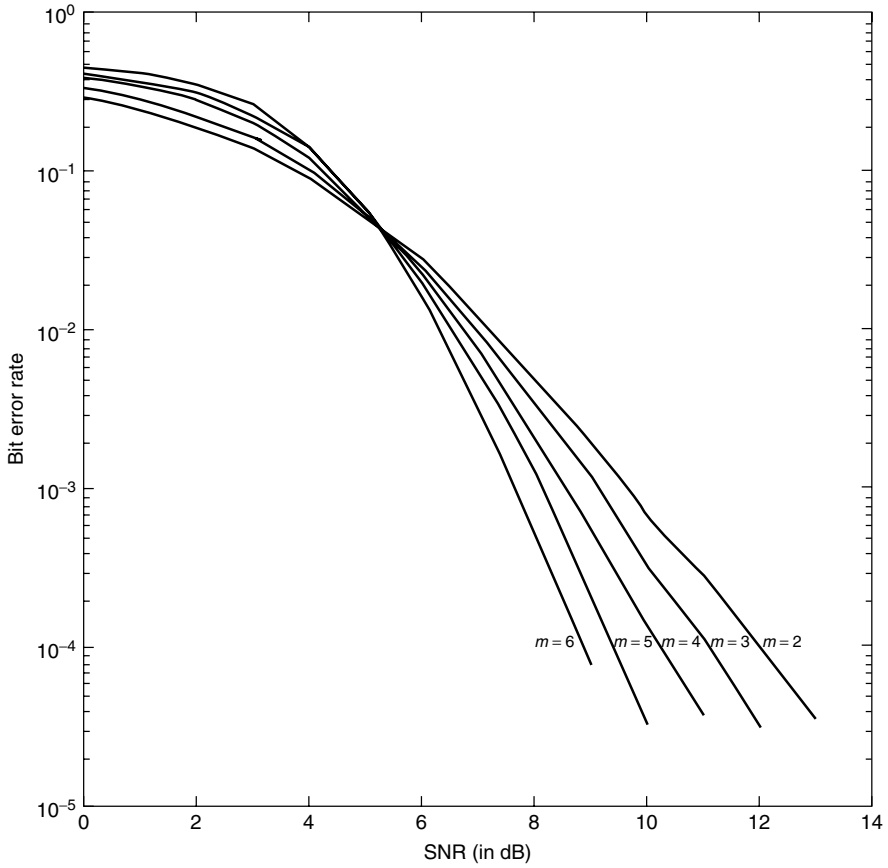


Figure 9. Simulated bit error rate for a BICM system that transmits 2 bits per symbol, which is obtained by employing a rate-$\frac{1}{2}$ convolutional encoder of maximal free distance in conjunction with a 16-QAM modulator, for various encoder memory sizes m.

of Caire et al. [2], which suggest that it is a Gray labeling of the signal points that maximizes the capacity of BICM. However, there are two explanations for this apparent conflict: (1) the model of Fig. 8b, which, as assumed by Caire et al. [2], does not take into account the constraint imposed by 4 bits of the encoder being combined for a single symbol, which is exploited by the iterative decoder; and (2) the signal-to-noise ratios (SNRs) at which the iterative decoder improves performance are quite a bit larger than capacity, indicating that capacity arguments might not be pertinent here.

It is tempting to equate iteratively decoded BICM with iteratively decoded serial concatenated codes, but one should be careful. In violation of the design rules for a serial concatenated code, the inner code in the BICM system is not recursive. Hence, the interleaver gain observed in various types of Turbo codes is not observed in iteratively decoded BICM. Perhaps a better analogy is to a little-known class of codes known as feedback-decoding trellis codes [18]. Much like in the case of BICM, feedback-decoding trellis codes use the knowledge of bits already decoded to aid in the demodulation of a constellation symbol combining bits from different sections of the same trellis.

5. SUMMARY AND SUGGESTED LITERATURE

In this article, the development of bit-interleaved coded modulation has been motivated from a history of coded modulation for the AWGN and perfectly interleaved Rayleigh fading channels. The designs for the encoder and various decoders for BICM have been described. Simulation results indicate that BICM is a strong competitor to traditional coded modulation techniques on perfectly interleaved Rayleigh fading channels and displays a sort of universality in the sense that it works well for a variety of different channels. In addition, iteratively decoded BICM can compete with Turbo code techniques for implementation on AWGN or Rayleigh fading channels.

For further information on this subject, the reader is encouraged to read in detail the work of Caire et al. [2]. The work of Li and Ritcey [11] is the authoritative work on the iterative decoding of BICM, but it is useful to understand the similar idea included in the concept of feedback-decoding trellis codes [18]. The work of Wachsmann et al. [14] on multilevel coding includes the useful extension of including BICM in that framework.

Acknowledgment

The author is indebted to Prof. Rick Wesel of UCLA, whose conversations and collaboration on topics of coded modulation have greatly contributed to the author's knowledge of the subject. In addition, the author would like to thank Prof. Jim Ritcey of the University of Washington for discussions on BICM with iterative decoding and for providing an advanced version of Ref. 11.

BIOGRAPHY

Dennis Goeckel split time between Purdue University and Sundstrand Corporation from 1987 to 1992, receiving his B.S.E.E. (with highest honors) from Purdue in 1992. From 1992 to 1995, he was a National Science Foundation Graduate Fellow at the University of Michigan, where he received his M.S.E.E. in 1993 and his Ph.D. in 1996, both in Electrical Engineering with a speciality in communications systems. In September 1996, he joined the Electrical and Computer Engineering Department at the University of Massachusetts, where he is currently an Associate Professor. Dr. Goeckel is the recipient of a 1999 CAREER Award from the National Science Foundation, and he is an Editor for the *IEEE Transactions on Wireless Communications*. His research interests are in the design of digital communication systems, particularly for wireless communication applications.

BIBLIOGRAPHY

1. E. Zehavi, 8-PSK trellis codes for a Rayleigh channel, *IEEE Trans. Commun.* **40**: 873–884 (May 1992).
2. G. Caire, G. Taricco, and E. Biglieri, Bit-interleaved coded modulation, *IEEE Trans. Inform. Theory* **44**: 927–945 (May 1998).
3. C. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (July 1948); **27**: 623–656 (Oct. 1948).
4. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **28**: 55–67 (Jan. 1982).
5. D. Divsalar and M. Simon, The design of trellis-coded MPSK for fading channels: Set partitioning for optimum code design, *IEEE Trans. Commun.* **36**: 1004–1011 (Sept. 1988).
6. C. Schlegel and D. Costello, Jr., Bandwidth efficient coding for fading channels: Code construction and performance analysis, *IEEE J. Select. Areas Commun.* **7**: 1356–1368 (Dec. 1989).
7. P. Örmeci, X. Liu, D. L. Goeckel, and R. D. Wesel, Adaptive bit-interleaved coded modulation, *IEEE Trans. Commun.* **49**: 1572–1581 (Sept. 2001).
8. S. Lin and D. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
9. S. Al-Semari and T. Fuja, I-Q TCM: reliable communication over the Rayleigh fading channel close to the cutoff rate, *IEEE Trans. Inform. Theory* **43**: 250–262 (Jan. 1997).
10. U. Fincke and M. Phost, Improved methods for calculating vectors of short length in a lattice, including a complexity analysis, *Math. Comput.* **44**(170): 463–471 (April 1985).
11. X. Li, A. Chindapol, and J. Ritcey, Bit-interleaved coded modulation with iterative decoding and 8PSK modulation, *IEEE Trans. Commun.* (in press).
12. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* 909–926 (May 1998).
13. D. L. Goeckel and G. Ananthaswamy, On the design of multidimensional signal sets for OFDM, *IEEE Trans. Commun.* **50**: 442–452 (March 2002).
14. U. Wachsmann, R. Fischer, and J. Huber, Multilevel codes: Theoretical concepts and practical design rules, *IEEE Trans. Inform. Theory* **45**: 1361–1391 (July 1999).
15. P. Schramm, Multilevel coding with independent decoding on levels for efficient communication on static and interleaved fading channels, *Proc. Personal, Indoor, and Mobile Radio Conf.* 1997, pp. 1186–1200.
16. U. Wachsmann, J. Huber, and P. Schramm, Comparison of coded modulation schemes for the AWGN and the Rayleigh fading channel, *Proc. Int. Symp. Information Theory*, 1998, p. 5.
17. C. Berrou and A. Glavieux, Near optimum limit error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* **44**: 1261–1271 (Oct. 1996).
18. G. Hellstern, Coded modulation with feedback decoding trellis codes, *Proc. IEEE Conf. Communications*, 1993, pp. 1071–1075.

BLIND EQUALIZATION TECHNIQUES

JITENDRA K. TUGNAIT
Auburn University
Auburn, Alabama

1. INTRODUCTION

Two major sources of impairments of digital communications signals as they propagate through analog channels (such as telephone, cable, and wireless radio) are multipath propagation and limited bandwidth, causing (linear) channel and signal distortions. Linear channel distortion leads to intersymbol interference (ISI) at the receiver which, in turn, may lead to high error rates in symbol detection. Equalizers are designed to compensate for these channel distortions. One may directly design an equalizer given the received signal, or one may first estimate the channel impulse response and then design an equalizer based on the estimated channel. The received signals are sampled at the baud (symbol) or higher (fractional) rate before processing them for channel estimation and/or equalization. Depending on the sampling rate, one has either a single-input/single-output (SISO) (baud rate sampling), or a single-input/multiple-output (SIMO) (fractional sampling), complex discrete-time equivalent baseband channel.

Traditionally, a training sequence (known to the receiver) is transmitted during startup (acquisition mode). In the operational stage, the receiver switches to a decision-directed mode where the previously equalized and detected symbols are used as a (pseudo)training sequence together with the received data to update the channel or the equalizer coefficients. The various issues involved and the tradeoffs among various competing approaches (linear, decision feedback, maximum-likelihood sequence estimation, least mean-square vs. recursive least-squares, baud rate vs. fractional rate, etc.) are fairly well understood and documented; see the well-known text by Proakis [21] and references cited therein. More recently, there has been much interest in blind (self-recovering) channel estimation and blind equalization where no training sequences

are available or used and the receiver starts up without any (explicit) cooperation from the transmitter. In point-to-multipoint networks, whenever a link from the server to one of the tributary stations is interrupted, it is clearly not feasible (or desirable) for the server to start sending a training sequence to reestablish a particular link. In broadcast applications such as FTTC (fiber-to-the-curb) and DSL (digital subscriber line), it is not desirable to require the transmitter to pause to train each client as it comes online. Transmission of periodic training sequences may incur costly overhead by diluting the transmission rate of the revenue-bearing content. It has also been argued [40] that a blind startup is more straightforward to implement than a startup that requires a training sequence; this eases interoperability issues among different manufacturers. In digital communications over fading/multipath channels, a restart is required following a temporary path loss due to a severe fade. During online transmission impairment monitoring, the training sequences are obviously not supplied by the transmitter.

As in the trained case, various approaches to blind channel estimation and equalization have been developed. When sampled at the baud rate, the received signal is discrete-time stationary and typically non-minimum-phase. When sampled at higher than baud rate (typically an integer multiple of baud rate), the signal is discrete-time scalar cyclostationary and equivalently, it may be represented as a discrete-time vector stationary sequence with an underlying SIMO model. With baud rate sampling, one has to exploit the higher-order statistics (HOS) of the received signal either implicitly (as in Refs. 12 and 26, where direct design of equalizers is considered) or explicitly (as in Refs. 13 and 33–36, where the focus is on first estimating the channel impulse response using higher-order cumulants of the received signal). Higher-order statistics provide an incomplete characterization of the underlying non-Gaussian process. Joint channel and data estimation using maximum-likelihood and related approaches (see Refs. 16 and 24 and references cited therein) exploit a complete (non-Gaussian) probabilistic characterization of the noisy signal. Computational complexity of these algorithms (explicit HOS and joint channel data estimation) is large when the ISI spans many symbols (as in telephone channels) but they are relatively simple when ISI span is short (as in mobile radio channels). However, they may suffer from local convergence problems.

When there is excess channel bandwidth, baud rate sampling is below the Nyquist rate leading to aliasing and depending on the symbol timing phase, in certain cases, causing deep spectral notches in sampled, aliased channel transfer function [11]. This renders the equalizer performance quite sensitive to symbol timing errors. Initially, in the trained case, fractional sampling was investigated to robustify the equalizer performance against timing error. More recently, in the blind context, it was discovered (see Ref. 29 and references cited therein) that oversampling provides some new information regarding the channel that can be exploited for blind channel estimation and equalization provided some technical conditions are satisfied (the "no common subchannel zeros" condition, also called channel disparity, for the underlying equivalent SIMO model). A similar SIMO model results if multiple sensors are used with or without fractional sampling. The work of Tong et al. [29] has spawned intense research activity in the use of second-order statistics for blind identification and equalization. It should be noted that the requisite technical conditions for applicability of these approaches are not always satisfied in practice; some examples are given in Ref. 34.

In this article, we will present a tutorial review of various approaches to single-user blind channel equalization and estimation. Our emphasis is on linear time-invariant channels; linear time-varying, as well as nonlinear channels are outside the scope of this article. The article is organized as follows. In Section 2 we present the relevant channel models and equalizer structures used later for discussion of blind equalization and channel estimation techniques. In Section 3, combined channel and symbol estimation methods are presented. Direct equalization and symbol estimation approaches without first or concurrently estimating the channel impulse response, are discussed in Section 4. In Section 5 various channel estimation approaches are presented. Commercial applications of blind equalization reported in the literature are briefly discussed in Section 6.

2. SYSTEMS MODELS

In this section we first describe the models that are used to characterize the wireless and mobile communications channels. Then we turn to a brief discussion of the various equalizer structures that are used to undo the signal distortions caused by the channel.

2.1. Channel Models

After some processing (e.g., matched filtering), the continuous-time received signals are sampled at the baud (symbol) or higher (fractional) rate before processing them for channel estimation and/or equalization. It is therefore convenient to work with an equivalent baseband discrete-time white-noise channel model [21, Sect. 10.1]. For a baud-rate sampled system, the equivalent baseband channel model is given by

$$y_k = \sum_{n=0}^{L} f_n I_{k-n} + w_k \tag{1}$$

where $\{w_k\}$ is a white Gaussian noise sequence with variance σ^2; $\{I_k\}$ is the zero-mean, independent and identically distributed (i.i.d.), information (symbol) sequence, possibly complex, taking values from a finite set; $\{f_k\}$ is an FIR (finite impulse response) linear filter (with possibly complex coefficients) that represents the equivalent channel; and $\{y_k\}$ is the (possibly complex) equivalent baseband received signal. A tapped delay line structure for this model is shown in Fig. 1.

The model (1) results in a single-input/single-output (SISO) complex discrete-time baseband-equivalent channel model. The output sequence $\{\hat{I}_k\}$ in Eq. (1) is discrete-time stationary. When there is excess channel

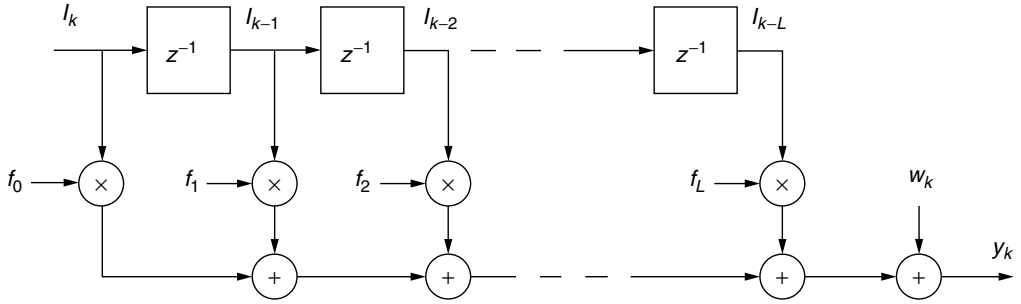


Figure 1. Tapped delay line model of the baud-rate channel.

bandwidth [bandwidth $> \frac{1}{2} \times$ (baud rate)], baud rate sampling is below the Nyquist rate leading to aliasing and depending on the symbol timing phase, in certain cases, causing deep spectral notches in sampled, aliased channel transfer function [11]. Linear equalizers designed on the basis of the baud-rate sampled channel response, are quite sensitive to symbol timing errors. Initially, in the trained case, fractional sampling was investigated to robustify the equalizer performance against timing errors. The model (1) does not apply to fractionally spaced samples, namely, when the sampling interval is a fraction of the symbol duration. The fractionally sampled digital communications signal is a cyclostationary signal [7] that may be represented as a vector stationary sequence using a time-series representation (TSR) ([7, Sec. 12.6]). Suppose that we sample at P times the baud rate with signal samples spaced T/P seconds apart where T is the symbol duration. Then a TSR for the sampled signal is given by

$$y_{ik} = \sum_{n=0}^{L} f_{in} I_{k-n} + w_{ik}; \quad (i = 1, 2, \ldots, P) \tag{2}$$

where now we have P samples every symbol period, indexed by i. Notice, however, that the information sequence I_k is still one "sample" per symbol. It is assumed that the signal incident at the receiver is first passed through a receive filter whose transfer function equals the square root of a raised-cosine pulse, and that the receive filter is matched to the transmit filter. The noise sequence in (2) is the result of the fractional rate sampling of a continuous-time, filtered white Gaussian noise process. Therefore, the sampled noise sequence is white at the symbol rate, but correlated at the fractional rate. Stack P consecutive received samples in the nth symbol duration to form a P vector $\mathbf{y}_k$ satisfying

$$\mathbf{y}_k = \sum_{n=0}^{L} \mathbf{f}_n I_{k-n} + \mathbf{w}_k \tag{3}$$

where $\mathbf{f}_n$ is the vector impulse response of the SIMO equivalent channel model given by

$$\mathbf{f}_n = [f_{1n} \quad f_{2n} \quad \cdots \quad f_{Pn}]^T \tag{4}$$

$$\mathbf{y}_k = [y_{1k} \quad y_{2k} \quad \cdots \quad y_{Pk}]^T \tag{5}$$

$$\mathbf{w}_k = [w_{1k} \quad w_{2k} \quad \cdots \quad w_{Pk}]^T \tag{6}$$

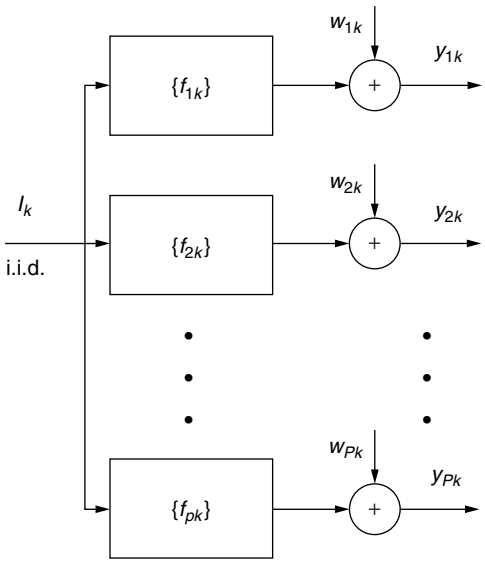


Figure 2. Block diagram of the fractionally sampled ($P \times$ baud-rate) channel.

[When $P = 2$, one way to look at the TSR model is to note that y_{1k} are "odd-numbered" fractionally spaced samples, y_{2k} are the "even-numbered" fractionally spaced samples and k indexes the baud (symbol); similarly for f_{ik}.] A block diagram of model (2) is shown in Fig. 2.

2.2. Equalizer Structures

The most common channel equalizer structure is a linear transversal filter. Given the baud-rate sampled received signal [see Eq. (1)] $\hat{I}_k$, the linear transversal equalizer output $\hat{I}_k$ is an estimate of I_k, given by

$$\hat{I}_k = \sum_{n=-N}^{N} c_n y_{k-n} \tag{7}$$

where $\{c_n\}_{n=-N}^{n=N}$ are the $(2N+1)$ tap-weight coefficients of the equalizer; see Fig. 3. As noted earlier, linear equalizers designed on the basis of the baud-rate sampled received signal, are quite sensitive to symbol timing errors [11]. Therefore, fractionally spaced linear equalizers (typically with twice the baud-rate sampling: oversampling by a factor of 2) are quite widely used to mitigate sensitivity to symbol timing errors. A fractionally

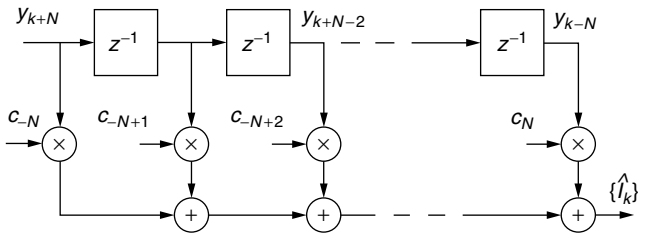


Figure 3. Structure of a baud-rate linear transversal equalizer.

spaced equalizer (FSE) in the linear transversal structure has the output

$$\hat{I}_k = \sum_{n=-N}^{N} \mathbf{c}_n^T \mathbf{y}_{k-n} = \sum_{n=-N}^{N} \left(\sum_{i=1}^{P} c_{in} y_{i(k-n)} \right) \tag{8}$$

where we have P samples per symbol, $\mathbf{y}_k$ and $\mathbf{c}_k$ are P-column vectors [cf. Eq. (3)], $\{\mathbf{c}_k\}$ are the $(2N+1)$ vector tap (or $P(2N+1)$ scalar tap) weight coefficients of the FSE, and the superscript T denotes the transpose operation. Note that the FSE outputs data at the symbol rate. Various criteria and cost functions exist to design the linear equalizers in both batch and recursive (adaptive) form; these are discussed later in this article. Figure 4 shows a block diagram of a generic FSE. [The transfer functions $F_i(z)$ and $C_i(z)$ are defined later in Eqs. (29) and (30).]

Linear equalizers do not perform well when the underlying channels have deep spectral nulls in the passband. Several nonlinear equalizers have been developed to deal with such channels. Two effective approaches are

- *The decision-feedback equalizer* (DFE), which is a nonlinear equalizer that employs previously detected symbols to eliminate the ISI due to the previously detected symbols on the current symbol to be detected. The use of the previously detected symbols makes the equalizer output a nonlinear function of the data. DFE can be symbol-spaced or fractionally spaced. Figure 5 is a block diagram of a DFE.
- *A maximum-likelihood sequence detector*, which estimates the information sequence to maximize the joint probability of the received sequence conditioned on the information sequence.

A detailed discussion may be found in the text by Proakis [21].

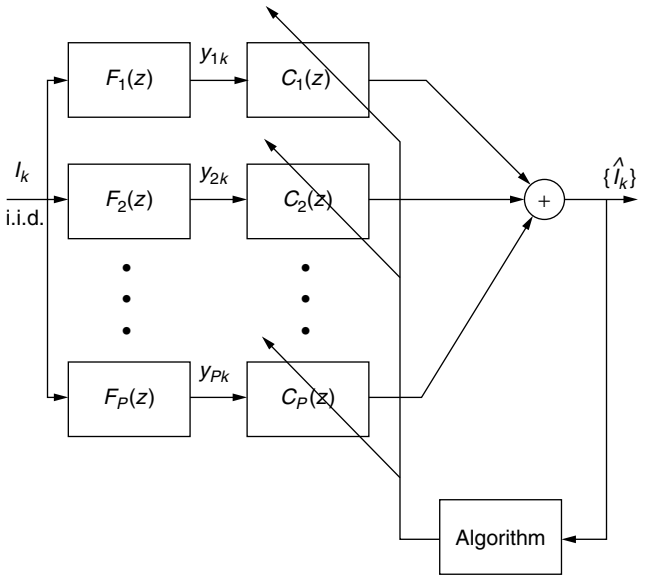


Figure 4. Block diagram of a fractionally spaced equalizer.

3. COMBINED CHANNEL AND SYMBOL ESTIMATION

In general, one of the most effective and popular parameter estimation algorithms is the maximum-likelihood (ML) method. The ML estimators can be derived in a systematic way. Perhaps more importantly, the class of ML estimators are optimal asymptotically. Not surprisingly, this class of algorithms has been applied to the blind equalization problem.

Let us consider the P-vector channel model given in Eq. (3). Suppose that we have collected M samples of the observation $Y = [\mathbf{y}_{M-1}^T, \ldots, \mathbf{y}_0^T]^T$. We then have the following linear model:

$$Y = \begin{pmatrix} I_{M-1}\mathcal{I}_P & I_{M-2}\mathcal{I}_P & \cdots & I_{M-L-1}\mathcal{I}_P \\ \vdots & \text{block} & \text{Hankel} & \text{matrix} \\ I_0\mathcal{I}_P & I_{-1}\mathcal{I}_P & \cdots & I_{-L}\mathcal{I}_P \end{pmatrix} \begin{pmatrix} \mathbf{f}_0 \\ \vdots \\ \mathbf{f}_L \end{pmatrix} + \begin{pmatrix} \mathbf{w}_{M-1} \\ \vdots \\ \mathbf{w}_0 \end{pmatrix} = \mathcal{H}(\mathbf{I})_{[MP]\times[P(L+1)]} F + W \tag{9}$$

where $\mathcal{I}_P$ is a $P \times P$ identity matrix; $\mathbf{I}$ and W are vectors consisting of samples of the input sequence $\{I_k\}$ and noise $\{\mathbf{w}_k\}$, respectively; F is the vector of the channel parameters; and a block Hankel matrix has identical block entries on its block antidiagonals.

Let θ be the vector of unknown parameters that may include the channel parameters F and possibly the entire or part of the input vector $\mathbf{I}$. Given the probability space that describes jointly the noise vector W and possibly the input data vector $\mathbf{I}$, we can then obtain, in principle, the probability density function (PDF) of the observation Y. As a function of the unknown parameter θ, the PDF of the observation $f(Y|\theta)$ is referred to as the *likelihood function*. The maximum likelihood estimator is defined by the following optimization

$$\hat{\theta} = \arg\max_{\theta \in \Theta} f(Y|\theta) \tag{10}$$

where Θ defines the domain of the optimization.

While the ML estimator is conceptually simple, and it usually has good performance when the sample size is sufficiently large, the implementation of ML estimator is sometimes computationally intensive. Furthermore, the optimization of the likelihood function in Eq. (10) is often hampered by the existence of local maxima. Therefore, it is desirable that effective initialization techniques are used in conjunction with the ML estimation. The simultaneous

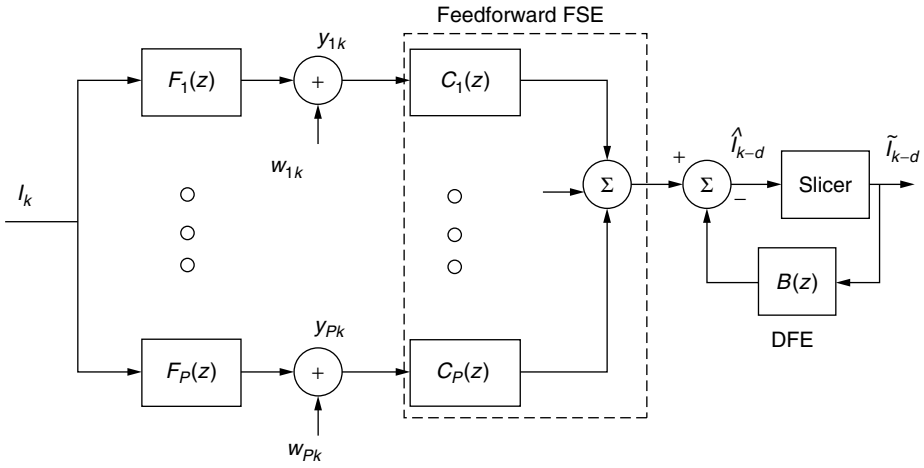


Figure 5. Feedforward and decision-feedback channel equalization filters.

estimation of the input vector and the channel appears to be ill-posed; how is it possible that the channel and its input can be distinguished using only the observation? The key in blind channel estimation is the utilization of qualitative information about the channel and the input. To this end, we consider two different types of maximum likelihood techniques based on different models of the input sequence.

3.1. Stochastic Maximum-Likelihood Estimation

While the input vector $\mathbf{I}$ is unknown, it may be modeled as a random vector with a known distribution. In such a case, the likelihood function of the unknown parameter $\theta = F$ can be obtained by

$$f(Y|F) = \int f(Y|\mathbf{I}, F) f(\mathbf{I}) d\mathbf{I} \tag{11}$$

where $f(\mathbf{I})$ is the marginal PDF of the input vector and $f(Y|\mathbf{I}, F)$ is the likelihood function when the input is known. Assume, for example, that the input data symbol I_k takes, with equal probability, a finite number of values. Consequently, the input data vector $\mathbf{I}$ also takes values from the signal set $\{\mathbf{I}_1, \ldots, \mathbf{I}_K\}$. The likelihood function of the channel parameters is then given by

$$\begin{aligned} f(Y|F) &= \sum_{i=1}^{K} f(Y|\mathbf{I}_i, F) \text{ Prob}(\mathbf{I} = \mathbf{I}_i) \\ &= C \sum_{i=1}^{K} \exp\left\{ -\frac{\|Y - \mathcal{H}(\mathbf{I}_i)F\|^2}{2\sigma^2} \right\} \end{aligned} \tag{12}$$

where C is a constant, $\|Y\|^2 := Y^H Y$, Y^H is the complex conjugate transpose of the complex vector Y, and the stochastic MLE is given by

$$\hat{F} = \arg\min_F \sum_{i=1}^{K} \exp\left\{ -\frac{\|Y - \mathcal{H}(\mathbf{I}_i)F\|^2}{2\sigma^2} \right\} \tag{13}$$

The maximization of the likelihood function defined in (11) is in general difficult because $f(Y|\theta)$ is nonconvex. The expectation–maximization (EM) algorithm can be applied to transform the complicated optimization to a sequence of quadratic optimizations. Kaleh and Vallet [16] first applied the EM algorithm to the equalization of communication channels with input sequence having finite alphabet property. By using a *hidden Markov model* (HMM), they developed a batch (offline) procedure that includes the so-called forward and backward recursions. Unfortunately, the complexity of this algorithm increases exponentially with the channel memory.

To relax the memory requirements and facilitate channel tracking, "online" sequential approaches have been proposed [17] for input with finite alphabet properties under a HMM formulation. Given the appropriate regularity conditions and a good initialization guess, it can be shown that these algorithms converge to the true channel value.

3.2. Deterministic Maximum-Likelihood Estimation

The deterministic ML approach assumes no statistical model for the input sequence $\{I_k\}$. In other words, both the channel vector F and the input source vector $\mathbf{I}$ are parameters to be estimated. When the noise is zero-mean Gaussian with covariance $\sigma^2 I$, the ML estimates can be obtained by the nonlinear least squares optimization

$$\{\hat{F}, \hat{\mathbf{I}}\} = \arg\min \|Y - \mathcal{H}(\mathbf{I})F\|^2. \tag{14}$$

Joint minimization of the likelihood function with respect to both the channel and the source parameter spaces is difficult. Fortunately, the observation vector Y is linear in both the channel and the input parameters individually. In particular, we have

$$Y = \mathcal{H}(\mathbf{I})F + W = \mathcal{T}(F)\mathbf{I} + W \tag{15}$$

where

$$\mathcal{T}(F) = \begin{pmatrix} \mathbf{f}_0 & \cdots & \mathbf{f}_L & & \\ & \ddots & & \ddots & \\ & & \mathbf{f}_0 & \cdots & \mathbf{f}_L \end{pmatrix} \tag{16}$$

is the so-called filtering matrix. We therefore have a separable nonlinear least squares problem that can be

solved sequentially:

$$\{\hat{F}, \hat{\mathbf{I}}\} = \arg\min_{\mathbf{I}}\{\min_{F} \|Y - \mathcal{H}(\mathbf{I})F\|^2\} \tag{17}$$

$$= \arg\min_{F}\{\min_{\mathbf{I}} \|Y - \mathcal{T}(F)\mathbf{I}\|^2\} \tag{18}$$

If we are interested only in estimating the channel, the minimization described above can be rewritten as

$$\hat{F} = \arg\min_{F} \| \underbrace{(I - \mathcal{T}(F)\mathcal{T}^{\dagger}(F))}_{\mathcal{P}(F)} Y\|^2 = \arg\min_{F} \|\mathcal{P}(F)Y\|^2 \tag{19}$$

where $\mathcal{P}(F)$ is a projection transform of Y into the orthogonal complement of the range space of $\mathcal{T}(F)$, or the noise subspace of the observation, and $\mathcal{T}^{\dagger}(F)$ denotes the pseudoinverse of $\mathcal{T}(F)$. Tong and Perreau have discussed algorithms of this type [28].

Similar to the HMM for statistical maximum-likelihood approach, the finite alphabet properties of the input sequence can also be incorporated into the deterministic maximum-likelihood methods. These algorithms, first proposed by Seshadri [24] and Ghosh and Weber [9], iterate between estimates of the channel and the input. At iteration k, with an initial guess of the channel $F^{(k)}$, the algorithm estimates the input sequence $\mathbf{I}^{(k)}$ and the channel $F^{(k+1)}$ for the next iteration by

$$\mathbf{I}^{(k)} = \arg\min_{\mathbf{I}\in\mathcal{S}} \|Y - \mathcal{T}(F^{(k)})\mathbf{I}\|^2 \tag{20}$$

$$F^{(k+1)} = \arg\min_{F} \|Y - \mathcal{H}(\mathbf{I}^{(k)})F\|^2 \tag{21}$$

where $\mathcal{S}$ is the (discrete) domain of $\mathbf{I}$. The optimization in (21) is a linear least-squares problem whereas the optimization in (20) can be achieved by using the Viterbi algorithm [21]. Seshadri [24] presented blind trellis search techniques. Reduced-state sequence estimation was proposed by Ghosh and Weber [9]. Raheli et al. proposed a per survivor processing technique [22].

The convergence of such approaches is not guaranteed in general. Interesting examples have been provided [5] in which two different combinations of F and $\mathbf{I}$ lead to the same cost $\|Y - \mathcal{H}(\mathbf{I})F\|^2$.

4. DIRECT EQUALIZATION AND SYMBOL ESTIMATION

In this section, we describe several types of approaches to the problem of direct input signal recovery under linear time-invariant channels. We first outline the basic principle of blind adaptive equalization based on implicit HOS criteria. Next, we explain the principle of some simple algorithms for blind symbol estimation exploiting second-order statistics. Finally, we discuss the method of symbol estimation via iterative least square criterion and some variations.

4.1. SISO Blind Equalization Based on HOS

In this subsection we consider baud-rate data and equalizers. In the case of known training sequence transmission, the linear equalizer tap $\mathbf{c}_n$ values are chosen to minimize the cost $E\{|\hat{I}_k - I_k|^2\}$ where $\{I_k\}$ is the training sequence. In the blind case, there is no training sequence. The key to designing a blind equalizer is to design rules of equalizer parameter adjustment. With the lack of training sequence, the receiver does not have access to the desired equalizer output I_k to adopt the traditional minimum mean-square-error criterion. Evidently, blind equalizer adaptation needs to minimize some special, non-mean-square-error (MSE)-type cost function, which implicitly involves higher order statistics of the channel output signal. The design of the blind equalizer thus translates into defining a *mean-cost function* $E\{\Psi(\hat{I}_k)\}$, where $\Psi(x)$ is a scalar function. Thus, the stochastic gradient descent minimization algorithm is easily determined by the derivative function $\psi(x) := \Psi'(x) := d\Psi(x)/dx$. Hence, a blind equalizer can either be defined by the cost function $\Psi(x)$, or equivalently, by its derivative $\psi(x)$ function. Ideally, the function $\Psi(\cdot)$ should be selected such that local minima of the mean cost correspond to a significant removal of ISI in the equalizer output $\hat{I}_k$.

Let

$$\mathbf{C} := [\,\mathbf{c}_{-N}^T \quad \mathbf{c}_{-N+1}^T \quad \cdots \quad \mathbf{c}_N^T\,]^T \tag{22}$$

Let $\mathbf{C}^{(k)}$ denote the value of $\mathbf{C}$ at the kth iteration. Then a stochastic gradient algorithm for the adaptation of $\mathbf{C}$ is given by

$$\mathbf{C}^{(k+1)} = \mathbf{C}^{(k)} - \alpha\nabla_{\mathbf{C}^{(k)}}\Psi(\hat{I}_k) \tag{23}$$

where $\nabla_{\mathbf{C}}\Psi$ denotes the gradient of Ψ with respect to the tap vector $\mathbf{C}$ and $\alpha > 0$ is the step-size parameter [21].

We now summarize several blind adaptation algorithms designed for feedforward equalizers.

4.1.1. Decision-Directed Algorithm. The simplest blind equalization algorithm is the decision-directed algorithm without training sequence. It minimizes the mean-square error between equalizer output $\hat{I}_k$ and the slicer output $\tilde{I}_{k-d}$. The performance of decision-directed algorithm depends on how close the initial parameters are to their optimum settings. The closer they are, the more accurate the slicer output is to the true channel input I_{k-d}. On the other hand, local convergence is highly likely if initial parameter values cause significant number of slicer errors [19].

4.1.2. Sato Algorithm and Some Generalizations. The first truly blind algorithm was introduced by Sato [23]. For M-level PAM channel input, this is defined by

$$\psi(x) = x - R_1\text{sgn}(x), \qquad \text{where} \qquad R_1 := \frac{E|I_k|^2}{E|I_k|} \tag{24}$$

The Sato algorithm was extended by Benveniste et al. [1] into a class of error functions given by

$$\psi_b(\hat{I}_k) = \psi_a(\hat{I}_k) - R_b\text{sgn}(\hat{I}_k), \qquad \text{where}$$
$$R_b := \frac{E\{\psi_a(I_k)I_k\}}{E|I_k|} \tag{25}$$

The generalization uses an odd function $\psi_a(x)$ whose second derivative is nonnegative for $x \geq 0$.

4.1.3. Constant-Modulus Algorithm and Extensions. The best known blind algorithms were presented elsewhere [12,31] with cost functions

$$\Psi_q(x) = \frac{1}{2q}\left(|x|^q - R_q\right)^2, \qquad \text{where}$$

$$R_q := \frac{E|I_k|^{2q}}{E|I_k|^q}, \quad q = 1, 2, \ldots \tag{26}$$

This class of *Godard algorithms* is indexed by the positive integer q. Using the stochastic gradient descent approach, equalizer parameters can be adapted accordingly.

For $q = 2$, the special Godard algorithm was developed as the *constant-modulus algorithm* (CMA) independently by Treichler and co-workers [31] using the philosophy of property restoral. For channel input signal that has a constant modulus $|I_k|^2 = R_2$, the CMA equalizer penalizes output samples $\hat{I}_k$ that do not have the desired constant modulus characteristics. The modulus error is simply $e_k = |\hat{I}_k|^2 - R_2$, and the squaring of this error yields the constant-modulus cost function that is the identical to the Godard cost function with $q = 2$.

This modulus restoral concept has a particular advantage in that it allows the equalizer to be adapted independent of carrier recovery. A carrier frequency offset of Δ_f causes a possible phase rotation of the equalizer output. Because the CMA cost function is insensitive to the phase of $\hat{I}_k$, the equalizer parameter adaptation can occur independently and simultaneously with the operation of the carrier recovery system. This property also allows CMA to be applied to analog modulation signals with constant amplitude such as those using frequency or phase modulation [31]. Practical implementations and theoretical properties of the CMA equalizers are discussed in the literature [15,32,40]. Blind DFE has also been considered [4]; in the absence of reliable initialization, blind DFEs can be unstable and can easily misconverge.

The methods of Shalvi and Weinstein [25] generalize CMA and are explicitly based on higher-order statistics of the equalizer output. Define the kurtosis of the equalizer output signal $\hat{I}_k$ as

$$K_{\hat{I}} := E|\hat{I}_k^4| - 2E^2|\hat{I}_k^2| - |E\{\hat{I}_k^2\}|^2 \tag{27}$$

The Shalvi–Weinstein algorithm maximizes $|K_{\hat{I}}|$ subject to the power constraint $E\{|\hat{I}_k|^2\} = E\{|I_k|^2\}$. Superexponential iterative methods have been presented [26] in which a superexponential convergence rate in the absence of noise has been established for the linear equalizer. A deconvolution-based approach can be also be found [2]. Werner et al. [40] discuss modifications of CMA, called the *multimodulus algorithm* (MMA) and generalized MMA, for high-order QAM (quadrature amplitude modulation) and CAP (carrierless amplitude and phase) signals.

4.2. SIMO Equalization and Symbol Estimation

We now consider fractionally sampled data and equalizers. Any adaptive blind equalization algorithm can be easily adopted for linear SIMO equalizers [18]. SIMO blind equalization may offer a convergence advantage given the subchannel diversity [15]. While algorithms such as CMA in SISO equalization may suffer from local convergence [15], CMA and the superexponential method [26] are shown to converge to complete ISI removal under noiseless channels [18]. Furthermore, there is a close relationship between CMA and the nonblind minimum mean-square-error (MMSE) equalizer [42].

Consider the FSE shown in Fig. 4. If the baud-rate "subchannel" transfer functions $F_i(z)$, $1 \le i \le P$, have no common zeros [i.e., there exists no complex number ρ for which $F_i(\rho) = 0$ for every $i, i = 1, 2, \ldots, P$], then there exist FIR "subequalizers" $C_i(z)$s such that

$$\sum_{i=1}^{P} C_i(z)F_i(z) = z^{-d} \tag{28}$$

where

$$C_i(z) := \sum_{n=-N}^{N} c_{in}z^{-n}, \quad 2N \ge L - 1 \tag{29}$$

$$F_i(z) := \sum_{n=0}^{N} f_{in}z^{-n} \tag{30}$$

and d is an integer $\ge -N$. This relation implies perfect equalization (i.e., complete removal of ISI), in the absence of noise, using FIR equalizers, which is not possible in the SISO (baud-rate) case.

If Eq. (28) is not satisfied, then perfect equalization is not possible (using FIR equalizers) and there may be local convergence problems [15].

4.3. Iterative Blind Symbol Estimation

The iterative channel and symbol estimation method, as summarized in Section 3.2, also allows direct channel input estimation. Iterative least-squares with enumeration (ILSE) and Iterative least-squares with projection (ILSP) methods both exploit the finite alphabet nature of the channel input signals. Given that elements in $\mathbf{I}$ come from $\mathcal{S}$, the task of implementing

$$\min_{F,\mathbf{I}\in\mathcal{S}} \|\mathcal{H}(Y) - \mathcal{T}(F)\mathcal{H}(\mathbf{I})\|^2 \tag{31}$$

can be iteratively implemented to improve the estimate in each step, as in Eqs. (20) and (21). ILSP simply replaces the complex symbol estimation step of (20) by a simpler projection [27]

$$\mathbf{I}^{(k)} = \text{proj}_{\mathcal{S}}\left(\mathcal{T}(F^{(k)})^{\dagger}Y\right). \tag{32}$$

5. BLIND CHANNEL ESTIMATION

Although the ML channel estimator discussed in Section 3 usually provides better performance, the computation complexity and the existence of local optima are the two major difficulties. Therefore, "simpler" approaches have also been investigated.

5.1. SISO Channel Estimation

For baud-rate data, second-order statistics of the data do not carry enough information to allow estimation of the channel impulse response as a typical channel is nonminimum-phase. On the other hand, higher-order statistics (in particular, fourth-order cumulants) of the baud-rate (or fractional rate) data can be exploited to yield the channel estimates to within a scale factor.

Given the mathematical model (1), there are two broad classes of approaches to channel estimation, the distinguishing feature among then being the choice of the optimization criterion. All of the approaches involve (more or less) a least-squares-error measure. The error definition differs, however, as follows:

- *Fitting Error*. Match the model-based higher-order (typically fourth-order) statistics to the estimated (data-based) statistics in a least-squares sense to estimate the channel impulse response, as in Refs. 35 and 36, for example. This approach allows consideration of noisy observations. In general, it results in a nonlinear optimization problem. It requires availability of a good initial guess to prevent convergence to a local minimum. It yields estimates of the channel impulse response. The estimator obtained by minimizing Eq. (44) is a fitting error estimate.
- *Equation Error*. This technique is based on minimizing an "equation error" in some equation that is satisfied ideally. The approaches of Refs. 13 and 39 (among others) fall in this category. In general, this class of approaches results in a closed-form solution for the channel impulse response so that a global extremum is always guaranteed provided the channel length (order) is known. These approaches may also provide good initial guesses for the nonlinear fitting error approaches. Quite a few of these approaches fail if the channel length is unknown. The estimator in Eq. (38) is an equation error estimate.

Further details may be found in Ref. 38 and references cited therein.

We now briefly consider the approach of Ref. 35 to illustrate the basic ideas.

5.1.1. Cumulant Matching. We wish to estimate the channel impulse response via a fitting error approach using fourth (and second)-order cumulants of the noisy data. Our main objective is to minimize the cost (44) discussed later in this section. Since this optimization problem requires a good initial guess, we first discuss a simple equation error approach coupled with an model order selection procedure.

Denote the fourth (joint) cumulant of the complex random variables $y^*_{k+\tau_1}$, $y_{k+\tau_2}$, $y^*_{k+\tau_3}$, and y_k as $C_4(\tau_1, \tau_2, \tau_3)$ given by (the superscript $*$ denotes the complex conjugation operation)

$$\begin{aligned} C_4(\tau_1, \tau_2, \tau_3) = {} & E\{y_k y^*_{k+\tau_1} y_{k+\tau_2} y^*_{k+\tau_3}\} - E\{y_k y^*_{k+\tau_1}\} \\ & \times E\{y_{k+\tau_2} y^*_{k+\tau_3}\} - E\{y_k y_{k+\tau_2}\} E\{y^*_{k+\tau_1} y^*_{k+\tau_3}\} \\ & - E\{y_k y^*_{k+\tau_3}\} E\{y_{k+\tau_2} y^*_{k+\tau_1}\} \end{aligned} \tag{33}$$

Then it can be shown that [γ_I = fourth cumulant (kurtosis) of I_k] for model (1) we have

$$C_4(\tau_1, \tau_2, \tau_3) = \gamma_I \sum_{k=0}^{L} f^*_{k+\tau_1} f_{k+\tau_2} f^*_{k+\tau_3} f_k \tag{34}$$

In particular, we have

$$C_4(L, \tau, \tau_1) = \gamma_I f^*_L f_\tau f^*_{\tau_1} f_0 \tag{35}$$

It then follows that

$$C_4(L, 0, \tau_1) f_\tau = C_4(L, \tau, \tau_1) \qquad \text{for} \qquad 0 \le \tau_1 \le L \tag{36}$$

Assuming that $f_L \neq 0$, this immediately leads to the least-squares solution

$$f_\tau = \frac{\sum_{\tau_1=0}^{L} C^*_4(L, 0, \tau_1) C_4(L, \tau, \tau_1)}{\sum_{\tau_1=0}^{L} |C_4(L, 0, \tau_1)|^2} \quad \text{for } 1 \le \tau \le L \tag{37}$$

In practice, true cumulants of the data are unknown. Therefore, we replace them with their consistent estimates. Let $\hat{C}_{4N}(i, j, k)$ denote an estimate of $C_4(i, j, k)$ obtained from Eq. (33) by replacing the moments in (33) by their respective sample averages, based on the N data samples; see Ref. 33 for more details. Then we have

$$\hat{f}_\tau = \frac{\sum_{\tau_1=0}^{L} \hat{C}^*_{4N}(L, 0, \tau_1) \hat{C}_{4N}(L, \tau, \tau_1)}{\sum_{\tau_1=0}^{L} |\hat{C}_{4N}(L, 0, \tau_1)|^2} \quad \text{for } 1 \le \tau \le L \tag{38}$$

In general, there is no guarantee that $f_L \neq 0$ in (1). If $f_L = 0$ in (1) (if, e.g., we overfit), then $C_4(L, i, k) = 0$ for every $0 \le i, k \le L$ rendering the estimates (37–38) useless. Therefore, we perform a search over all possible values of true order L of the FIR model (1); that is, we search over the range $0 \le L \le \overline{L}$, where $\overline{L}$ is an upper bound on the model order such that the true order is known to be less than or equal to $\overline{L}$. Denote the ith coefficient of an MA(L) model (moving-average model of order L) as $f_{i,L}$ so that $f_{0,L} := 1$ for every L and $f_{i,L} := 0$ for $L + 1 \le i \le \overline{L}$ for $L < \overline{L}$. Estimate $f_{i,L}$ by $\hat{f}_{i,L}(N)$ as

$$\hat{f}_{i,L}(N) = \frac{\sum_{\tau_1=0}^{L} \hat{C}^*_{4N}(L, 0, \tau_1) \hat{C}_{4N}(L, \tau, \tau_1)}{\sum_{\tau_1=0}^{L} |\hat{C}_{4N}(L, 0, \tau_1)|^2 + \Delta}, 1 \le i \le L \tag{39}$$

where $\Delta > 0$ is a "small" number. Define a correlation coefficient as

$$\rho_{L,\overline{L}}(N) = \frac{|\sum_{i=0}^{\overline{L}} \sum_{k=0}^{i} \sum_{l=0}^{\overline{L}} \hat{C}_{4N}(i, l, k) \tilde{C}^*_4(i, l, k|\theta_L)|}{\hat{P} P_{\theta_L}} \tag{40}$$

where the lags in (40) range over the nonredundant lag region for complex MA($\overline{L}$) models, $\theta_L := [\hat{f}_{1,L}(N), \ldots, \hat{f}_{L,L}(N)]^T$, the normalized ($\gamma_I = 1$) theoretical fourth-order cumulants corresponding to the parameter vector θ_L are given by

$$\tilde{C}_4(i,l,k|\theta_L) = \sum_{m=0}^{L} \hat{f}_{m,L}(N)\hat{f}^*_{m+i,L}(N)\hat{f}_{m+l,L}(N)\hat{f}^*_{m+k,L}(N) \quad (41)$$

$$P_{\theta_L} := \sqrt{\sum_{i=0}^{\overline{L}} \sum_{k=0}^{i} \sum_{l=0}^{\overline{L}} |\tilde{C}_4(i,l,k|\theta_L)|^2 + \Delta} \quad (42)$$

and

$$\hat{P} := \sqrt{\sum_{i=0}^{\overline{L}} \sum_{k=0}^{i} \sum_{l=0}^{\overline{L}} |\hat{C}_{4N}(i,l,k)|^2} \quad (43)$$

Thus it is easy to see that the preceding correlation coefficient is a measure of fit between the data-based cumulants and the theoretical cumulants obtained from the fitted model.

The estimation of the FIR parameters proceeds as follows. Perform the computations (37)–(43) for $0 \le L \le \overline{L}$. Pick that value of L as the correct FIR order that leads to a maximum correlation coefficient (40); denote it by $\hat{L}_{\overline{L}}$. Repeat (39) with $L = \hat{L}_{\overline{L}}$ and $\Delta = 0$ yielding the desired FIR parameter estimates noting that $\hat{f}_{i,\hat{L}_{\overline{L}}}(N) := 0$ for $\hat{L}_{\overline{L}} + 1 \le i \le \overline{L}$. To justify this procedure asymptotically, let L_0 be the true order such that $0 \le L_0 \le \overline{L}$. As $N \to \infty$, it follows from Ref. 33 that $\hat{C}_{4N}(j,i,k) \to C_4(j,i,k)$ (with probability 1) for any i, j, k. If $L < L_0$, then clearly $\rho_{L,\overline{L}}(N) < 1$ for large N because $f_{L_0,L_0} \ne 0$ by assumption whereas $f_{L_0,L} = 0$, also by assumption. If $L > L_0$, then $\hat{C}_{4N}(L,i,k) \to 0$ w.p.1 as $N \to \infty$ so that $\hat{f}_{i,L}(N) \to 0$ w.p.1 for $1 \le i \le L$, leading to $\rho_{L,\overline{L}}(N) \to \delta < 1$. When $L = L_0$, then $\hat{f}_{i,L}(N) \to \tilde{f}_{i,L}$ as $N \to \infty$, such that as $\Delta \to 0$, $\tilde{f}_{i,L} \to f_{i,L_0}$. Therefore, for Δ small enough, $\rho_{L_0,\overline{L}}(N) \approx 1$ for large N. Thus, asymptotically, correct model order will be selected.

Using the preceding estimates as the initial guess, the next step is to refine the channel estimates by minimizing a quadratic cumulant matching criterion [33]. Let $\overline{L}$ denote the upper bound on the FIR model order as before. Let θ denote the vector of all unknown system parameters given by $\theta := (f_0, f_1, \ldots, f_{\overline{L}}, \gamma_I, \sigma_I)$, $f_{i_0} := 1$ for some $0 \le i_0 \le \overline{L}$, where γ_I and σ_I^2 are the fourth cumulant and the second cumulant (variance) of the information sequence I_k (where one of the channel coefficients has been arbitrarily fixed at 1.0). Choose θ to minimize

$$\sum_{\tau_1=0}^{\overline{L}} \sum_{\tau_3=0}^{\tau_1} \sum_{\tau_2=0}^{\overline{L}} |\hat{C}_{4N}(\tau_1,\tau_2,\tau_3) - C_4(\tau_1,\tau_2,\tau_3|\theta)|^2 + \lambda \sum_{\tau=0}^{\overline{L}} |\hat{R}_N(\tau) - R(\tau|\theta)|^2 \quad (44)$$

where $\hat{C}_{4N}(-)$ denotes the data-based cumulant estimates, $C_4(-|\theta)$ denotes the theoretical cumulants obtained from the hypothesized model, $\hat{R}_N(-)$ denotes the data-based correlation estimates, $R(-|\theta)$ denotes the theoretical correlations obtained from the hypothesized model, and the weighting factor λ is designed to make the cost function invariant to any scale changes. The factor λ is chosen as [33]

$$\lambda := \lambda_0 \frac{\sum_{\tau_1=0}^{\overline{L}} \sum_{\tau_3=0}^{\tau_1} \sum_{\tau_2=0}^{\overline{L}} |\hat{C}_4(\tau_1,\tau_2,\tau_3)|^2}{\sum_{\tau=0}^{\overline{L}} |\hat{R}(\tau)|^2} \quad (45)$$

where $\lambda_0 > 0$ determines the relative weighting between the correlations and the fourth-order cumulants.

The initial guess for minimization of (44) is obtained from the linear estimator. If the selected order in the linear approach is $\hat{L}_{\overline{L}}$, then we set $m_0 = \lfloor (\overline{L} - \hat{L}_{\overline{L}})/2 \rfloor$, $f_n = 0$ for $0 \le n \le m_0 - 1$, $f_{i+m_0} = \hat{f}_{i,\hat{L}_{\overline{L}}}(N)$ for $0 \le i \le \hat{L}_{\overline{L}}$, and $f_n = 0$ for $1 + m_0 + \hat{L}_{\overline{L}} \le n \le \overline{L}$; that is, we "center" the result of the linear approach to obtain the initial guess.

The estimators of the correlation and the fourth-order cumulant functions obtained via appropriate sample averaging of data are strongly consistent [33]. By the preceding results [such as (38)] and the cost function (44) (see also Ref. 33), it follows that the parameter estimator minimizing (44) is strongly consistent provided that $\overline{L} \ge$ true length of the channel.

5.2. SIMO Channel Estimation

Here we concentrate on second-order statistical methods. For single-input (SIMO) multiple-output vector channels the autocorrelation function of the observation is sufficient for the identification of the channel impulse response up to an unknown constant [29,34], provided the various subchannels have no common zeros. This observation led to a number of techniques under both statistical and deterministic assumptions of the input sequence [28]. By exploiting the multichannel aspects of the channel, many of these techniques lead to a constrained quadratic optimization

$$\hat{F} = \arg \min_{\|F\|=1} F^H Q(Y) F \quad (46)$$

where $Q(Y)$ is a positive definite matrix constructed from the observation. Asymptotically (either as the sample size increases to infinity or the noise variance approaches to zero), these estimates converge to true channel parameters.

5.2.1. The Cross-Relation Approach. Here we present a simple yet informative approach [41] that illustrates the basic idea. Suppose that we have only two channels with finite impulse responses f_{1n} and f_{2n}, respectively. If there is no noise, the received signals from the two channels satisfy

$$y_{1n} = f_{1n} * I_n, y_{2n} = f_{2n} * I_n \quad (47)$$

where $*$ is the linear convolution. Consequently, we must have

$$y_{1n} * f_{2n} = y_{2n} * f_{1n} \quad (48)$$

Since the convolution operation is linear with respect to the channel and y_{in} $(i = 1, 2)$ are available, Eq. (48) is equivalent to solving a homogeneous linear equation

$$R\tilde{F} = 0 \tag{49}$$

where R is constructed from the M received data samples

$$R = Y_2 - Y_1 \tag{50}$$

$$Y_j := \begin{pmatrix} y_{jL} & y_{j(L-1)} & \cdots & y_{j0} \\ y_{j(L+1)} & y_{jL} & \cdots & y_{j1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{j(M-1)} & y_{j(M-2)} & \cdots & y_{j(M-L-1)} \end{pmatrix} \tag{51}$$

$$\tilde{F} := (f_{10} \quad f_{11} \quad \cdots \quad f_{1L} \quad f_{20} \quad \cdots \quad f_{2L})^T \tag{52}$$

It can be shown that under certain identifiability conditions [28] (which include knowledge of L and no common subchannel zeros), the null space of R has dimension 1, which means that the channel can be identified up to a constant. When there is noise, the channel estimator can be obtained from a constrained quadratic optimization

$$\hat{\tilde{F}} = \arg\min_{\|\tilde{F}\|=1} \tilde{F}^H R^H R \tilde{F} \tag{53}$$

which implies that $\hat{\tilde{F}}$ is the eigenvector corresponding to the smallest eigenvalue of $Q = R^H R$.

Hua [14] has shown that the cross-relation method combined with the ML approach offers performance close to the Cramer–Rao lower bound. The main problem with this method is that the channel length L needs to be accurately known (in addition to the no-common-subchannel-zeros condition).

5.2.2. Noise Subspace Approach. Alternatively, one can also exploit the subspace structure of the filtering matrix. We now consider a method proposed by Moulines et al. [20]. Define the $M \times [M + L]$ filtering matrix

$$\mathcal{T}_{M+L}(\mathbf{f}_l) = \begin{pmatrix} f_{l0} & \cdots & f_{lL} & & \\ & \ddots & & \ddots & \\ & & f_{l0} & \cdots & f_{lL} \end{pmatrix} \tag{54}$$

and the $[PM] \times [M + L]$ multichannel filtering matrix

$$\mathcal{T}_{M+L}(F) = \left(\mathcal{T}^T_{M+L}(\mathbf{f}_1) \quad \mathcal{T}^T_{M+L}(\mathbf{f}_2) \quad \cdots \quad \mathcal{T}^T_{M+L}(\mathbf{f}_P)\right)^T \tag{55}$$

Define $(M \geq L)$

$$\mathbf{Y}_n = \left(\mathbf{Y}^T_{1n} \quad \mathbf{Y}^T_{2n} \quad \cdots \quad \mathbf{Y}^T_{Pn}\right)^T$$

where

$$\mathbf{Y}_{in} = \left(y_{in} \quad y_{i(n-1)} \quad \cdots \quad y_{i(n-M+1)}\right)^T \tag{57}$$

Then the correlation matrix $\mathbf{R} = E\{\mathbf{Y}_n \mathbf{Y}_n^H\}$ has an eigenvalue decomposition (EVD)

$$\mathbf{R} = \sum_{k=1}^{PM} \lambda_k \, \mathbf{q}_k \, \mathbf{q}_k^H \tag{58}$$

where λ_ks are in the descending order of magnitude. It can be shown that the range space of $\mathbf{R}$ (signal subspace), also the range space of $\mathcal{T}_{M+L}(F)$, is spanned by the eigenvector $\mathbf{q}_k$ values for $k = 1, 2, \ldots, L + M$ whereas the noise subspace (orthogonal complement of the range space) is spanned by the remaining $\mathbf{q}_k$ values for $k = L + M + 1, L + M + 2, \ldots, PM$.

Define $\mathbf{g}_k = \mathbf{q}_{L+M+k+1}$ for $k = 0, 1, \ldots, PM - L - M - 1$. It then follows that

$$\mathcal{T}^H_{M+L}(F)\mathbf{g}_k = 0 \quad k = 0, 1, \ldots, PM - L - M - 1 \tag{59}$$

The vectors $\mathbf{g}_K$ values can be estimated from data via estimated correlation matrix $\mathbf{R}$ and its EVD. Partition the $PM-$ vector $\mathbf{g}_k$ as

$$\mathbf{g}_k = \left(\mathbf{g}^T_{1k} \quad \cdots \quad \mathbf{g}^T_{Pk}\right)^T \tag{60}$$

to conform to $\mathcal{T}_{M+L}(F)$, where $\mathbf{g}_{ik}$ is $M \times 1$. For a given k, define the $[L + 1] \times [L + M]$ matrix $\mathcal{T}_{M+L}(\mathbf{g}_{lk})$ just as $\mathcal{T}_{M+L}(\mathbf{f}_l)$ in (54) except for replacing $\mathbf{f}_l$ with $\mathbf{g}_{lk}$, and similarly define $\mathcal{T}_{M+L}(\mathbf{g}_k)$ by mimicking $\mathcal{T}_{M+L}(F)$ in (55). It has been shown by [20] that

$$\mathcal{T}^H_{M+L}(F)\mathbf{g}_k = 0 = \mathcal{T}^H_{M+L}(\mathbf{g}_k)F \tag{61}$$

It has been further shown [20] that under the knowledge of L and no common subchannel zeros, the channel F can be estimated (up to a scale factor) by the optimization problem

$$\hat{F} = \arg\min_{\|F\|=1} F^H \mathcal{Q} F \quad \text{where}$$

$$\mathcal{Q} := \sum_{k=0}^{PM-L-M-1} \mathcal{T}_{M+L}(\mathbf{g}_k)\mathcal{T}^H_{M+L}(\mathbf{g}_k) \tag{62}$$

The solution is given by the eigenvector corresponding to the smallest eigenvalue of $\mathcal{Q}$.

As with the cross-relation approach, the noise subspace method requires that the channel length L be accurately known in addition to the channel satisfying the no-common-subchannel-zeros condition. A detailed development of this class of methods may be found in Ref. 10, Chaps. 3 and 5.

5.2.3. Multistep Linear Prediction. More recently, the problem of blind channel identification has been formulated as problems of linear prediction [6,8,37] and smoothing [30]. Define the signal (noise-free) part of (3) as

$$\mathbf{s}_k = \sum_{n=0}^{L} \mathbf{f}_n I_{k-n} \tag{63}$$

with s_{ik} denoting the i-th component of $\mathbf{s}_k$. By (28) with $d = -N$, there exists a causal FIR filter of length $M \leq L-1$ such that

$$I_k = \sum_{n=0}^{M} \sum_{i=1}^{P} \tilde{c}_{in} s_{i(k-n)} \tag{64}$$

Using (63) and (64), we have

$$\mathbf{s}_k = \mathbf{e}_{k|k-1} + \hat{\mathbf{s}}_{k|k-1} \tag{65}$$

where

$$\mathbf{e}_{k|k-1} := \mathbf{f}_0 I_k \tag{66}$$

and

$$\hat{\mathbf{s}}_{k|k-1} := \sum_{n=1}^{L} \mathbf{f}_n I_{k-n} = \sum_{i=1}^{L_e} \mathbf{A}_i \mathbf{s}_{k-i} \tag{67}$$

such that

$$E\{\mathbf{e}_{k|k-1}\mathbf{s}_{k-l}^H\} = 0 \,\forall\, l \geq 1 \tag{68}$$

That is, by the orthogonality principle, $\hat{\mathbf{s}}_{k|k-1}$ is the one-step ahead linear prediction (of finite length) of $\mathbf{s}_k$, and $\mathbf{e}_{k|k-1}$ is the corresponding prediction error (linear innovations). Existence of $L_e \leq L-1$ in (67) can be established. The predictor coefficient $\mathbf{A}_i$ values can be estimated from data (after removal of noise effects); therefore, one can calculate $E\{\mathbf{e}_{k|k-1}\mathbf{e}_{k|k-1}^H\}$ from data-based correlation estimates. By (66), we obtain

$$E\{\mathbf{e}_{k|k-1}\mathbf{e}_{k|k-1}^H\} = E\{|I_k|^2\}\mathbf{f}_0\mathbf{f}_0^H \tag{69}$$

a rank 1 matrix. Equation (69) allows estimation of $\mathbf{f}_0$ up to a scale factor (the estimate equals the eigenvector of $E\{\mathbf{e}_{k|k-1}\mathbf{e}_{k|k-1}^H\}$ corresponding to the largest eigenvalue). Once we have a scaled estimate of $\mathbf{f}_0$, we can estimate the remaining channel coefficients using (63) with $\{\mathbf{s}_k\}$ as output and $\|\mathbf{f}_0\|^{-2}\mathbf{f}_0^H\mathbf{e}_{k|k-1}(= I_k e^{j\alpha})$ as input (where α is arbitrary).

The approach described above can be extended by using multistep linear prediction. It can be shown that

$$\mathbf{s}_k = \mathbf{e}_{k|k-2} + \hat{\mathbf{s}}_{k|k-2} \tag{70}$$

where

$$\mathbf{e}_{k|k-2} := \mathbf{f}_0 I_k + \mathbf{f}_1 I_{k-1} \tag{71}$$

and

$$\hat{\mathbf{s}}_{k|k-2} := \sum_{n=2}^{L} \mathbf{f}_n I_{k-n} = \sum_{i=2}^{L_e+1} \mathbf{A}_{2i} \mathbf{s}_{k-i} \tag{72}$$

such that

$$E\{\mathbf{e}_{k|k-2}\mathbf{s}_{k-l}^H\} = 0 \,\forall\, l \geq 2 \tag{73}$$

By the orthogonality principle, $\hat{\mathbf{s}}_{k|k-2}$ is the two-step-ahead linear prediction (of finite length) of $\mathbf{s}_k$, and $\mathbf{e}_{k|k-2}$ is the corresponding prediction error. Define

$$\mathbf{E}_k := \left((\mathbf{e}_{k+1|k-1} - \mathbf{e}_{k+1|k})^T \quad \mathbf{e}_{k|k-1}^T\right)^T \tag{74}$$

so that we have

$$\mathbf{E}_k = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_0 \end{pmatrix} I_k \tag{75}$$

By (75), we have

$$E\{\mathbf{E}_k\mathbf{E}_k^H\} = E\{|I_k|^2\} \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_0 \end{pmatrix} \begin{pmatrix} \mathbf{f}_1^H & \mathbf{f}_0^H \end{pmatrix} \tag{76}$$

a rank 1 matrix. That is, we can estimate $\mathbf{f}_0$ and $\mathbf{f}_1$ simultaneously up to the same scale factor. By adding larger step predictors, one can estimate the entire channel impulse response simultaneously. An advantage over one-step predictor approach is that the results are not unduly influenced by any estimation errors in estimating the leading coefficient $\mathbf{f}_0$.

The multistep linear prediction approach was proposed by Ding [6] in a different form and by Gesbert and Duhamel [8] in the form given above. Both of them assumed FIR channels with known channel length and no common subchannel zeros. Tugnait [37] extended the approach of Gesbert and Duhamel [8] by allowing common subchannel zeros, IIR (infinite impulse response) channels and unknown channel length. It has been shown [37] that minimum-phase common subchannel zeros pose no problems for the multistep linear prediction approach, and in the presence of nonminimum-phase common subchannel zeros, the multistep linear prediction approach yields a minimum-phase equivalent version of these zeros. It is also worth noting that linear prediction approaches (both single-step and multistep) are robust against overdetermination of channel length, unlike the cross-relation and noise subspace approaches.

6. COMMERCIAL APPLICATIONS

The commercial implementations and applications of blind equalizers reported in the literature are all based on CMA/Godard FIR equalizers (typically FSEs with twice the baud-rate sampling) and its variations in the acquisition stage followed by a decision-directed implementation in the operational stage. Treichler et al. [32] describe a variety of digitally implemented demodulators ranging from digital signal processor (DSP) chip-based designs used for voiceband modems (modulation types up to 128-QAM, baud rate up to 3500 baud) to very-large-scale-integration (VLSI)-based designs for digital microwave radio (modulation types up to 128-QAM, symbol rates up to 40 Mbaud). The intended applications of Treichler's designs [32] include high-speed voiceband modems, digital cable modems, and high-capacity digital microwave radios.

Werner et al. [40] describe successful laboratory experimental results with a 51.84-Mbp/s 16-CAP (12.92-Mbaud) transceiver prototype used for FTTC and VDSL (very-high-rate DSL).

Reports of the performance of other blind equalizers and channel estimators are based on computer simulations (or "controlled real data"). Promising simulation results have been reported on the application of blind equalization in the popular wireless GSM cellular system [3] using a

higher-order statistical deconvolution method [2] where the estimated channel is used in conjunction with MLSE (ML sequence estimator) for symbol estimation. Boss et al. [3] report that their HOS-based approach, using only 142 data samples per frame, incurs an SNR loss of 1.2-1.3 dB only while it saves the 22% overhead in the GSM data rate caused by the transmission of training sequences. (Thus, on the average, the Boss et al. HOS-based approach [2] requires 1.2–1.3 dB higher SNR than the conventional GSM system to achieve the same bit error rate.)

Acknowledgments
This work was prepared in part under the support of the National Science Foundation under Grant CCR-9803850.

BIOGRAPHY

Jitendra K. Tugnait received his B.Sc.(Hons.) degree in electronics and electrical communication engineering from the Punjab Engineering College, Chandigarh, India, in 1971, M.S. and the E.E. degrees from Syracuse University, Syracuse, New York, and a Ph.D. degree from the University of Illinois, Urbana-Champaign Urbana, Illinois, in 1973, 1974, and 1978, respectively, all in electrical engineering. From 1978 to 1982 he was an assistant professor of Electrical and Computer Engineering at the University of Iowa, Iowa City, IA. He was with the Long Range Research Division of the Exxon Production Research Company, Houston, TX, from 1982 to 1989 working on geophysical signal processing problems. He joined the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, in September 1989 as a professor. His research interests are in statistical signal processing, wireless and wireline digital communications, blind channel estimation and equalization for single and multiuser systems, and system identification. Dr. Tugnait has published over 95 journal and 120 conference articles. He was elected a fellow of IEEE in 1994. He is a past associate editor of the *IEEE Transactions on Signal of Processing* and of the *IEEE Transactions on Automatic Control*.

BIBLIOGRAPHY

1. A. Benveniste, M. Goursat, and G. Ruget, Robust identification of a nonminimum phase system: blind adjustment of a linear equalizer in data communications, *IEEE Trans. Autom. Control* **AC-25**: 385–399 (June 1980).
2. D. Boss, B. Jelonek, and K. D. Kammeyer, Eigenvector algorithm for blind MA system identification, *Signal Process.* **66**: 1–26 (April 1998).
3. D. Boss, K.-D. Kammeyer, and T. Petermann, Is blind channel estimation feasible in mobile communication systems? A study based on GSM, *IEEE J. Select. Areas Commun.* **SAC-16**: 1480–1492 (Oct. 1998).
4. R. A. Casas et al., Current approaches to blind decision feedback equalization, in G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 1: *Trends in Channel Estimation and Equalization*, Prentice-Hall, Upper Saddle River, NJ, 2001, Chap. 11, pp. 367–415.
5. K. M. Chugg, Blind acquisition characteristics of PSP-based sequence detectors, *IEEE J. Select. Areas Commun.* **SAC-16**: 1518–1529 (Oct. 1998).
6. Z. Ding, Matrix outer-product decomposition method for blind multiple channel identification, *IEEE Trans. Signal Process.* **45**: 3054–3061 (Dec. 1997).
7. W. A. Gardner, *Introduction to Random Processes: With Applications to Signals and Systems*, 2nd ed., McGraw-Hill, New York, 1989.
8. D. Gesbert and P. Duhamel, Robust blind channel identification and equalization based on multi-step predictors, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processes*, Seattle, WA, April 1997, pp. 3621–3624.
9. M. Ghosh and C. L. Weber, Maximum-likelihood blind equalization, *Opt. Eng.* **31**: 1224–1228 (June 1992).
10. G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 1: *Trends in Channel Estimation and Equalization*, Prentice-Hall, Upper Saddle River, NJ, 2001.
11. R. D. Gitlin and S. B. Weinstein, Fractionally-spaced equalization: An improved digital transversal equalizer, *Bell Syst. Tech. J.* **60**: 275–296 (Feb. 1981).
12. D. N. Godard, Self-recovering equalization and carrier tracking in two-dimensional data communication systems, *IEEE Trans. Commun.* **COM-28**: 1867–1875 (Nov. 1980).
13. D. Hatzinakos and C. L. Nikias, Blind equalization using a tricepstrum based algorithm, *IEEE Trans. Commun.* **COM-39**: 669–681 (May 1991).
14. Y. Hua, Fast maximum likelihood for blind identification of multiple FIR channels, *IEEE Trans. Signal Process.* **SP-44**: 661–672 (March 1996).
15. C. R. Johnson, Jr., et al., Blind equalization using the constant modulus criterion: A review, *Proc. IEEE* **86**: 1927–1950 (Oct. 1998).
16. G. K. Kaleh and R. Vallet, Joint parameter estimation and symbol detection for linear or non linear unknown dispersive channels, *IEEE Trans. Commun.* **COM-42**: 2406–2413 (July 1994).
17. V. Krishnamurthy and J. B. Moore, On-line estimation of hidden Markov model parameters based on Kullback-Leibler information measure, *IEEE Trans. Signal Process.* **SP-41**: 2557–2573 (Aug. 1993).
18. Y. Li and Z. Ding, Global convergence of fractionally spaced Godard adaptive equalizers, *IEEE Trans. Signal Process.* **SP-44**: 818–826 (April 1996).
19. O. Macchi and E. Eweda, Convergence analysis of self-adaptive equalizers, *IEEE Trans. Inform. Theory* **IT-30**: 162–176 (March 1983).
20. E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue, Subspace-methods for the blind identification of multichannel FIR filters, *IEEE Trans. Signal Process.* **SP-43**: 516–525 (Feb. 1995).
21. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
22. R. Raheli, A. Polydoros, and C. K. Tzou, Per-survivor processing: A general approach to MLSE in uncertain environments, *IEEE Trans. Commun.* **COM-43**: 354–364 (Feb.–April 1995).

23. Y. Sato, A method of self-recovering equalization for multilevel amplitude modulation, *IEEE Trans. Commun.* **COM-23**: 679–682 (June 1975).

24. N. Seshadri, Joint data and channel estimation using fast blind trellis search techniques, *IEEE Trans. Commun.* **COM-42**: 1000–1011 (March 1994).

25. O. Shalvi and E. Weinstein, New criteria for blind deconvolution of nonminimum phase systems (channels), *IEEE Trans. Inform. Theory* **IT-36**: 312–321 (March 1990).

26. O. Shalvi and E. Weinstein, Super-exponential methods for blind deconvolution, *IEEE Trans. Inform. Theory* **IT-39**: 504–519 (March 1993).

27. S. Talwar, M. Viberg, and A. Paulraj, Blind separation of synchronous co-channel digital signals using an antenna array — Part I: Algorithms, *IEEE Trans. Signal Process.* **SP-44**: 1184–1197 (May 1996).

28. L. Tong and S. Perreau, Multichannel blind channel estimation: From subspace to maximum likelihood methods, *Proc. IEEE* **86**: 1951–1968 (Oct. 1998).

29. L. Tong, G. Xu, and T. Kailath, A new approach to blind identification and equalization of multipath channels, *IEEE Trans. Inform. Theory* **IT-40**: 340–349 (March 1994).

30. L. Tong and Q. Zhao, Joint order detection and blind channel estimation by least squares smoothing, *IEEE Trans. Signal Process.* **SP-47**: 2345–2355 (Sept. 1999).

31. J. R. Treichler and M. G. Agee, A new approach to multipath correction of constant modulus signals, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-31**: 349–472 (April 1983).

32. J. R. Treichler, M. G. Larimore, and J. C. Harp, Practical blind demodulators for high-order QAM signals, *Proc. IEEE* **86**: 1907–1926 (Oct. 1998).

33. J. K. Tugnait, Identification of linear stochastic systems via second- and fourth-order cumulant matching, *IEEE Trans. Inform. Theory* **IT-33**: 393–407 (May 1987).

34. J. K. Tugnait, On blind identifiability of multipath channels using fractional sampling and second-order cyclostationary statistics, *IEEE Trans. Inform. Theory* **IT-41**: 308–311 (Jan. 1995).

35. J. K. Tugnait, Blind estimation and equalization of digital communication FIR channels using cumulant matching, *IEEE Trans. Commun.* **COM-43**(Pt. III): 1240–1245 (Feb.–April 1995).

36. J. K. Tugnait, Blind equalization and estimation of FIR communications channels using fractional sampling, *IEEE Trans. Commun.* **COM-44**: 324–336 (March 1996).

37. J. K. Tugnait, Multistep linear predictors-based blind equalization of FIR/IIR single-input multiple-output channels with common zeros, *IEEE Trans. Signal Process.* **SP-47**: 1689–1700 (June 1999).

38. J. K. Tugnait, Channel estimation and equalization using higher-order statistics, in G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 1: *Trends in Channel Estimation and Equalization*, Prentice-Hall, Upper Saddle River, NJ, 2001, Chap. 1, pp. 1–39.

39. J. Vidal and J. A. R. Fonollosa, Adaptive blind equalization using weighted cumulant slices, *Int. J. Adapt. Control Signal Process.* **10**(2–3): 213–238 (March–June 1996).

40. J.-J. Werner, J. Yang, D. D. Harman, and G. A. Dumont, Blind equalization for broadband access, *IEEE Commun. Mag.* **37**: 87–93 (April 1999).

41. G. Xu, H. Liu, L. Tong, and T. Kailath, A least-squares approach to blind channel identification, *IEEE Trans. Signal Process.* **SP-43**: 2982–2993 (Dec. 1995).

42. H. Zeng, L. Tong, and C. R. Johnson, Jr., Relationships between CMA and Wiener receivers, *IEEE Trans. Inform. Theory* **IT-44**: 1523–1538 (July 1998).

BLIND MULTIUSER DETECTION

XIAODONG WANG
Columbia University
New York, New York

1. INTRODUCTION

Code-division multiple access (CDMA) implemented with direct-sequence spread-spectrum (DSSS) modulation is emerging as a popular multiple-access technology for personal, cellular, and satellite communication services. Multiuser detection techniques can substantially increase the capacity of CDMA systems. Since the early 90s, a significant amount of research has addressed various multiuser detection schemes [33]. Considerable attention has been focused on adaptive multiuser detection [10]. For example, methods for adapting the linear decorrelating detector that require the transmission of training sequences during adaptation have been proposed [5,20,21]. An alternative linear detector, the linear minimum mean-square error (MMSE) detector, however, can be adapted either through the use of training sequences [1,18,19,24], or in the blind mode, with the prior knowledge of only the signature waveform and timing of the user of interest [9,38]. Blind adaptation schemes are especially attractive for the downlinks of CDMA systems, since in a dynamic environment, it is very difficult for a mobile user to obtain the accurate information of other active users in the channel, such as their signature waveforms; and the frequent use of training sequence is certainly a waste of channel bandwidth. There are primarily two approaches to blind multiuser detection, namely, the direct matrix inversion (DMI) approach and the subspace approach. In this article, we present batch algorithms and adaptive algorithms under both approaches. We first treat the simple synchronous CDMA channels and present the main techniques for blind multiuser detection. We then generalize these methods to the more general asynchronous CDMA channels with multipath effects.

2. LINEAR RECEIVERS FOR SYNCHRONOUS CDMA

2.1. Synchronous CDMA Signal Model

We start by introducing the most basic multiple-access signal model, namely, a baseband, K-user, time-invariant, synchronous, additive white Gaussian noise (AWGN) system, employing periodic (short) spreading sequences, and operating with a coherent BPSK (binary phase shift keying) modulation format. The waveform received by a given user in such a system can be modeled as

$$r(t) = \sum_{k=1}^{K} A_k \sum_{i=0}^{M-1} b_k[i] s_k(t - iT) + n(t) \qquad (1)$$

where M is the number of data symbols per user in the data frame of interest; T is the symbol interval; A_k, $\{b_k[i]\}_{i=0}^{M-1}$, and $s_k(t)$ denote, respectively, the received complex amplitude, the transmitted symbol stream, and the normalized signaling waveform of the kth user; and $n(t)$ is the baseband complex white Gaussian ambient channel noise with power spectral density σ^2. It is assumed that for each user k, $\{b_k[i]\}_{i=0}^{M-1}$ is a collection of independent equiprobable ± 1 random variables and that the symbol streams of different users are independent. The user signaling waveform is of the form

$$s_k(t) = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} c_{j,k}\psi(t - jT_c), \qquad 0 \le t < T \tag{2}$$

where N is the processing gain; $\{c_{j,k}\}_{j=0}^{N-1}$ is a signature sequence of ± 1 values assigned to the kth user; and $\psi(\cdot)$ is a chip waveform of duration $T_c = T/N$ and with unit energy, that is, $\int_0^{T_c} \psi(t)^2 dt = 1$.

At the receiver, the received signal $r(t)$ is filtered by a chip-matched filter and then sampled at the chip rate. The sample corresponding to the jth chip of the ith symbol is given by

$$r_j[i] \triangleq \int_{iT+jT_c}^{iT+(j+1)T_c} r(t)\psi(t - iT - jT_c)\, dt \tag{3}$$
$$j = 0, \ldots, N-1; \quad i = 0, \ldots, M-1$$

The resulting discrete-time signal corresponding to the ith symbol is then given by,

$$\boldsymbol{r}[i] = \sum_{k=1}^{K} A_k b_k[i]\boldsymbol{s}_k + \boldsymbol{n}[i] \tag{4}$$
$$= \boldsymbol{S}\boldsymbol{A}\boldsymbol{b}[i] + \boldsymbol{n}[i] \tag{5}$$

with

$$\boldsymbol{r}[i] \triangleq \begin{bmatrix} r_0[i] \\ r_1[i] \\ \vdots \\ r_{N-1}[i] \end{bmatrix}, \quad \boldsymbol{s}_k \triangleq \frac{1}{\sqrt{N}} \begin{bmatrix} c_{0,k} \\ c_{1,k} \\ \vdots \\ c_{N-1,k} \end{bmatrix},$$
$$\boldsymbol{n}[i] \triangleq \begin{bmatrix} n_0[i] \\ n_1[i] \\ \vdots \\ n_{N-1}[i] \end{bmatrix}$$

where $n_j[i] = \int_{jT_c}^{(j+1)T_c} n(t)\psi(t - iT - jT_c)dt \sim \mathcal{N}_c(0, \sigma^2)$ is a complex Gaussian random variable with independent real and imaginary components; $\boldsymbol{n}[i] \sim \mathcal{N}_c(\boldsymbol{0}, \sigma^2\boldsymbol{I})$; $\boldsymbol{S} \triangleq [\boldsymbol{s}_1 \cdots \boldsymbol{s}_K]$; $\boldsymbol{A} \triangleq \mathrm{diag}(A_1, \ldots, A_K)$; and $\boldsymbol{b}[i] \triangleq [b_1[i] \cdots b_K[i]]^T$.

2.2. Linear MMSE Detector

Suppose that we are interested in demodulating the data bits of a particular user, say user 1, $\{b_1[i]\}_{i=0}^{M-1}$, based on the received waveforms $\{\boldsymbol{r}[i]\}_{i=0}^{M-1}$. A *linear* receiver for this purpose is a vector $\boldsymbol{w}_1 \in \mathbb{C}^N$, such that the desired user's data bits are demodulated according to

$$z_1[i] = \boldsymbol{w}_1^H \boldsymbol{r}[i] \tag{6}$$

$$\hat{b}_1[i] = \mathrm{sign}\,\{\Re(A_1^* z_1[i])\} \tag{7}$$

In case that the complex amplitude A_1 of the desired user is unknown, we can resort to differential detection. Define the differential bit as

$$\beta_1[i] \triangleq b_1[i]b_1[i-1] \tag{8}$$

Then, using the linear detector output (6), the following differential detection rule can be applied:

$$\hat{\beta}_1[i] = \mathrm{sign}\,\{\Re(z_1[i]z_1[i-1]^*)\} \tag{9}$$

Substituting (4) into (6), the output of the linear receiver $\boldsymbol{w_1}$ can be written as

$$z_1[i] = A_1(\boldsymbol{w}_1^H\boldsymbol{s}_1)b_1[i] + \sum_{k=2}^{K} A_k(\boldsymbol{w}_1^H\boldsymbol{s}_k)b_k[i] + \boldsymbol{w}_1^H\boldsymbol{n}[i] \tag{10}$$

In (10), the first term contains the useful signal of the desired user; the second term contains the signals from other undesired users — the so-called multiple-access interference (MAI); and the last term contains the ambient Gaussian noise. The simplest linear receiver is the conventional matched-filter, where $\boldsymbol{w}_1 = \boldsymbol{s}_1$. It is well known that such a matched-filter receiver is optimal only in a single-user channel (i.e., $K = 1$). In a multiuser channel (i.e., $K > 1$), this receiver may perform poorly since it makes no attempt to ameliorate the MAI, a limiting source of interference in multiple-access channels.

The linear minimum mean-square error (MMSE) detector is designed to minimize the total effect of the MAI and the ambient noise at the detector output. Specifically, it is given by the solution to the following optimization problem:

$$\boldsymbol{w}_1 = \arg\min_{\boldsymbol{w}\in\mathbb{C}^N} E\left\{\|A_1 b_1[i] - \boldsymbol{w}^H\boldsymbol{r}[i]\|^2\right\} \tag{11}$$

Denote $|\boldsymbol{A}| \triangleq \mathrm{diag}(|A_1|, \ldots, |A_K|)$ and $\boldsymbol{R} \triangleq \boldsymbol{S}^T\boldsymbol{S}$. The solution to (11) is given by [33]

$$\boldsymbol{w}_1 = \boldsymbol{S}\left(\boldsymbol{R} + \sigma^2|\boldsymbol{A}|^{-2}\right)^{-1}\boldsymbol{e}_1 \tag{12}$$

where $\boldsymbol{e}_1$ denotes the first unit vector in $\mathbb{C}^K$.

3. BLIND MULTIUSER DETECTION: DIRECT METHODS

It is seen from (12) that the linear MMSE detector $\boldsymbol{w}_1$ is expressed in terms of a linear combination of the signature sequences $\boldsymbol{S}$ of all K users. Recall that for the matched-filter receiver, the only prior knowledge required is the

desired user's signature sequence $\boldsymbol{s}_1$. In the downlink of a CDMA system, the mobile receiver typically has knowledge only of its own signature sequence, and not of those of the other users. Hence it is of interest to consider the problem of *blind* implementation of the linear detector, that is, without the requirement of knowing the signature sequences of the interfering users. This problem is relatively easy for the linear MMSE detector. To see this, consider again the definition (11). Directly solving this optimization problem, we obtain the following alternative expression for the linear MMSE detector:

$$\begin{aligned}\boldsymbol{w}_1 &= \arg\min_{\boldsymbol{w}\in\mathbb{C}^N} \boldsymbol{w}^H \underbrace{E\left\{\boldsymbol{r}[i]\boldsymbol{r}[i]^H\right\}}_{\boldsymbol{C}_r}\boldsymbol{w} - 2\boldsymbol{w}^H\Re\{A_1^* \underbrace{E(\boldsymbol{r}[i]b_1[i])}_{A_1\boldsymbol{s}_1}\} \\ &= |A_1|^2\boldsymbol{C}_r^{-1}\boldsymbol{s}_1 \end{aligned} \tag{13}$$

where by (5)

$$\boldsymbol{C}_r \triangleq E\left\{\boldsymbol{r}[i]\boldsymbol{r}[i]^H\right\} = \boldsymbol{S}|\boldsymbol{A}|^2\boldsymbol{S}^T + \sigma^2\boldsymbol{I} \tag{14}$$

is the autocorrelation matrix of the receiver signal. Note that $\boldsymbol{C}_r$ can be estimated from the received signals by the corresponding sample autocorrelation. Note also that the constant $|A_1|^2$ in (13) does not affect the linear decision rule (7) or (9). Hence (13) leads straightforwardly to the following blind implementation of the linear MMSE detector — the so-called direct matrix inversion (DMI) blind detector.

- *Compute the detector*:

$$\hat{\boldsymbol{C}}_r \triangleq \frac{1}{M}\sum_{i=0}^{M-1}\boldsymbol{r}[i]\boldsymbol{r}[i]^H$$

$$\hat{\boldsymbol{w}}_1 = \hat{\boldsymbol{C}}_r^{-1}\boldsymbol{s}_1$$

- *Perform differential detection*:

$$z_1[i] = \hat{\boldsymbol{w}}_1^H\boldsymbol{r}[i]$$

$$\hat{\beta}_1[i] = \text{sign}\,\{\Re(z_1[i]z_1[i-1]^*)\}, \quad i = 1,\ldots,M-1$$

This algorithm is a *batch* processing method; that is, it computes the detector only once on the basis of a block of received signals $\{\boldsymbol{r}[i]\}_{i=0}^{M-1}$; and the estimated detector is then used to detect all the data bits of the desired user $\{b_1[i]\}_{i=0}^{M-1}$ contained in the same signal block. In what follows, we consider the *adaptive* implementation of the blind linear MMSE detector.

The idea is to perform sequential (i.e., online) blind detector estimation and data detection; that is, suppose that at time $(i-1)$, a detector $\boldsymbol{w}_1[i-1]$ is used for detecting $b_1[i-1]$. At time i, a new signal $\boldsymbol{r}[i]$ is received and is then used to update the detector to obtain $\boldsymbol{w}_1[i]$. The updated detector is used to detect the data bit $b_1[i]$. Hence the blind detector is sequentially updated at the symbol rate. In order to develop such an adaptive algorithm, we need an alternative characterization of the linear MMSE detector. Consider the following constrained optimization problem:

$$\boldsymbol{w}_1 = \arg\min_{\boldsymbol{w}\in\mathbb{C}^N} E\left\{\|\boldsymbol{w}^H\boldsymbol{r}[i]\|^2\right\}, \quad \text{s.t.} \quad \boldsymbol{w}^H\boldsymbol{s}_1 = 1 \tag{15}$$

The solution to (15) is given by

$$\boldsymbol{w}_1 = \alpha\boldsymbol{C}_r^{-1}\boldsymbol{s}_1 \tag{16}$$

where $\alpha = (\boldsymbol{s}_1^T\boldsymbol{C}_r^{-1}\boldsymbol{s}_1)^{-1}$. Comparing this solution with (13), it is seen that the two differ only by a positive scaling constant. Since such a scaling constant will not affect the linear decision rule (7) or (9), (15) constitutes an equivalent definition of the linear MMSE detector. We next consider the adaptive implementation of the linear MMSE detector basis of the on the least mean-square (LMS) algorithm. Note that $\boldsymbol{w}_1$ can be decomposed into two orthogonal components

$$\boldsymbol{w}_1 = \boldsymbol{s}_1 + \boldsymbol{x}_1 \tag{17}$$

with

$$\boldsymbol{x}_1 \triangleq \boldsymbol{P}\boldsymbol{w}_1 = \boldsymbol{P}\boldsymbol{x}_1 \tag{18}$$

where $\boldsymbol{P} \triangleq \boldsymbol{I} - \boldsymbol{s}_1\boldsymbol{s}_1^T$ is a projection matrix that projects any signal in $\mathbb{C}^N$ onto the orthogonal space of $\boldsymbol{s}_1$. Using this decomposition, the constrained optimization problem (15) can then be converted to the following unconstrained optimization problem:

$$\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\mathbb{C}^N} E\left\{\|(\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x})^H\boldsymbol{r}[i]\|^2\right\} \tag{19}$$

The LMS algorithm for adapting the weights $\boldsymbol{x}_1$ based on the cost function (19) is then given by [8]

$$\boldsymbol{x}_1[i+1] = \boldsymbol{x}_1[i] - \frac{\mu}{2}g(\boldsymbol{x}_1[i]) \tag{20}$$

where μ is the step size, and the stochastic gradient $g(\boldsymbol{x}_1[i])$ is given by

$$\begin{aligned} g(\boldsymbol{x}_1[i]) &\triangleq \frac{d}{dx}\|(\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x})^H\boldsymbol{r}[i]\|^2|\boldsymbol{x} = \boldsymbol{x}_1[i] \\ &= 2[\boldsymbol{r}[i] - (\boldsymbol{s}_1^T\boldsymbol{r}[i])\boldsymbol{s}_1][(\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x}_1[i])^H\boldsymbol{r}[i]]^* \end{aligned} \tag{21}$$

Substituting (21) into (20), we obtain the following LMS implementation of the blind linear MMSE detector. Suppose that at time i, the estimated blind detector is $\boldsymbol{w}_1[i] = \boldsymbol{s}_1 + \boldsymbol{x}_1[i]$. The algorithm performs the following steps for data detection and detector update:

- *Compute the detector output*:

$$z_1[i] = (\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x}_1[i])^H\boldsymbol{r}[i]$$

$$\hat{\beta}_1[i] = \text{sign}\,\{\Re(z[i]z[i-1]^*)\}$$

- *Update*:

$$\boldsymbol{x}_1[i+1] = \boldsymbol{x}_1[i] - \mu z[i]^*[\boldsymbol{r}[i] - (\boldsymbol{s}_1^T\boldsymbol{r}[i])\boldsymbol{s}_1]$$

This algorithm is initialized as $\boldsymbol{x}_1[0] = \mathbf{0}$. The adaptive approach outlined above was first proposed in [9], and is termed the *minimum-output-energy* (MOE) detector.

4. BLIND MULTIUSER DETECTION: SUBSPACE METHODS

In this section, we discuss another approach to blind multiuser detection, which is based on estimating the signal subspace spanned by the user signature waveforms. This approach, first proposed in [38], offers a number of advantages over the direct methods discussed in the previous section.

Assume that the spreading waveforms $\{\boldsymbol{s}_k\}_{k=1}^K$ of K users are linearly independent. The eigendecomposition of the signal autocorrelation matrix $\boldsymbol{C}_r$ in (14) can be written as

$$\boldsymbol{C}_r = \boldsymbol{U}_s \boldsymbol{\Lambda}_s \boldsymbol{U}_s^H + \sigma^2 \boldsymbol{U}_n \boldsymbol{U}_n^H \tag{22}$$

where $\boldsymbol{\Lambda}_s = \text{diag}(\lambda_1, \ldots, \lambda_K)$ contains the largest K eigenvalues of $\boldsymbol{C}_r$; $\boldsymbol{U}_s = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]$ contains the K orthonormal eigenvectors corresponding to the largest K eigenvalues in $\boldsymbol{\Lambda}_s$; $\boldsymbol{U}_n = [\boldsymbol{u}_{K+1}, \ldots, \boldsymbol{u}_N]$ contains the $(N-K)$ orthonormal eigenvectors corresponding to the smallest eigenvalue σ^2 of $\boldsymbol{C}_r$. It is easy to see that range $(\boldsymbol{S})$ = range $(\boldsymbol{U}_s)$. The column space of $\boldsymbol{U}_s$ is called the *signal subspace* and its orthogonal complement, the *noise subspace*, is spanned by the columns of $\boldsymbol{U}_n$. The linear MMSE detector can be expressed in terms of the signal subspace parameters $\boldsymbol{U}_s$ and $\boldsymbol{\Lambda}_s$ as [38]

$$\boldsymbol{w}_1 = \alpha \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \boldsymbol{s}_1 \tag{23}$$

with $\alpha = (\boldsymbol{s}_1^T \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \boldsymbol{s}_1)^{-1}$.

Since the decision rules (7) and (9) are invariant to a positive scaling, the subspace linear multiuser detector given by (23) can be interpreted as follows. First the received signal $\boldsymbol{r}[i]$ is projected onto the signal subspace to get $\boldsymbol{y}[i] \triangleq \boldsymbol{U}_s^H \boldsymbol{r}[i] \in \mathbb{C}^K$, which clearly is a sufficient statistic for demodulating the K users' data bits. The spreading waveform $\boldsymbol{s}_1$ of the desired user is also projected onto the signal subspace to obtain $\boldsymbol{p}_1 \triangleq \boldsymbol{U}_s^H \boldsymbol{s}_1 \in \mathbb{C}^K$. The projection of the linear multiuser detector in the signal subspace is then a signal $\boldsymbol{c}_1 \in \mathbb{C}^K$ such that the detector output is $z_1[i] \triangleq \boldsymbol{c}_1^H \boldsymbol{y}[i]$, and the data bit is demodulated as $\hat{b}_1[i] = \text{sign}\,\{\Re(A_1^* z_1[i])\}$ for coherent detection, and $\beta_1[i] = \text{sign}\,\{\Re(z_1[i] z_1[i-1]^*)\}$ for differential detection. According to (23), the projection of the linear MMSE detector in the signal subspace is given by

$$\boldsymbol{c}_1 = \begin{bmatrix} \frac{1}{\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_K} \end{bmatrix} \boldsymbol{p}_1 \tag{24}$$

Thus, it is obtained by projecting the spreading waveform of the desired user onto the signal subspace, followed by scaling the kth component of this projection by a factor of $1/\lambda_k$.

Since the autocorrelation matrix $\boldsymbol{C}_r$, and therefore its eigencomponents, can be estimated from the received signals, we see that the abovementioned subspace method indeed leads to a blind implementation of the linear MMSE detector. Finally we summarize the subspace blind multiuser detector as follows:

- *Compute the detector*:

$$\begin{aligned} \hat{\boldsymbol{C}}_r &\triangleq \frac{1}{M} \sum_{i=0}^{M-1} \boldsymbol{r}[i] \boldsymbol{r}[i]^H \\ &= \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H + \hat{\boldsymbol{U}}_n \hat{\boldsymbol{\Lambda}}_n \hat{\boldsymbol{U}}_n^H \\ \hat{\boldsymbol{w}}_1 &= \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H \boldsymbol{s}_1 \end{aligned}$$

- *Perform differential detection*:

$$\begin{aligned} & z_1[i] = \hat{\boldsymbol{w}}_1^H \boldsymbol{r}[i], \\ & \hat{\beta}_1[i] = \text{sign}\,\{\Re(z_1[i] z_1^*[i-1])\}, \quad i = 1, \ldots, M-1 \end{aligned}$$

It is seen from the discussion above that the linear MMSE detector is obtained as long as the signal subspace components are identified. The classic approach to subspace estimation is through batch eigenvalue decomposition (ED) of the sample autocorrelation matrix, or batch singular value decomposition (SVD) of the data matrix, which is computationally too expensive for adaptive applications. Modern subspace tracking algorithms are recursive in nature and update the subspace in a sample-by-sample fashion. Various subspace tracking algorithms exist in the literature [e.g., 4,6,7,28,32,41], with different computational complexity and tracking performance. Among the low-complexity subspace tracking algorithms are the PASTd algorithm [41], and the more recently developed NAHJ algorithm [25,26]. Both algorithms have a complexity of $O(NK)$ when tracking K subspace components in a N-dimensional space; but the NAHJ algorithm has a far superior performance.

The adaptive blind multiuser detector based on subspace tracking sequentially estimates the signal subspace components, and forms the closed-form detector from these estimates. Specifically, supposedly at time $(i-1)$, the estimated signal subspace rank is $K[i-1]$ and the components are $(\boldsymbol{U}_s[i-1], \boldsymbol{\Lambda}_s[i-1])$. Then at time i, the adaptive detector performs the following steps to update the detector and to estimate the data:

- *Update the signal subspace*: Using a particular signal subspace tracking algorithm, update the signal subspace rank $K[i]$ and the signal subspace components $(\boldsymbol{U}_s[i], \boldsymbol{\Lambda}_s[i])$.
- *Form the detector and perform differential detection*:

$$\begin{aligned} \boldsymbol{w}_1[i] &= \boldsymbol{U}_s[i] \boldsymbol{\Lambda}_s[i]^{-1} \boldsymbol{U}_s[i]^H \boldsymbol{s}_1, \\ z_1[i] &= \boldsymbol{w}_1[i]^H \boldsymbol{r}[i], \\ \hat{\beta}_1[i] &= \text{sign}\,\{\Re(z_1[i] z_1[i-1]^*)\} \end{aligned}$$

Simulation Example

This example compares the performance of the subspace blind adaptive multiuser detector using the NAHJ

subspace tracking algorithm [26], with that of the LMS MOE blind adaptive multiuser detector. It assumes a synchronous CDMA system with seven users ($K = 7$), each employing a gold sequence of length 15 ($N = 15$). The desired user is user 1. There are two 0-dB and four 10-dB interferers. The performance measure is the output signal-to-interference-plus-noise ratio (SINR). The performance is shown in Fig. 1. It is seen that the subspace blind detector significantly outperforms the LMS MOE blind detector in terms of both convergence rate and steady-state SINR.

5. BLIND MULTIUSER DETECTION IN MULTIPATH CHANNELS

In the previous sections, we have focused primarily on the synchronous CDMA signal model. In a practical wireless CDMA system, however, the user signals are asynchronous. Moreover, the physical channel exhibits dispersion due to multipath effects that further attenuate the user signals. In this section, we address blind multiuser detection in such channels. As will be seen, the principal techniques developed in the previous section can be applied to this more complicated situation as well.

5.1. Multipath Signal Model

We now consider a more general multiple-access signal model where the users are asynchronous and the channel exhibits multipath distortion effects. Let the multipath channel impulse response of the kth user be

$$g_k(t) = \sum_{l=1}^{L} \alpha_{l,k}\delta(t - \tau_{l,k}) \tag{25}$$

where L is the total number of paths in the channel; $\alpha_{l,k}$ and $\tau_{l,k}$ are, respectively, the complex path gain and the delay of the kth user's lth path, $\tau_{1,k} < \tau_{2,k} < \cdots < \tau_{L,k}$. The received continuous-time signal in this case is given by

$$\begin{aligned} r(t) &= \sum_{k=1}^{K}\sum_{i=0}^{M-1} b_k[i]\{s_k(t - iT) \star g_k(t)\} + n(t) \\ &= \sum_{k=1}^{K}\sum_{i=0}^{M-1} b_k[i] \sum_{l=1}^{L} \alpha_{l,k} s_k(t - iT - \tau_{l,k}) + n(t) \end{aligned} \tag{26}$$

where $\star$ denotes convolution.

At the receiver, the received signal $r(t)$ is filtered by a chip-matched filter and sampled at the chip rate. Let

$$\iota \triangleq \max_{1\le k\le K} \left\{ \left\lceil \frac{\tau_{L,k} + T_c}{T} \right\rceil \right\} \tag{27}$$

be the maximum delay spread in terms of symbol intervals. Denote $r_q[i] \triangleq \int_{iT+qT_c}^{iT+(q+1)T_C} r(t)\psi(t - iT - qT_c)dt$ and $n_q[i] = \int_{iT+qT_c}^{iT+(q+1)T_c} n(t)\psi(t - iT - qT_c)dt$, for $q = 0, \ldots, N-1; i = 0, \ldots, M-1$. Denote further

$$\underbrace{\underline{r}[i]}_{P\times 1} \triangleq \begin{bmatrix} r_0[i] \\ \vdots \\ r_{N-1}[i] \end{bmatrix}, \quad \underbrace{\underline{b}[i]}_{K\times 1} \triangleq \begin{bmatrix} b_1[i] \\ \vdots \\ b_K[i] \end{bmatrix},$$

$$\underbrace{\underline{n}[i]}_{P\times 1} \triangleq \begin{bmatrix} n_0[i] \\ \vdots \\ n_{N-1}[i] \end{bmatrix}$$

$$\underbrace{\underline{H}[m]}_{N\times K} \triangleq \begin{bmatrix} h_1[mN] & \cdots & h_K[mN] \\ \vdots & \vdots & \vdots \\ h_1[mN+N-1] & \cdots & h_K[mN+N-1] \end{bmatrix},$$

$$m = 0, \ldots, \iota$$

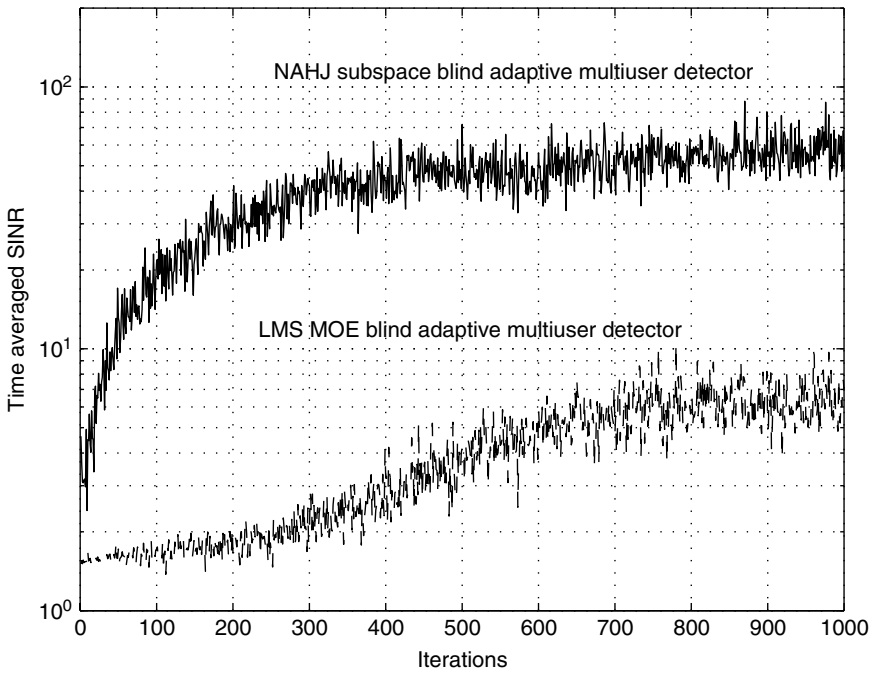


Figure 1. Performance comparison between the subspace blind adaptive multiuser detector using the NAHJ subspace tracking algorithm, and the LMS MOE blind adaptive multiuser detector.

where $\{h_k[j]\}_j$ is the composite signature waveform of the kth user, which will be discussed later. We then have the following discrete-time signal model [37]

$$\underline{r}[i] = \underline{H}[i] \star \underline{b}[i] + \underline{n}[i]. \tag{28}$$

By stacking Q successive sample vectors, we further define the quantities

$$\underbrace{\boldsymbol{r}[i]}_{NQ\times 1} \triangleq \begin{bmatrix} \underline{r}[i] \\ \vdots \\ \underline{r}[i+Q-1] \end{bmatrix}, \quad \underbrace{\boldsymbol{n}[i]}_{NQ\times 1} \triangleq \begin{bmatrix} \underline{n}[i] \\ \vdots \\ \underline{n}[i+Q-1] \end{bmatrix},$$

$$\underbrace{\boldsymbol{b}[i]}_{K(Q+\iota)\times 1} \triangleq \begin{bmatrix} \underline{b}[i-\iota] \\ \vdots \\ \underline{b}[i+Q-1] \end{bmatrix}$$

$$\underbrace{\boldsymbol{H}}_{NQ\times K(Q+\iota)} \triangleq \begin{bmatrix} \underline{H}[\iota] & \cdots & \underline{H}[0] & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \underline{H}[\iota] & \cdots & \underline{H}[0] \end{bmatrix}$$

where Q, called the "smoothing factor," is given by $Q = \lceil (N+K)/(N-K) \rceil \iota$; Note that for such Q, the matrix $\boldsymbol{H}$ is a "tall" matrix: $NQ \geq K(Q+\iota)$. We can then write (28) in a matrix forms as

$$\boldsymbol{r}[i] = \boldsymbol{H}\, \boldsymbol{b}[i] + \boldsymbol{n}[i] \tag{29}$$

5.2. Linear MMSE Multiuser Detector

Suppose that we are interested in demodulating the data bits of user 1. We can then write (28) as

$$\underline{r}[i] = \underline{H}^1[0] b_1[i] + \sum_{m=1}^{\iota} \underline{H}^1[m] b_1[i-m] + \sum_{k=2}^{K} \sum_{m=0}^{\iota} \underline{H}^k[m] b_k[i-m] + \underline{n}[i] \tag{30}$$

where $\underline{H}^k[m]$ denotes the kth column of $\underline{H}[m]$. In (30), the first term contains the data bit of the desired user at time i; the second term contains the previous data bits of the desired user, namely, intersymbol interference (ISI); and the last term contains the signals from other users, namely, multiple-access interference (MAI). Hence compared with the synchronous model considered in the previous sections, the multipath channel introduces ISI, which, together with MAI, must be contended with at the receiver. It is seen that the augmented signal model (29) is very similar to the synchronous signal model (5). We proceed to develop the linear receiver for this system.

A linear receiver for user 1 is a (NQ)-dimensional complex vector $\boldsymbol{w}_1 \in \mathbb{C}^{NQ}$, which is correlated with the received signal $\boldsymbol{r}[i]$ in (29), to compute the ith bit of this user, according to the following rule:

$$z_1[i] = \boldsymbol{w}_1^H \boldsymbol{r}[i] \tag{31}$$

$$\hat{\beta}_1[i] = \text{sign}\,\{\Re(z_1[i] z_1[i-1]^*)\} \tag{32}$$

The linear MMSE detector has the form of (32) with the weight vector chosen to minimize the output mean-square error (MSE):

$$\boldsymbol{w}_1 = \arg \min_{\boldsymbol{w}\in\mathbb{C}^{NQ}} E\left\{ \|b_1[i] - \boldsymbol{w}^H \boldsymbol{r}[i]\|^2 \right\} = \boldsymbol{C}_r^{-1} \overline{\boldsymbol{h}}_1 \tag{33}$$

where

$$\boldsymbol{C}_r = E\{\boldsymbol{r}[i]\boldsymbol{r}[i]^H\} = \boldsymbol{H}\boldsymbol{H}^H + \sigma^2 \boldsymbol{I} \tag{34}$$

$$\overline{\boldsymbol{h}}_1 \triangleq E\{\boldsymbol{r}[i] b_1[i]\} = \boldsymbol{H}[:, KQ+1] \tag{35}$$

$$= [\underbrace{h_1[0], \ldots, h_1[N-1], \ldots, h_1[\iota N], \ldots, h_1[\iota N + N - 1]}_{\boldsymbol{h}_1^T}, \underbrace{0, \ldots, 0}_{[N(Q-\iota-1)]0\text{s}}]^T$$

($\boldsymbol{H}[:, m]$ denotes the mth column of $\boldsymbol{H}$.)

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{NQ}$ be the eigenvalues of $\boldsymbol{C}_r$ in (34). Assuming that the matrix $\boldsymbol{H}$ has full column rank $r \triangleq K(Q+\iota)$, the signal component of the covariance matrix $\boldsymbol{C}_r$, namely, $(\boldsymbol{H}\boldsymbol{H}^H)$, has rank r. Therefore we have

$$\lambda_i > \sigma^2 \quad \text{for} \quad i = 1, \ldots, r$$
$$\lambda_i = \sigma^2 \quad \text{for} \quad i = r+1, \ldots, NQ$$

By performing an eigendecomposition of the matrix $\boldsymbol{C}_r$, we obtain

$$\boldsymbol{C}_r = \boldsymbol{U}_s \boldsymbol{\Lambda}_s \boldsymbol{U}_s^H + \sigma^2 \boldsymbol{U}_n \boldsymbol{U}_n^H \tag{36}$$

where $\boldsymbol{\Lambda}_s = \text{diag}(\lambda_1, \ldots, \lambda_r)$ contains the r largest eigenvalues of $\boldsymbol{C}_r$ in descending order and $\boldsymbol{U}_s = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_r]$ contains the corresponding orthonormal eigenvectors; $\boldsymbol{U}_n = [\boldsymbol{u}_{r+1} \cdots \boldsymbol{u}_{NQ}]$ contains the $(NQ-r)$ orthonormal eigenvectors that correspond to the eigenvalue σ^2. It is easy to see that range $(\boldsymbol{H})$ = range $(\boldsymbol{U}_s)$. The column space of $\boldsymbol{U}_s$ is the signal subspace and the noise subspace is spanned by the columns of $\boldsymbol{U}_n$. The linear MMSE detector given by (33) can be expressed in terms of the abovementioned signal subspace components as

$$\boldsymbol{w}_1 = \alpha \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \overline{\boldsymbol{h}}_1 \tag{37}$$

with $\alpha \triangleq (\overline{\boldsymbol{h}}_1^H \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \overline{\boldsymbol{h}}_1)^{-1}$.

5.3. Blind Channel Estimation

It is seen from the preceding discussion that, unlike the synchronous case where the linear MMSE detector can be written in closed form once the signal subspace components are identified, in multipath channels, the composite signature waveform of the desired user, $\overline{\boldsymbol{h}}_1$, is needed to form the blind detector. It is essentially the channel distorted original spreading waveform $\boldsymbol{s}_1$. We next address the problem of blind channel estimation. It can be shown [37] that for each k, $1 \leq k \leq K$

$$h_k[n] = \sum_{j=0}^{N-1} c_{j,k} f_k[n-j], \quad n = 0, 1, \ldots, (\iota+1)N - 1 \tag{38}$$

with

$$f_k[m] \triangleq \frac{1}{\sqrt{N}} \sum_{l=1}^{L} \alpha_{l,k} \int_0^{T_c} \psi(t)\psi(t - \tau_{l,k} + mT_c), \quad m = 0, 1, \ldots, \mu - 1 \tag{39}$$

where the length of the discrete-time channel response $\{f_k[m]\}$ satisfies

$$\mu = \left\lceil \frac{\tau_{L,k}}{T_c} \right\rceil = \left\lceil \frac{\tau_{L,k}}{T} \cdot \frac{T}{T_c} \right\rceil \le \iota N \tag{40}$$

Denote

$$\boldsymbol{h}_k = \begin{bmatrix} h_k[0] \\ \vdots \\ h_k[(\iota+1)N-1] \end{bmatrix}_{(\iota+1)N\times 1}, \quad \boldsymbol{f}_k = \begin{bmatrix} f_k[0] \\ \vdots \\ f_k[\mu-1] \end{bmatrix}_{\mu\times 1}$$

and

$$\boldsymbol{\Xi}_k = \begin{bmatrix} c_{0,k} & & & \\ c_{1,k} & c_{0,k} & & \\ \vdots & c_{1,k} & \ddots & \\ \vdots & \vdots & \ddots & c_{0,k} \\ c_{N-1,k} & \vdots & & c_{1,k} \\ & c_{N-1,k} & & \vdots \\ & & \ddots & \vdots \\ & & & c_{N-1,k} \end{bmatrix}_{(\iota+1)N\times\mu}$$

Then (38) can be written in a matrix form as

$$\boldsymbol{h}_k = \boldsymbol{\Xi}_k \boldsymbol{f}_k \tag{41}$$

Recall that when the ambient channel noise is white, through an eigendecomposition on the autocorrelation matrix of the received signal, the signal subspace and the noise subspace can be identified. In order to estimate the desired user's composite signature waveform, it suffices to estimate the corresponding channel $\boldsymbol{f}_1$, which in turn can be estimated by exploiting the orthogonality between the signal subspace and the noise subspace [3,15,29,37]. Specifically, since $\boldsymbol{U}_n$ is orthogonal to the column space of $\boldsymbol{H}$, and $\overline{\boldsymbol{h}}_1$ is in the column space of $\boldsymbol{H}$ [cf. (35)], we have

$$\boldsymbol{U}_n^H \overline{\boldsymbol{h}}_1 = \boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1 = \boldsymbol{0} \tag{42}$$

where

$$\overline{\boldsymbol{h}}_1 = \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{0}_{(Q-\iota-1)N\times 1} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\Xi}_1 \\ \boldsymbol{0}_{(Q-\iota-1)N\times\mu} \end{bmatrix}}_{\overline{\boldsymbol{\Xi}}_k} \boldsymbol{f}_1 \tag{43}$$

From this relationship, we can obtain an estimate of the desired user's channel response $\boldsymbol{f}_1$, by computing the minimum eigenvector of the matrix $(\overline{\boldsymbol{\Xi}}_1^T \boldsymbol{U}_n \boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_1)$. The condition for the channel estimate obtained in such a way to be unique is that the matrix $(\boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_k)$ has rank $(\mu - 1)$, which necessitates this matrix to be tall: $[Nm - K(m + \iota)] \ge \mu$. Since $\mu \le \iota N$ [cf. (40)], we therefore choose the smoothing factor Q to satisfy

$$NQ - K(Q + \iota) \ge \iota N \ge \mu \tag{44}$$

That is, $Q = \lceil [(N-K)/(N+K)] \cdot \iota \rceil$. On the other hand, the condition (44) implies that for fixed Q, the total number of users that can be accommodated in the system is $\lceil [(Q-\iota)/(Q+\iota)] \cdot N \rceil$. Moreover, if the received signal is sampled at a multiple (p) of the chip rate, the total number of users that can be accommodated in the system is $\lceil [(Q-\iota)/(Q+\iota) \cdot Np \rceil$ [34].

Finally, we summarize the batch algorithm for blind linear multiuser detection in multipath CDMA channels as follows:

- *Estimate the signal subspace*:

$$\hat{\boldsymbol{C}}_r \triangleq \frac{1}{M} \sum_{i=0}^{M-1} \boldsymbol{r}[i]\boldsymbol{r}[i]^H = \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H + \hat{\boldsymbol{U}}_n \hat{\boldsymbol{\Lambda}}_n \hat{\boldsymbol{U}}_n^H$$

- *Estimate channel*:

$$\boldsymbol{f}_1 = \text{min-eigenvector } (\overline{\boldsymbol{\Xi}}_1^T \boldsymbol{U}_n \boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_1)$$
$$\overline{\boldsymbol{h}}_1 = \overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1$$
$$\hat{\boldsymbol{w}}_1 = \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H \overline{\boldsymbol{h}}_1$$

- *Perform differential detection*:

$$z_1[i] = \hat{\boldsymbol{w}}_1^H \boldsymbol{r}[i]$$
$$\hat{\beta}_1[i] = \text{sign } \{\Re(z_1[i] z_1[i-1]^*)\}, \quad i = 1, \ldots, M-1$$

5.4. Adaptive Receiver Structure

We next consider an adaptive blind multiuser receiver in multipath CDMA channels based on the subspace linear MMSE detector. First, we address an adaptive implementation of the blind channel estimator discussed above. Suppose that the signal subspace $\boldsymbol{U}_s$ is known. Denote by $\boldsymbol{z}[i]$ the projection of the received signal $\boldsymbol{r}[i]$ onto the noise subspace:

$$\boldsymbol{z}[i] \triangleq \boldsymbol{r}[i] - \boldsymbol{U}_s \boldsymbol{U}_s^H \boldsymbol{r}[i] \tag{45}$$
$$= \boldsymbol{U}_n \boldsymbol{U}_n^H \boldsymbol{r}[i] \tag{46}$$

Since $\boldsymbol{z}[i]$ lies in the noise subspace, it is orthogonal to any signal in the signal subspace. In particular, it is orthogonal to $\overline{\boldsymbol{h}}_1 = \overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1$. Hence $\boldsymbol{f}_1$ is the solution to the following constrained optimization problem:

$$\min_{\boldsymbol{f}_1 \in \mathbb{C}^\mu} E\left\{ \|(\overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1)^H \boldsymbol{z}[i]\|^2 \right\}, \quad \text{s.t.} \quad \|\boldsymbol{f}_1\| = 1 \tag{47}$$

Standard adaptive algorithms can be used to sequentially update $\boldsymbol{f}_1$ on the basis of the proceeding optimization criterion [26].

The block diagram of the subspace blind adaptive receiver is shown in Fig. 2. The received signal $\boldsymbol{r}[i]$ is

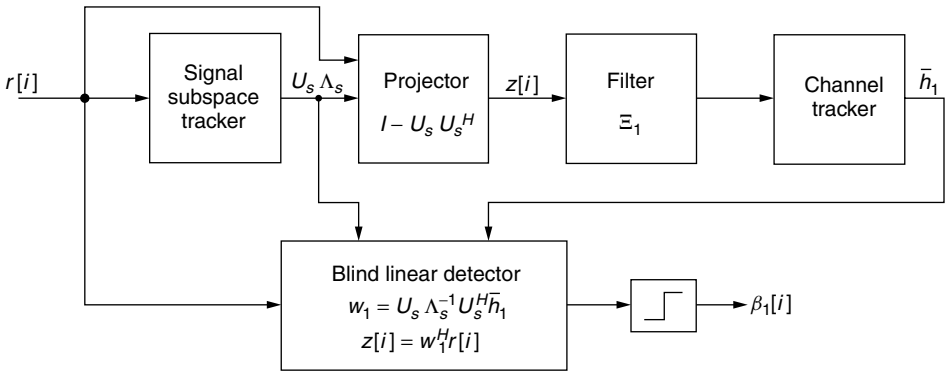


Figure 2. Diagram of a subspace blind adaptive receiver in multipath CDMA channels.

fed into a subspace tracker that sequentially estimates the signal subspace components ($\boldsymbol{U}_s$, $\boldsymbol{\Lambda}_s$). The signal $\boldsymbol{r}[i]$ is then projected onto the noise subspace to obtain $\boldsymbol{z}[i]$, which is in turn passed through a linear filter that is determined by the signature sequence $\boldsymbol{s}_1$ of the desired user. The output of this filter is fed into a channel tracker that estimates the channel state $\boldsymbol{f}_1$. Finally, the blind linear MMSE detector $\boldsymbol{w}_1$ is constructed in closed form, based on the estimated signal subspace components and the channel state. The adaptive receiver algorithm is summarized as follows. Suppose that at time $(i-1)$, the estimated signal subspace rank is $K[i-1]$ and the components are ($\boldsymbol{U}_s[i-1]$, $\boldsymbol{\Lambda}_s[i-1]$). The estimated channel vector is $\boldsymbol{f}_1[i-1]$. Then at time i, the adaptive detector performs the following steps to update the detector and to estimate the data:

- *Update the signal subspace*: Using a particular signal subspace tracking algorithm, update the signal subspace rank $K[i]$ and the subspace components ($\boldsymbol{U}_s[i]$, $\boldsymbol{\Lambda}[i]$).
- *Update the channel*: Using a particular adaptive algorithm, update the channel estimate $\boldsymbol{f}_1[i]$.
- *Form the detector and perform differential detection*:

$$\boldsymbol{w}_1[i] = \boldsymbol{U}_s[i]\boldsymbol{\Lambda}_s[i]^{-1}\boldsymbol{U}_s[i]^H\overline{\boldsymbol{\Xi}}_1\boldsymbol{f}_1[i],$$
$$z_1[i] = \boldsymbol{w}_1[i]^H\boldsymbol{r}[i],$$
$$\hat{\beta}_1[i] = \text{sign}\left\{\Re(z_1[i]z_1[i-1]^*)\right\}.$$

Simulation Example. We next give a simulation example on the performance of the blind adaptive receiver in an asynchronous CDMA system with multipath channels. The processing gain $N = 15$ and the spreading codes are Gold codes of length 15. Each user's channel has $L = 3$ paths. The delay of each path $\tau_{k,l}$ is uniform on $[0, 10T_c]$. Hence, the maximum delay spread is one symbol interval, namely, $\iota = 1$. The fading gain of each path in each user's channel is generated from a complex Gaussian distribution and is fixed for all simulations. The path gains in each user's channel are normalized so that each user's signal arrives at the receiver with the same power. The smoothing factor is $Q = 2$. The received signal is sampled at twice the chip rate ($p = 2$). Hence, the total number of users that this system can accommodate is 10. Figure 3 is shows the

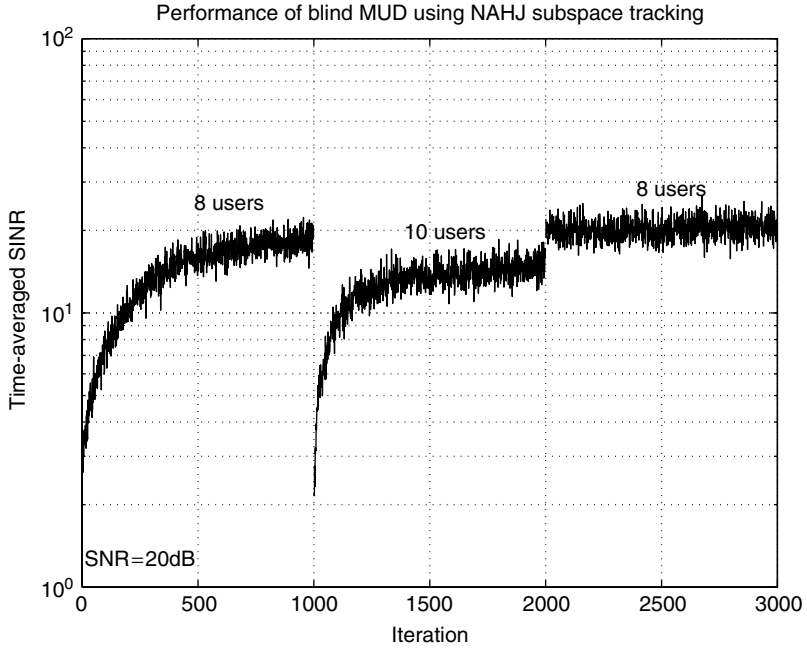


Figure 3. Performance of the subspace blind adaptive multiuser detector (MUD) in an asynchronous CDMA system with multipath.

performance of the subspace blind adaptive receiver using the NAHJ subspace tracking algorithm [26], in terms of SINR. During the first 1000 iterations there are 8 total users. At iteration 1000, 2 new users are added to the system. At iteration 2000, one additional known user is added and three existing users vanish. We see that this blind adaptive receiver can closely track the dynamics of the channel.

6. CONCLUDING REMARKS

In this article, we have presented signal processing techniques for blind multiuser detection in CDMA systems. The main objective is to perform interference suppression and signal detection in a CDMA downlink environment, with the prior knowledge of only the spreading waveform of the desired user. We have presented two approaches to blind multiuser detection — the direct matrix inversion (DMI) method and the subspace method. Note that in addition to what we discussed here, both approaches have been extended to address a number of other channel impairments. For example, under the DMI framework, techniques have been proposed to combat narrowband interference [22,23], channel dispersion [30,31] fading channels [2,14,35,42], and synchronization [16,17]. Moreover, within the subspace framework, extensions have been made to fading channels [27,36] and antenna array spatial processing [38], for blind adaptive joint channel/array response estimation, multiuser detection, and equalization. Another salient feature of the subspace approach is that it can be combined with the M-regression techniques to achieve blind adaptive robust multiuser detection in non-Gaussian ambient noise channels [39]. Furthermore, an analytical performance assessment of the DMI blind detector and the subspace blind detector is given in the literature [11–13,40]. Finally, we remark that in the CDMA uplink, typically the base station receiver has the knowledge of the spreading waveforms of all users within its cell, but not that of the users from other cells. *Group-blind* multiuser detection techniques have been developed [34] to address such scenarios.

Acknowledgments

This work is supported in part by the NSF grant CAREER CCR-9875314 and the NSF grant CCR 9980599. The author would like to thank Dr. Daryl Reynolds for providing the two simulation examples.

BIOGRAPHY

Xiaodong Wang received a B.S. degree in electrical engineering and applied mathematics (with the highest honor) from Shanghai Jiao Tong University, Shanghai, China, in 1992; an M.S. degree in electrical and computer engineering from Purdue University, West Lafayette, Indiana, in 1995; and a Ph.D degree in electrical engineering from Princeton University, New Jersey, in 1998. From July 1998 to December 2001 he was an assistant professor in the Department of Electrical Engineering, Texas A&M University. In January 2002, he joined the Department of Electrical Engineering, at Columbia University, New York, as an assistant professor.

Dr. Wang's research interests fall in the general areas of computing, signal processing, and communications. He has worked in the areas of digital communications, digital signal processing, parallel and distributed computing, nanoelectronics and quantum computing, and has published extensively in these areas. His current research interests include multiuser communications theory and advanced signal processing for wireless communications. He received the 1999 NSF CAREER Award, and the 2001 IEEE Communications Society and Information Theory Society Joint Paper Award. He currently serves as an associate editor for the *IEEE Transactions on Communications*, the *IEEE Transactions on Signal Processing*, and the *IEEE Transactions on Wireless Communications*.

BIBLIOGRAPHY

1. A. Abdulrahman, D. D. Falconer, and A. U. Sheikh, Decision feedback equalization for CDMA in indoor wireless communications, *IEEE J. Select. Areas Commun.* **12**(4): 698–706 (May 1994).
2. A. N. Barbosa and S. L. Miller, Adaptive detection of DS/CDMA signals in fading channels, *IEEE Trans. Commun.* **COM-46**(1): 115–124 (Jan. 1998).
3. S. E. Bensley and B. Aazhang, Subspace-based channel estimation for code-division multiple-access communication systems, *IEEE Trans. Commun.* **COM-44**(8): 1009–1020 (Aug. 1996).
4. C. H. Bischof and G. M. Shroff, On updating signal subspaces, *IEEE Trans. Signal Process.* **40**(1): 96–105 (Jan. 1992).
5. D.-S. Chen and S. Roy, An adaptive multiuser receiver for CDMA systems, *IEEE J. Select. Areas Commun.* **12**(5): 808–816 (June 1994).
6. P. Comon and G. H. Golub, Tracking a few extreme singular values and vectors in signal processing, *Proc. IEEE* **78**(8): 1327–1343 (Aug. 1990).
7. R. D. DeGroat, Noniterative subspace tracking, *IEEE Trans. Signal Process.* **40**(3): 571–577 (March 1992).
8. S. Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice-Hall, 1996.
9. M. Honig, U. Madhow, and S. Verdú, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **IT-41**(4): 944–960 (July 1995).
10. M. Honig and H. V. Poor, Adaptive interference suppression, in H. V. Poor and G. W. Wornell, eds., *Wireless Communications: A Signal Processing Perspective*, Prentice-Hall, Upper Saddle River, NJ, 1998, pp. 64–128.
11. A. Høst-Madsen and X. Wang, Performance of blind and group-blind multiuser detection, *IEEE Trans. Inform. Theory* **48**(6): (June 2002).
12. A. Høst-Madsen and X. Wang, Performance of blind and group-blind multiuser detectors, *Proc. 38th Annual Allerton Conf. Communications, Computing and Control*, Monticello, IL, Oct. 2000.
13. A. Høst-Madsen and X. Wang, Performance of blind multiuser detectors, *Proc. 10th Int. Symp. Information Theory and Its Applications* (*ISITA'00*), Honolulu, HI, Nov. 2000.

14. H. C. Huang and S. Verdú, Linear differentially coherent multiuser detection for multipath channels, *Wireless Pers. Commun.* **6**(1–2): 113–136 (Jan. 1998).

15. H. Liu and G. Xu, A subspace method for signal waveform estimation in synchronous CDMA systems, *IEEE Trans. Commun.* **COM-44**(10): 1346–1354 (Oct. 1996).

16. U. Madhow, Blind adaptive interference suppression for the near-far resistant acquisition and demodulation of direct-sequence CDMA signals, *IEEE Trans. Signal Process.* **45**(1): 124–136 (Jan. 1997).

17. U. Madhow, Blind adaptive interference suppression for CDMA, *Proc. IEEE* **86**(10): 2049–2069 (Oct. 1998).

18. U. Madhow and M. Honig, MMSE interference suppression for direct-sequence spread-spectrum CDMA, *IEEE Trans. Commun.* **COM-42**(12): 3178–3188 (Dec. 1994).

19. S. L. Miller, An adaptive direct-sequence code-division multiple-access receiver for multiuser interference rejection, *IEEE Trans. Commun.* **COM-43**(2–4): 1556–1565 (Feb.–April 1995).

20. U. Mitra and H. V. Poor, Adaptive receiver algorithms for near-far resistant CDMA, *IEEE Trans. Commun.* **COM-43**(2–4): 1713–1724 (Feb.–April 1995).

21. U. Mitra and H. V. Poor, Analysis of an adaptive decorrelating detector for synchronous CDMA channels, *IEEE Trans. Commun.* **COM-44**(2): 257–268 (Feb. 1996).

22. H. V. Poor and X. Wang, Code-aided interference suppression in DS/CDMA communications. Part II: Parallel blind adaptive implementations, *IEEE Trans. Commun.* **COM-45**(9): 1112–1122 (Sept. 1997).

23. H. V. Poor and X. Wang, Blind adaptive suppression of narrowband digital interferers from spread-spectrum signals, *Wireless Pers. Commun.* **6**(1–2): 69–96 (Jan. 1998).

24. P. B. Rapajić and B. S. Vučetić, Adaptive receiver structures for asynchronous CDMA systems, *IEEE J. Select. Areas Commun.* **12**(4): 685–697 (May 1994).

25. D. Reynolds and X. Wang, Adaptive group-blind multiuser detection based on a new subspace tracking algorithm, *IEEE Trans. Commun.* (in press).

26. D. Reynolds and X. Wang, Group-blind multiuser detection based on subspace tracking, *Proc. 2000 Conf. Information Sciences and Systems*, Princeton, NJ, March 2000.

27. Y. Song and S. Roy, Blind adaptive reduced-rank detection for DS-CDMA signals in multipath channels, *IEEE J. Select. Areas Commun.* **17**(11): 1960–1970 (Nov. 1999).

28. G. W. Stewart, An updating algorithm for subspace tracking, *IEEE Trans. Signal Process.* **40**(6): 1535–1541 (June 1992).

29. M. Torlak and G. Xu, Blind multiuser channel estimation in asynchronous CDMA systems, *IEEE Trans. Signal Process.* **45**(1): 137–147 (Jan. 1997).

30. M. K. Tsatsanis, Inverse filtering criteria for CDMA systems, *IEEE Trans. Signal Process.* **45**(1): 102–112 (Jan. 1997).

31. M. K. Tsatsanis and G. B. Giannakis, Blind estimation of direct sequence spread spectrum signals in multipath, *IEEE Trans. Signal Process.* **45**(5): 1241–1252 (May 1997).

32. D. W. Tufts and C. D. Melissinos, Simple, effective computation of principal eigenvectors and their eigenvalues and application to high resolution estimation of frequencies, *IEEE Trans. Acoust. Speech Signal Process.* **34**(5): 1046–1053 (Oct. 1986).

33. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, Cambridge, UK, 1998.

34. X. Wang and A. Høst-Madsen, Group-blind multiuser detection for uplink CDMA, *IEEE J. Select. Areas Commun.* **17**(11): 1971–1984 (Nov. 1999).

35. X. Wang and H. V. Poor, Adaptive joint multiuser detection and channel estimation in multipath fading CDMA channels, *ACM/Baltzer Wireless Networks*, 1998, pp. 453–470.

36. X. Wang and H. V. Poor, Blind adaptive multiuser detection in multipath CDMA channels based on subspace tracking, *IEEE Trans. Signal Process.* **46**(11): 3030–3044 (Nov. 1998).

37. X. Wang and H. V. Poor, Blind equalization and multiuser detection for CDMA communications in dispersive channels, *IEEE Trans. Commun.* **COM-46**(1): 91–103 (Jan. 1998).

38. X. Wang and H. V. Poor, Blind multiuser detection: A subspace approach, *IEEE Trans. Inform. Theory* **44**(2): 677–691 (March 1998).

39. X. Wang and H. V. Poor, Robust multiuser detection in non-Gaussian channels, *IEEE Trans. Signal Process.* **47**(2): 289–305 (Feb. 1999).

40. X. Wang, J. Zhang, and A. Høst-Madsen, Blind and group-blind multiuser detection: Effect of estimation error and large system performance. Invited talk at the 2001 IEEE Communication Theory Workshop, Borrego Springs, CA, April 2001.

41. B. Yang, Projection approximation subspace tracking, *IEEE Trans. Signal Process.* **44**(1): 95–107 (Jan. 1995).

42. L. J. Zhu and U. Madhow, Adaptive interference suppression for direct sequence CDMA over severely time-varying channels, *Proc. IEEE Globecom'97*, Nov. 1997.

BLUETOOTH RADIO SYSTEM

JAAP C. HAARTSEN
Ericsson Technology
Licensing AB
Emmen, The Netherlands

1. INTRODUCTION

Progress in microelectronics and VLSI technology has fostered the widespread use of computing and communication devices for commercial usage. The success of consumer products such as notebooks, laptops, personal digital assistants (PDAs), cell phones, cordless phones, and their peripherals has been based on continuous cost and size reduction. Information transfer between these devices has been cumbersome mainly relying on cables or infrared. Although infrared transceivers are inexpensive, they have limited range, are sensitive to directions and to objects in the propagation path, and can in principle be used only between two devices. By contrast, radio transceivers have much larger range, can propagate through various materials and around objects, and can connect many devices simultaneously. A new universal radio interface has been developed enabling electronic devices to communicate wirelessly via *short-range radio* connections. The *Bluetooth technology* — which has gained the support from leading manufacturers in the telecom, PC

and consumer industry — eliminates the need for wires, cables, and the corresponding connectors, and paves the way for completely new devices and applications. The Bluetooth technology provides a solution for access to information and personal communication by enabling connectivity between devices in proximity of each other, allowing each device to keep its inherent function based on its user interface, form factor, cost, and power constraints. Radio technology will allow this connectivity to occur without any explicit user interaction. The Bluetooth technology has been optimized with respect to low-power, small-size, and low-cost, enabling singe-chip radios that can be embedded into these personal devices.

2. AD HOC COMMUNICATIONS

Most radio systems in use today rely on a fixed infrastructure. Cellular phone systems like GSM, IS136, or IS95 [1] obtain regional coverage by applying a wired backbone network using a multitude of base stations placed at strategic positions to provide local cell coverage. The mobile users apply mobile terminals to access this public land mobile network (PLMN); the terminals maintain a connection to the network via a radio link to the base stations. There is a strict separation between the fixed base stations and the mobile terminals. Once registered to the mobile network, the terminals remain locked to the control channels on the radio interface and connections can be established and released according to the control protocols. Channel access, channel allocation, traffic control, and interference minimization is taken care of by intelligent centers in the network that also coordinate the activity of the base stations. The basic architecture of a *mobile system* is shown in Fig. 1. The mobile network, which handles all mobility issues, is strictly separated from the fixed public switched telephone network (PSTN). Gateways between the PLMN and PSTN provide a smooth connection between the mobile and fixed telephony world. Alternatively, the base stations can directly be connected to the public switched network as shown in Fig. 2. In this case, the radio interface forms a *wireless extension* of the wired network. Because of a lack of coordination between the base stations, radio functions like channel access, channel allocation, traffic control, and interference mitigation must now be dealt with by the base stations independently. Still in the wireless extensions, there

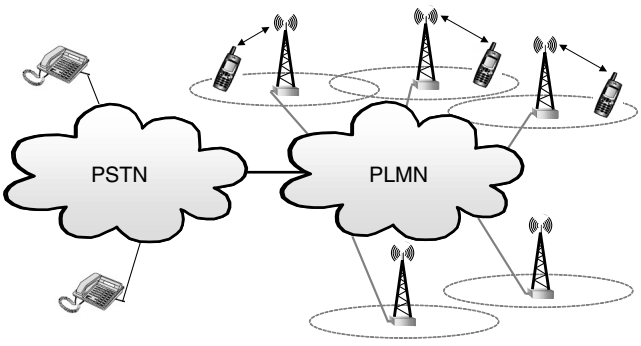


Figure 1. Mobile system architecture.

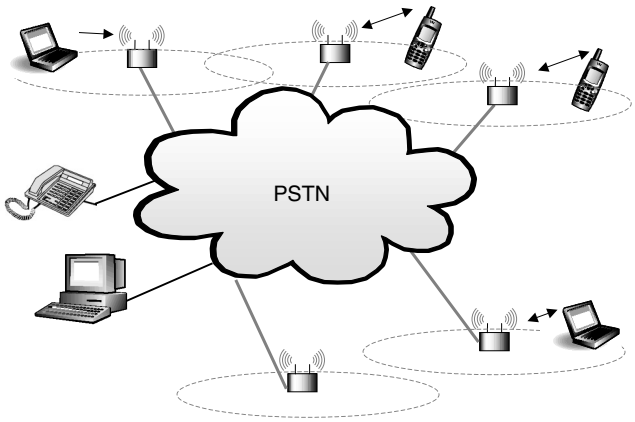


Figure 2. Wireless extension architecture.

are cells defined by the base stations and there is a strict separation between the fixed base stations and the mobile terminals. Mutual interference can reasonably be controlled by proper selection of channels. Residential and office cordless phones for example based on DECT [2] provide wireless extensions to the PSTN, whereas, for example, WLAN 802.11 or Hiperlan2 [3] provide wireless extensions to the Ethernet or ATM networks.

In *ad hoc systems*, there is no distinction between radio units; that is, there are no separate base stations or terminals. Ad hoc connectivity is based on *peer* communications. There is no wired infrastructure to support the connectivity between the portable units; there is no central controller for the units to rely on for making connections nor is there support for coordination of communications. Some WLAN systems do have an ad hoc mode where terminals can make connections without the intervention of a base station. However, in these ad-hoc scenarios, a single channel is created to which all units in range are connected as illustrated in Fig. 3. Base-station-like functions are shared among the mobile terminals. By contrast, Bluetooth is based on device-to-device connections where only two or a few mobile units share the same channel. For the scenarios envisioned by Bluetooth, it is highly likely that a large number of independent connections coexist in the same area without any mutual coordination; that is, tens of ad hoc links must share the same radio spectrum at the same location in an uncoordinated fashion. This will be indicated as a *scatter ad-hoc environment* (see Fig. 4).

Ad hoc radio systems have been in use for some time, for example, walkie-talkie systems used by the military, the police, the fire brigade and by rescue teams in general. However, the Bluetooth system is the first commercial ad-hoc radio system envisioned to be used on a large scale and widely available to the public.

3. SPECTRAL COEXISTENCE

3.1. Unlicensed Radio Band

The lack of a geographically fixed infrastructure (i.e., ad hoc networks can be considered floating) necessitates

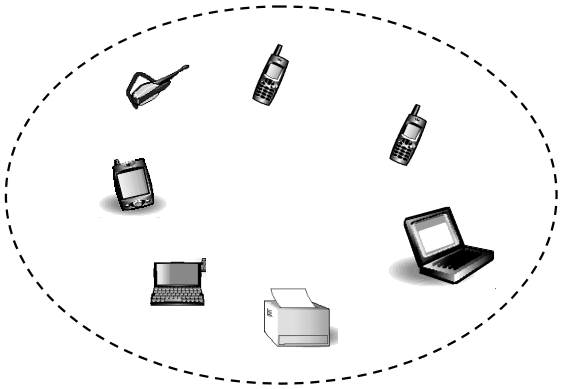


Figure 3. Single ad hoc network.

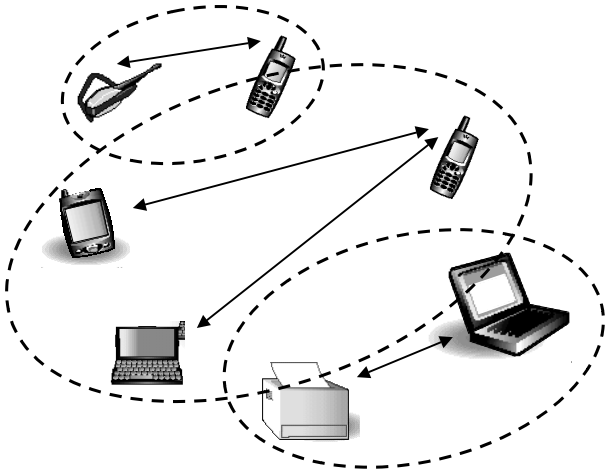


Figure 4. Scatter ad hoc network.

the deployment of a radioband that is globally available. Indeed, a user can setup an ad hoc connection anywhere in the world without the interaction of an operator or another third party. In addition, consumer-targeted applications necessitate the deployment of a radioband that is unlicensed. The most suitable band for these ad hoc applications is the industrial–medical–scientific (ISM) band ranging from 2400 to 2483.5 MHz. This band was formerly reserved for some professional user groups but has recently been opened worldwide for commercial use. The operating rules have been set by regulatory bodies such as the FCC in the United States [4], the CEPT in Europe [5], and ARIB in Japan [6]. Although the rules per region may differ slightly, their scope is to enable a fair access to the radioband by any user. The regulations generally specify the spreading of the transmitted signal energy and the maximum allowable transmit power. For a system to operate globally, a radio concept has to be found that satisfies all regulations simultaneously. The Bluetooth standard, therefore, satisfies the minimum denominator of all the requirements.

Radio propagation at 2.45 GHz provides reasonable coverage with relatively low transmit power. Current state-of-the-art radio technology allows highly integrated radio transceivers to operate with low current consumption which can be manufactured at low cost, a prerequisite for *embedded radio systems*.

3.2. Spectrum Sharing

The consequence of an unlicensed and open band is the abundance of different (radio) systems encountered in this band. Applications range from garage-door openers to microwave ovens. Also the wireless LAN systems based on 802.11 can be found in this band. The extent and nature of the interference in the 2.45-GHz ISM band cannot be predicted. With high probability, the different systems sharing the same band will not be able to communicate. Coordination is, therefore, not possible. A larger problem pose the high-power transmitters covered by the FCC part 18 rules that, for example, include microwave ovens and lighting devices. These devices fall outside the power and spreading regulations of Part 15 but still coexist in the 2.45-GHz ISM band. In addition to interference from external sources, couser interference resulting from other Bluetooth users in close proximity must be taken into account in the scatter ad hoc scenario.

Interference resistance can be obtained by interference suppression or interference avoidance. Suppression can be obtained by coding or by direct-sequence spreading. However, the dynamic range of the interfering and intended signals in a scatter ad hoc environment can be huge. Taking into account the distance ratios and the power differences of uncoordinated transmitters, near : far ratios in excess of 50 dB are no exception. With the desired user rates in the order of 1 Mbps (megabits per second) and beyond, practically attained coding and processing gains are inadequate. Instead, interference avoidance is more attractive as the desired signal is transmitted at locations in frequency and or time where interference is low or absent. Avoidance in time can be an alternative if the interference concerns a pulsed jammer and the desired signal can be interrupted. This requires coordination in time though. Avoidance in frequency is more practical. Since the 2.45-GHz band provides about 80 MHz of bandwidth and most radio transmissions are band-limited, with high probability parts of the radio spectrum can be found where there is no dominant interference. Filtering in the frequency domain provides the suppression of the interferers at other parts of the radio band. The filter suppression can easily arrive at 50 dB or more.

3.3. Frequency Hop Spread Spectrum

The selection of the multiple access scheme for ad hoc radio systems is driven by the lack of coordination and by the regulations in the ISM band. FDMA is attractive for ad hoc systems since channel orthogonality relies only on the accuracy of the crystal oscillators in the radio units. Combined with an adaptive channel allocation scheme, interference can be avoided. Unfortunately, pure FDMA does not fulfill the spreading requirements set in the ISM band by the FCC rules Part 15 [4]. TDMA requires a strict timing synchronization for channel orthogonality. For multiple collocated ad hoc connections, maintaining a common timing reference

becomes rather cumbersome. CDMA offers the best properties for ad hoc radio systems since it provides spreading and can deal with uncoordinated systems. DSCDMA is less attractive because of the *near–far* problem, which requires coordinated power control or excessive processing gain. In addition, as in TDMA, direct-sequence orthogonality requires a common timing reference. For higher user rates, DSCDMA requires rather high chip rates which are less attractive because of the wide bandwidth (interference resistance) and the higher current consumption. FHCDMA (Frequency-hopped CDMA) combines a number properties, which makes it the best choice for ad hoc radio systems. On average the signal can be spread over a wide frequency range, but instantaneously only a small bandwidth is occupied avoiding most of the potential interference in the ISM band. The hop carriers are orthogonal in frequency, and the interference on adjacent hops can effectively be suppressed by filtering. The hop sequences will not be orthogonal, though, but narrowband and couser interference is experienced as short interruptions that can be overcome with measures at higher-layer protocols. Frequency hopping was originally introduced in World War II for the remote control of torpedoes. The robustness of FH made it an ideal candidate for reliable transmission in a hostile environment. It was a Hollywood actress, Hedy Lamarr, who invented the procedure together with her pianist George Antheil [7].

Frequency hopping enables low-cost and low-power radio implementations. Since the instantaneous bandwidth is relatively narrow, conventional radio technology can be used. Also, truly single-chip solutions become feasible [8].

4. BLUETOOTH RADIO INTERFACE

4.1. Bluetooth Protocol Stack

The Bluetooth protocol stack [9] has been drafted along the OSI layered architecture but renamed (see Fig. 5). The four lower layers are the Bluetooth-specific protocols. At the RF layer, all radio-related operations are defined like modulation, frequency generation, filtering, spectral shaping, and so on. At the baseband layer, operations on the packet level are defined, such as error correction, encryption, and retransmissions. The Link Manager Protocol (LMP) takes care of control functions like authentication, setup of connections, traffic scheduling, link supervision, and power management tasks. The Logical Link Control and Adaptation Protocol (L2CAP) is an intermediate layer between the Bluetooth-specific protocols and more general protocols. It handles the multiplexing of higher-layer protocols and the segmentation and reassembly of large packets. Real-time traffic such as audio bypasses the L2CAP and LMP layers and streams into the baseband layer directly. Yet, the audio stream is controlled (non-real-time) by the LM. For more information about application oriented profiles, the user is referred to Ref. 10 or 11.

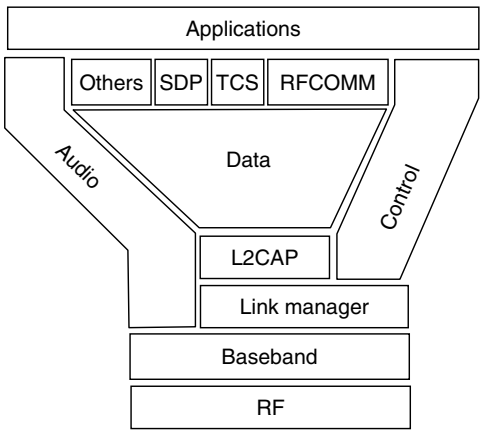


Figure 5. The Bluetooth protocol stack.

4.2. RF Layer

Bluetooth deploys FHCDMA. Each channel makes use of a different hop code or hop pattern. The radios hop over 79 carriers. These carriers have been defined at a 1-MHz spacing in the 2.45-GHz ISM band. The nominal dwell time is 625 μs. The Bluetooth radios hop with a nominal rate of 1600 hops/s. In the time domain, the channel is divided into time slots. The dwell time of 625 μs corresponds to a single slot. To simplify implementation, full-duplex communications is achieved through time-division duplex (TDD). This means that a unit alternately transmits and receives. Separation of transmission and reception in time effectively prevents crosstalk between the transmit and receive operations in the radio transceiver, which is essential if a one-chip implementation is desired. Since transmission and reception take place at different time slots, transmission and reception also take place at different hop carriers. Figure 6 illustrates the FHTDD channel applied in Bluetooth. Note that different ad hoc links will make use of different hopping sequences and will have misaligned slot timing.

The instantaneous bandwidth of the transmitted spectrum is limited to 1 MHz. For robustness, a binary modulation scheme is used. With the abovementioned bandwidth restriction, the data rates are limited to about 1 Mbps. For FH systems and support for bursty data traffic, a noncoherent detection scheme is most appropriate. Bluetooth uses a Gauss-shaped FSK modulation with a nominal modulation index of $h = 0.3$. Logical ones are sent as positive frequency deviations, logical zeros as negative

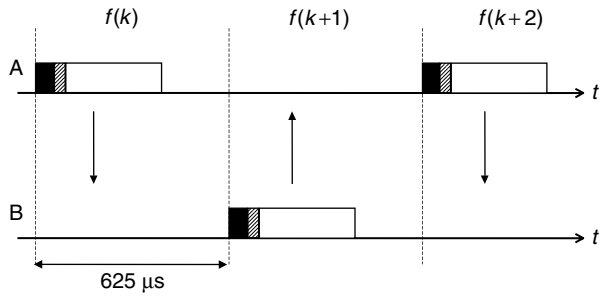


Figure 6. The slotted FH/TDD channel in Bluetooth.

frequency deviations. GFSK provides a constant envelope modulation, which is insensitive to nonlinear operations. Transmission can be achieved with a class C amplifier, while demodulation can simply be accomplished by a limiting FM discriminator. GFSK allows the implementation of low-cost and low-power radio units.

4.3. Baseband Layer

The baseband layer handles all crucial communication procedures such as connection setup, hop pattern selection and hop synchronization, traffic scheduling, and all operations at the packet level. Since most of the baseband operations occur in real time, the baseband processes have been optimized to enable dedicated hardware implementations, thus preserving power consumption. Basically, for a Bluetooth unit in idle mode, no CPU has to be activate.

4.3.1. Connection Setup. A critical design issue in ad hoc radio systems is the connection establishment. How do units find each other, and how do they make connections? In Bluetooth, three elements have been defined to support the connection establishment: *scan*, *page*, and *inquiry*. A unit that is in idle mode wants to "sleep" most of the time to save power. However, in order to allow connections to be made, the unit frequently has to listen as to whether other units want to connect. In ad hoc systems, there is no common control channel that a unit can lock to in order to listen for page messages as is common in conventional (cellular) radio systems. In Bluetooth, a unit periodically "wakes up" to listen for a page message. This page message consists of an access code which is derived from the unit's identity. When a Bluetooth receiver wakes up to scan, it opens a sliding correlator that is matched to the access code derived from its own identity. The scan window is a little longer than 10 ms. Every time the unit wakes up, it scans at a different hop carrier. This is required by the regulations that do not permit a fixed wakeup frequency; it also provides the necessary interference immunity. The Bluetooth wakeup hop sequence covers only 32 carriers and is cyclic. All 32 carriers in the wakeup sequence are unique and they span about 64 MHz of the 80 MHz available. The wakeup sequence is pseudorandom and unique for each Bluetooth device; like the access code, the sequence is derived from the unit's identity. The phase in the sequence is determined by a free-running native clock in the unit. It will be understood that a tradeoff has to be made between idle mode power consumption and the response time; increasing the sleep time T will reduce power consumption but prolong the time before an access can be made. The unit that wants to connect has to resolve the frequency–time uncertainty; it does not know when the idle unit will wake up and on what carrier frequency. The burden of resolving this uncertainty is deliberately placed at the paging unit because this will require power consumption. Since a radio unit will most of the time be in idle mode, the paging unit should take the power burden. First we assume that the paging unit knows the identity of the unit that it wants to connect to. It then knows the wakeup sequence and can also generate the access code that serves as the page message. The paging unit then transmits the access code repeatedly at different frequencies; every 1.25 ms, the paging units transmits two access codes and listens twice for a response (see Fig. 7). The access code is transmitted consecutively on different carrier frequencies selected from the wakeup sequence. In a 10-ms period, 16 different frequencies are visited, which covers half of the wakeup sequence. The paging unit transmits the access code on these 16 frequencies cyclically for the duration of the sleep period T of the idle unit. If the idle unit wakes up on any of these 16 frequencies, it will receive the access code and a connection setup procedure follows. However, since the paging unit does not know the clock that the idle unit is using, the idle unit can equally well wake up in any of the 16 remaining frequencies in the 32-hop wakeup sequence. Therefore, if the paging unit does not receive a response from the idle unit after a time corresponding to the sleep time T, it will transmit the access code repeatedly; the response time therefore amounts to twice the sleep time T. When the idle unit receives the page message, it notifies the paging unit by returning a message that again is the access code derived from the idle unit's identity. Thereafter the paging unit transmits a control packet that contains all of the pager's information (e.g., identity and clock). This information is then used by both the paging unit and the idle unit to establish a FH channel.

The above-described paging process assumed that the paging unit had no knowledge at all of the clock in the idle unit. However, if the units have met before, the paging unit will have an estimate of the clock in the idle unit. When units connect, they exchange their clock information and the time offset between their free-running native clocks is stored. This offset is only accurate during the connection; when the connection is released, the offset information becomes less reliable as a result of clock drifts. The reliability of the offset is inversely proportional to the time elapsed since the last connection. Yet, the paging unit can exploit the offset information to estimate the clock of the idle unit. Suppose, that the clock estimate of the idle unit in the paging unit is k'. If $f(m)$ represents the frequency in the wakeup sequence at time m, the paging unit will assume the idle unit will wake up in $f(k')$. But since in 10 ms it can cover 16 different frequencies, it will also transmit the access code on a few frequencies before and after $f(k')$ or $f(k'-8), f(k'-7), \ldots, f(k'), f(k'+1), \ldots, f(k'+7)$. As a result, the clock estimate in the paging unit can be off by -8 or $+7$ while it still covers the wake-up frequency of the unit in idle mode. With a free-running clock accuracy of ± 250 ppm, the clock estimate k' is still useful at least 5 h after the last connection. In that case, the average response time is reduced to half the sleep time T.

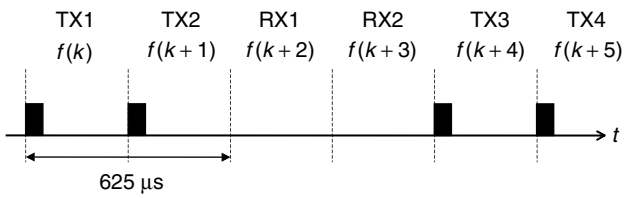


Figure 7. Transceiving routine for the paging unit.

To establish a connection, the identity of the recipient is required to determine the page message and the wakeup sequence. If the identity is not known, a unit that desires to make a connection may broadcast an inquiry message that induces recipients to return their identity and clock information. With the inquiry procedure, the inquirer can determine which units are within range and what their characteristics are. The inquiry message is again an access code, but derived from a reserved identity (the inquiry address). Idle units also listen to the inquiry message according to a 32-hop inquiry sequence. Units that receive the inquiry message return a control packet that includes their identity and clock information. For the return of the control packet a random backoff mechanism is used to prevent multiple recipients to transmit simultaneously. (In determining the frequencies of the second half of the sequence, the paging unit takes into account that the clock in the idle unit also progresses. The remaining half will therefore have one frequency in common with the first half.)

4.3.2. Hop Selection Mechanism. To allow for many overlapping hop channels (scatter ad hoc networking), a large number of pseudorandom hopping sequences have been defined. Note that no extra effort has been taken to make the sequences orthogonal. With only 79 carriers to hop over, the number of orthogonal sequences is rather limited. Bluetooth applies a special hop selection mechanism. The hop selection mechanism uses an identity and clock as inputs, and a carrier frequency as output (see Fig. 8). The identity selects a particular hop sequence while the clock points at a particular phase in this hop pattern. As the clock progresses at a rate of 1600 ticks/s, 1600 hops/s are produced by the output. By changing either the identity or the clock, instantaneously another carrier frequency is produced. The selection box contains only combinatorial logic and is memoryless. Prestored sequences are not feasible because of the large number of sequences required. The hop sequence is very long, and at a rate of 1600 hops/s, it takes more than 23 h to arrive at the exact phase in the sequence again. This feature prevents repetitions in the interference pattern when several hopping channels are collocated. Repetitive interference is detrimental for real-time services such as voice. Any segment of 32 consecutive hops in the sequence spans about 64 MHz of spectrum. By spreading as much as possible over a short time interval, maximal interference immunity is obtained. This is most important for real-time services.

We will now have a closer look at the selection scheme in Fig. 8. In the first block, the identity selects a 32-hop subsequence with pseudorandom properties. The least significant part of the clock hops through this sequence with a rate of 1600 hops/s. The first block thus provides an index in a 32-hop segment. The segments are mapped on the 79-hop carrier list. The carrier list is constructed in such a fashion that even-numbered hops are listed in the first half of the list, whereas the odd-numbered hops are listed in the second half of the list. An arbitrary segment of 32 consecutive list elements spans about 64 MHz. For the paging and inquiry procedures, the mapping of the 32-hop

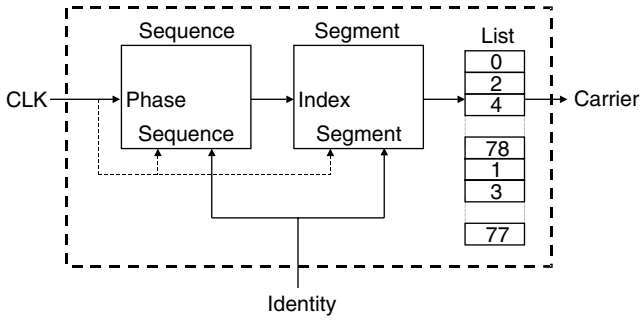


Figure 8. Carrier selection mechanism. Dashed line indicates that more significant clock part is used in connection mode only.

segment on the carrier list is fixed. When the clock runs, the same 32-hop sequence and the same 32 hop carriers will repeatedly be used. However, different identities will map to different segments and different sequences. So the wakeup hop sequences of different units are well randomized. During the connection, the more significant part of the clock affects both the sequence selection and the segment mapping; after 32 hops (one segment) the sequence is altered, and the segment is shifted in the forward direction by half its size (16 hops). Segments, each 32 hops in length, are concatenated, and the random selection of the index changes for each new segment; the segments slide through the carrier list and on average, all carriers are visited with equal probability. Changing the clock and/or identity will directly change the sequence and the segment mapping.

4.3.3. Packet-Based Communications. The Bluetooth system uses packet-based transmission; the information stream is fragmented into packets. In each TDD slot, only a single packet can be sent. All packets have the same format, starting with an access code, followed by a packet header and ending with the user payload (see Fig. 9). The access code has pseudorandom properties. All packets exchanged on the channel are preceded by the same access code. Only if the access code matches the access code corresponding to the channel will the packet be accepted by the recipient. This prevents packets sent in one FH channel falsely being accepted by units of another FH channel that happens to land on the same hop carrier. In the receiver, the access code is matched against the anticipated code in a sliding correlator. The packet header contains link control information: a 3-bit MAC address to separate the units on the channel, a 1-bit ACK/NAK for the retransmission scheme, a 4-bit packet type code to define 16 different payload types, and an 8-bit header error check (HEC) code which is a cyclic redundancy check (CRC). The packet header is limited to 18 information bits in order to restrict the overhead.

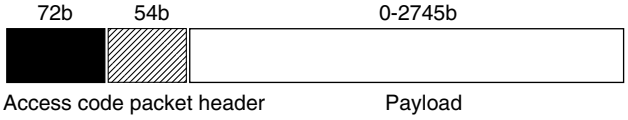


Figure 9. Format of packets exchanged on the FH/TDD channel.

The header is further protected by a rate-$\frac{1}{3}$ forward error control (FEC) coding.

Bluetooth defines four control packets:

1. *The ID or identification packet* — this packet only consists of the access code and is used for signaling.
2. *The NULL packet* — this packet only has an access code and a packet header and is used if link control information carried by the packet header has to be conveyed.
3. *The POLL packet* — this packet is similar to the NULL packet, and can be used in a polling procedure.
4. *The FHS packet* — this is a FH synchronization packet and is used to exchange real-time clock and identity information between the units. The FHS packet contains all the information to get two units hop synchronized.

The remaining 12 type codes are used to define packets for synchronous and asynchronous services. These 12 types are divided into 3 segments; segment 1 specifies packets that fit into a single slot, segment 2 specifies 3-slot packets, and segment 3 specifies 5-slot packets. Multislot packets have been introduced to increase the user data rate. They are sent on a single carrier frequency (see also Fig. 10). Note that packets can cover only an odd number of TDD slots, which guarantees that the TX/RX timing is maintained. The payload length is variable and depends on the amount of user data available. However, the maximum length is limited by the minimum switching time between RX and TX, which is specified at 200 μs. This switching time seems large, but allows the use of open-loop VCOs for direct modulation and provides time for packet processing between RX and TX.

4.3.4. Physical Links. The Bluetooth link supports both synchronous services such as voice traffic and asynchronous services such as bursty data traffic. Two physical link types have been defined:

1. The synchronous connection-oriented (SCO) link
2. The asynchronous connection-less (ACL) link

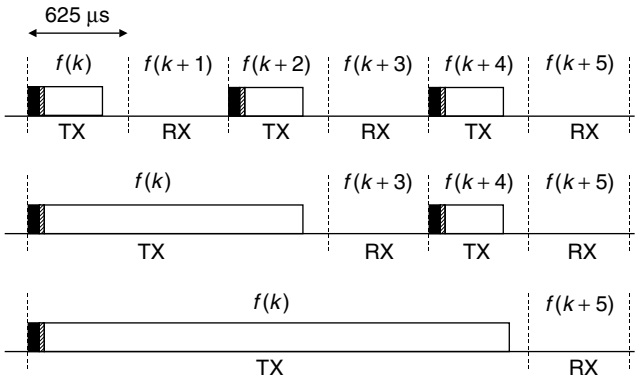


Figure 10. Characteristics of single-slot, three-slot, and five-slot packets.

The *SCO link* is a point-to-point link between two units on the FH channel. The link is established by reservation of duplex slots at regular intervals. This link can be used to carry synchronous, real-time traffic and provides quality of service. The *ACL link* is a point-to-multipoint link which is used to carry asynchronous, best-effort traffic.

Different packet types have been defined for the SCO link and the ACL link. The ACL links support payloads with or without a FEC coding scheme. In addition, on these links single-slot, 3-slot, and 5-slot packets are available. The maximum user rate that can be obtained over the ACL link is 723.2 kbps. In that case, a return link of 57.6 kbps can still be supported. If propagation conditions change, link adaptation can be applied on the ACL link by changing the packet length and the FEC coding. For the SCO link, only single-slot packets have been defined. The payload length is fixed and may choose between two different FEC schemes and no FEC. The SCO link supports a full-duplex link with a user rate of 64 kbps in both directions.

4.3.5. Error Correction Schemes. In Bluetooth, two different FEC schemes are supported. The rate-$\frac{1}{3}$ FEC code merely uses a 3-bit repeat coding with majority decisions at the recipient. With the repeat coding, extra gain is obtained because of the reduction in the instantaneous bandwidth. As a result, intersymbol interference (ISI) introduced by the receive filtering is decreased. This code is used for the packet header, and can additionally be applied on the payload of the SCO link packets. For the rate-$\frac{2}{3}$ FEC code, a (15,10) shortened Hamming code is used. Error trapping can be applied for decoding. This code can be used in the payload of both SCO link and ACL link packets. The applied FEC codes are very simple and fast to encode and decode, which is a requirement because of the limited processing time between RX and TX.

On the ACL link, an automatic retransmission query (ARQ) scheme is applied. In this scheme, a packet retransmission is carried out if the reception of the previous packet is not acknowledged. Each ACL payload contains a CRC to check for errors. To minimize complexity, overhead, and wasteful retransmissions, Bluetooth has implemented a fast-ARQ scheme where the sender is notified of the packet reception in the first packet in the return path after the transmission. The ACK/NAK information is piggybacked in the packet header of the return packet. There is only the RX/TX switching time for the recipient to determine the correctness of the received packet and creating the ACK/NAK field in the header of the return packet. In addition, the ACK/NAK field in the header of the packet received indicates whether the previously sent payload was correctly received, and thus determines whether a retransmission is required or the next packet can be sent. This process is illustrated in Fig. 11. Due to the short processing time, decoding is preferably carried out on the fly while the packet is received. The simplicity of the FEC coding schemes speed up the processing.

4.3.6. Low-Power Modes. In the Bluetooth system, special attention has been paid to the reduction of current consumption. In the idle mode, the unit scans for only about 10 ms every T seconds where T can range from 1.28

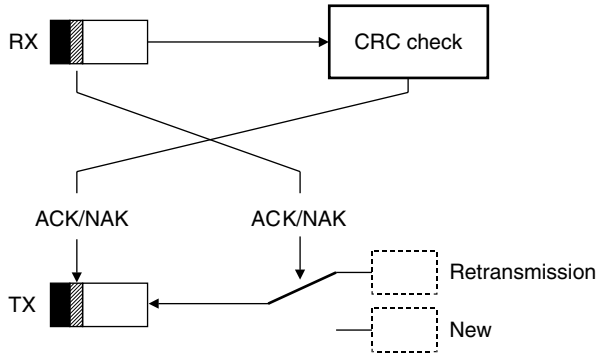


Figure 11. ARQ routine: received ACK/NAK information decides on retransmission; received payload determines returned ACK/NAK.

to 3.84s. So the duty cycle is well below 1%. Additionally, a PARK mode has been defined where the duty cycle can be reduced even more. However, the PARK mode can be applied only after the channel has been established. Units can then be parked; they listen to the channel at only a very low duty cycle. The unit only has to listen to the access code and the packet header (126 μs, excluding some guard time to account for drift) to resynchronize its clock and to decide whether it can return to sleep. Since there is no uncertainty in time and frequency (the parked unit remains locked to the channel, in much the same way as cordless and cellular phones are locked to their base stations), a much lower duty cycle is obtainable. Another low-power mode during connection is the SNIFF mode, in which a unit now and then listens to the channel, but is still considered to be active. Finally, a HOLD mode has been defined. A unit can enter the HOLD mode for a predefined time duration. During this time duration, it is not active on the channel. When the HOLD timeout expires, the unit automatically returns to the channel and is active again.

4.4. Link Manager

The link manager is involved in non-real-time, supervisory, and control operations. It takes care of attachment and detachments of units (note that real-time connection setup routines like page and scan are carried out at the baseband layer), sets up SCO and ACL links, and initiates low-power modes such as HOLD, SNIFF, and PARK. The LMP also takes care of security operations; it runs authentication procedures to prevent unauthorized usage of the connections and initiates the encryption routines. In addition, it is responsible for the key distribution. After the unit has been attached and been authenticated and logical links have been established, the link manager is involved in link supervision; it checks whether the device is still in coverage range of other units, and is responsible for selecting the proper packet type depending on the link quality and the required quality of service. Also adaptive power control is handled at the LMP level. The link manager is responsible for traffic scheduling. Finally, the link manager is used to configure (optional) parameters in the baseband layer both at connection setup and during the connection.

LMP messages are control messages and fit into the payload of a single-slot baseband packet using a rate-$\frac{2}{3}$ FEC. The LMP PDU consists of a header and a body. The header contains an 7-bit opcode specifying the LMP message, and a 1-bit transaction ID. The body may contain additional parameters used by the LMP message.

4.5. Logical Link and Adaptation Protocol

The Logical Link Control and Adaptation Protocol (L2CAP) forms an interface layer between the Bluetooth LMP layer on one hand and Bluetooth-independent higher layers on the other hand. The L2CAP handles the multiplexing of different logical links over the ACL link. It also effects segmentation and reassembly of datagram packets provided by the higher layers, for example, IP packets, onto the baseband packets. The logical links are connection-oriented with an 16-bit destination address in the L2CAP header.

5. BLUETOOTH NETWORKING

Bluetooth has been optimized to allow a large number of uncoordinated communications to take place in the same area. Unlike other ad hoc solutions where all units in the same range share the same channel, Bluetooth has been designed to allow a large number of independent channels, each channel serving only a limited number of participants. With the considered narrowband modulation scheme, a single FH channel in the ISM band only supports a gross bit rate of 1 Mbps. This capacity has to be shared among all participants on the channel. In the user scenarios targeted by Bluetooth, it is highly unlikely that all units in range need to share all information among all of them. By using a large number of independent 1-Mbps channels to which only the units are connected that really want to exchange information, the 80 MHz is exploited much more effectively.

In Bluetooth, a FH channel is designated as a *piconet*. On a single piconet, up to eight units can participate. Each piconet has its own, unique hop sequence. The particular sequence is determined by the identity of the unit that controls the FH channel, which is called the "master." The native clock of the master unit defines the phase in the hopping sequence. All other participants on the hopping channel are "slave"; they use the master identity to select the same hop sequence. Each Bluetooth radio unit has a free-running system or native clock. There is no common timing reference, but when a piconet is established, the slaves add an offset to their native clocks to synchronize to the master. These offsets are released again when the piconet is cancelled, but can be stored for later use. Different channels have different masters and therefore also different hop sequences and phases. Bluetooth is based on peer communications. The master/slave role is only attributed to a unit for the duration of the piconet. When the piconet is released, the master and slave roles are cancelled. Each unit can become a master or a slave. By definition, the unit that establishes the piconet becomes the master.

The piconet can be considered as a small network. In addition to defining the piconet, the master also controls

the traffic on the piconet and takes care of access control. The access code preceding the packets exchanged on the channel is derived from the master's identity and must match in the slave before it accepts any packet. Channel access is completely contention-free. The short dwell time of 625 μs allows the transmission of only a single packet. The master implements a centralized control in a star configuration; communication only between the master and one or more slaves is possible. The time slots are alternately used for master transmission and slave transmission. In the master transmission, the master includes the MAC address of the slave for which the information is intended. In order to prevent collisions on the channel due to multiple slave transmissions, the master applies a polling technique; for each slave-to-master slot the master decides which slave is allowed to transmit. This decision is performed at a slot-per-slot basis: only the slave addressed in the master-to-slave slot directly preceding the slave-to-master slot is allowed to transmit in the slave-to-master slot. If the master has information to send to a specific slave, this slave is polled implicitly and can return information. If the master has no information to send, it has to poll the slave explicitly with a short poll packet. Since the master schedules the traffic both in forward and return link, intelligent scheduling algorithms have to be used that take into account the slave's traffic characteristics. The master control prevents collisions among participants of the same piconet. Independent, collocated piconets may interfere when they occasionally use the same hop carrier. A type of ALOHA is applied — information is transmitted without checking for a clear carrier (listen before talk). If the information is received incorrectly, it is retransmitted at the next transmission opportunity at a different carrier frequency (for ACL links only). Because of the short dwell time, collision avoidance schemes are less appropriate for the FH radio. For each hop, different contenders are encountered. Backoff mechanisms would therefore be less efficient.

In the piconet, the master can support SCO and ACL links. The SCO links form point-to-point links between a master and a single slave. SCO links are defined by a forward and return slot pair reserved at a fixed interval. On the remaining slots, the master can address each slave via an ACL link. The traffic over the ACL link is scheduled by the master via the polling mechanism. The slotted structure of the piconet channel allows an effective mixing of the synchronous (SCO) and asynchronous (ACL) links. However, it is a centralized control and all traffic has to flow via the master. An example of a channel with SCO and ACL links is illustrated in Fig. 12.

A multiple of piconets colocated in the same area is indicated as a *scatternet*. These piconets all operate independently, each controlled by its own master. Because Bluetooth uses packet-based communication over slotted links, it is possible to interconnect different piconets. This means that units can participate in different piconets. However, since the radio can only tune to a single-carrier frequency at one time, at any instant in time a unit can communicate in one piconet only. But a unit can jump from one piconet to another piconet by adjusting the piconet channel parameters (i.e., the master identity and the master clock), thus applying time-division multiplexing (TDM) to be virtually present in several piconets. A unit can also change role when jumping from one piconet to another piconet. For example, a unit can be master in one piconet at one instant in time, and be a slave in a different piconet at another instant in time. A unit can also be slave in different piconets. However, by definition, a unit cannot be master in different piconets, since the master parameters specify the piconet FH channel. The hop selection mechanism has been designed

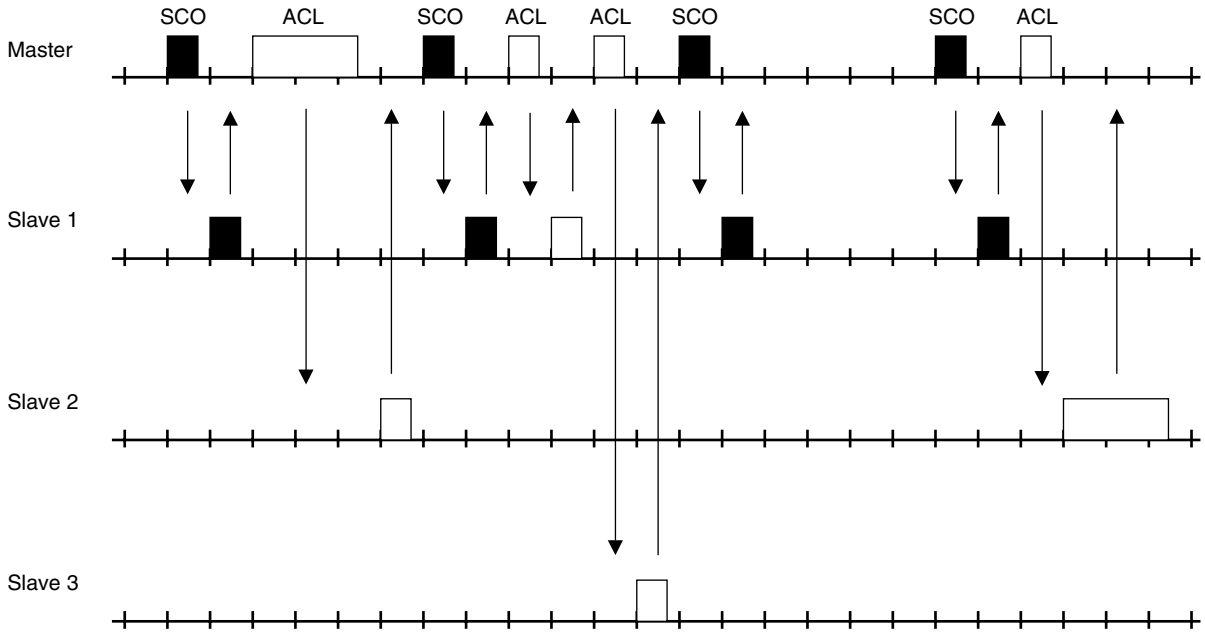


Figure 12. Master traffic scheduling and mixing SCO and ACL links.

to allow for interpiconet communications; by changing the identity and clock input to the selection mechanism, instantaneously a new frequency for the new piconet is selected. In order to make the jumps between the different piconets feasible, guard time has to be included in the traffic scheduling to account for the slot misalignment of the different piconets. In Bluetooth, the low-power modes HOLD and SNIFF can be applied to temporary leave one piconet and visit another piconet. Since all piconets are hopping independently without coordination of hopping or timing, traffic scheduling and routing in a scatternet with interpiconet communications becomes a challenge (12).

6. SECURITY

Although Bluetooth is mainly intended for short-range connectivity between personal devices, some basic security elements are included to prevent unauthorized usage and eavesdropping. At connection establishment, an authentication process is carried out to verify the identities of the units involved. The authentication process uses a conventional challenge–response routine. The claimant transmits its claimed 48-bit identity to the verifier. The verifier returns a challenge in the form of a 128-bit random number. This random number, the claimant address, and a 128-bit common secret key form the inputs to a computational secure hash function based on SAFER+ that produces a 32-bit signed response. The signed response produced by the claimant is sent to the verifier which compares this result with its own signed response. Only if the two calculated responses are the same will the challenger continue with the connection establishment. The authentication can be uni- or bidirectional.

In addition to authentication, encryption is applied. Although the pseudorandomness of the FH channel gives some protection against a casual eavesdropper, it provides no privacy in the cryptographic sense. Therefore, the payload of each packet is encrypted. Encryption is based on stream ciphering; the payload bits are modulo-2 added to a binary key stream. The binary key stream is generated by a second hash function that is based on linear feedback shift registers (LFSRs). When encryption is enabled, the master sends a random number to the slave. Before the transmission of each packet, the LFSR is initialized by a combination of this random number, the master identity, an encryption key, and the slot number. Since the slot number changes for each new packet, the initialization is new for each packet.

The central element in the security process is the 128-bit key. This key is a secret key residing in the Bluetooth hardware and is not accessible by the user. The secret key is generated during an initialization phase, also called *pairing*. Two units that want to authenticate each other and establish secure links have to be paired; that is, they have to be provided with the same secret key. An initialization phase initiated by the user is required to pair two devices. To authorize pairing, the user has to enter an identical PIN in both devices. For devices without a user interface (e.g., headsets), initialization is possible only during a short time window, such as after the user has pressed an initialization key. Once the pairing has been carried out, the 128-bit key resides in the devices and can be used for automatic authentication without user interaction [13].

Bluetooth provides a limited number security elements at the lowest level. More advanced security procedures (public keys, certificates to name only a few) can be implemented at higher layers, but are not part of the protocol.

7. CONCLUSION

The Bluetooth technology is a new radio technology for providing short-range connectivity between personal devices. FHCDMA is applied to allow many independent connections to coexist within the same area, efficiently sharing the unlicensed ISM band at 2.45 GHz. The FH channel forms a piconet between a master and one or more slaves. The hopping and the traffic scheduling in this FH channel is controlled by the master. Key elements in the Bluetooth technology are low cost, low power, and small size, enabling single-chip radio transceivers to be embedded in millions of personal devices in the near future.

BIOGRAPHY

Jaap C. Haartsen received his M.S. and Ph.D. degrees with honors in electrical engineering from the Delft University of Technology, the Netherlands, in 1986 and 1990, respectively. He joined Ericsson in 1991 and has worked in the area of wireless technology at Ericsson sites in the United States, Sweden, and the Netherlands. In Sweden, he laid the foundations for the Bluetooth radio concept. He played an active role in the creation of the Bluetooth Special Interest group and served as chair for the SIG air protocol specification group from 1998 till 2000. In April 2001, he became chief scientist of Ericsson Technology Licensing AB, an Ericsson company, fully dedicated to Bluetooth IP. In May 2000, he was appointed as adjunct professor at the Twente University of Technology, the Netherlands, in the area of mobile radio communications. He has authored numerous papers and holds over 40 patents in the area of wireless communications. His areas of interest are wireless system architectures, radio technology, ad-hoc radio communications, and short-range communications.

BIBLIOGRAPHY

1. D. J. Goodman, *Wireless Personal Communications Systems*, Addison-Wesley, Reading, MA, 1997.
2. ETSI Radio Equipment and Systems (RES), *Digital European Cordless Telecommunications (DECT), Common interface Part 1: Overview*, ETS 300 175-1, 1996.
3. R. van Nee et al., New high-rate wireless LAN standards, *IEEE Commun. Mag.* **37**: 82–88 (1999).
4. Federal Communications Commission, CFR47, *Part 15—Radio Frequency Devices*, 1999.
5. ETSI Radio Equipment and Systems (RES), *Wideband Data Transmission Systems: Technical Characteristics and Test*

Conditions for Data Transmission Equipment Operating in the 2.4 GHZ ISM Band and using Spread Spectrum Modulation Techniques, ETS 300 328, 1994.

6. ARIB Std., *Low Power Data Communication/Wireless LAN System*, RTC STD-33, 1998.
7. H. Howe Jr., A starlet's secret life as inventor, *Microwave J.* **42**: 70–74 (1999).
8. J. C. Haartsen and S. Mattisson, Bluetooth—A new low-power radio interface providing short-range connectivity, *Proc. IEEE* **88**: 1651–1661 (2000).
9. *Specification of the Bluetooth System*, Version 1.1, Bluetooth Special Interest Group, *www.bluetooth.com*.
10. B. A. Miller and C. Bisdikian, *Bluetooth Revealed*, Prentice-Hall, Upper Saddle River, NJ, 2001.
11. N. J. Muller, *Bluetooth Demystified*, McGraw-Hill, New York, 2001.
12. M. Frodigh, P. Johansson, and P. Larsson, Wireless ad-hoc networking—the art of networking without a network, *Ericsson Rev.* **77**: 248–263 (2000).
13. J. Persson and B. Smeets, Bluetooth security—an overview, *Information Security Technical Report*, Elsevier Advanced Technology, 2000, Vol. 5, pp. 32–43.

BROADBAND WIRELESS ACCESS

HIKMET SARI
Juniper Networks
Paris, France

1. INTRODUCTION

The telecommunications network has undergone major improvements to fulfill the ever-increasing need of end users for higher data rates. The multigigabit routers and optical transmission lines, which are now installed in the core and edge networks, have tremendously increased the data rates that can be transmitted. The speed bottleneck, which sets a limit on the services that can be offered to the end users, is the access network that connects them to the edge and core networks. The best-known access network is the twisted-pair copper cable, which serves virtually all homes and businesses. These cables were traditionally used to carry voice services and low-speed data communications using voiceband modems. They are now used to offer digital subscriber line (DSL) services, which are available in different forms. High-speed DSL (HDSL) uses two or three twisted pairs to offer symmetric 2-Mbps (megabits per second) data services, while the more recently developed asymmetric DSL (ADSL) technology offers a 6–8-Mbps data rate downstream and several hundreds of kbit/s upstream. Higher data rates will be offered in the near future using very high-speed DSL (VDSL) as the fiber nodes get closer to the end users and twisted-pair portion of the traditional telephone network gets shorter and shorter.

Similarly, coaxial cable networks were traditionally used for broadcast TV services only. Since the access network was opened to competition in the early 1990s, cable operators upgraded their cable plants and turned them into bidirectional networks in order to offer high-speed data services to the subscribers. Numerous cable operators today offer a variety of services using either proprietary technologies or industry standards like the data-over-cable system interface specification (DOCSIS). Upstream transmission in cable networks employs the frequency spectrum of 5–42 MHz in the DOCSIS standard and 5–65 MHz in its European version known as Euro-DOCSIS.

In countries with a well-developed telecommunications infrastructure, there has been little need in the past for fixed wireless access. This type of systems were essentially deployed in developing countries with a large population that is not served by the twisted-pair telephone cable. Those wireless access systems, however, are narrowband, and can only carry telephony and low bit-rate data services. The emergence of broadband wireless access (BWA) is very closely related to the more recent deregulation of the world telecommunications market. This deregulation has created a new environment in which new operators can compete with incumbent operators that often are former state-owned monopolies. Wireless access networks are very appealing to new operators without an existing wired infrastructure, because they not only are rapidly deployed but also involve a low initial investment, which is determined by the initial customer base. Once in place, wireless networks are easily upgraded to accommodate additional subscribers as the customer base grows. This is a very attractive feature with respect to wired networks where most of the investment needs to be made during the initial deployment phase.

Most frequencies available for BWA are at millimeter-wave frequencies between 20 and 45 GHz. Dedicated frequency bands for these applications have recently become available in Europe, North America, Asia Pacific, and other regions. After an extensive field trial period, BWA systems operating at millimeter-wave frequencies are currently in the field, but their commercial deployment remains small scale. These cellular radio networks, which are commonly referred to as *local multipoint distribution service* (LMDS) networks, are intended to offer integrated broadband services to residential and business customers. LMDS networks are particularly suited for urban or suburban areas with high user density, because the cell capacity is typically in the range of the STM-1 data rate (155 Mbps) and the cell coverage is only 2–5 km. This characteristic makes them attractive for business customers, whereas DSL and cable access systems are essentially for residential subscribers.

Although not as much as in the millimeter-wave frequency range, there are also some frequency bands below 11 GHz for BWA applications. These include the 2.5-GHz microwave multipoint distribution service (MMDS) band in the United States, the 3.5-GHz band in Europe, and the 10-GHz band in a limited number of countries. Below 11 GHz, there are also some license-exempt frequency bands that may be used for wireless access. These frequency bands are not specific to BWA and will not be specifically covered here.

The article gives a state-of-the-art review of broadband wireless access systems and discusses the current trends

and ongoing standardization work. First, in the next section, we give a brief introduction to millimeter-wave BWA (LMDS) networks. In Section 3, we present an analysis of the intercell interference that limits the frequency reuse in these networks. Next, in Section 4, we outline the current standardization work by the IEEE 802.16 and the ETSI BRAN groups for next generation BWA systems at millimeter-wave bands. Finally, in Section 5, we discuss BWA at microwave frequencies below 11 GHz, which is essentially focused on residential applications. Some conclusions are given in Section 6.

2. AN OVERVIEW OF LMDS NETWORKS

LMDS was originally used to designate the 28-GHz band in the United States, but it is often used today to designate BWA systems operating at all millimeter-wave frequencies above 20 GHz. LMDS networks are cellular, each cell serving a number of fixed subscribers located in its coverage area, which has a radius of 2–5 km. The base station (BS) is connected to the backbone network through a backhaul point-to-point link, which can be a fiber or a radio link.

The network topology resembles that of mobile cellular radio systems, but LMDS systems have several distinctive features. The main of those is that since users are at fixed locations, each user is assigned to a predetermined BS (typically the nearest BS). Furthermore, fixed wireless access systems employ narrowbeam directional subscriber antennas pointed to the serving BS during installation. The increased gain in the direction of the BS reduces network interference and increases cell coverage.

Another major difference concerns propagation. While mobile radio systems are subjected to shadowing and severe multipath propagation, LMDS systems are based on clear line-of-sight (LoS) between the BS and the fixed users, and are virtually free of multipath propagation. Signal attenuation during normal propagation conditions is proportional to the square of the distance, and what truly limits the cell coverage is "rain fading," which further attenuates the transmitted signal by several dB or several tens of dB per kilometer. Because of this phenomenon and the limited power that can be generated at low cost at millimeter-wave frequencies, the cell radius in LMDS networks is in the range of 2–5 km depending on the climatic zone, the available transmit power, and the required availability objectives.

Although LMDS systems can be based on hexagonal cell patterns that are commonly used in mobile radio systems [1], rectangular cell patterns with 90° cell sectoring have become very popular in LMDS network design [2,3] and will be considered throughout this article. Each sector in this cell pattern is served by a 90° sector antenna, and different channels are assigned to the different sectors. The channel bandwidth differs from region to region; In Europe and other countries that follow the CEPT channeling, the channel spacing is of the form $112/2^n$ MHz, where n is an integer. The typical channel spacing for BWA in this region is 28 MHz, unless the operator does not have sufficient bandwidth allocation, in which case 14 or 7 MHz channels are used. In North America, the channel bandwidth is of the form $80/2^n$ MHz, with a typical bandwidth of 20 MHz.

With a simple quaternary phase shift keying (QPSK) modulation, a 28-MHz CEPT channel is sufficient to transmit a useful data rate of 16×2 Mbps. The total bit rate per cell is then 64×2 Mbps and can be used to serve for example 64 business customers with a 2-Mbps leased line each. This example is only to give an idea of the cell capacity. The number of subscribers per cell may be several hundreds or several thousands, and such a large number of users can still be accommodated by dynamically sharing the available resources between them. Since some users may be idle while some users are requesting high instantaneous data rates, there is a substantial statistical multiplexing gain, which makes it possible to accommodate a large number of subscribers.

3. NETWORK INTERFERENCE ISSUES

Because of the lack of industry standards, first-generation BWA systems are based on proprietary solutions. In fact, technical specifications for LMDS systems were developed by the Digital Video Broadcasting (DVB) Project [4] and the Digital AudioVisual Council (DAVIC) [5] in the 1994–1996 time period, but these specifications were primarily intended for digital TV broadcasting and two-way communications with low interactivity. Most proprietary systems today use pieces of the DVB/DAVIC standards, but they do not follow these standards completely. Also, most of them, as well as the DVB and DAVIC specifications, are based on time-division multiplexing (TDM) on the downstream channel (from BS to subscribers), time-division multiple access (TDMA) on the upstream channel (from subscribers to BS), and frequency-division duplexing (FDD). This means that separate channels are used for downstream and upstream transmissions.

Since bandwidth is a limited and costly resource, the frequency reuse factor and the achievable cell capacity are crucial to the deployment of LMDS networks. These factors are strongly impacted by intercell interference. In this article, we discuss intercell interference, assuming that the network is based on rectangular cell pattern with 90° sectoring and that a separate channel is assigned to each sector, which means that four channels are used to cover each cell. But the same channels are reused in all cells as shown in Fig. 1. Note that channel assignment between neighboring cells follows a mirror-image rule in the horizontal, vertical, and diagonal directions.

The BSs, which are designated by heavy dots in Fig. 1, are located on a rectangular grid. The solid lines represent the sector borders, and the dotted lines indicate the cell boundaries. The labels A, B, C, and D represent the channels used in different sectors. As it is indicated in Refs. 2 and 3, the mirror-image assignment of these channels eliminates interference between adjacent cells.

However, the second-nearest cells have the same channel assignment as the cell at hand and interfere with it. Let 2Δ designate the distance between two adjacent BSs in the horizontal and vertical directions. Now suppose that a user is located on the border of two sectors at a distance

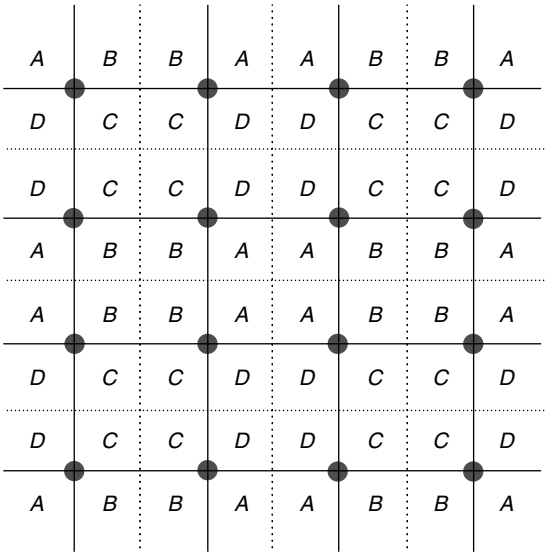


Figure 1. Rectangular cell pattern with 90° sectors.

δ from the serving BS. This user's antenna will also be pointed toward a second-nearest BS that is at a distance $4\Delta + \delta$. Assuming that all BS's transmit the same signal power and that the signal attenuation is proportional to the squared distance, which is a common assumption in line-of-sight microwave and millimeter-wave radio systems, the downstream signal-to-interference ratio (SIR) for this user is

$$\mathrm{SIR(dB)} = 20 \cdot \log\left(\frac{4\Delta + \delta}{\delta}\right) \tag{1}$$

This expression, which is valid for $0 < \delta \leq \Delta$, achieves its minimum value for $\delta = \Delta$. The corresponding SIR is 14 dB. In writing (1), we have assumed that BSs further than the second-nearest BS are not in clear LoS with the user of interest; specifically, their signals are blocked by buildings, trees, or other obstacles.

As it is shown in Ref. 2, the worst-case SIR of 14 dB is also valid for the upstream channel when automatic transmit power control is used. But the similarity of downstream and upstream channels in terms of interference is limited to the value of the worst-case SIR. On the downstream channel, the SIR is a function of the user position, and only in a very small part of the cell, the users are subjected to strong interference. Using a common subscriber antenna radiation diagram with a beamwidth of 5°, we have plotted in Fig. 2 the SIR distribution within a sector, where the BS is located in its upper left corner. Specifically, the figure shows the boundaries of the regions corresponding to an SIR higher than a given value. Notice that only in very small regions located at the other 3 corners, the SIR is lower than 15 dB. Furthermore, the SIR is higher than 30 dB in virtually half of the cell.

Figure 2 indicates that if the system design requires an SIR value higher than 15 dB, three small regions will not be covered. Coverage will be even smaller if the system design requires an SIR higher than 20 or 25 dB. This means that a bandwidth-efficient modulation scheme that

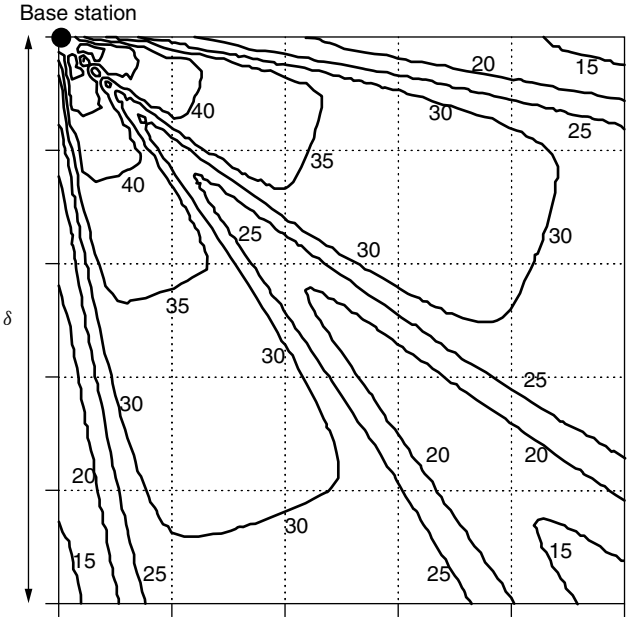


Figure 2. SIR distribution within a sector corresponding to a subscriber antenna beamwidth of 5°.

requires a high SIR value will not be usable if full cell coverage is required. But the figure also suggests that while users at unfavorable positions (regions of low SIR values) must use a low-level modulation scheme such as QPSK, users in more favorable locations can use higher-level quadrature amplitude modulation (QAM) schemes such as 16-QAM or 64-QAM, at least during normal propagation conditions. This adaptive modulation and coding concept is now used in international standards and will be discussed in Section 4.

The situation is quite different on the upstream channel, because in this direction, all users get the same amount of interference; that is, the SIR is not a function of the user position. Consequently, the user-dependent modulation concept makes little sense for the upstream channel, but the adaptive modulation concept can still be used to adapt the modulation to propagation conditions.

4. CURRENT STANDARDIZATION

While first-generation LMDS systems are today in the field, standardization activities are now at a very advanced stage at both the IEEE and the ETSI to define technical specifications for future BWA systems. The groups that are carrying out this work for millimeter-wave frequency bands are the IEEE 802.16.1 task group and the HIPERACCESS group of ETSI BRAN. Specification work by both groups covers the physical (PHY) layer and the medium access control (MAC) layer functions. At the time of this writing, the IEEE 802.16 group has already issued its draft technical specifications [6] for BWA systems at frequencies between 11 and 60 GHz. As for the HIPERACCESS group of the ETSI, which started its specification work later than the IEEE 802.16.1 task group, it intends to complete its specifications by mid-2002.

There is a significant level of commonality between the IEEE 802.16.1 and the ETSI BRAN HIPERACCESS (draft) specifications concerning basic choices for PHY layer functions. This includes the following [7]:

- The transmission technique is based on single-carrier transmission. The reason for this is that BWA systems at millimeter-wave frequencies suffer very little multipath propagation because of the small cell size and directive subscriber antennas used. This does not give much motivation for using orthogonal frequency-division multiplexing (OFDM) which is appealing for strong intersymbol interference (ISI) channels [8]. In addition, the strong sensitivity of OFDM to oscillator phase noise and transmit power amplifier nonlinearity makes this technique rather undesirable for systems operating at millimeter-wave frequencies, where high transmit power and low phase noise incur substantial cost for the outdoor radio unit.
- As in the earlier DVB/DAVIC specifications, TDM and TDMA have been adopted for the downstream channel and the upstream channel, respectively. This choice can be justified by the relative maturity of TDMA with respect to code-division multiple access (CDMA) that has been adopted in third-generation digital mobile radio standards [9].
- To increase cell capacity with respect to pure QPSK, the IEEE and ETSI specifications include adaptive modulation and coding. The idea is to use the most bandwidth-efficient modulation and coding schemes that are compatible with the signal-to-noise ratio (SNR) and the interference level affecting user signals. This is a function of the user position on the one hand (on the downstream channel), and the instantaneous fade level on the other hand. The candidate modulation schemes are 4-QAM (QPSK), 16-QAM, and 64-QAM for the downstream channel, and 4-QAM and 16-QAM for the upstream channel. To have an adaptation with a finer granularity in terms of signal-to-interference plus noise ratio (SINR), both specifications also allow adaptively changing the coding scheme.

Adaptive modulation and coding substantially increase the cell capacity for a given level of quality of service. Assuming that the SIR required is 12 dB for QPSK, 19 dB for 16-QAM, and 25 dB for 64-QAM, and using a subscriber antenna beamwidth of 6°, it was shown [10] that an adaptive modulation that combines these three signal formats on the downstream channel achieves an increase of cell capacity by a factor of 2.7 with respect to QPSK. Since all users are subjected to the same level of interference on the upstream channel, it was proposed [10] that the channel be split in two parts and each subchannel be assigned to a specific region of the sector of interest. This assignment can be done in such a way that the level of interference is significantly reduced for some subscribers. Using this subchanneling concept along with an adaptive modulation involving the QPSK and the 16-QAM signal formats, a capacity improvement by a factor of 1.4 was achieved on the upstream channel. These results indicate that adaptive modulation substantially increases the cell capacity, although to a lesser extent on the upstream channel.

One way to increase capacity on the upstream channel is to use an adaptive antenna at the BS. Indeed, if the BS employs a steered narrowbeam antenna, only the users near the sector borders in the horizontal and vertical directions and those near the diagonal will be subjected to strong upstream interference, and the situation becomes similar to that on the downstream channel. Users located outside these regions will be subjected to a smaller level of interference and can use a 16- or 64-QAM modulation. The upstream cell capacity then becomes similar to that of the downstream channel. One difficulty in applying this concept is that adaptive antenna technology is not yet mature for microwave and millimeter-wave frequencies. Adaptive antennas appear as an option in current standards, for future evolutions.

5. BWA AT LOWER FREQUENCY BANDS

Both the IEEE 802.16 Group and ETSI BRAN first put a priority on the definition of system specifications for BWA systems operating at millimeter-wave frequencies, but they later turned their attention toward licensed frequency bands between 2 and 11 GHz. The respective task groups of the IEEE and the ETSI that are in charge of defining technical specifications for BWA at frequencies below 11 GHz are the IEEE 802.16.3 task group and the HIPERMAN group of ETSI BRAN, respectively. The IEEE 802.16.3 group is already at an advanced technical specifications phase, and the ETSI HIPERMAN group has recently completed the functional requirements phase and entered the technical specifications phase.

In many aspects, BWA at lower frequencies is quite similar to LMDS, but it also has two basic distinctive features. The first concerns the traffic model. Whereas LMDS systems are essentially intended for small-business applications, frequencies below 11 GHz are primarily for residential subscribers where the major application is high-speed Internet access. The implication of this is that traffic at lower frequency bands is highly asymmetric, and most of the traffic is on the downlink from the BS to subscribers. This feature has a strong impact on both the PHY layer and the data-link (DLC) layer. The second distinctive feature is that due to larger cell sizes, smaller subscriber antenna directivity, and non-LoS propagation, lower-frequency bands are subjected to multipath propagation (and a significant level of ISI), which must be compensated.

One solution for BWA at lower microwave frequencies is to use a single-carrier transmission technology as for millimeter-waves. The only additional requirement in this case is to use an adaptive equalizer that is capable of handling the multipath propagation encountered in this kind of network. Another solution consists of using the OFDM technology, which has been adopted in the IEEE 802.11a and ETSI HIPERLAN/2 specifications for wireless local area networks (wireless LANs) at 5 GHz [11,12].

At the time of this writing, The ETSI BRAN group is still examining proposals and has not made a final decision regarding the technology to use for fixed BWA systems at lower microwave frequencies, but the IEEE 802.16.3 task group of the IEEE has already made major decisions and released a baseline document for the physical (PHY) layer. Failing to agree on a single standard, this group decided to include both an OFDM-based PHY and a single-carrier PHY layer specifications. Furthermore, the OFDM-based PHY comprises an OFDM/TDMA mode and an orthogonal frequency-division multiple access (OFDMA) [13] mode, which means that the forthcoming IEEE 802.16a actually include three different transmission and multiple access technologies. In the following subsections, we will briefly describe them.

5.1. Single-Carrier Transmission

To operate on channels that suffer from strong multipath propagation, single-carrier systems must use an adaptive equalizer. When OFDM was first proposed for the Digital Audio Broadcasting (DAB) and Digital Video Broadcasting (DVB) in Europe in the late 1980s and the early 1990s, it was assumed that single-carrier transmission does not give adequate performance on difficult radio channels, particularly for mobile reception. In the 1993–1995 time period, a series of articles were published by the present author [e.g., 8] suggesting that the common perception that single-carrier transmission does not give adequate performance on difficult radio channels is a result of constraining them to use a time-domain equalizer. After making the observation that single-carrier systems with a time-domain equalizer have an inherent limitation due to convergence and tracking problems when the number of taps is large, it was next suggested that a single-carrier system with frequency-domain equalization (SC/FDE) closely resembles an OFDM system while avoiding its well-known problems:

1. Its high peak-to-average power ratio (PAPR), which makes it very sensitive to the transmit high-power amplifier (HPA) nonlinearity
2. Its high sensitivity to the local oscillator phase noise

A schematic block diagram of the basic transmit and receive functions in OFDM and SC-FDE is given in Fig. 3. As can be seen in this figure, there is a strong resemblance between an OFDM system and an SC/FDE system. The frequency-domain equalizer in the latter system gives it the possibility to compensate for long channel impulse responses without facing the convergence problems that are inherent to single-carrier systems with time-domain equalization (SC/TDE). Indeed, under the minimum mean-square error (MMSE) criterion, the optimum coefficients of a linear transversal time-domain equalizer are the solution of the matrix equation

$$C = A^{-1}V \tag{2}$$

where A is the autocorrelation matrix of the input signal vector X_k, and V is the cross-correlation of the input signal vector X with the transmitted symbol a_k [14]. The conventional least mean squares algorithm for coefficient adaptation at time k is

$$C_{k+1} = C_k - \alpha X_k^* e_k \tag{3}$$

where α is the step-size parameter that controls convergence, and e_k is the equalizer output error at time k [14]. Without any mathematics, it can easily be seen that the equalizer coefficients do not converge independently of each other, because their adaptation is driven by the same error signal e_k and also the components of the vector X_k are not independent.

Now consider a frequency-domain equalizer with N taps. The DFT operator that forms the first stage of the equalizer gives N signal samples denoted $(Y_1, Y_2, \ldots, Y_N)$. These samples are the inputs to a complex multiplier bank whose coefficients are denoted $(D_1, D_2, \ldots, D_N)$. The coefficient values which minimize signal distortion are given by

$$D_n = \frac{H_n^*}{|H_n|^2} \tag{4}$$

where H_n denotes the channel transfer function at frequency f_n. A better criterion is the MMSE criterion, which minimizes the combined effect of channel distortion and additive noise. The optimum coefficients in the MMSE sense are

$$D_n = \frac{H_n^*}{|H_n|^2 + \gamma} \tag{5}$$

where γ is the inverse of the signal-to-noise ratio (SNR). Clearly, the optimum coefficients of a frequency-domain equalizer are independent of each other, and therefore convergence occurs at speeds much higher than is possible in time-domain equalizers. The consequence of this is that a frequency-domain equalizer can employ a large number of taps and compensate for long impulse response channels while converging fast and tracking rapid channel variations.

An important feature of the SC/FDE system concept proposed [8] is that it employs a cyclic prefix (similar to OFDM) so as to make the linear convolution of the channel look like the circular convolution performed by the frequency-domain equalizer. The articles published by this author in which SC/FDE was shown to be an attractive alternative to OFDM stimulated further research on the subject and led to the rebirth of frequency-domain equalization which had long been ignored in the literature.

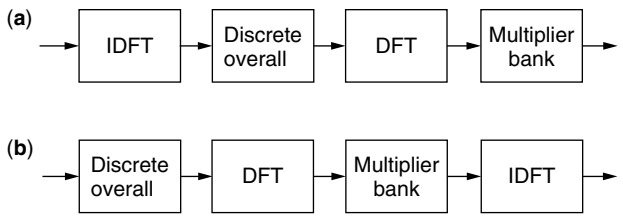


Figure 3. Transmit and receive block diagram in OFDM (**a**) and SC-FDE (**b**).

5.2. OFDM

The basic idea in OFDM is to split the channel bandwidth into a large number of subchannels such that the channel frequency response is essentially flat over the individual subchannels. This is performed using an inverse discrete Fourier transform (DFT) at the transmitter and a forward DFT at the receiver. More specifically, the transmitter of an OFDM system with N carriers includes a serial-to-parallel (S/P) converter, an inverse DFT operator of size N, and a parallel-to-serial (P/S) converter which serializes the DFT output before sending it to subsequent filtering and frequency upconversion stages.

The way OFDM compensates for frequency-selective fading is substantially different from the way single-carrier transmission handles this phenomenon. Since the N symbols of each DFT block are transmitted at different frequencies and the individual subchannels are very narrow, the symbols transmitted at faded frequencies (located on a deep notch of the channel frequency response) cannot be detected reliably. OFDM systems must therefore resort to channel coding in order to protect the symbols transmitted at faded frequencies, whereas single-carrier systems can operate on frequency-selective channels without channel coding. Operation of OFDM systems is best explained using a simple example. Suppose that the channel impulse response has a strong attenuation at K frequencies. Then, the K symbols per DFT block transmitted at these frequencies will be in error with a large probability. A block code whose length is equal to the DFT block length and error correction capability exceeds K symbols will correct these errors, and the resulting OFDM system will be efficient on that channel. OFDM systems can also use convolutional coding. In that case, the code must have a large Hamming distance (the minimum length of error events) and an interleaver must be included in order to distribute the effect of fading on transmitted symbols.

A cyclic prefix is inserted between OFDM symbols at the transmitter so that the linear convolution of the channel becomes a circular convolution for the transmitted symbols. This requires that the cyclic prefix be larger than the channel impulse response length. The two important parameters of an OFDM system are the number of carriers and the length of the cyclic prefix. The prefix represents overhead, and its length is dictated by the maximum length of the channel impulse response to be compensated. In order to limit the loss in throughput due to the cyclic prefix, the DFT block length (the number of carriers) must be increased. But increasing the number of carriers increases complexity on the one hand and the sensitivity to timing variations of the channel on the other hand. In the IEEE 802.16a specifications, the number of carriers is 256, and the prefix can have up to 64 samples (a quarter of an OFDM symbol length).

5.3. OFDMA

OFDM transmission on multiple access channels is often used with TDMA, and the resulting combination is referred to as OFDM/TDMA. (This is the case in the IEEE 802.11a and HIPERLAN/2 standards.) In this scheme, the base station assigns time slots to different users, and the signal transmitted within a time slot is an OFDM signal. For convenience, a time slot is an integer multiple of an OFDM symbol.

The third transmission mode included in the IEEE 802.16a specifications is OFDMA [13]. In this technique, the N symbols per DFT block are not all assigned to the same user, but instead they are partitioned into M subsets of N/M symbols, and resource assignment is performed subset by subset. This means that resources can be allocated to M users during the same OFDM symbol period.

OFDMA has several interesting features with respect to OFDM/TDMA. First, it reduces the granularity of the bursts allocated to different users thereby increasing the efficiency of the MAC protocol. Next, it increases the cell range in the upstream direction by concentrating the power available from the transmit amplifier on a subset of carriers. (Every division by a factor of 2 of the number of carriers used per subscriber is equivalent to increasing the transmit amplifier power by 3 dB.) This means that an OFDMA system based on splitting the total number of carriers N by 16 and allocating a single subset to users will increase the cell coverage by as much as 12 dB. The cell range can also be increased in the downstream direction by allocating a transmit power to each set of carriers that is function of the distance to the user to which this set is allocated.

6. CONCLUSIONS

Its ease of deployment and the low initial investment involved makes BWA the most attractive broadband access technology for new operators without an existing infrastructure. Millimeter-wave BWA (LMDS) is mostly suited for business subscribers in high-density urban or suburban areas, and BWA at lower microwave frequencies is essentially suited for residential subscribers. After briefly discussing the cell capacity, frequency planning, and interference issues in current LMDS networks based on proprietary technologies, this article summarizes the current status of standardization work by the ETSI BRAN and the IEEE 802.16 Groups for both millimeterwave frequencies and lower microwave frequencies between 2 and 11 GHz. Because of the LoS propagation that characterizes this type of networks, standards for BWA at millimeterwave frequencies are based on single-carrier transmission. But BWA below 11 GHz is subjected to strong multipath propagation, and the IEEE 802.16a standard for this type of networks has three different modes: SC/FDE, OFDM, and OFDMA.

BIOGRAPHY

Hikmet Sari received his Diploma (M.S.) and Ph.D. in telecommunications engineering from the ENST, Paris, France, and the *Habilitation* degree from the University of Paris XI, France. From 1980 to 2000 he held research and management positions at the Philips Research Laboratories, SAT, and Alcatel Paris, France. In May 2000, he joined Pacific Broadband Communications (PBC)

as chief scientist. He is now with Juniper Networks, which acquired PBC in December 2001. He has published over 130 technical papers and holds over 25 patents. In 1995, he was elevated to the IEEE fellow grade and received the Andre Blondel Medal from the SEE (France). He was an editor of the IEEE Transactions on Communications from 1987 to 1991, a guest editor of the European Transactions on Telecommunications (ETT) in 1993, and a guest editor of the *IEEE JSAC* in 1999. Presently, he is an associate editor of the *IEEE Communications Letters* and a distinguished lecturer of the IEEE Communications Society.

BIBLIOGRAPHY

1. T. S. Rappaport, *Wireless Communications: Principles and Practice*, IEEE Press, New York, and Prentice-Hall, Englewood Cliffs, NJ, 1996.
2. H. Sari, Broadband radio access to homes and businesses: MMDS and LMDS, *Comput. Networks* **31**: 379–393 (Feb. 1999).
3. G. LaBelle, LMDS: A broadband wireless interactive access system at 28 GHz, in M. Luise and S. Pupolin, eds., *Broadband Wireless Communication*, Springer-Verlag, Berlin, 1998, pp. 364–377.
4. ETS 300 748, *Digital Video Broadcasting* (*DVB*)*: Framing Structure, Channel Coding, and Modulation for MVDS at 10 GHz and above*, ETSI, October 1996.
5. DAVIC 1.1 Specifications, Part 8, *Lower-Layer Protocols and Physical Interfaces*, Revision 3.3, Geneva, September 1996.
6. *Air Interface for Fixed Broadband Wireless Access Systems*, IEEE 802.16.3 task group, Sept. 2000.
7. ETSI Website: *www.etsi.org*
8. H. Sari, G. Karam, and I. Jeanclaude, Transmission techniques for digital terrestrial TV broad-casting, *IEEE Commun. Mag.* **33**: 100–109 (Feb. 1995).
9. F. Adachi, M. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communications systems, *IEEE Commun. Mag.* **36**(9): 56–69 (Sept. 1998).
10. J. P. Balech and H. Sari, Advanced modulation techniques for broadband wireless access systems, *Proc. 7th Eur. Conf. Fixed Radio Systems and Networks* (ECRR 2000), Dresden, Germany, Sept. 2000, pp. 159–164.
11. P802.11a/D6.0, *LAN/MAN Specific Requirements*, Part 2: *Wireless MAC and PHY Specifications — High Speed Physical Layer in the 5 GHz Band*, IEEE 802.11, May 1999.
12. DTS/BRAN030003-1, *Broadband Radio Access Networks HIPERLAN Type 2 Functional Specification*, Part 1: *Physical Layer*, ETSI, Sophia Antipolis, Sept. 1999.
13. H. Sari and G. Karam, Orthogonal frequency-division multiple access and its application to CATV networks, *Eur. Trans. Telecommun. Related Technol.* (*ETT*) **9**(6): 507–516 (Nov.–Dec. 1998).
14. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

C

CABLE MODEMS

DONALD G. MCMULLIN
Broadcom Corporation
Irvine, California

1. INTRODUCTION

Since the inception of the Internet as a high-speed data connection between universities in the early 1970s, the search for a low-cost broadband last-mile delivery system has been pursued. The fiberoptic backbone is capable of sustaining terabits of data throughput, but the last mile connection has traditionally been limited to, at best, about 28 kb (kilobits) and more recently 56 kbits. Realizing that the needed bandwidth for these high-speed data links could be supplied by the television cable plant, in the mid 1970s the FCC mandated that all new cable television trunk lines and drop lines be installed as two-way-ready. Two-way amplifiers were installed allowing both downstream and upstream data traffic to occupy selected frequency spectra on a single coaxial cable. New head-end cable equipment was installed, and the cable operators began to deploy broadband Internet access over the cable infrastructure (Fig. 1a,b). In the event that a cable plant had not been upgraded for two-way operation, the Telco modem (or cable downstream and telephone upstream) has been successfully deployed. Typical bandwidth usage models require a broadband downstream channel, since users nominally request large amounts of data from the Internet server. The return path (or upstream) bandwidth can be reduced, since users rarely transmit large amounts of data upstream. In fact, the limited upstream traffic has allowed for further bandwidth efficiency by utilization of a time-division multiple access (TDMA) scheme for two-way cable modem implementation. This method allows multiple users to transmit data on the same IF carrier frequency, but at different times. This is known as *burst-mode transmission*, and is contrasted to the subscriber modem receiver downstream data, which are supplied as a continuous bitstream. Exceptions to this limited upstream bandwidth are applications requiring two-way videoconferencing, and these have just recently (at the time of writing) been addressed in new specifications.

In 1995 efforts were made by the newly formed Multimedia Cable Network Systems (MCNS) organization and the IEEE 802.14 committee to define and establish standards for transmission of IP data over existing cable lines. Both of these bodies eventually dissolved into what is known today as the Data over Cable Service Interface Specification (DOCSIS) standard. The lower four layers of the data protocol are primarily what DOCSIS 1.0/1.1 defines [1] and are outlined as follows:

Layer 1 — PHY (physical layer): defines upstream and downstream modulation schemes, 64/256-QAM downstream and QPSK/16-QAM upstream

Layer 2 — MPEG2: defines the data packet organization and FEC (forward error correction) codes

Layer 3 — MAC (media access control): defines the data processing protocols between cable modem (CM) at the customer premise, and the head-end (HE) equipment, also known as the cable modem termination system (CMTS) residing at the central office

Layer 4 — BPI (Baseline PrIvacy): sets the key codes for encryption to provide security on the shared cable network

The structure of the downstream payload data has a unique packet ID (PID), service ID (SID), and destination address (DA) embedded in the data packets. The PID is used to identify a "data type" packet as opposed to digital video information. The downstream SID identifies the security association of each packet and the DA identifies packets that belong to a particular user. Packets are framed in standard MPEG-2 format. This allows the data channels to occupy the already defined digital video channel spacing and decoder technology. MPEG-2 defines what is specified as "well known" packet identifiers, and for cable modem data traffic this hex value is 0x1FFE.

Thus, as the packet parser contained in the cable modem MAC looks at each PID inserted in each MEG packet received, it will proceed to the next level of decoding of the SID only if it finds a PID indicating that this is a data channel. If there are no payload data (actual data to receive), then a "null packet" will be transmitted consisting of the hexadecimal (hex) value 0xFF for all payload data bytes, enabling the downstream to remain locked to the QAM channel and decoding MPEG packets at all times. A brief description of the MPEG-2 packet structure will be presented later in this article. A block diagram of the cable modem is shown in Fig. 2 and will be discussed in detail later in this article.

For both upstream and downstream data to coexist on a single cable, a means to separate the frequency spectra is necessary. For the North American standard, the downstream data services reside with the already established downstream video channels occupying the 54–860-MHz band (using 6-MHz channel spacing). The upstream data are placed in the unused frequency bands from 5 to 42 MHz (Fig. 3). A diplex filter is used to mitigate crosstalk between the respective frequency allocations. The diplex filter consists of a HI-PASS section for the downstream channels and a LO-PASS section for the upstream channels. As mentioned earlier, upstream return channels are burst mode and symbol rates are assigned during the logon process on the basis of requested/available bandwidth. Thus, the head end can allocate bandwidth in accordance with the demands as more users logon and require more channel capacity. This results in a slow degradation in system performance, in contrast to a telephone modem, whereby when no more switch ports are available, the user cannot establish a connection at all. Additionally, when

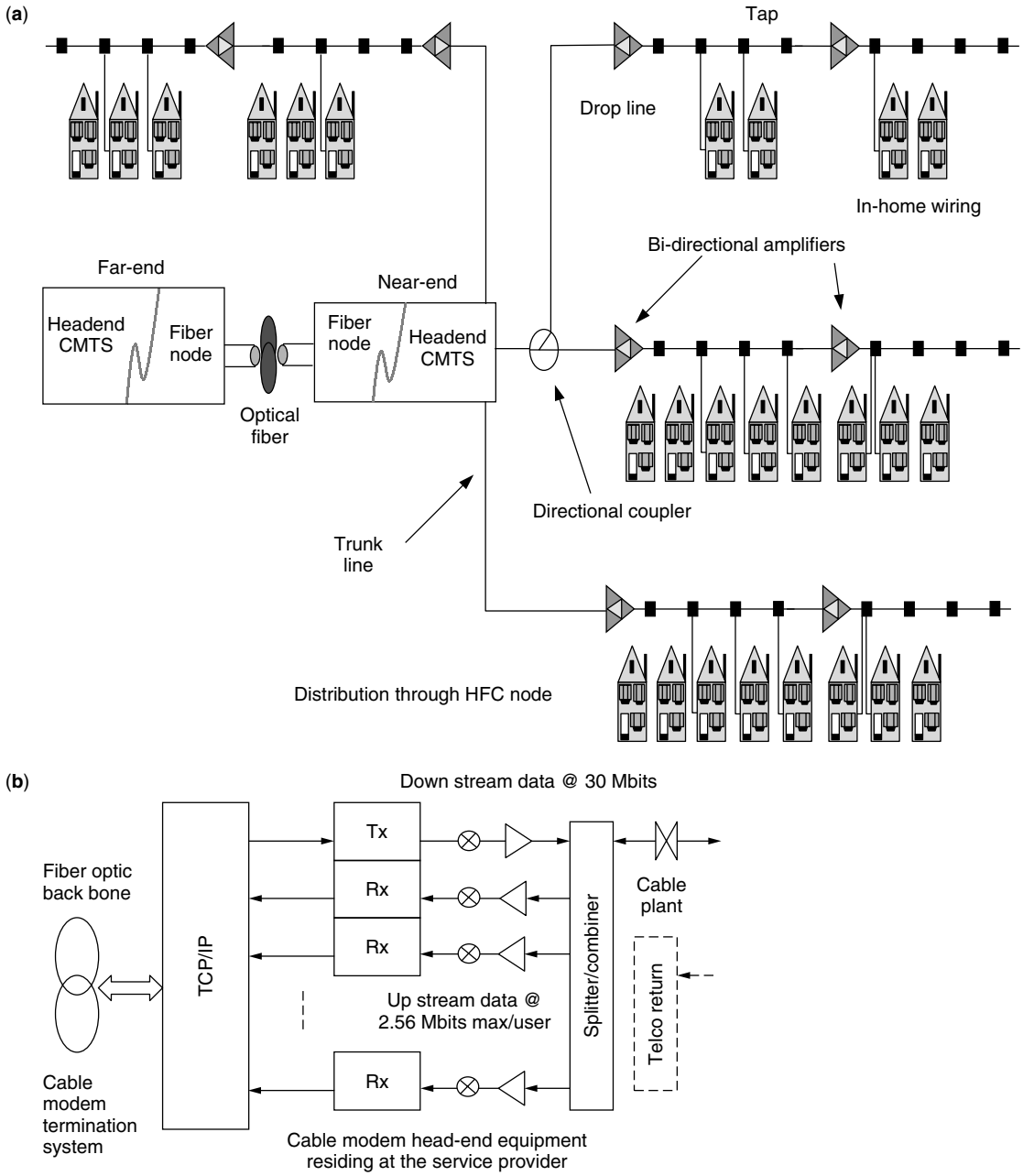


Figure 1. Block diagrams of (**a**) HFC plant and (**b**) CMTS.

the cable system bandwidth reaches an unacceptable level of performance, a new RF downstream or IF upstream frequency is assigned to some of the users, and the system data throughput can be restored dynamically without disruption of service or any knowledge by the users. Typical loading currently is about 200 or so users per downstream channel and optimum channel loading has been established by historical usage models for telephone lines.

2. PHYSICAL-LAYER: MODULATION FORMAT

Why use quadrature amplitude modulation (QAM) for transmission and reception? The search for a compact efficient means to transmit and receive data has led to the implementation of QAM for cable applications. The simplest form of QAM is called quadrature phase shift keying (QPSK) and, as the term implies, this form of modulation takes advantage of the orthogonal nature of an I (in phase) and Q (quadrature — 90° phase shift) coordinate system. As has been shown by Euler, Parseval, and others, an orthogonal system allows the encoding of two distinct and independent sets of information that can be combined, transmitted, and demodulated without interaction or distortion to each other. This allows a second degree of freedom, raising to the power of two the capacity of any transmission system. In its simplest form, QPSK has been used for satellite communications for many years, dating back to the early 1950s. Only until relatively

recent advances in semiconductor technology and system integration have higher-order QAM modulation formats become commonplace. Still, in a high-noise environment a constant vector magnitude modulation scheme such as QPSK or 8-PSK is far preferred. For this type of constellation, all symbols (or data points) lie on a circle; thus the magnitude is constant and it is only necessary to detect the phase difference between each transmitted symbol to complete the demodulation process. This facilitates robust reception of data even in environments with high levels of noise.

The transmission medium drives the choice of modulation scheme and for a "wired" or cable system, the design constraints are very different from those in a wireless system. The fundamental specification driving the choice of modulation is the obtainable system SNR, and secondary to this are the expected multipath reflections, which can be stationary (cable) or time-varying (wireless). The well-known Shannon–Hartley capacity theorem [2] states

$$C = W \log_2 \left(1 + \frac{S}{N}\right)$$

where C = system capacity
S = signal power
W = system bandwidth
N = noise power

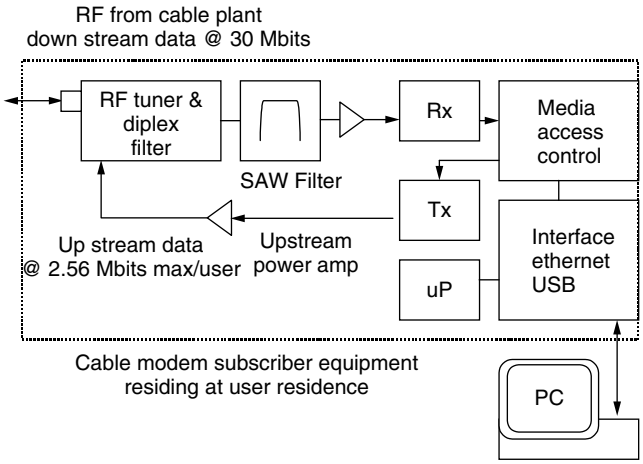


Figure 2. Block diagram of CM.

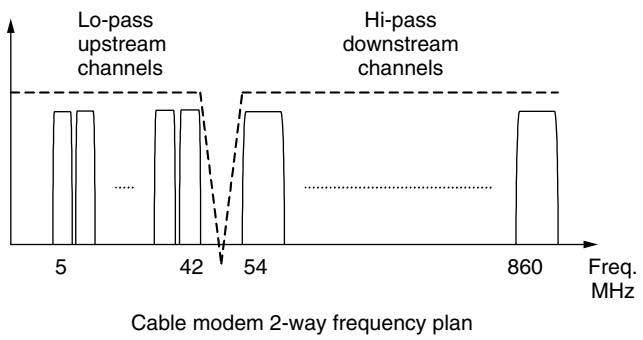


Figure 3. Cable modem frequency plan.

Rearrangement and normalizing yields the channel capacity C/W in bits per second per hertz, which defines the maximum number of bits per symbol that can be transmitted for a given SNR (Fig. 4). Although the Shannon–Hartley equation does not explicitly set a limit for the error probability, achievable SNR has a large effect in determining the QAM receiver bit error rate (BER). This, in turn, will dictate the reliability and absolute data rate of the communications link. NTSC analog video requires an SNR of ~50 dB (peak signal voltage/Rms noise voltage) and linearity, HD2 and HD3, need to be suppressed below −60 dBc. Cable plants with achievable SNRs of 40 dB (RMS power/RMS noise) allow for feasible deployment of modulation orders as high as 256-QAM (8 bits per symbol). For a symbol rate of 5 Mbaud, this would correspond to a bit rate of 40 Mbps (megabits per second). Today most North American cable operators routinely deploy downstream QAM orders of 64-QAM at approximately 5 Mbaud for a downstream data rate of 30 Mbps.

3. PHYSICAL LAYER DOWNSTREAM: RF AND AFE REQUIREMENTS

The downstream RF front end for a cable modem begins with a TV tuner designed to downconvert and filter the unwanted adjacent channels of the RF frequency spectrum of 54–860 MHz to an IF frequency of 43.75 MHz for NTSC systems or 36.125 MHz for European PAL systems. Cable service providers usually place the digital video and data channels together above 400 MHz, although they can reside at any frequency in the spectrum mentioned earlier. Traditional TV tuners (so-called single-conversion tuners) take the RF input and mix it with an LO (local oscillator) offset by the IF frequency to place the incoming RF signal at the desired IF (43.75 MHz in the NTSC case). Since these tuners were designed for off-air reception, the LO leakage back into the antenna was of little concern because of the limited range of radiation. However, this LO leakage was found to be of great concern when the RF was connected to a cable plant. The normal worst case would be for all 200 modem users to tune to the same RF channel, and since these LOs are at the same frequency but not correlated, they would add in an root sum of squares (RSS) sense at a frequency offset of 43.75 MHz

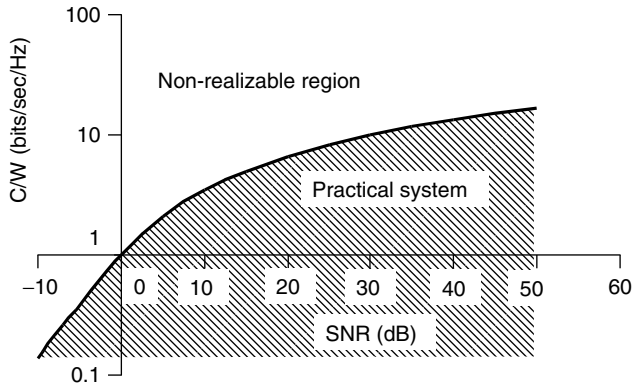


Figure 4. Channel capacity.

from the tuned RF channel. This would result in significant LO power leaking to the adjacent RF channel, which might be in use by other viewers of analog television or digital video/data, which could result in corruption of that signal. To circumvent the LO leakage problem, designers have increased the reverse isolation of the low-noise amplifier (LNA) at the very front end of the tuner or switched to what is commonly called a "double-conversion tuner architecture." Basic operation of a double-conversion tuner is to first upconvert the entire spectrum to a much higher frequency (typically 1–2 GHz), process the channel selection using a filter and then downconvert the selected channel back to the required 43.75-MHz IF. In this way, the frequency plan places the mixer LOs out of the desired frequency spectrum and thus mitigating the LO leakage issue altogether. As with any RF system, the first-stage NF (noise figure) will dominate the obtainable receiver system SNR. For RF input levels of −15 to +15 dBmV, as specified by DOCSIS, the tuner must have a NF of 10 dB or less to meet this specification with a reasonable margin of 2.5 dB above the FEC limit (Fig. 5). In addition to the noise and input level, the QAM receiver carrier recovery loop is sensitive to the phase noise contribution of the tuner, and typical values of −85 dBc at a 10-kHz offset are required. Single conversion tuners have been accepted into cable plant operations due to the low phase noise, low cost, and good NF. However, more recent advances in silicon tuner technology have introduced inexpensive double-conversion tuners in standard bulk CMOS process technology, which promise the possibility of integration onto the cable modem chip, thus further simplifying the cable modem design.

Following the downconversion to the 43.75-MHz IF, a surface acoustic wave (SAW) bandpass filter is placed in the signal path to eliminate any residual power from the adjacent channels that may not have been attenuated by the tuner and also to band-limit the noise. Stop band attenuation, passband ripple and group delay variation are all important design constraints for this filter. Virtually all QAM receivers employ an adaptive equalizer as part of the digital processing, and this can relax some the SAW filter requirements. Equalizers are unable to compensate any signal that is not correlated to the input such as AWGN. However, any symbol-spaced equalizer can correct for correlated linear distortions produced by the cable link or RF/AFE/ADC as long as they occur within the equalizer length time interval and the equalizer has sufficient dynamic range to compensate for them. This is a very powerful result, and we will see how this will affect the design requirements of the ADC and the AFE. In a typical QAM system, symbol rates of 5 Mbaud equate to symbol periods of 200 ns, and equalizer dynamic ranges of ≥10 dB allow the use of inexpensive SAW filters, which typically have passband ripple as large as 2 dB and the group delays of ≥50 nS. The equalizer will correct for these distortions, which create intersymbol interference (ISI), resulting in degraded QAM performance, and will attempt to produce a flat channel response. Generally, these inexpensive SAW filters have significant insertion loss, as much as 20 dB, and thus a fixed-gain amplifier is needed to compensate for this attenuation. The driving specification for this amplifier is low noise. Any distortion that produces ISI will again be compensated by the equalizer. An automatic gain control (AGC) amplifier is included in the tuner RF front end and is closed around this IF amplifier; thus gain drift will be automatically compensated. These two facts allow for consideration of a low-cost open-loop design for the IF amplifier; however, a significant gain is required on the order of 34–36 dB. In current implementations, this amplifier drives the input of the single-chip cable modem QAM receiver directly. An internal analog programmable gain amplifier (PGA) provides additional AGC range and the ADC samples the 43.75-MHz IF in a subsample mode; thus sample rate less than the input IF frequency usually in the range of 20–30 megasamples per second. Since the input frequency is slewing faster than the sample clock, aperture jitter of the sample hold is an important consideration.

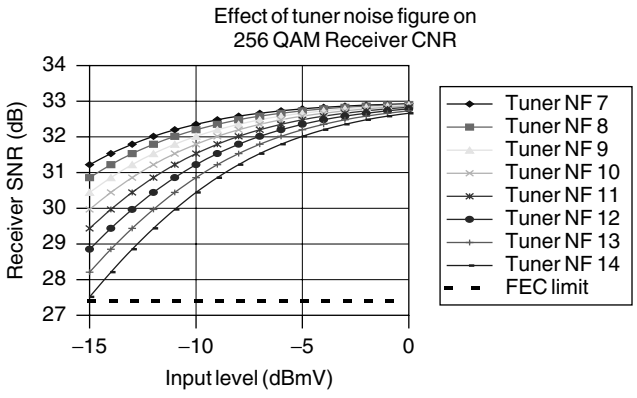


Figure 5. SNR versus input level for 256-QAM.

Traditional specifications for the ADC refer to the effective number of bits (ENOB) as the figure of merit. But for communication systems, more information is needed to optimize QAM receiver performance and reduce ADC complexity. ENOB is a metric of the combined SNR and distortion components (linearity) of an ADC. As has been shown in much communication literature, the SNR for a given ADC sampling at sub-Nyquist (nonoversampled) can be found from the following equation:

$$\text{SNR} = 6.02 \times N + 1.76\,\text{dB}$$

where N is the number of bits for the ADC.

Thus, for a 10-bit ADC, the maximum theoretical SNR that can be expected is 61.96 dB. Typical values of 59–60 dB SNR (peak signal/RMS noise) are common place for integrated ADCs in a bulk CMOS process. For a QAM receiver application, the obtainable SNR is the most important component of the ADC performance; however, the required distortion is the more interesting ADC performance parameter and must be broken into the two constituent parts [integral nonlinearity (INL) and differential nonlinearity (DNL)]. DNL can be described as the worst-case code-to-code variation in an ADC. What this means is that, as the input signal transverses the quantization levels of the ADC that contains DNL, there will be instantaneous "spikes" or distortions at the code

transitions that are not linear. Since a QAM signal has already been randomized (a scrambler is requisite for reliable transmission), these instantaneous code errors are pseudo-non-correlated and will produce a "white Gaussian noise–like" distribution. In the case of INL it is quite a different matter. INL can be described as the average nonlinear component composed of all the cells in the ADC. Generally, as the input signal swings from minimum to maximum full scale, this type of distortion will produce a second-, third-, or higher-order correlated component, and for a QAM-modulated signal the result can introduce ISI. As we will see later in this article, the ISI component can be compensated by an adaptive equalizer, and thus the DNL contributes more to degrade the performance of an ADC in a QAM system than does the INL, as mentioned earlier. An additional ADC parameter that will have an effect on receiver performance, especially when operated in the subsample mode, is the aperture jitter of the sample hold. Aperture jitter is the instantaneous amplitude error produced by imperfect sample time periods of the ADC. Typical methods for specifying this value have been based on single-tone analysis and assuming that zero crossing of a SIN wave is the worst-case slew, using the slope here for the worst-case measurement. Observing QAM constellations currently in use, it can be seen that there are no symbol decisions at zero crossing, and thus this measurement is far too pessimistic for these types of systems. The proper method for measurement of this parameter is to look at the slope of the slew on an eye diagram at the symbol decision points as shown in Fig. 6, and calculate the worst-case aperture error in this region.

Additional spreading of integral nonlinearities and distortion can be found by examination of the QAM signal itself. As mentioned earlier, in order to transmit and receive a QAM-modulated signal in a robust manner, a guaranteed level of randomization must exist in the signal; that is, the distribution of the modulated symbols must be equally likely. This is accomplished by inserting a pseudo-random bit sequence (PRBS) logically xORed (exclusive-ORed) with the data on the transmitter side, and an identical PRBS pattern on the receiver side to reverse the process. The effect this has on the original signal, and more importantly the distortion products, is to distribute them over the symbol rate bandwidth which makes them appear as additive noise components (Figs. 7 and 8). This is a very important result for a QAM system and again illustrates how important the noise contribution is to system performance and how integral distortion products are actually spread and less significant. Of particular importance related to distortion are the intermodulation distortion (IMD) products because these will appear as sidebands and can behave as adjacent-channel interference to the desired signal of the receiver. In a real HFC cable plant, many impairments inhibit the maximum SNR and BER that can be achieved and must be minimized or compensated in order to maintain a reliable data link. Figures 9–13 serve to illustrate the types of impairments that will be encountered in a typical cable environment and the effect that these imperfections will have on the receiver performance.

After the input signal has been properly downconverted and sampled by the ADC, the QAM receiver performs a complex (real/imaginary) digital down conversion to baseband (DC) for carrier recovery and symbol rate conversion. The most common form of quadrature direct digital frequency synthesis (QDDFS) is performed by using separate SIN and COS lookup tables (stored in read-only-memory) and what is known as a numerically controlled oscillator (NCO) as a mixer LO. The digitally sampled data from the ADC are split into two paths, and each set of data is input to separate mixers using the appropriate SIN (Q component) and COS (I component) driving the LO. At this time, a slight frequency offset can be added to each of these LO's to remove any gross systematic carrier frequency offset induced from components in the RF/AFE or any source that contributes a constant frequency offset. Following the quadrature downconversion, baseband processing begins with lowpass filtering to remove the image created from this conversion and then Nyquist matched filtering and timing recovery. It is well known in communications theory that in order to maximize the AWGN performance and minimize ISI, a set of identically matched filters at the transmitter and receiver must be used. The necessity of this filter is to ensure that the slew-rate-dependent properties (both amplitude and phase) of the modulation and transmission link are compensated. The convolution of the transmit and receive filters, ignoring the transmission link for the moment, gives the overall system response, which has the Nyquist property of zero ISI at symbol-spaced sampling instants (as seen in an eye diagram). Each matched filter is thus referred to as "square-root Nyquist." In addition, finite excess bandwidth must be provided beyond the ideal filter responses. The excess bandwidth damps the time-domain response of the filter and reduces sensitivity to timing recovery errors. Since it takes a finite amount of time to transverse from one symbol to the next, an additional amount of system bandwidth is required to ensure that each symbol can be transmitted and decoded properly. The most common filter mask (frequency response) for the matched filter is referred to as the *square-root raised-cosine filter*, and the mathematical form of the impulse response can be found from the following equation:

$$g(t) = \left[\frac{\sin(\pi t/T)}{(\pi t/T)}\right]\left[\frac{\cos(\alpha\pi t/T)}{(1-4(\alpha t/T)^2)}\right]$$

where α is excess bandwidth and T is 1/symbol rate. Note that when $\alpha = 0$, representing 0 excess bandwidth, this equation collapses to the following familiar form:

$$g(t) = \frac{\sin(\pi t/T)}{(\pi t/T)} = \text{sinc}\left(\frac{\pi t}{T}\right)$$

After the I (in-phase) and Q (quadrature) samples have been filtered, they are passed to the timing recovery loop, sometimes called the *baud / symbol loop*. The simplest form of timing recovery uses an I and Q zero-crossing detector to drive a simple integrator that controls a variable oscillator, and thus can lock and track any instantaneous changes in the input data timing. Generally this loop can be modeled as a phase-locked loop (PLL) with a second-order loop filter (integral and linear terms) and including an additional constant offset term equal to the desired symbol rate

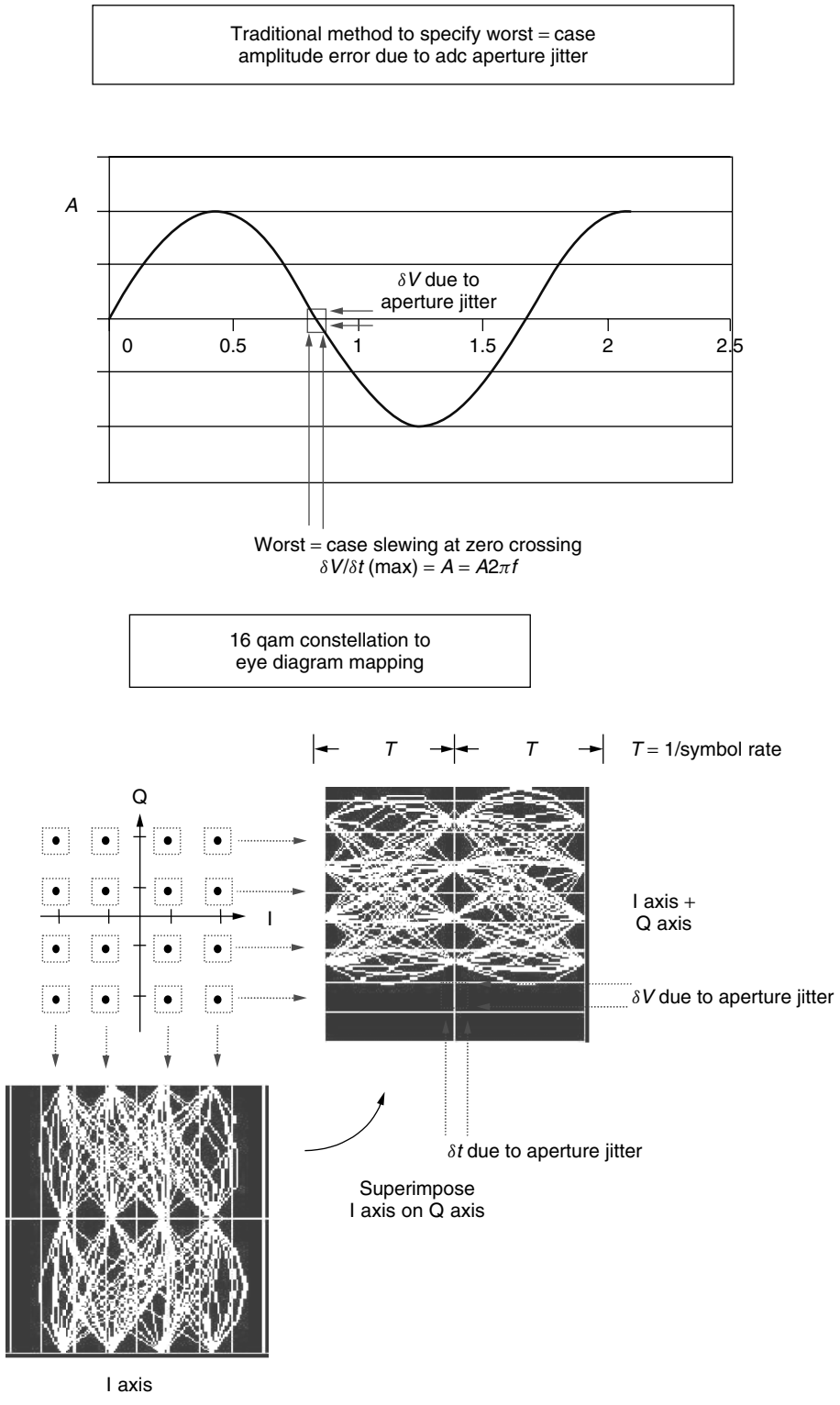


Figure 6. Aperture jitter in ADC applications for QAM.

of the receiver. If a difference term were included, this would form the popular proportional, integral, differential (PID) controller found in many modern control systems. Once the symbol timing has been recovered, the basic I/Q constellation will be formed but will need to be rotationally stabilized. This process is completed using the derotator or carrier recovery loop. Since the exact location of each ideal constellation point is known for a given QAM constellation, and the I/Q information from the receiver has been decoded, it is a relatively simple task to compute the phase difference from the received points and add it back in to compensate any rotational errors which have been introduced. Again this is done using a second order PLL structure similar to the timing recovery loop. An adaptive

equalizer, usually consisting of feedforward (FFE) and decision feedback (DFE) taps and implemented using a least-mean square (LMS) algorithm compensates for channel distortions and coax cable multipath reflections (Fig. 14). In many cases the output of this equalizer is used as an input to the carrier recovery loop, thus providing a corrected soft decision to drive the convergence of that loop, which dramatically improves the performance under impaired channel conditions.

At this point the QAM demodulation is complete and all that is left is to slice and demap the constellation points (soft decisions) back into a bitstream, derandomize and decode the forward error correction (FEC) blocks, and resolve the MPEG framing. A concatenated FEC consisting of inner trellis-coded modulation followed by an outer Reed–Solomon code are specified by DOCSIS. Also specified are various packet interleaving options that distribute any clustered errors over a number of packets, providing immunity to burst noise. MPEG-2 framing defines each packet to consist of 188 bytes with the first byte (or sync byte) to be the hex value 0x47. In addition, every 8th packet shall have the sync byte inverted (hex value 0xB8) in order to facilitate acquisition and lock retention of the packet stream.

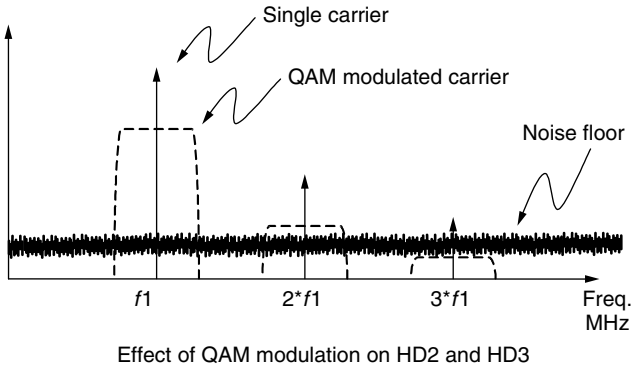


Figure 7. Distortion in a QAM modulation system.

4. PHYSICAL LAYER UPSTREAM: IF REQUIREMENTS

The upstream burst modulator consists of digital I/Q data stream that utilizes quadrature direct digital frequency synthesis (QDDFS) to upconvert to the desired IF frequency for transmission from the CM back to the head end (CMTS) and the Internet server. A high-speed DAC (with sample rate typically $\geq$200 MHz) is used to convert this digital IF to an analog voltage. The modulation format is either QPSK or 16-QAM for all current modems, with 64-QAM already defined in the next generation of the DOCSIS 2.0 specification. Since the upstream data are transmitted in a burst mode, a means for packet synchronization is necessary. This is accomplished by the addition of a preamble at the beginning of each packet, which allows the receiver (residing at the head-end equipment) to synchronize before the actual payload data

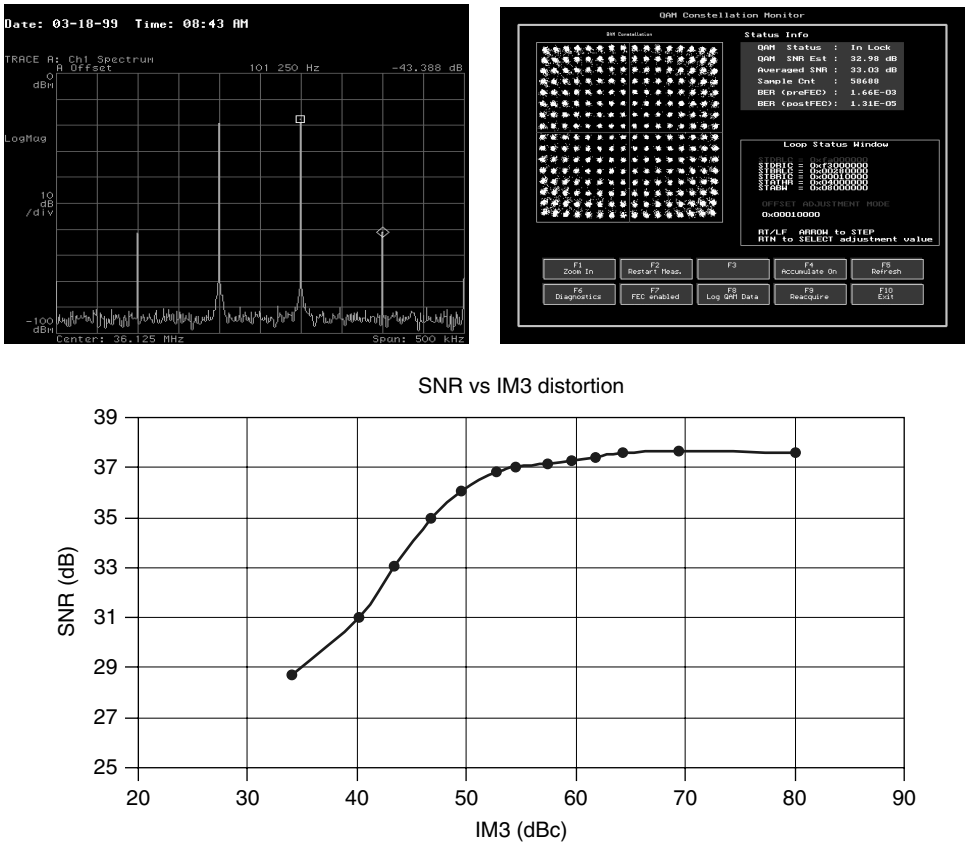


Figure 8. Effects of IMD distortion on 256-QAM.

No added C/N

Added C/N = 21 dB

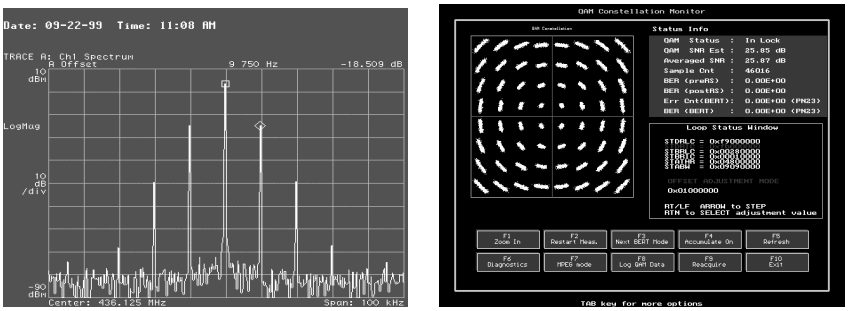
Sources ofAWGN (additive white Gaussian noise)

- Tuner noise figure
- Broadband noise in amplifiers
- ADC sample/hold aperture jitter
- Round-off errors in digital truncation
- Nyquist filter mismatch (appears like AWGN)

Methods to improve AWGN performance

- Reduce tuner noise figure to <9 dB
- Low noise amplifiers in RF/IF signal chain
- Direct clocking of ADC sample/hold
- Ensure matched α for

Figure 9. Effect of broadband AWGN on 64-QAM performance.



Added Φ_n = 10 kHz FM, 5 kHz Deviation
Carrier loop BW = 10 kHz

Sources of Φ_n

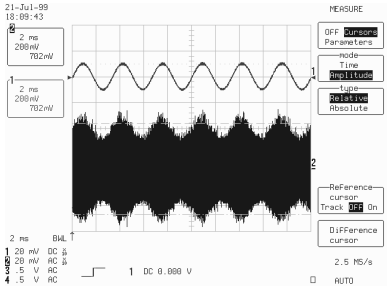
- Tuner LO
- Phase modulation of ADC sample/hold clock
- Poor supply decoupling of RF and AFE

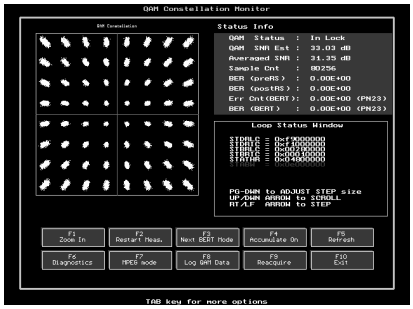
Methods to improve Φ_n performance

- Tuner LO Φ_n < –85 dBc/Hz @ 10 kHz offset
- Optimize carrier recovery loop BW (tradeoff between AWGN and phase noise effects)
- Low-noise narrowband PLL's
- Optimize power supply decoupling

Figure 10. Effects of phase noise (ϕ_n) on 64-QAm performance.

are demodulated. A simple pattern of 0xCCCCC0D is appended to the data, which corresponds to I/Q zero crossings (CCCCC) plus a unique word (0D) and has been determined to be adequate for locking a quadrature system (Fig. 15). Unlike a continuous receiver, if the burst receiver is not able to lock to the preamble, then the entire packet is lost, creating packet errors that are much more severe than bit errors. The basic burst modulator functional block diagram begins with a set of first-in first-out (FIFO) data buffers, allowing the front-end digital data for the next burst event to be loaded asynchronously while the analog IF output is transmitting the current burst, thus forming





Added AM: 20% modulation (200 mVpp/1Vpp); frequency = 130 Hz

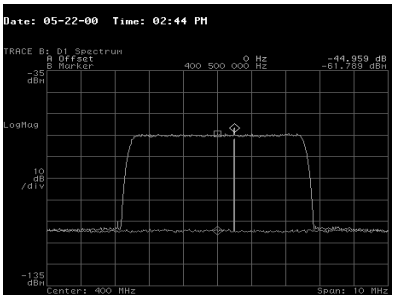
Sources of AM

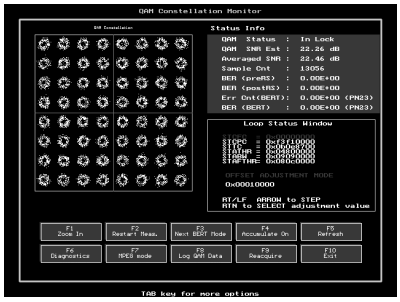
- 2nd harmonic of 60 Hz power line (120 Hz)
- Low frequency power supply ripple
- AGC loop bandwidth set incorrectly
- Incorrectly terminated grounding for RF/AFE

Methods to improve AM

- Improve filtering for power supply
- LC or ferrite in DC supply
- AGC loop BW and dominant pole set higher than frequency of AM impairment

Figure 11. Effects of AM on 64-QAM performance.





Added RFI = –24 dBc @ 401.0 MHz; 400 MHz
center frequency of RF input

Sources of RFI

- Fixed tone from digital clocks/oscillators
- Mixing products of adjacent NTSC channels
- CSO and CTB from loading (130 RF channels)
- Incorrectly terminated grounding for RF/AFE
- Ingress into cable

Methods to improve RFI

- Improve filtering for RF/AFE/digital power supply
- Change FFE main tap location in equalizer
- Narrowband notch filters
- Add high frequency ferrites to isolate grounds

Figure 12. Effects of RFI on 64-QAM performance.

an effective pipelined transmission. As data are pulled out of this FIFO, they are passed to the randomizer and the FEC, which scramble the symbol bits and encode them for transmission. Current implementations for the FEC are a programmable Reed–Solomon (RS) code with various Galois field selections and T values (number of correctable bytes) ranging from 1 to 10. This block calculates the FEC code parity bytes that are appended to the end of each burst and used by the receiver to correct for errors generated in the transmission link. After the preamble is

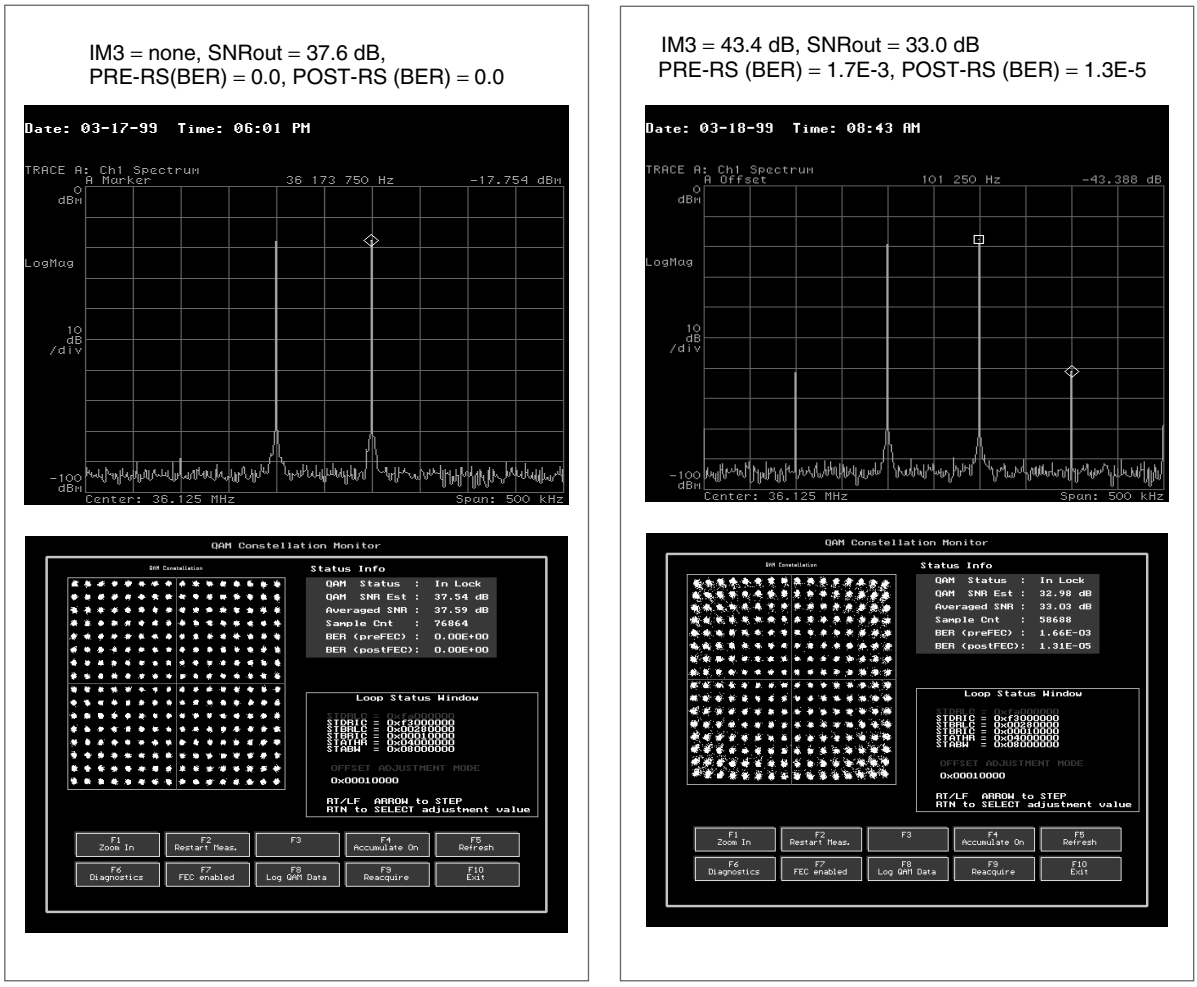
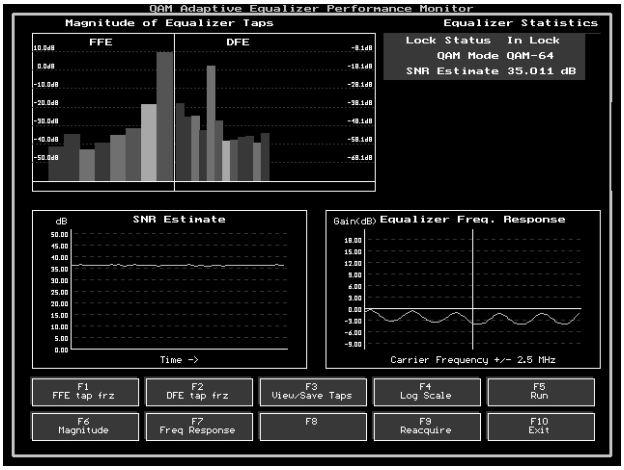


Figure 13. Effects of IM3 on 256-QAM performance.

inserted to indicate the beginning of a burst, the symbol mapping block performs a bitwise grouping into separate I and Q bitstreams. Each of these data paths is passed to matched square-root Nyquist filters that shape the necessary excess bandwidth to ensure that ISI remains at a minimum. Currently used alpha values are 0.25 (25%) for the DOCSIS upstream. In addition, on the head-end side, the receiver operates in the burst mode. To minimize preamble overhead, the receiver is not required to converge an adaptive equalizer on each burst, since each burst may come from a different transmitter having different channel characteristics. Thus an alternate means to provide echo cancellation must be derived. The DOCSIS 1.1 specification addresses this problem by defining a complex (real and imaginary) preequalizer residing in the upstream modulator that can effectively predistort the transmitted signal to cancel any echoes that will be generated in the transmission path for that particular modem. This preequalizer is located just prior to the Nyquist filtering. Coefficients for each modem's preequalizer are sent via the downstream channel from the head-end receiver on the basis of the received channel response for each particular modem. Following the Nyquist filters, the I/Q symbols are processed by a variable interpolating filter that upsamples the signal from the symbol rate to the DAC sample rate. From there the data enters the QDDFS, which consists of a structure similar to the downstream receiver whereby SIN and COS lookup tables are used as mixer LOs. These digital mixers upconvert and combine the digital data that are subsequently input to the DAC to create the analog IF output frequency.

The analog portion of the upstream design begins with an image reject filter following the QDDFS and DAC. The DAC will produce an image of the desired IF frequency (sample rate−IF frequency) and this image must be attenuated so as to not over drive the input to the power amplifier or leak into the upstream output. Typically, a high-order analog Chebyshev or elliptical filter is used for this purpose. The filter specifications can be relaxed with higher-frequency sample rates resulting in a higher-image frequency. Noise and distortion are very important for the upstream IF path because they must not be allowed to interfere with the lower downstream channels beginning at 54 MHz. With this in mind, a fully differential topology is preferred, allowing the cancellation of HD2, leaving HD3 to contend with. There is some help from the diplex filter at the output to attenuate the third-order product since it must have a sharp stopband attenuation to keep

64 QAM DFE real tap #5



64 QAM, 5 Mbaud, delay = 1uS, Phase = 320 degree, attenuation 10 dB

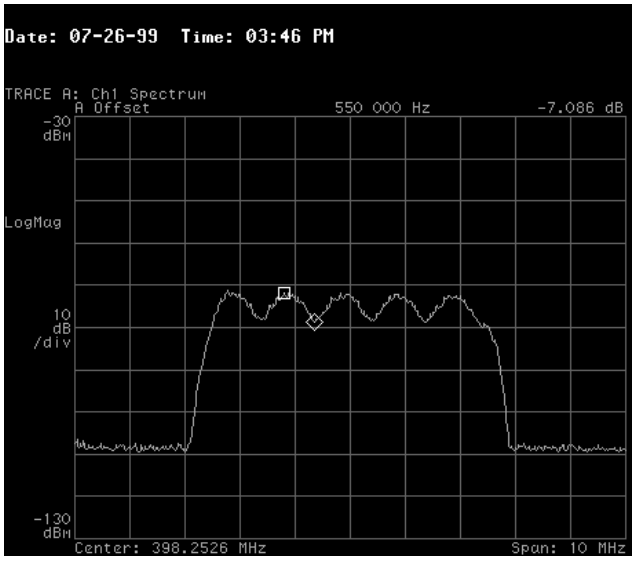


Figure 14. Multipath reflections.

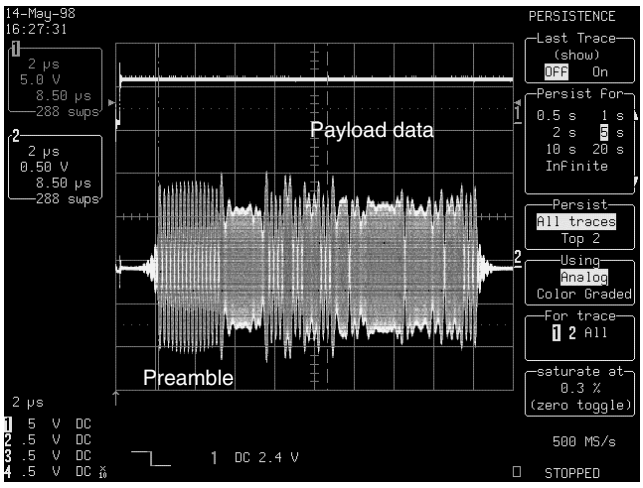


Figure 15. Burst upstream data packet.

any spurious and noise out of the downstream channels. Additionally, as with the downstream receiver, the intermodulation products must be kept to a minimum. The power amplifier that drives the upstream data back through 75-Ω cable must supply +8 to +58 dBmV of signal level implying a variable attenuation of 50 dB. Common practice is to design the burst modulator with an analog programmable gain amplifier (PGA) as fine gain control (25 dB in 0.4-dB increments) and design the power amplifier to supply the remaining 25 dB in coarse steps (6 dB). In addition to the needed large signal and low distortion, the amplifier must have very low noise (DOCSIS spec is −59 dBmV/Hz), which necessitates disabling it in between data bursts. This in turn leads to the potential turn-on/off glitch, which has been specified to be less than 100 mV integrated over 200 ns. To make matters even more difficult, the on/off impedance match must maintain a 75-Ω termination to ensure proper cable termination and return loss under all conditions, and it is desirable to use a single +3.3-V-DC supply or common voltage required for the modem chip. Taking all of this into consideration makes for a very challenging design (Fig. 16). The leading cable modem chip designers have integrated this amplifier onto the cable modem chip, thus extending the state of the art and providing further cost reductions and simplicity for modem product designers.

5. MAC LAYER: DATA PROTOCOLS

The cable modem MAC acts as the data parser and decoder to enable the DOCSIS point-to-multipoint communication system. As pointed out earlier in this text, DOCSIS employs a continuous downstream signal and a TDMA burst upstream signal, and the MAC acts as the basic controller between the modem and head-end equipment residing at the service provider. The DOCSIS specification defines many different packet types and usage codes, called interval usage codes (IUCs), but by far the three most essential MAC messages are the SYNC (synchronization), upstream channel descriptor (UCD), and minislot usage information (MAP).

The basic process flow of channel acquisition begins with the downstream receiver scanning all RF channels and obtaining QAM and FEC lock in search for MPEG packets containing the well-known PID for a DOCSIS data channel (0x1FFE). Once a DOCSIS channel has been

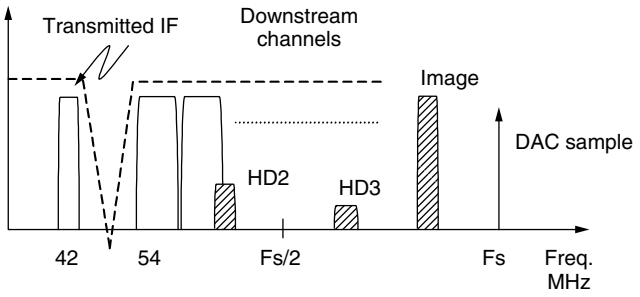


Figure 16. Burst upstream frequency spectrum.

found, the cable modem (CM) begins looking for the three MAC messages that are regularly sent from the head end (CMTS). The first necessary step is to synchronize the CM with the CMTS and all other modems in the system. This is accomplished by the CMTS, which sends a periodic SYNC message containing a 32-bit timestamp over the downstream channel. The CM receives the SYNC message and locks the frequency of its local clock so that it matches the time stamp. This process may require many SYNC messages before the CM's local clock is adequately tracking the CMTS reference clock. For upstream TDMA data transfers, the concept of minislots (a convenient partitioning of time) is used, instead actual number of bytes of information to transmit, which facilitates bandwidth allocation when switching modulation types. Once the modem has determined the common time reference, the next message required is the UCD. The UCD instructs the CM to adjust a number of upstream parameters such as the transmitter frequency, modulation type, symbol rate, minislot size, preamble pattern, and selection of a burst profile to use for further communication. This is the initial setup needed to establish basic two-way connectivity with the CMTS. The final step in channel acquisition is the processing of the bandwidth allocation in the MAP message, which corresponds to the upstream described in the UCD. This message designates the minislot information and is used to establish at what time and for how long the modem can transmit, with the SYNC message providing the time reference for these transmissions. The MAP messages assign burst type (via IUCs) and burst duration (via minislots) to upstream SIDs. The upstream SIDs are used for bandwidth allocation and security associations. The initial signon process uses a special time slot called "initial maintenance," denoted by IUC 3. At this point in the process, the modem has established (1) a time reference, (2) initial upstream transmission configuration, and (3) knowledge of when and for how long to transmit.

While the previous steps provide the modem with a notion of relative time, the CM still needs to know the exact time. The SYNC messages that provide the CM with its notion of time incur propagation delays as they travel from the CMTS to the CM. This propagation delay will vary depending on the position of each CM on the cable plant. Thus, each CM has a relative notion of time through frequency locking to the SYNC messages, but not an exact notion due to propagation delay. A process called "ranging" adjusts each CMs notion of time to be the exact notion required for TDMA operation with the CMTS. The ranging process begins with the modem sending the head end a ranging request. A number of problems may prevent the CMTS from issuing a ranging response message acknowledging the modem. The CM ranging request could collide with another modem which is initiating the logon process or the transmit power of the CM may be too low for the head end to receive it. Therefore, the ranging request will be repeated with appropriate time backoff and power adjustment until eventually the head end will acknowledge with a ranging response and will send the CM a dedicated ranging opportunity called "station maintenance," denoted IUC 4 in a new bandwidth allocation MAP. The CM will now automatically transmit a ranging request in any station maintenance slots reserved for it. At this point the CM and the CMTS enter an interactive mode whereby fine adjustments are made to the transmit frequency (must be with ± 10 Hz of commanded), transmit power (must be within ± 2 dB of commanded), time offset (must be within 1 μs of commanded) and multipath reflections (preequalizer tap adjustments). This may take many fine adjustments with the final outcome of calibrating out any time and amplitude variations for each modem's roundtrip data. Once the CMTS detect the CM is properly ranged, the CMTS sends a ranging response message with a ranging complete notation. The CM then uses request regions denoted IUC 1 to send up a request for the bandwidth required to transmit its nonranging packet. The head end will respond with the bandwidth allocation MAP, granting the modem the bandwidth requested, and the CM MAC can now send its first nonranging information to the CMTS.

Next the CM needs to establish IP layer connectivity. This is accomplished by use of the dynamic host configuration protocol (DHCP), which will assign the CM an IP address and form the IP link between the modem and the DHCP sever. When the modem has terminated connectivity, the IP address will be relinquished back to the pool, and DHCP can reallocate it to another IP user. Registration of the modem begins with the CM downloading a configuration file using trivial file transfer protocol (TFTP) and establishment of a service identification (SID). Only after a number of file checks and authorization confirmation will the CM be allowed to transmit "real" data onto the cable system. At this point the modem is able to transmit and receive data, but one final step in basic connectivity must be performed. Since the cable protocol is a shared medium, a means to protect and secure data transfers is necessary. This is accomplished by what is known as baseline privacy (BPI). Each modem is uniquely identified with a 48-bit MAC address, which can only obtain BPI encryption keying information it is authorized to access. BPI uses the Cipher Block Chaining mode of the data encryption standard (DES) algorithm to encrypt data in both upstream and downstream data paths. The CM uses RSA, a public-key encryption algorithm (proprietary to RSA Data Security, Inc.) to obtain authorization and encryption keys from the head end and to support periodic encryption key changes. The cable operators determine how often new encryption keys are sent. This final step relinquishes the cable modem to "surf the Net" at typical downstream data rates of 1–2 Mbps.

6. CONCLUSIONS

This article presents an overview of the cable modem system and descriptions of some of the key components found in cable modem equipment. A detailed discussion of the four DOCSIS layers (PHY, MPEG, MAC, BPI) is examined. Current video delivery to most homes in the United States is via cable, and as more interactive services are offered, there will be increasing emphasis on providing simultaneous high-speed data available to these users. The bandwidth is available from the existing cable

plants to provide this growth. Cable modems have shown increases in speed of 1000 times over telephone modems, and nearly all housing developments in the United States have a cable infrastructure already in place. Increased levels of integration have dramatically reduced the cost of cable modems, enabling explosive growth and accelerated deployment for the near future.

Acknowledgments
The author wishes to acknowledge the contributions of Dr. Charles Reames, Lisa Denney, Bruce Currivan, Dr. Thomas Kolze, and Dr. Henry Samueli for valuable advice and criticism of this text.

BIOGRAPHY

Donald McMullin (M'87) received his B.S. in electrical engineering from California State University Northridge. He joined the Electro-optical and Data Systems Group of Hughes Aircraft Company, El Segundo CA, in 1988 where he worked on forward looking infrared night vision systems and Aided Target Recognition computers. In 1992 he joined the Advanced Circuit Technology Center of Hughes Aircraft and specialized in full custom analog chip design for high performance data converters. In 1996 Mr. McMullin joined Broadcom Corp. and focused his attention on QAM receivers/modulators and broadband cable data transmission systems. He is currently the manager of hardware development for cable products at Broadcom Corp. Donald holds 3 patents for amplifier design topologies intended for data converter applications and has 4 patents pending in the area of communication design.

BIBLIOGRAPHY

1. DOCSIS (Data-over-Cable Service Interface Specification), *Radio Frequency Interface Specification*, SP-RFIv1.1-I08-020301.
2. B. Sklar, *Digital Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
3. J. Min and H. Samueli, Frequency-agile TDMA system for upstream cable-modem applications and B. Currivan, Cable modem physical layer specification and design, in *Cable Modems: Current Technologies and Applications*, International Engineering Consortium, Chicago, 1999.

CARRIERLESS AMPLITUDE–PHASE MODULATION

BURTON R. SALTZBERG
Middletown, New Jersey

Carrierless amplitude–phase modulation (CAP) is a variation of quadrature amplitude modulation (QAM), in which explicit modulation and demodulation is omitted. In virtually every aspect, the performance, analysis, and most of the implementation techniques of QAM are applicable to CAP. While QAM has been the preferred modulation technique for a wide variety of applications for many decades, CAP has been used for digital communications only in recent years, largely for transmission over wire-pair channels.

To illustrate conceptually how QAM may evolve into CAP, we first show a standard QAM system in Fig. 1. A datastream of user bits is first assembled into pairs of symbols, each of which is chosen from an alphabet representing some number of bits. The symbol pair may be considered to be a two-dimensional symbol, or a complex quantity. The mapper chooses a point in two-dimensional, or complex, space for each possible symbol value. The set of such points is the signal constellation. The real and imaginary values of those points are denoted as the I and Q components. Each component is lowpass-filtered, as in pulse amplitude modulation (PAM), and input to a modulator that multiplies that component by one of two sinusoidal carriers. The carriers are at the same frequency and differ in phase by 90°. The modulated signals are added and presented to the transmission channel. At the receiver, the line signal is demodulated by the same two carriers to form a pair of baseband signals. An essential component of the receiver is a means of reconstructing the carriers. The baseband filters may include an equalizer. The detection process is then completed and the bit stream reconstituted.

It has long been recognized that the order of filtering and modulation may be interchanged in either the transmitter, receiver, or both. In fact, passband filtering and equalization are quite common in QAM receivers. Figure 2 shows the case in which this interchange is done in both the transmitter and the receiver. Each pair of lowpass filters is replaced by a Hilbert pair of bandpass filters. The lowpass filter with impulse response $f(t)$ is replaced by the following pair:

$$f_1(t) = f(t)\cos\omega_c t$$
$$f_2(t) = f(t)\sin\omega_c t$$

The filters form a Hilbert pair, in that

$$f_2(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f_1(u)}{u-t}\,du$$

or, more understandably in terms of the Fourier transforms, as

$$F_2(f) = j\,\mathrm{sgn}(f)F_1(f)$$

which says the two frequency responses are equal in amplitude and differ in phase by 90° over the entire frequency range. The important property in this application is the orthogonality of the Hilbert pair:

$$\int_{-\infty}^{\infty} f_1(t)f_2(t)\,dt = 0$$

Examination of Fig. 2 reveals that the only function served by the modulator is to multiply the complex data symbols by $e^{j2\pi f_c t}$, and the demodulator is to multiply by $e^{-j2\pi f_c t}$. This amounts to a rotation and a counterrotation by the same quantity. The final and key step in forming the CAP system is simply to eliminate this rotation and counterrotation as shown in Fig. 3. If $f_c T$ is an integer,

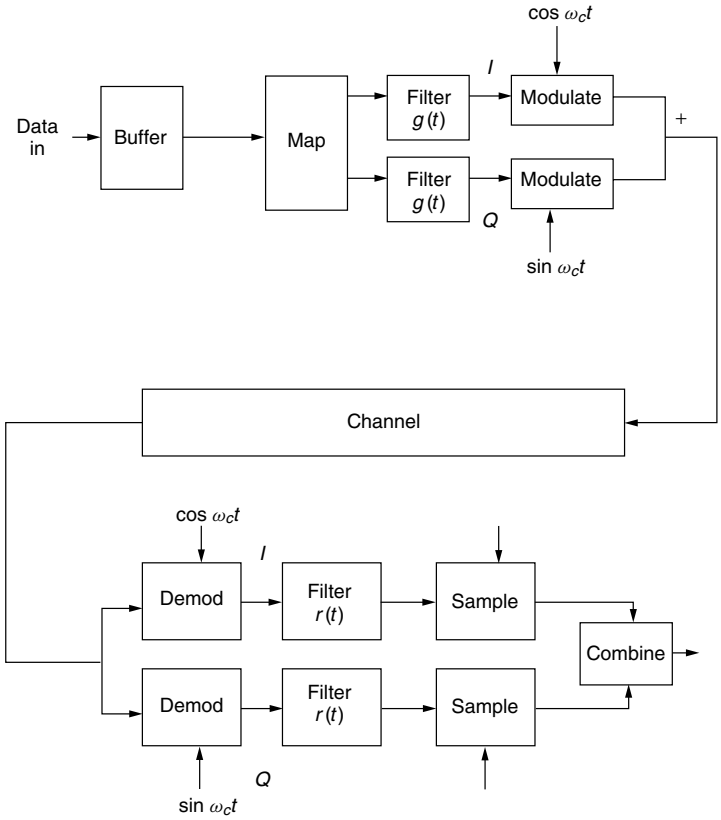


Figure 1. A standard QAM system.

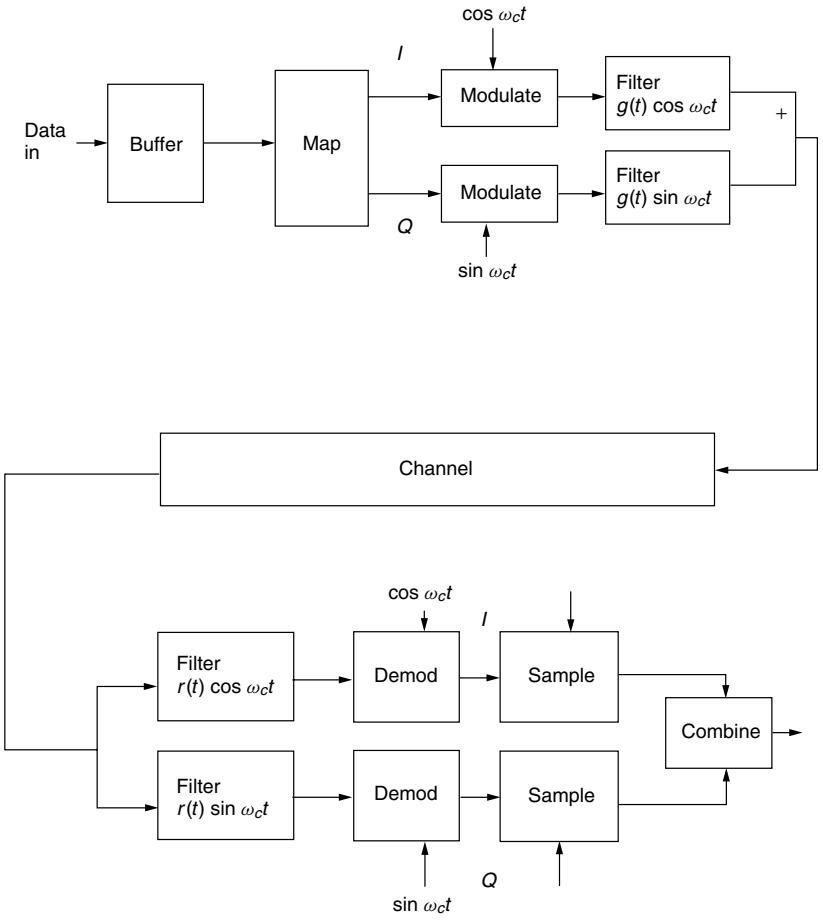


Figure 2. A QAM system with bandpass filtering.

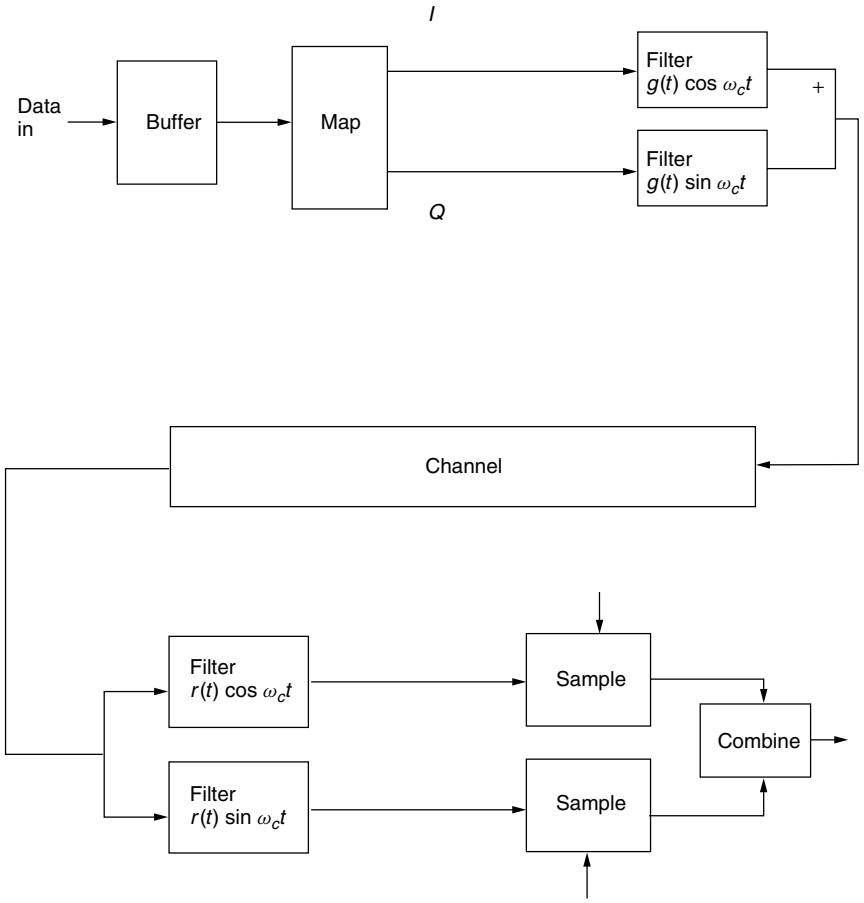


Figure 3. A CAP system.

where T is the symbol duration, then the QAM and CAP systems are totally identical. If f_cT is an integral multiple of $\frac{1}{4}$, and the constellation is symmetric, then they are again identical except for coding of the symbols. In general, for any value of f_cT, the corresponding QAM and CAP systems are virtually identical except for a unitary rotation of the constellation.

CAP modulation may be explained in terms of PAM rather than QAM. Consider a PAM system using bandpass filtering in place of the usual lowpass filtering. Such a system would require twice the bandwidth of a standard PAM system. However, if two such systems are superimposed on the same frequency band, bandwidth efficiency is achieved in that the same bandwidth is required for the same bit rate. The two passband signals may be superimposed and separated at the receiver if the bandpass filters form a Hilbert pair. This is illustrated by Fig. 4.

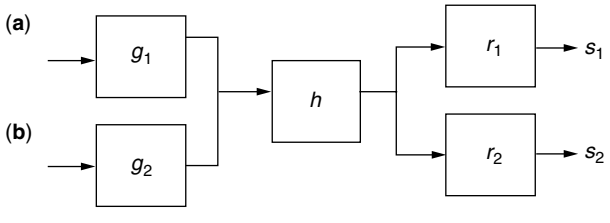


Figure 4. Basic model of a CAP system.

Let the convolution of each of the transmit filters, the channel, and each of the receive filters be denoted by an overall response

$$z_{ij}(t) = g_i(t) * h(t) * r_j(t).$$

Then the received signals are

$$s_1(t) = \sum_k a_k z_{11}(t - kT) + \sum_k b_k z_{21}(t - kT)$$

$$s_2(t) = \sum_k a_k z_{12}(t - kT) + \sum_k b_k z_{22}(t - kT)$$

Intersymbol interference is eliminated if the usual Nyquist condition is met for each subchannel:

$$z_{11}(t - kT), z_{22}(t - kT) = 1 \quad \text{for} \quad k = 0$$

$$\text{and 0 for all integers} \quad k \neq 0$$

Interchannel interference is eliminated if

$$z_{12}(t - kT), z_{21}(t - kT) = 0 \quad \text{for all integer } k$$

Under these conditions, $s_1(kT) = a_k$ and $s_2(kT) = b_k$, as desired.

A standard two-dimensional equalizer, such as is used in QAM systems, can adapt $r_1(t)$ and $r_2(t)$ so that these conditions are met. As in a passband QAM system, the

sampling rate must be at least twice the highest frequency in the line signal spectrum. Unlike the QAM equalizer, the error signal used for adaptation is not remodulated, since no modulation is present.

All techniques used in QAM systems may be directly applied in CAP systems. These include constellation shaping, trellis coding in various dimensions, decision feedback equalization, Tomlinson filtering, and sequence detection.

For channels that introduce frequency offset or phase noise, due to effects such as Doppler shift and additional stages of modulation and demodulation, some subsystem similar to carrier recovery is required at the receiver. The simplification in CAP over QAM is therefore no longer present. For this reason CAP has been applied primarily over channels in which frequency offset and phase noise are not present, in particular the wire-pair channel. CAP has been widely applied to various digital subscriber line (DSL) and local-area network systems.

BIOGRAPHY

Burton R. Saltzberg received a Sc.D. degree from New York University in 1964. He is a consultant in digital communications with several companies, and presents short courses to international audiences. He was with Bell Laboratories from 1957 through early 1996. His most recent position there, which he held for several years, was technical manager of the Data Theory Group. For most of his career, Dr. Saltzberg was engaged in research in communication theory and in analysis and initiation of new data communications offerings over a wide variety of channels. He has published extensively in this field, and was issued 29 U.S. patents. He is a fellow of the IEEE and received the IEEE Communications Society Armstrong Achievement Award in 1991.

BIBLIOGRAPHY

1. T. Starr, J. M. Cioffi, and P. Silverman, *Understanding Digital Subscriber Line Technology*, Prentice-Hall, Upper Saddle River, NJ, 1999.
2. A. K. Aman, R. L. Cupo, and N. A. Zervos, Combined trellis coding and DFE through Tomlinson precoding, *IEEE J. Select. Areas Commun.* **9**: 876–884 (1991).
3. G. H. Im et al., 51.84 Mb/s 16-CAP ATM LAN standard, *IEEE J. Select. Areas Commun.* **13**: 620–633 (1995).
4. G. H. Im and J. J. Werner, Bandwidth-efficient digital transmission over unshielded twisted-pair wiring, *IEEE J. Select. Areas Commun.* **13**: 1643–1655 (1995).

CARRIER-SENSE MULTIPLE ACCESS (CSMA) PROTOCOLS

LEONIDAS GEORGIADIS
Aristotle University of Thessaloniki
Thessaloniki, Greece

1. INTRODUCTION

Communication of information between two or more parties takes place over a variety of physical media called *channels*. Such channels can be simple twisted pair cables, coaxial and optical cables, or free space. Sometimes the channel is dedicated to a specific transmitter–receiver pair. This may be the case when two parties establish a telephone conversation over a dedicated cable. Channels of this type are called *point-to-point*. There are situations, however, where multiple users need to have access to the same channel. The most familiar one is human speech communication. When a number of humans are located in the same room, they all use the same channel, the atmosphere, for their conversation exchange. Computer local area networks (LANs) is another example: a common approach in this case is to attach a number of computers to the same cable as in Fig. 1. Hence, each computer can listen to the transmission of every other computer attached to the same cable. For a third example, consider Satellite communication. As shown in Fig. 2, a number of terminals need to communicate between each other but because of physical obstacles they cannot all listen to each other directly. Instead, each terminal first sends the information to the satellite through the upstream channel. The satellite listens to the upstream channel, receives the information, and then retransmits to the downstream channel, to which all terminal can listen. Hence the upstream channel needs to be accessed by all terminals. Channels of this type are called *multiple-access*.

If the terminals in a multiple-access channel are left unchecked so that they can transmit information whenever they need to do so, then it may be possible for two or more terminals to attempt to use the channel at the same time. In such a situation, the concurrently transmitted messages interfere with each other and generally cannot be received correctly by the intended receivers. Hence, a fundamental issue in multiple-access channel communication is how to coordinate the transmitting terminals in order to avoid or recover from the interference that may result because of concurrent transmissions. The mechanisms by which this is achieved are termed *multiple-access protocols*.

The simplest way to address the coordination problem in multiple-access communication is to avoid concurrent transmissions altogether. To be more specific, we must make certain assumptions about the manner in which transmission of information takes place. First, as is very common today, we assume that all information, whether sound, picture, or text, is transformed to a sequence of bits, 0 or 1, and that each terminal needs to transmit this sequence of bits to the receiving terminal — the receiver knows how to recover the original information from the received sequence of bits. Let us assume further that the

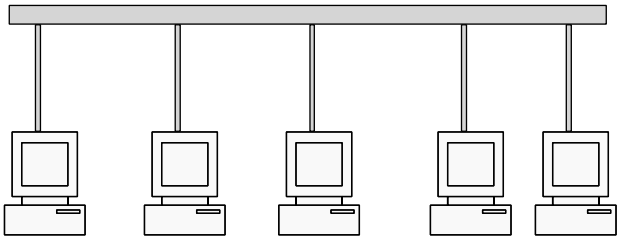


Figure 1. A local area network.

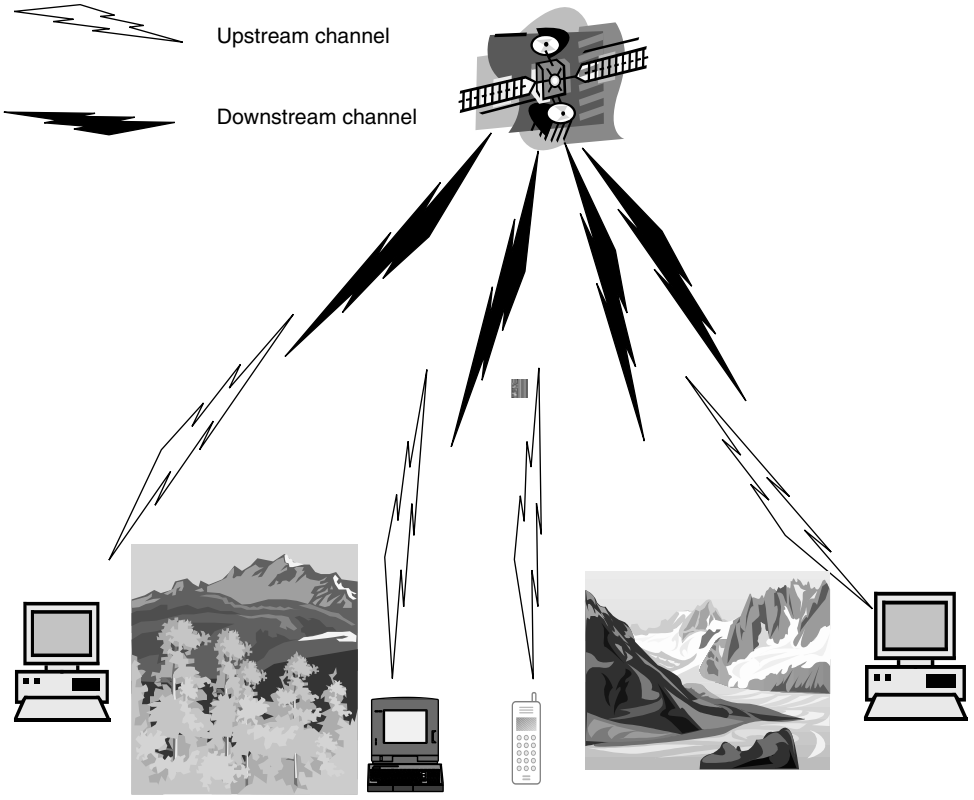
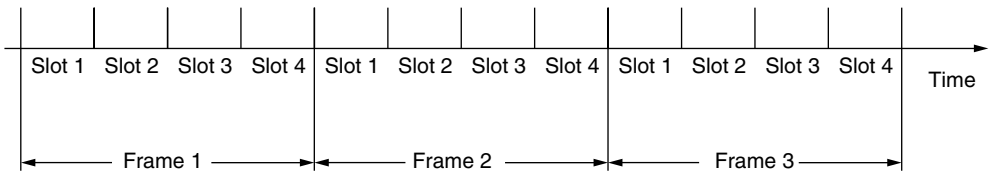


Figure 2. Satellite communications.

sequence of bits is subdivided into groups called *packets*, and that the transmitter needs to transmit one packet at a time to the receiver. All packets contain the same number of bits B. If bits can be transmitted over the channel at a rate of C bits per second (bits/s or bps), then each packet takes $T = B/C$ seconds to be transmitted. We refer to T as the "length" of the packet.

We are now ready to describe the protocol by which access to the channel is free of concurrent transmissions. We divide time into fixed intervals of length T called time *slots* (see Fig. 3). Hence each slot fits exactly one packet. Let the number of terminals that can potentially use the channel be n. We group the time slots into *frames*, where each frame contains n consecutive time slots. Terminal i is allowed to transmit in the ith time slot of each frame. The protocol just described is called *time-division multiple-access* (TDMA) protocol. Since slots are allocated exclusively to each user, no interference occurs and packets are transmitted successfully. Note that there is still a possibility that the packet may be received in error because of noise that naturally exists on the channel, but this is a lower-level issue that is addressed by methods that are beyond the scope of the current discussion.

The TDMA protocol in effect divides the channel into n point-to-point channels. While simple, the protocol has some serious disadvantages. First, if a terminal does not have packets to transmit, the slots allocated exclusively to it cannot be used by anybody else, even if other terminals have a large number of packets to transmit and could use these slots. The second disadvantage is related to packet delays. Since the time interval between two successive slots during which terminal i can transmit is n time slots, a packet generated randomly at a terminal will take on the average $n/2$ time slots to be transmitted, a delay that can be very large if the number of terminals in the system



The channel is accessed by $n = 4$ terminals. Terminal i may transmit in slot i of each frame.

Figure 3. The TDMA protocol.

is large. This will happen regardless of whether the rest of the terminals have packets to transmit.

The disadvantages of TDMA are due to the fact that a terminal can transmit only during the slots allocated to it, even if other terminals are inactive. What if we dispense with the idea of allocating slots exclusively to transmitters? In fact, what if we take the exact opposite approach and allow a terminal to transmit in any slot when it has packets to transmit? In this case, if no other terminal has packets to transmit, then the given terminal can transmit a large number of packets with very low delay. However, if more than one terminals wish to send packets in the same slot, then a "collision" will occur and no message will be received correctly. In this case one must specify how the terminals will react and cooperate in order to make sure that the packets are eventually delivered to their intended destinations. The simplest idea is to instruct the terminals to retransmit their collided packets. However, if two or more terminals pick again and again the same slot for retransmission, the packets will continue to collide and will never be transmitted successfully. There are various methods to avoid this situation. We will concentrate on the most prevalent method encountered in practice: randomized retransmissions. If collisions occur, then the terminals whose packets collided pick randomly some future time slot for retransmission. Hence, while collision may again occur, it is hoped that eventually the transmitting terminals will each pick different slots for transmission and thus their packets will be delivered successfully to their intended destination.

The algorithm just described comes by the name ALOHA protocol and will be described in more detail in Section 2. This algorithm constitutes the basis for the development of carrier-sense multiple-access (CSMA) protocol, which takes advantage of certain channel features and transmitter capabilities in order to provide improved performance. The CSMA protocol will be described in Section 3.

2. THE ALOHA PROTOCOL

The ALOHA protocol was designed by Abramson [1] to provide radio communication between several terminals scattered at various places over the islands of Hawaii. The terminals were sending their data packets to a central station over a common channel (the upstream channel). The central station was then retransmitting the packets to another channel (the downstream channel) that could be listened to by all the terminals. The situation is similar to the one described in Fig. 2. Collisions could occur at the upstream channel if two or more terminals were attempting to transmit their packets. If this happened, the central computer was informing all the terminals that a collision had occurred.

There are two versions of the ALOHA protocol: slotted and unslotted. Slotted ALOHA requires time to be divided in time slots and terminals to transmit their packets at the beginning of each slot. Unslotted ALOHA permits the stations to transmit their packets at any time. The retransmission policy in case of collision is essentially the same for both protocols. In the next two sections we examine these two variants of the ALOHA protocol. Unslotted ALOHA was the precursor of slotted, but it will be more instructive and simpler to concentrate on the slotted ALOHA first.

2.1. Slotted ALOHA

Let us provide a model for this protocol. As in Section 1, the channel is divided into time slots. Terminals are synchronized to transmit their packets at the beginning of a time slot. At the end of each time slot, terminals that transmitted their packets during that slot are informed whether there was a successful transmission or a collision in the slot. If the packet that a terminal transmitted collides with another packet, then the terminal attempts a retransmission in the next slot with probability p and defers for the end of the next slot with probability $1 - p$. In case of deferral, at the end of the next slot the terminal reattempts transmission with probability p and defers with probability $1 - p$. Figure 4 shows an example of the operation of the ALOHA protocol. At slot 1 three terminals attempt to transmit their packets and there is a collision. Hence all three terminals will attempt to retransmit their packets. No terminal chooses to retransmit at slot 2, which is thus idle. Terminals a and b attempt to retransmit at slot 3, and hence there is again a collision. Terminal b is the only one attempting retransmission at slot 4 and its transmission is successful. The transmissions from terminals a and c collide again in slot 6, but they eventually pick different slots for retransmission and their packets are transmitted successfully in slots 8 and 9. Note that other terminals may become active (i.e., they may generate a new packet for transmission) while the retransmission process takes place. These terminals may cause additional collisions. For example, if terminal d generates a new packet and attempts to transmit it in slot 8, an additional collision will occur. In the network designed by Abramson, all terminals (not only those that transmitted their packets) can listen to the downstream channel and hence can be

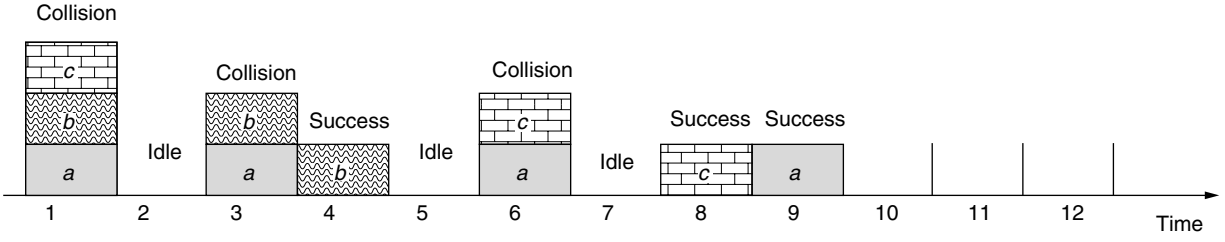


Figure 4. The operation of the ALOHA protocol.

informed about the status of the transmission at the end of the current slot, that is, whether there was no transmission, a successful transmission, or a collision during the slot. However, the ALOHA protocol does not make use of this extra information that a terminal can have.

The protocol just described has the desirable property that packets are not delayed at all if only one terminal needs to transmit at a given time slot. What happens, however, when two or more terminals attempt transmission? As the example in Fig. 4 shows, in this case collisions occur that are followed by retransmission attempts. This results in two inefficiencies; slots may (1) be wasted because of collisions or (2) remain idle even though some terminals have packets to transmit; the latter will happen if all packets that attempt retransmission are deferring in the current slot and no new packets are generated. It is therefore important to know the useful information that can come out of the channel. An appropriate measure for this information is the average number of successfully transmitted packet, S, per slot. We refer to S as the *throughput* of the channel.

Next we provide a method for evaluating S. We need to first make an assumption regarding the statistics of new packets generation process: the number of new packets, K, generated for transmission during a time slot, is a Poisson random variable with rate λ packets/slot. That is, the probability that $K = k$ is given by

$$\Pr(K = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$

This model of packet generation is called the "infinite population model" because it implies that the number of terminals in the system is potentially infinite (the probability that K is any large number is nonzero) and that each terminal generates packets infrequently, so that packet queues are not formed at the terminals. It is used because it is simple, a good approximation when the number of terminals is large, and provides some important insights.

Two sets of terminals may attempt transmission at the beginning of a time slot: (1) those that generate new packets and (2) those whose generated packets have collided in some previous slot and attempt retransmission. In case 2 we say that the packets are "backlogged." Assume that the system can reach steady state and let M be the random number of packets (newly generated and backlogged) transmitted in a given slot in steady state. Denote by G packets/slot the average value of M. Since M includes both newly generated and backlogged packets, it follows that $G > \lambda$. Observe that a successful transmission occurs only when $M = 1$. Indeed, if $M \neq 1$, then either the slot will be idle (if $M = 0$) or there will be a collision in the slot (if $M \geq 2$). Therefore, by the definition of S we have

$$S = 0\Pr(M \neq 0) + 1\Pr(M = 1) = \Pr(M = 1)$$

Hence, if we knew the statistics of M, then we would be able to evaluate S. The exact evaluation of the statistics of M is complicated. To simplify the situation, we make the additional assumption that M is a Poisson random variable. Since the rate of M is G, we have from (1)

$$S = \Pr(M = 1) = e^{-G}G \tag{2}$$

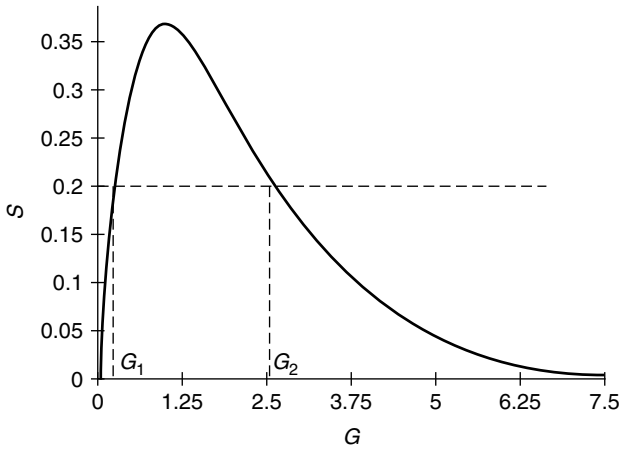


Figure 5. The throughput of the ALOHA protocol.

In Fig. 5 we plot S as a function of G given by Eq. (2). It can be shown that the maximum value of S is $1/e \approx 0.368$ and is obtained at $G = 1$. Hence the maximum channel throughput of the slotted ALOHA protocol is 0.368 packets/slot. A conspicuous feature of the plot in Fig. 5 is that a given channel throughput is achieved for *two* values of G: a small, G_1 and a large G_2. The small value implies that the number of backlogged packets is small while for the large value this number is large. Clearly we would prefer to operate the system at the value G_1, but why do two values appear and what is their meaning?

There are two flaws with the analysis presented above: (1) the existence of steady state is assumed and (2) the probability distribution of M is assumed to be Poisson. For the infinite Poisson model, both these assumptions turn out to be invalid! However, the derived bound on the achievable throughput is still correct. A more detailed analysis of the system for a finite number of users, which is beyond the scope of this presentation, reveals that indeed the throughput of the system is at most $1/e$. Moreover it can be shown that the system behaves qualitatively as follows. There are long periods of time during which the number of backlogged packets in the system remains small and the system operates well, inducing small packet delays. However, from time to time a large increase in the number of backlogged packets in the system will occur and system performance in terms of throughput and delay will degrade. Fortunately, it can also be shown that the time interval for the transition from the "good" state to the "bad" state is generally very large. Hence this instability phenomenon of transiting from good to bad states is seldom a severe problem in real systems.

2.2. Unslotted ALOHA

In the previous section we assumed that the terminals are all synchronized to begin transmission of their packets at the beginning of each slot. If this feature is unavailable, the protocol can be easily modified to still operate. Indeed, the users can be allowed to transmit their new packets at packet generation time. If a collision occurs, then the terminal attempts a retransmission at a later randomly chosen time.

Let us evaluate the performance of the unslotted ALOHA system. We adopt the infinite population model and the notation of Section 2.1. Taking into account the cautionary statements at the end of Section 2.1, let us assume the existence of steady state and that $M(\gamma)$, the number of terminals that attempt transmission in any time interval of length γT, is a Poisson random variable with rate γG. If terminal a begins transmission at time t (see Fig. 6), its transmission will be successful if no other packet begins transmission in the interval $[t-T, t+T]$. Since this interval has length $2T$, the probability that no packet (other than terminal a's packet) is transmitted in the interval $[t-T, t+T]$ is $P_s = P(M(2)=0) = e^{-2G}$. We can interpret P_s as the proportion of attempted packet transmissions that are successful. Now, the rate (average number of packets per time T) by which packet transmissions are attempted is G, and a proportion P_s of these transmissions are successful. Hence the rate of successful transmissions is

$$S = GP_s = Ge^{-2G} \tag{3}$$

where it can be seen that the maximum throughput is $1/(2e)$ and is obtained for $G = \frac{1}{2}$.

We see that the throughput of the unslotted ALOHA is half the throughput of the slotted one. However, unslotted ALOHA does not require terminal synchronization. In any case, from the previous discussion we see that the throughput of both systems is much lower than one. Throughput 1 could be achieved if the terminals could be scheduled for transmission so that collisions are avoided. On the other hand, we have seen that the ALOHA protocol is very simple and distributed in the sense that the terminals operate independently of each other and require very small amount of feedback information to make their decisions. Moreover, the protocol induces very small packet delays when the system is lightly loaded. The question arises as to whether the throughput of the ALOHA protocol can be improved while maintaining its desirable features. These considerations lead to the development of CSMA protocols, which we discuss in the next section.

3. CSMA PROTOCOLS

In this section we present the versions of CSMA protocols that have found wide application. One can think of the CSMA protocol as an evolution of ALOHA where certain terminal capabilities are exploited in order to attain improved performance. It turns out that in real systems the required terminal capabilities depend on the transmission media, that is, whether communication takes place over wires — twisted pair, coaxial, or optical — or through radiowaves in the atmosphere — wireless communication. Accordingly, we first discuss the CSMA and CSMA/CD protocols that are appropriate for wired communications and next examine the CSMA/CA protocol, which is designed for wireless communications.

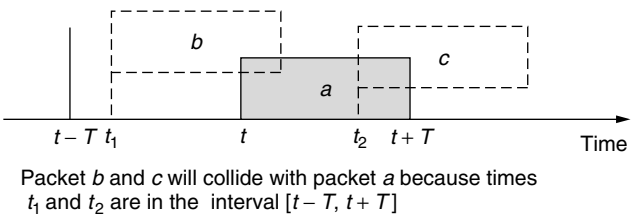


Figure 6. Possibility of collisions in the unslotted ALOHA protocol.

3.1. The CSMA and CSMA/CD Protocols

As we saw in the previous sections, the throughput loss of the ALOHA protocol is due to the fact that slots are wasted due to collisions or remain idle while there are terminals having packets ready for transmission. Let us see whether we can improve this situation while maintaining the desirable features of the ALOHA system. The throughput of the system can be improved if

1. The likelihood of a collision is reduced.
2. The time wasted transmitting garbled data when a collision occurs is reduced.

Consider the possibility of reducing collisions first. Let us assume that a terminal is able to listen to the channel and detect possible ongoing transmissions — busy channel. The ALOHA protocol can then be modified as follows. In case the terminal finds the channel busy, it defers transmission for a random time, or else it transmits its own packet. The protocol just described is called *carrier-sense multiple-access protocol*. The term "carrier sense" signifies the capability of the terminal to listen to the channel and ascertain whether it is busy.

At first sight it seems that with CSMA we succeed in avoiding collisions altogether. Indeed, if all terminals transmit their packets only when the channel is not busy and pick a random retransmission time if they find the channel busy, then it seems that a collision will occur only when two or more terminals begin transmission simultaneously, an event that is quite unlikely. However, the situation is not as rosy as it seems, due to the finite time it takes for a signal to propagate from one terminal to another. Consider the example in Fig. 7. Assume that it

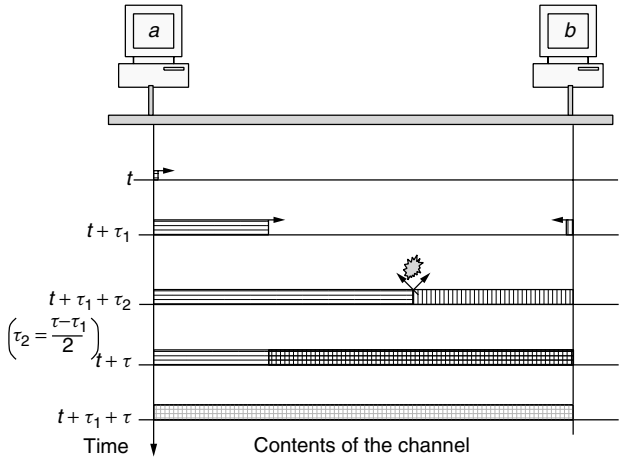


Figure 7. Collision occurrence in CSMA protocol.

takes τ seconds for a signal to be transferred from terminal a to b and vice versa. At time t terminal a senses that the channel is free and starts transmitting a packet. At time $\tau_1 < \tau$ terminal b senses the channel and finds it also free, although the packet from terminal a is well on its way on the channel. Terminal b starts transmitting its own packet, and $(\tau - \tau_1)/2$ seconds later the two packets begin to collide.

From the previous discussion we see that collisions will still occur with the CSMA protocol. However, we expect that the likelihood of a collision will indeed be reduced if the maximum signal propagation delay between two terminals in the system is small relative to the length of a packet. Indeed, this is the case. It can be shown that the throughput of the CSMA protocol is approximately, for small τ/T:

$$S_{\text{CSMA}} \approx \frac{1}{1 + 2(\tau/T)^{1/2}} \tag{4}$$

When $\tau \ll T$, the previous formula shows that S approaches one successful packet per packet duration time, that is, the maximum possible.

Let us examine Eq. (4) more closely. If the length of the packet is B bits and the transmission rate at the channel is C bps, then $T = B/C$. Therefore, we can rewrite (4) as

$$S_{\text{CSMA}} \approx \frac{1}{1 + 2(\tau C/B)^{1/2}} \tag{5}$$

The channel propagation time, τ, is independent of C and B. Therefore, if the network is extended to cover a wider area and as a result τ increases, then the throughput will be reduced. Assume next that we upgrade the channel to a higher transmission rate while maintaining the same arrangement of terminals (i.e., keep τ the same). What will happen to the channel throughput? We need to be careful here since throughput has been defined as the average number of successful packet transmissions per packet length T, and T changes as C varies and B remains constant. An appropriate measure in this case is the average number of successfully transmitted bits per second. This latter measure S^U_{CSMA} is simply related to S_{CSMA}, namely

$$S^U_{\text{CSMA}}(\text{bps}) = \frac{SB}{T} = S_{\text{CSMA}}C \approx \frac{C^{1/2}}{1/C^{1/2} + 2(\tau/B)^{1/2}} \tag{6}$$

where we see that the channel throughput in bits per second increases with C; however, the increase is proportional to $C^{1/2}$ and not C. In fact, *the throughput per channel transmission rate*, S^U_{CSMA}/C, is equal to S_{CSMA}, which decreases as C increases. Also, as seen from (6), for constant C, S^U_{CSMA} increases as the packet length B increases. These considerations should be taken into account when deploying networks operating with the CSMA protocol.

We now turn our attention to the possibility of reducing the time wasted to collisions. Assume that a terminal is able to continue listening to the channel while it transmits its own packet. In case it detects that collision occurred, it interrupts its own transmission and attempts retransmission at a later time. Hence, in general, if a collision occurs, a time interval smaller that the packet duration time will be wasted. In the example of Fig. 7, terminals b and a will detect the collision at times $t + \tau$ and $t + \tau_1 + \tau$, respectively. The CSMA protocol where nodes are interrupting their transmissions when a collision is detected comes by the term *CSMA/CD protocol* (where CD stands for collision detection). The throughput of the CSMA/CD protocol for τ/T small is given approximately by

$$S_{\text{CSMA/CD}} \approx \frac{1}{1 + 5(\tau/T)} \tag{7}$$

Figure 8 shows the throughput of the CSMA and CSMA/CD protocols for various values of $\beta = \tau/T$. We see that both protocols can achieve much higher throughput than the original ALOHA system when β is small. In fact, the throughput can be close to 1. We also see that for the same β, CSMA/CD can achieve significantly better throughput that CSMA. This improvement if due, of course, to the fact that less time is wasted in collisions in CSMA/CD systems than in CSMA.

Up to now we have specified that in case a terminal encounters a collision, it attempts a retransmission at some later random time. What is a good method of selecting such a random time? We discuss here one method that has found wide application. Intuitively, the random retransmission time, R, should depend on the number of backlogged users — the larger the number of backlogged users, the more spread out the distribution of R should be so that the likelihood of avoiding new collisions is reduced. Of course, R should not be too spread out because then terminals will attempt retransmissions rarely and a large portion of time will be left unused. In fact this intuition is correct and can be shown that if the number of backlogged terminals is known and the choice of R is based on this number, the instabilities of the CSMA protocol can be eliminated. However, in real systems the number of backlogged users is seldom known. As an alternative, a terminal may try to obtain an estimate of the number of backlogged users based on its retransmission history. This estimate should increase as the number of collisions encountered during the attempt to transmit a packet increases. Hence the distribution of R should become more spread out as the number of such collisions increases.

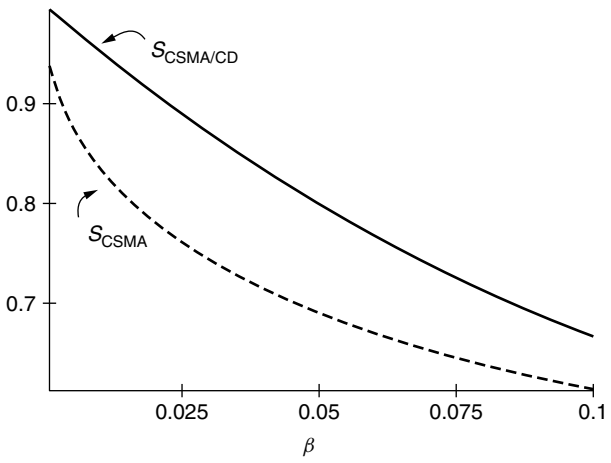


Figure 8. Comparison of CSMA and CSMA/CD protocols.

The previous discussion justifies the following retransmission strategy. If a terminal encounters k collisions during the attempt to transmit a packet, then it attempts a retransmission at time R that is uniformly distributed in the time interval $(0, A2^k)$, where A is a constant. There are various variants of this strategy; however, the main characteristic of all of them is that the "spreading" of R increases exponentially with k. For this reason, this retransmission strategy is known as *exponential backoff*.

3.1.1. Applications of CSMA/CD Protocol. The foremost application of the CSMA protocol is in the technology that connects computer terminals located within a company, an institution, university campus, or other facility using wires. Such a technology is known as *local-area network* (LAN) technology. Several LAN technologies have appeared, but the first and by far the most prevalent one is the Ethernet technology, also known as the IEEE 802.3 LAN technology.

The Ethernet technology was developed in the mid-1970s by Bob Metcalfe and David Boggs. Since then, although it faced challenges by several alternative LAN technologies (Token Ring, FDDI, ATM), it still dominates the marketplace. One of the reasons for this success is that the hardware required for its deployment became very cheap, which, in turn, is due to the large production volume and to the simplicity of the multiple-access protocol used for communication, which is the CSMA/CD protocol with exponential backoff. Moreover, the Ethernet technology proved capable of adapting itself to user demands for increased transmission rates. Currently, Ethernet LANs run at speeds of 10 Mbps, 100 Mbps, and even 1 Gbps.

3.2. The CSMA/CA Protocol

The distributed nature of the CSMA protocol and the low delays it induces when the number of active terminals is small make it a very attractive candidate for wireless communication. However, certain restrictions in such an environment do not permit the direct implementation of the protocol.

Let us recall that in order to be able to implement the CSMA/CD protocol, each terminal needs to be able to perform the following functions:

1. The terminal must be able to listen to the channel and hear whether one or more of the other terminals in the channel are attempting a transmission–carrier sensing capability.
2. The terminal must be able to listen to the channel while transmitting and detect whether its transmission collided with the transmission of some other terminals — collision detection capability.

The collision detection capability implies that a terminal must be able to transmit and receive at the same time, which in a wireless environment can be expensive and is often avoided. Hence, the transmitting terminal may not able to even ensure the correct delivery of its packet. Moreover, as we will see below, even if the collision detection capability exists, it is still possible that a transmitting station does not detect a collision while it is transmitting a packet, but the transmission collides at the receiver. This lack of collision detection capability can be remedied by having the receiver inform the transmitter that the transmitted packet has been correctly received. To do this, the receiving terminal, on correct reception of a packet, sends a short acknowledgment packet back to the transmitter. This packet is referred to as the *ACK message*.

Regarding the carrier-sensing capability of the terminals, while possible, it is not always sufficient to ensure with high probability that the channel is free of transmissions. To understand this problem, we must expand on the special restrictions imposed in a wireless environment. A characteristic of wireless transmission is that terminal a can deliver reliably information to b only if b is within a specified distance from a. Now consider the situation in Fig. 9, where we assume that transmissions are symmetric in the sense that if terminal a can deliver information to b, then b can deliver information to a. The transmission from terminal a can reach b but not c. The transmission from c can reach b but not a. Using the standard CSMA protocol in this environment, certain collisions can still be avoided by sensing the channel. For example, if b is transmitting to a, c can sense the ongoing transmission. However, assume that while a transmits to b, c receives a packet for transmission. If c listens to the channel, it will not hear a's transmission and therefore, if the standard CSMA protocol is employed, a collision will occur. This problem is known as the "hidden terminal" problem. Note that in this case, even if a is able to detect collisions, it will not be able to realize that a collision occurred since it cannot hear c's transmission — as we saw, the latter problem is remedied by the use of the ACK message. Because of the retransmission policy of the basic CSMA protocol, the system can still operate in this environment in spite of the increased number of collisions; however, system throughput may decrease dramatically if packet sizes are large. In fact, plain carrier sensing is not always desirable in this environment. To clarify this point, consider again the situation in Fig. 9. Suppose that b is sending data to a and c wishes to send data to terminal d. If c senses the channel, it will find it busy and therefore will defer transmission. However, since c's transmission cannot reach a, c could in fact deliver its packet to d without colliding with b's

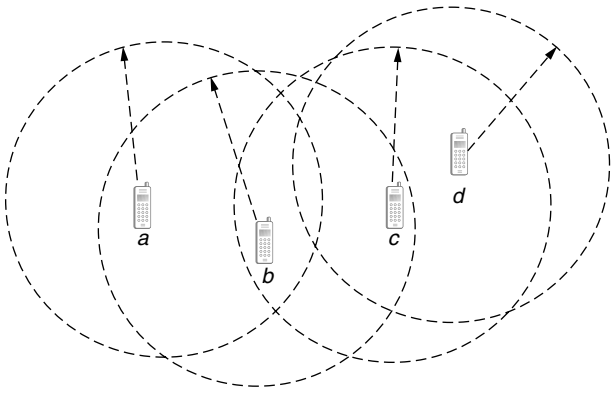


Figure 9. The hidden- and exposed-terminal problems.

transmission. As a result, plain carrier sensing in this case results in reduced utilization of the system. This problem is known as the "exposed terminal" problem.

We next provide a mechanism to address the above-mentioned problems. Two control signals are introduced. These control signals are short messages (compared to packet sizes) that are exchanged between the transmitter and the receiver before the initiation of packet transmission. The first control signal is sent by the transmitter to the receiver and indicates that the transmitter is "*r*equesting *t*o *s*end (RTS)" a packet. The receiver, on correct reception of the RTS, replies that "it is *c*lear *t*o *s*end (CTS)" the packet. Both RTS and CTS signals include a field indicating how long the packet transmission and the accompanied ACK message will last. The terminals now act as follows:

- If a terminal listens to a CTS signal, it waits until the end of the ongoing transmission; this is known since it is included in the CTS signal. It then waits for a random amount of time and attempts to initiate its own transmission process.

Let us see how this rule resolves the hidden-terminal problem. Assume for the moment that the transmission of the CTS and RTS signals is instantaneous, and let us return to the situation in Fig. 9, where a needs to transmit a packet to b. Terminal a sends an RTS to b, and b replies with a CTS signal. Terminal c receives the CTS signal and knows that a transmission has been initiated, so it defers its own transmission. Hence the hidden-terminal problems is alleviated. In effect, the exchange of CTS and RTS messages act as a virtual carrier sensing mechanism.

In fact, the RTS and CTS signals can also be used to address the exposed-terminal problem. Assume that we add the following rule.

- If a terminal listens to an RTS signal but not a CTS signal, it goes ahead with its own transmission, if any.

In Fig. 9 assume that b sends an RTS to a and a replies with a CTS. Terminal c hears the RTS from b but not the CTS from a, and so it knows that its own transmission will not interfere with the b-to-a transmission. Hence it can start its own transmission at any time. Therefore the exposed-terminal problem is avoided.

We assumed above that CTS and RTS signals are instantaneous. Of course, as described in Section 3.1, in a real system transmissions do not take place instantaneously and therefore one cannot assume that the RTS and CTS signals will be received correctly always and free of collisions. However, by now we know that by imposing appropriate retransmission rules the system can deal with occasional loss of RTS or CTS signals. RTS and CTS signals are useful if packet sizes are large. For small packet sizes, it is preferable to go ahead with the packet transmission rather than incurring the overhead of RTS–CTS message exchange.

The modified CSMA system whose principles of operation were described above, comes by the name CSMA/CA, where CA stands for *collision avoidance*. The acronym signifies that collisions are sought to be avoided, but, as we saw, they are not avoided altogether. Because of the retransmission policy of the CSMA system, collisions that may occur are not detrimental; in case of collision, the ACK message or RTS CTS messages will not be received and the transmitting terminal will defer its transmission for a later time. However, if the propagation delays are relatively large and the system is heavily loaded, collisions may degrade the performance of the system.

3.2.1. Applications of CSMA/CA Protocol. The principles of the CSMA/CA protocol have been applied to the specification of the MAC protocol for wireless local-area networks (WLANs), known as the IEEE 802.11 standard.[1] Originally the transmission rates of IEEE 802.11 were 1 and 2 Mbps. The IEEE 802.11b extension to this standard specified 5.5 and 11 Mbps transmission rates, while there is ongoing work that will increase the rate to 20 Mbps. There is currently a great interest in the development of WLAN technologies that support not only high data rates but also multimedia communication such as video, audio, and videoconference communication. The support for multimedia communication imposes additional requirements to the network, such as low packet delays and low packet loss. Networks that are able to provide such support are said to provide quality of service (QoS). CSMA networks were not designed originally to provide QoS. There is a large amount of ongoing works that either attempt to adapt the CSMA protocol to these additional requirements or investigate the feasibility of other approaches.

4. TO PROBE FURTHER

The literature on the ALOHA and the various variants of the CSMA protocols is huge and is still expanding. We do not attempt to provide a detailed account of all the works that contributed to the development of these protocols. Instead we provide some key references to which the interested reader may turn either for a more in-depth study or for a more comprehensive account of related work.

The book by Rom and Sidi [2] provides an in-depth analysis of the ALOHA, CSMA, CSMA/CD, and various other multiple-access protocols. A nice and detailed exposition of the subject can also be found in the book by Bertsekas and Gallager [3]. Very readable accounts of the protocols can be found in the books by Tanenbaum [4] and Kurose and Ross [5]. Information on the IEEE 802.3 and IEEE 802.11 standards and related activities can be found in their Website [6].

BIOGRAPHY

Leonidas Georgiadis received the Diploma degree in electrical engineering from Aristotle University, Thessaloniki, Greece, in 1979, and his M.S. and Ph.D. degrees both in electrical engineering from the University of

[1] Currently the standard does not incorporate a mechanism for dealing with the exposed terminal problem.

Connecticut, in 1981 and 1986, respectively. From 1981 to 1983 he was with the Greek army.

From 1986 to 1987 he was research assistant professor at the University of Virginia, Charlottesville. In 1987, he joined IBM T. J. Watson Research Center, Yorktown Heights, as a research staff member. Since October 1995, he has been with the Telecommunications Department of Aristotle University, Thessaloniki, Greece. His interests are in the area of high-speed networks, scheduling, congestion control, mobile communications, modeling, and performance analysis.

Professor Georgiadis is a senior member of IEEE Communications Society. In 1992, he received the IBM Outstanding Innovation Award for his work on goal-oriented scheduling for multi-class systems.

BIBLIOGRAPHY

1. N. Abramson, The Aloha system — another alternative for computer communications, *Proc. Fall Joint Comput. Conf. AFIPS Conf.*, 1970, p. 37.
2. R. Rom and M. Sidi, *Multiple Access Protocols Performance and Analysis*, Springer-Verlag, 1990.
3. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 2nd ed., 1992.
4. A. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, 1996.
5. J. F. Kurose and K. W. Ross, *Computer Networking, A Top-Down Approach Featuring the Internet*, Addison-Wesley, 2001.
6. *http://standards.ieee.org/getieee802/*.

CDMA/IS95

JHONG SAM LEE
J.S. Lee Associates, Inc.
Rockville, Maryland

LEONARD E. MILLER
Wireless Communications Technologies Group, NIST
Gaithersburg, Maryland

The IS95 cellular telephone standard [1] was designed as a second-generation (digital) system. Like the U.S. analog (first-generation) cellular system, known as the *Advanced Mobile Phone System* (AMPS) [2], the IS95 system uses frequency-division duplexing (FDD), with a 25-MHz bandwidth in each direction over the frequency allocations shown in Fig. 1. The cellular band is further divided equally between two service providers, known as the "A" (wire) and the "B" (nonwire) carriers, as illustrated in Fig. 1. In AMPS, each channel occupies 30 kHz of bandwidth in a frequency-division multiple access (FDMA) system, using analog frequency modulation waveforms. The frequencies that are used in one cell area are reused in another cell area at a distance such that mutual interference gives a carrier-to-interference power ratio of no less than 18 dB. Given this performance requirement and the fact that in the mobile radio environment the attenuation of carrier power usually is proportional to the fourth power of the distance from the emitter to a receiver, the analog cellular system utilizes seven-cell clusters, implying a frequency reuse factor of 7 [2]. The resulting capacity of AMPS is then just one call per 7×30 kHz $= 210$ kHz of spectrum in each cell, and in the total of 12.5 MHz allocated there can be no more than 60 calls per cell.

In 1988, the Cellular Telecommunications Industry Association (CTIA) stipulated requirements for the second-generation digital cellular system technology. The key requirements included a 10-fold increase in call capacity over that of AMPS, a means for call privacy, and compatibility with the existing analog system. The compatibility requirement arose from the fact that the second-generation system must operate in the same band as AMPS.

Proposed in 1989, the first U.S. standard for a second-generation cellular system was published in 1992 as IS54 [3] and adopted a time-division multiple access (TDMA) technology. The IS54 TDMA digital cellular system employs digital voice produced at 10 kbps (8 kbps plus overhead) and transmitted with $\pi/4$ differentially encoded quadrature phase shift keying ($\pi/4$ DQPSK) modulation. Because the IS54 system permits 30 kHz/10 kbps $= 3$ callers per 30-kHz channel spacing, the increase of capacity over AMPS is only a factor of 3 (180 calls per cell).

In 1990, Qualcomm, Inc. proposed a digital cellular telephone system based on code-division multiple access (CDMA) technology [4], which was adopted in 1993 as a second U.S. digital cellular standard, designated IS95 and later known as cdmaOne. Using spread-spectrum (SS) transmission techniques, the IS95 system provides a very high capacity. The full title of the IS95 CDMA standard is *Mobile Station-Base Station Compatibility Standard for*

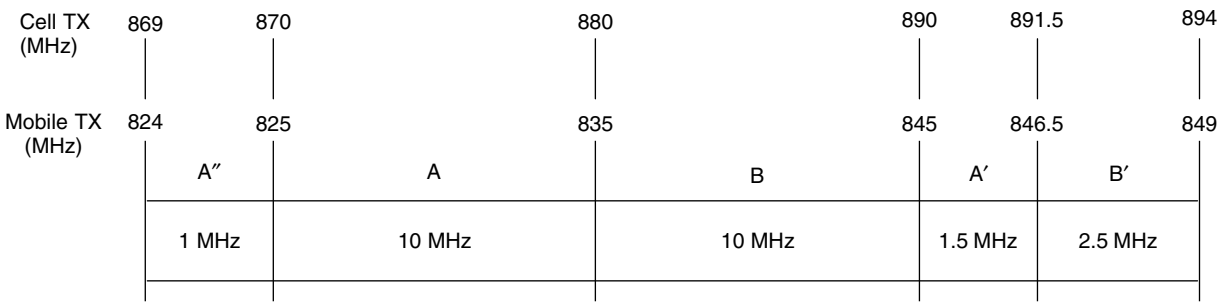


Figure 1. Cellular bands in the United States.

Dual-Mode Wideband Spread Spectrum Cellular System, indicating that the document is a common air interface (CAI) that does not specify the details of how the system is to be implemented.

1. WHAT IS CDMA?

Spread-spectrum techniques involve the transmission of a signal in a bandwidth substantially greater than the information bandwidth to achieve a particular operational advantage. How SS signals can be processed to yield gains for interference rejection can be understood by calculating the jamming margin for a SS system. Let the following parameters be defined:

S = received power for the desired signal in watts
J = received power for undesired signals in watts (jamming, other multiple access users, multipath, etc.)
$R = 1/T_b$ = data rate (data signal bandwidth in Hz)
W = spread bandwidth in Hz
E_b = received energy per bit for the desired signal in W · s (watt-seconds)
N_0 = equivalent noise spectral power density in W/Hz

Then the ratio of the equivalent "noise" power J to S is

$$\frac{J}{S} = \frac{N_0 W}{E_b/T_b} = \frac{WT_b}{E_b/N_0} = \frac{W/R}{E_b/N_0}$$

When E_b/N_0 is set to the value required for acceptable performance of the communications system, then the ratio J/S bears the interpretation of a jamming margin:

$$J/S = \text{tolerable excess of interference over desired signal power} = \frac{W/R}{(E_b/N_0)_{\text{req}}}$$

or

$$\text{Margin (dB)} = \frac{W}{R}(\text{dB}) - \left(\frac{E_b}{N_0}\right)_{\text{req}} (\text{dB})$$

The quantity W/R is called the SS *processing gain* (PG). For example, if the information bandwidth is $R = 9600$ Hz, corresponding to digital voice, the transmission bandwidth is $W = 1.2288$ MHz, and the required SNR is 6 dB, then the PG equals $128 = 21.1$ dB and the jamming margin equals $32 = 15.1$ dB. The communicator can achieve an SNR of at least 6 dB even in the face of interference (jamming) power in excess of 32 times the signal power, due to the PG. In a CDMA communications system, the cochannel communicators, occupying the same bandwidth simultaneously, account for the interference (jamming) power. If every user supplies the identical amount of signal power to the base station antenna through a perfect power control scheme, regardless of location, then for this example 32 other multiple-access users can be accommodated by a CDMA system.

The capacity of a CDMA system is proportional to the PG of the system. This fact may be illustrated as follows, assuming first that the system in question is isolated from all other forms of outside interference (i.e., a single cell): The carrier power equals $C = S = R \times E_b$ and, analogous to the jamming power, the interference power at the base station receiver may be defined as $I = W \times N_0$, where W is the transmission bandwidth and N_0 is the interference power spectral density. Thus a general expression for the carrier-to-interference power ratio for a particular mobile user at the base station is given by

$$\frac{C}{I} = \frac{RE_b}{WN_0} = \frac{E_b/N_0}{W/R}$$

where the PG is W/R. Let M denote the number of mobile users. If power control is used to ensure that every mobile has the same received power at the base station, then the interference power caused by the $M - 1$ interferers is $I = C(M - 1)$. Substituting for I in the previous equation, neglecting thermal noise, and solving for M, the capacity for a CDMA system is found to be

$$M = \frac{W/R}{E_b/N_0} + 1 \approx \frac{W/R}{E_b/N_0}$$

Thus, the capacity of a CDMA system is proportional to the PG. This PG is based on the fact that in the CDMA receiver the (multiple) interfering users' signals continue to have the bandwidth W, while the (single) selected user's signal is despread by the removal of the spreading code. The receiver then can perform filtering at the selected user's despread (baseband) bandwidth, which admits only a small fraction of the energy received from the interfering user signals.

It is possible in a digital telephone system to exploit pauses in speech to reduce the transmission rate or to use intermittent transmissions with an effective duty cycle of 40–50%. If the duty cycle of a speech traffic channels in the CDMA system is denoted by the variable α, then effectively the data rate is αR instead of R.

If the base station employs directional antennas that divide the cell into sectors, each antenna will receive only a fraction of the interference. In practice, the coverage areas of the receiving antennas overlap by approximately 15%. Standard implementations divide the cell into three sectors, providing an effective capacity increase of $G = 3 \times 0.85 = 2.55$.

Signals originating in other cells must be taken into account when determining the capacity of a particular "home" cell; such interference is, of course, diminished by the attenuation incurred by the interferers in propagating to the home cell. Simulations performed by Qualcomm have shown that interference from other cells accounts for only about 35% of that received at the base station [5–7]. On the basis of this information, the equation for CDMA capacity may be modified to include a reuse efficiency F_e to account for other-cell interference.

Taking into account voice duty cycle, antenna gain, and other-cell interference, the equation for CDMA capacity becomes

$$M \approx \frac{W/R}{E_b/N_0} \times \frac{GF_e}{\alpha}$$

For example, using realistic parameters for the IS95 system ($W/R = 128$, $E_b/N_0 = 7$ dB, $\alpha = 0.5$, $G = 2.55$, and $F_e = 0.65$), the capacity of the system is 85 users per cell. The achievement of this capacity over the mobile radio channel in practice depends on many factors [5,8,9].

2. THE IS95 SYSTEM

In the IS95 system, the mobile stations communicate with base stations over "forward" (base-to-mobile) and "reverse" (mobile-to-base) radio links, also sometimes called "downlink" and "uplink," respectively. As suggested in Fig. 2, the radiocommunications over these links are organized into different channels: *pilot, synchronization, paging*, and *traffic* channels for the forward link; and *access* and *traffic* channels for the reverse link.

Unlike an FDMA cellular system, a CDMA cellular system does not require the use of "clusters" of cells to enforce a minimum reuse distance between cells using the same frequency channels in order to control the amount of cochannel interference. Each CDMA cell uses the identical spectrum and employs pseudorandom noise (PN) code SS modulation and utilizes the resulting PG to overcome interference. The PN code signaling rate, known as the "chip rate," was initially chosen by Qualcomm to be 1.2288 megachips per second (Mchips/s), which is an integer multiple (128) of the maximum digital voice bit rate of 9600 bps, thereby requiring a spectrum occupancy of about 1.23 MHz. The particular choice of the chip rate in IS95 was presumably dictated in part by a desire to operate a single CDMA channel in the 1.5-MHz band designated as "A" in Fig. 1; Gilhousen argues that the bandwidth selected is a good choice in terms of the characteristics of the mobile channel (such as coherence bandwidth) and the complexity of a "RAKE" multipath receiver designed for that channel [10].

2.1. System Synchronization

Each base station of the IS95 system is required to maintain a clock that is synchronized to Universal Coordinated Time (UTC), indirectly through synchronization to GPS time signals. The known beginning of the GPS timecount (time 00:00:00 on Jan. 6, 1980) is traceable to UTC and is required to be aligned in a particular fixed way with the PN codes used by every base station in the CDMA system — two "short" PN codes having 26.66-ms periods, and a "long" PN code that has a period that is over 41 days long, all running at the 1.2288-Mchip/s rate. The synchronization of time standards among the CDMA base stations is necessary because each base station transmits on the same center frequency and uses the same two short PN codes to spread its (forward-link) waveform, the different base station signals being distinguished at a mobile receiver only by their unique short PN code starting positions (phase offsets), as illustrated in Fig. 3.

The system time reference for a particular mobile is established when it acquires a certain base station signal, usually that from the nearest base station, affiliates with that base station, and reads the synchronization message broadcast by that base station. The message contains information that enables the mobile unit to synchronize its long PN code and time reference with those of that particular base station.

2.2. Forward-Link Summary

The forward-link channel structure consists of the transmission of up to 64 simultaneous, distinct channels that are orthogonally multiplexed onto the same RF carrier. One of these channels is a pilot signal that is transmitted continuously, to be received by the mobiles as a coherent phase reference. Another of these channels is a continuously transmitted synchronization channel that is used to convey system information to all users in the cell. Up to

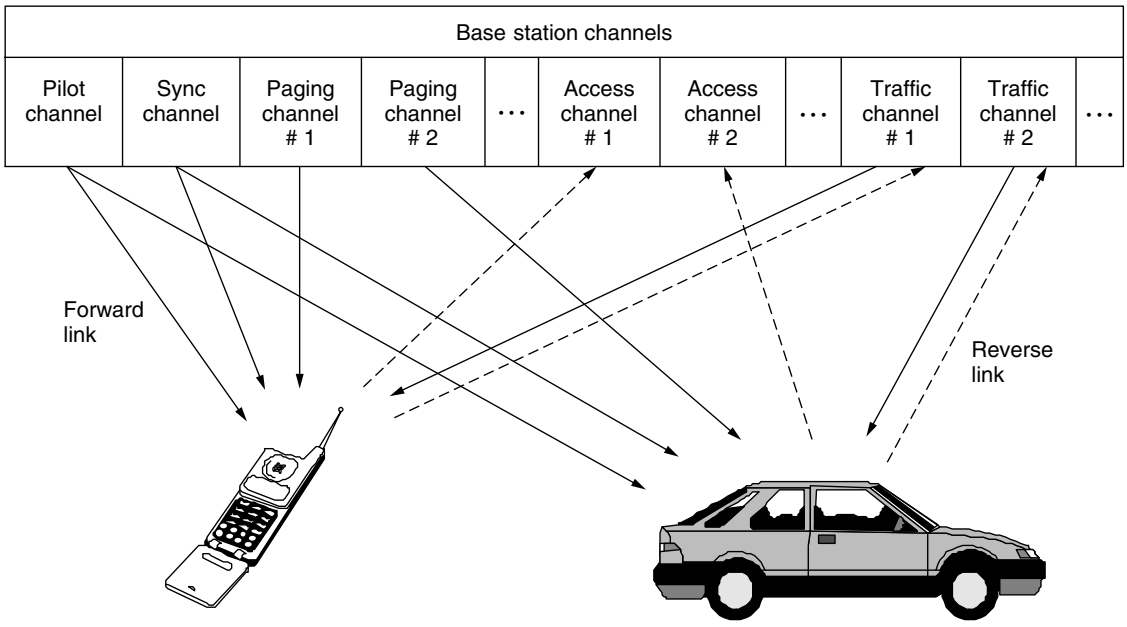


Figure 2. IS95 forward- and reverse-link channels.

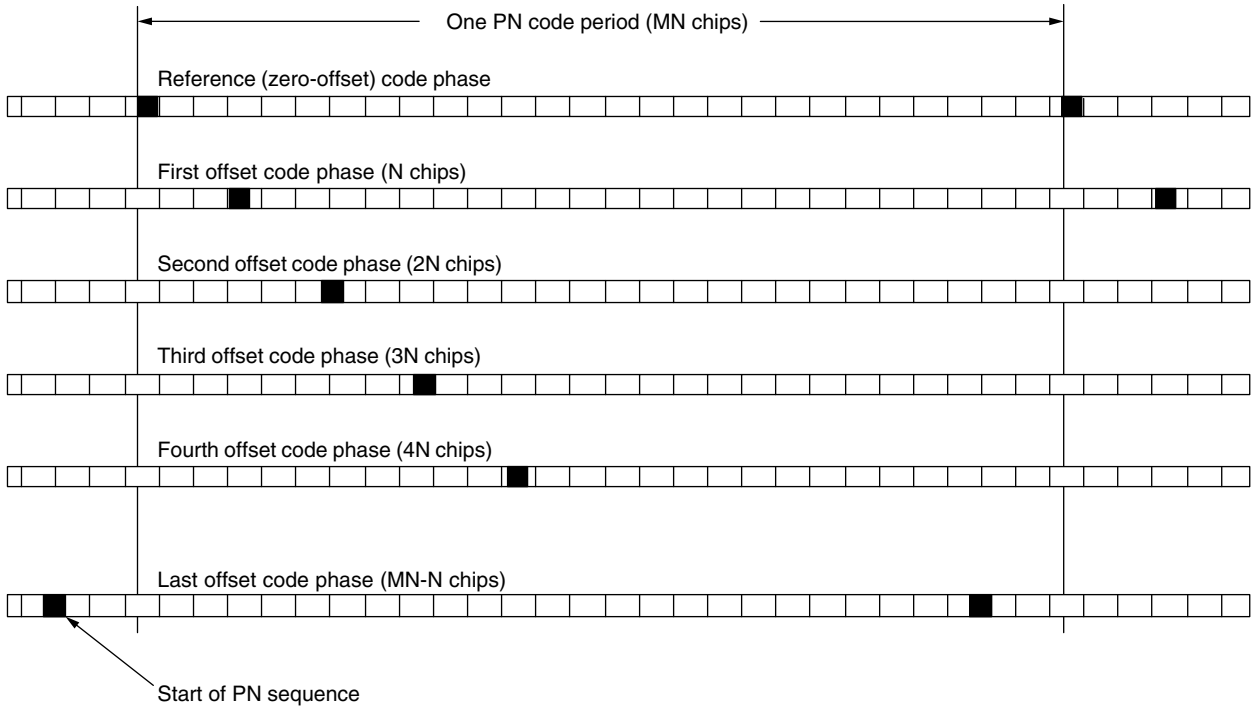


Figure 3. Short PN code offsets are assigned to different base stations.

seven paging channels are used to signal incoming calls to mobiles in the cell and to convey channel assignments and other signaling messages to individual mobiles. The remainder of the channels are designated as traffic channels, each transmitting voice and data to an individual mobile user.

The major features of the IS95 forward-link CAI are as follows:

- *Multiplexing*. The forward-link channelization is based on an orthogonal code-division multiplexing scheme using an orthogonal set of "subcarrier" digital waveforms known as *Walsh functions* [5,11].
- *Interference Rejection*. The forward-link waveform is modulated by direct-sequence PN code SS techniques to isolate the signals received from a particular base station and to discriminate against signals received from other base stations.
- *Modulation*. The forward-link waveform features modulation of I (cosine) and Q (sine) RF carriers by different PN-coded bipolar ($\pm$) baseband digital data streams, thereby generating a form of quaternary phase shift keying (QPSK).
- *Pulseshaping*. The shape of the baseband digital pulses in the I and Q output channels is determined by a finite impulse response (FIR) filter that is designed to control the spectrum of the radiated power for minimal adjacent-frequency interference [5].
- *PN Chip Rate*. The PN code chip rate, which is 1.2288 Mchips/s, is 128 times the maximal source data rate of 9.6 kbps, thereby providing a PG of 21 dB.
- *Effective Bandwidth*. For the PN chip rate and FIR filter spectrum control specified, the energy of the IS95 forward-link signal is almost entirely contained in a 1.25-MHz bandwidth.
- *Voice Coding*. A variable-rate vocoder is specified, with data rates 1200, 2400, 4800, and 9600 bps depending on voice activity. In 1996, a Personal Communications Services (PCS) version of the system specified in IS95 was released [12] with data rates of 14.4 kbps and fractions thereof; this set of rates also became available in the cellular standard revision known as IS95A.
- *Error Control Coding*. The forward link uses rate $\frac{1}{2}$ constraint length 9 convolutional coding, with Viterbi decoding.
- *Interleaving*. To protect against burst error patterns (a distinct possibility on the fading mobile communications channel), the forward link interleaves code symbols before transmission, using a 20-ms span.

The baseband data rate from each channel being multiplexed varies; the highest rate is 19.2 kilosymbols per second (ksps). The polarity of each channel's baseband data symbol multiplies an assigned 64-chip Walsh sequence that is repeated at the 19.2-ksps rate. Thus, the orthogonally multiplexed combination of forward-link channels forms a baseband data stream with a rate of 64×19.2 ksps $= 1.2288$ Mchips/s (see also Table 1). The orthogonal multiplexing operations on the forward link are shown in Figs. 4 and 5. Each channel is modulated by a channel-specific Walsh sequence, denoted $H_i, i = 0, 1, \ldots, 63$. The IS95 standard assigns H_0 for the pilot channel, H_{32} for the synchronization channel, H_1–H_7 for

Table 1. Forward Traffic Channel Modulation Parameters

Parameter	Value				Units
Data rate	9600	4800	2400	1200	bps
PN chip rate	1.2288				Mchips/s
Code rate	$\frac{1}{2}$				Bits/code symbol
Code repetitions	1	2	4	8	Modulation symbol/code symbol
Modulation symbol rate	19,200				sps
Code symbol energy	$E_b/2$	$E_b/4$	$E_b/8$	$E_b/16$	
PN chips/modulation symbol	64				N/A
PN chips/bit	128	256	512	1,024	

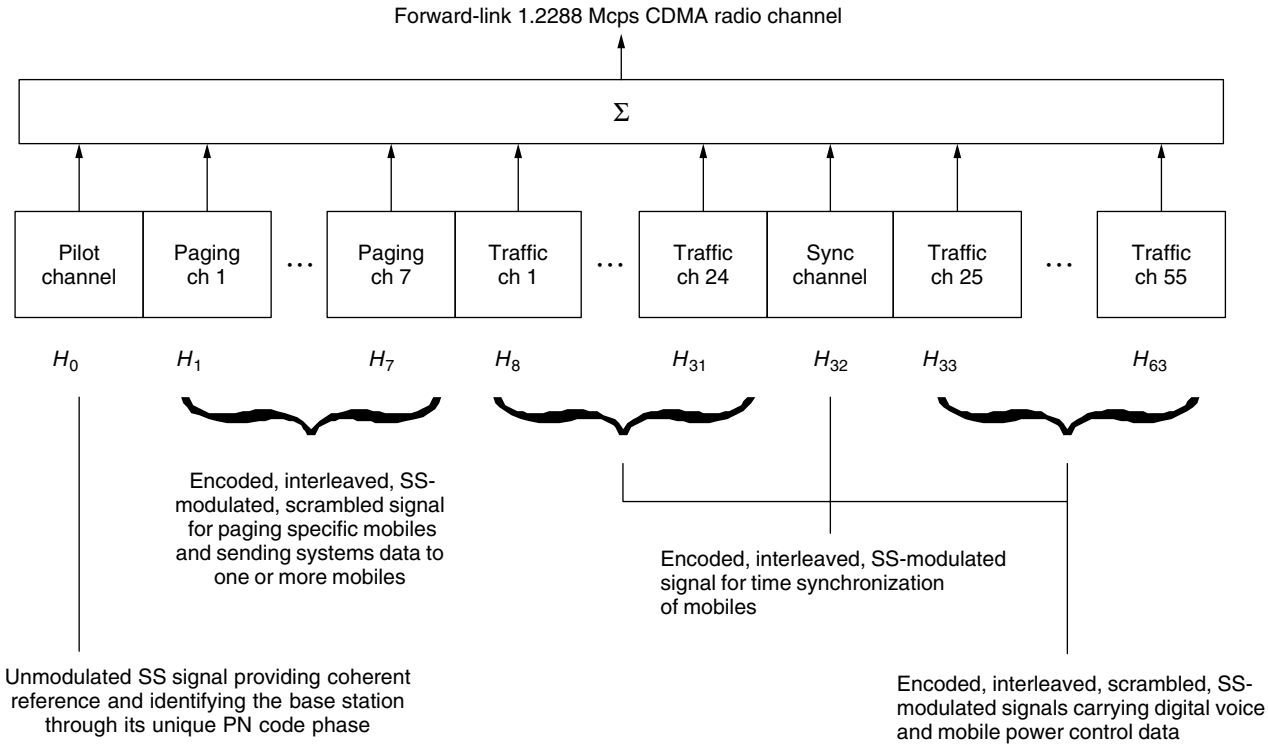


Figure 4. Forward-link channel assignments.

the paging channels, and the remainder of the H_i to the traffic channels. The multiplexed data stream for each channel is combined separately with two different short PN codes that are identified with I- and Q-quadrature carrier components.

The I- and Q-channel PN codes may be denoted by $\mathrm{PN}_I(t, \theta_i)$ and $\mathrm{PN}_Q(t, \theta_i)$, respectively, and are generated by 15-stage linear feedback shift registers (LFSRs). The parameter θ_i denotes the PN code offset phase assigned to a particular base station. Thus, unlike conventional QPSK, which assigns alternate baseband symbols to the I and Q quadratures, the IS95 system assigns the same data to both quadrature channels, each of which pseudorandomly preserves or inverts the data polarity. It is common to speak of these operations as "quadrature spreading." The two short distinct PN codes are maximum-length sequences and are lengthened by the insertion of one chip per period in a specific location in the PN sequence. Thus, these modified short PN codes have periods equal to the normal sequence length of $2^{15} - 1 = 32{,}767$ plus one chip, or 32,768 chips. At a rate of 1.2288 Mchips/s, the I and Q sequences repeat every 26.66 ms, or 75 times every 2 s.

The synchronization channel is demodulated by all mobiles (mobile units) and contains important system information conveyed by the sync (synchronized) channel message, which is broadcast repeatedly. This signal identifies the particular transmitting base station and conveys long PN code synchronization information, at a rate of 1.2 kbps, or 32 sync channel data bits per 26.66-ms sync channel frame and 96 bits per 80-ms sync channel "superframe." The sync channel frame length is equal to one period of the short PN codes, and the sync channel frames are in fact aligned with those periods so that, once the mobile has acquired a particular base station's pilot signal, the mobile automatically knows the basic timing structure of the sync channel.

After synchronization has been accomplished, the mobile can receive the paging channel and transmit on

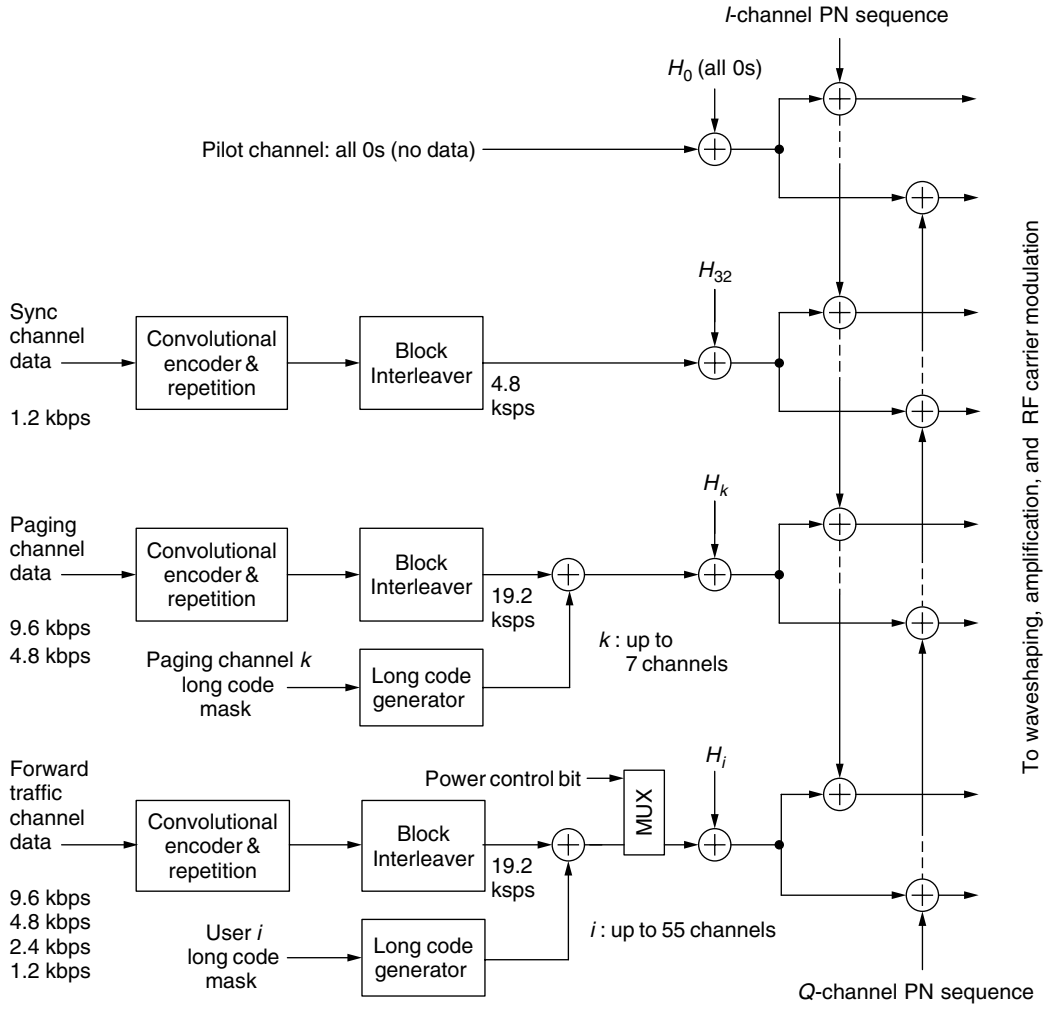


Figure 5. Forward-link multiplexing operations.

the access channel. Because all channels except the pilot and sync channels are scrambled using the long PN code, the sync channel message is necessary to correctly demodulate any other channel. The beginning of the paging and traffic channel frames coincide with the start of a sync channel superframe.

The paging channels are used to alert the mobile to incoming calls, to convey channel assignments. Paging information is generated at either 9.6 or 4.8 kbps. The information is convolutionally encoded, repeated, and interleaved. The repetition of the code symbols is adapted to the data rate to fix the rate of symbols being interleaved at 19.2 ksps. Other than for the sync channel, the data frames for the IS95 channels are 20-ms in length, and the interleaving is performed on a frame-by-frame basis. All the paging channel modulation symbols are transmitted at the same power and baseband data rate for a given CDMA system.

Unlike the sync channel, the encoded and interleaved paging channel symbols are scrambled (multiplied by a random sequence at the same rate) with a 42-stage long-code PN sequence running at 1.2288 Mchips/s that is decimated to a 19.2-ksps rate by sampling every 64th PN code chip. The long PN code is generated by a 42-stage shift register, with a period of $2^{42} - 1 \approx 4.4 \times 10^{12}$ chips (lasting over 41 days at 1.2288 Mchips/s). A phase offset of the original long PN code sequence that is unique to the particular paging channel and base station is obtained by combining the outputs of the shift register stages selected by a 42-bit mask. Details of the use of masks to shift PN codes and the effect of decimation on the codes are given in Ref. 5.

In IS95, each active paging channel has a number of periodically recurring message slots (e.g., 2048 slots) available for transmitting pages and other base-to-mobile messages. When a message is queued up for a particular mobile, the base station, using a hash function, pseudorandomly selects one of the paging channels and pseudorandomly selects one of the message slots in that paging channel for transmission to the particular mobile. The mobile knows exactly which paging channel and message slot to monitor for possible messages because the pseudorandom selection is based on its own identification number and known system parameters. The purpose of the hash function is to distribute the message traffic evenly among the paging channels and message slots. Details of the hash functions used in IS95 are given in Ref. 5.

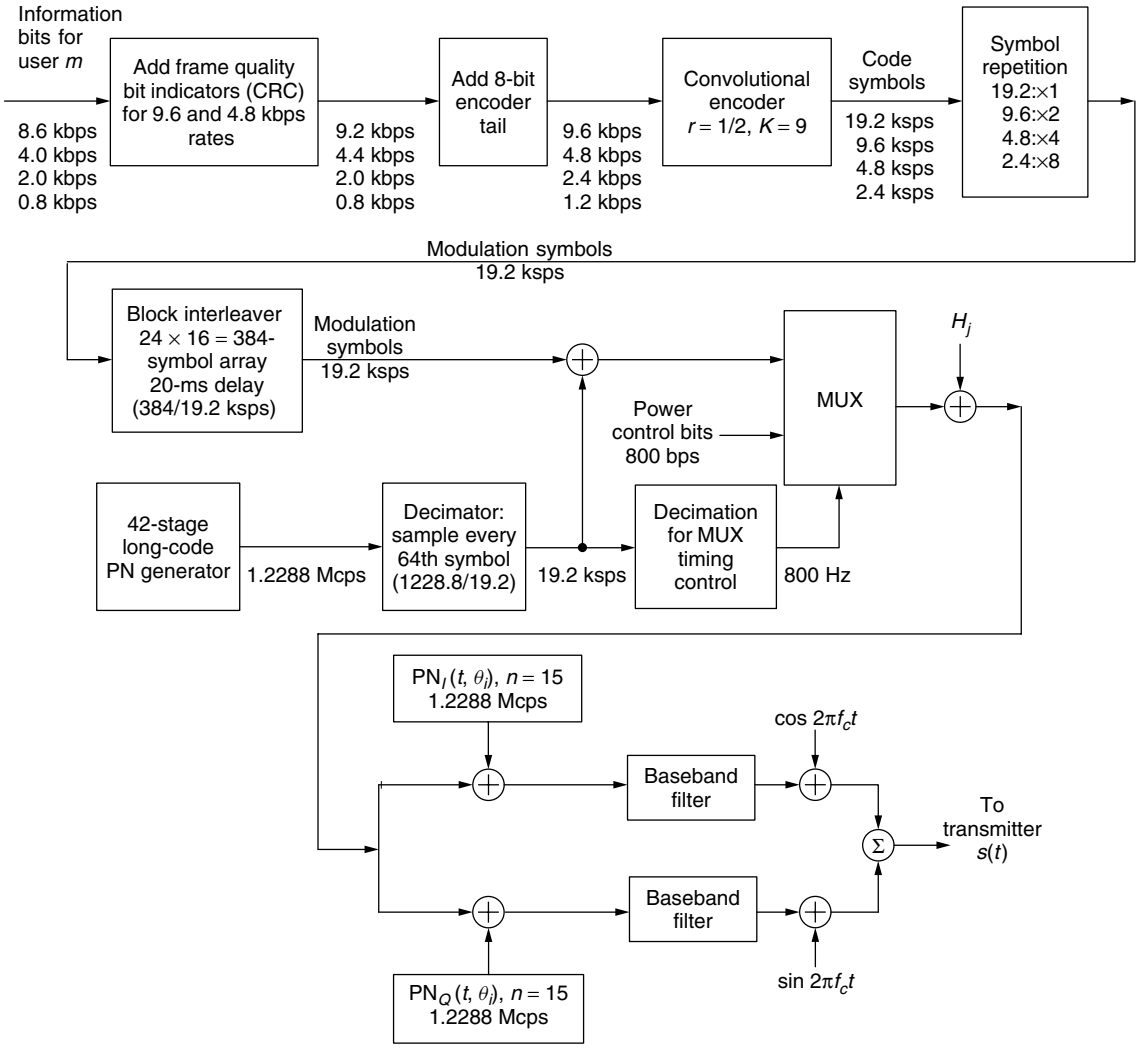


Figure 6. Forward traffic channel modulation.

A block diagram for the forward traffic channel modulation is given in Fig. 6. As shown in the diagram, voice data for the mth user is encoded on a frame-by-frame basis using a variable-rate voice coder, which generates data at 8.6, 4.0, 2.0, or 0.8 kbps depending on voice activity, corresponding respectively to 172, 80, 40, or 16 bits per 20-ms frame. A cyclic redundancy check (CRC) error-detecting code calculation is made at the two highest rates, adding 12 bits per frame for the highest rate and 8 bits per frame at the second highest rate. At the mobile receiver, which one of the possible voice data rates is being received is determined in part from performing similar CRC calculations, which also provide frame error reception statistics for forward-link power control purposes.

In anticipation of convolutional coding on a block basis (code symbols in one frame not affecting those in adjacent frames), a convolutional encoder "tail" of 8 bits is added to each block of data to yield blocks of 192, 96, 48, or 24 bits per frame, corresponding to the data rates of 9.6, 4.8, 2.4, and 1.2 kbps going into the encoder. Convolutional encoding is performed using a rate $\frac{1}{2}$, constraint length 9 code, resulting in coded symbol rates of 19.2, 9.6, 4.8, and 2.4 ksps.

Coded symbols are repeated as necessary to give a constant number of coded symbols per frame, giving a constant symbol data rate of 19.2 ksps (i.e., 19.2 ksps × 1, 9.6 ksps × 2, 4.8 ksps × 4, 1.2 ksps × 8). The 19.2 ksps × 20 ms = 384 symbols within the same 20-ms frame are interleaved to combat burst errors due to fading.

Each traffic channel's encoded voice or data symbols are scrambled to provide voice privacy by a different phase offset of the long PN code, decimated to yield 19.2 kchips/s. The scrambled data are punctured (overwritten) at an average rate of 800 bps by symbols that are used to control the power of the mobile station.

Note from Fig. 6 that, regardless of the data rate, the modulated channel symbol rate must be 19.2 ksps. This is accomplished by means of code symbol repetition for rates less than the 9.6-ksps data rate.

2.3. Reverse-Link Summary

The IS95 reverse link channel structure consists of access channels and traffic channels. To reduce interference and

save mobile power, a pilot channel is not transmitted on the reverse link. A mobile transmits on either an access or a traffic channel but never both at the same time.

The major features of the IS95 reverse link CAI are as follows:

- *Multiple Access*. The reverse-link channelization is based on a conventional SS PN CDMA scheme in which different mobile users are distinguished by distinct phase offsets of the 42-stage-long PN code, which serve as user addresses.
- *Quadrature Spreading*. In addition to the long PN code, the reverse-link datastream is direct-sequence modulated in quadrature by the same two short PN codes as on the forward link; each mobile station in each cell uses the reference or zero-offset phases of these two codes.
- *Modulation*. The reverse link waveform features 64-ary orthogonal modulation using sequences of 64 chips to represent six binary data symbols. The quadrature modulation of I (cosine) and Q (sine) RF carriers by the two different PN-coded bipolar ($\pm$) baseband digital datastreams, with the Q-quadrature stream delayed by half a PN chip, generates a form of offset quaternary phaseshift keying (OQPSK).
- *Pulseshaping*. The shape of the baseband digital pulses in the I and Q output channels is determined by a FIR filter that is designed to control the spectrum of the radiated power for minimal adjacent-frequency interference.
- *PN Chip Rate*. The PN code chip rate, which is 1.2288 Mchips/s, is 128 times the maximal source data rate of 9.6 kbps.
- *Acquisition*. The base station's acquisition and tracking of mobile signals is aided by the mobile's transmission of a preamble containing no data.
- *Voice Coding*. A variable-rate vocoder is specified, with data rates 1200, 2400, 4800, and 9600 bps depending on voice activity in a particular 20-ms frame. The transmission duty cycle of the reverse link signal during a call is proportional to the data rate.
- *Error Control Coding*. The reverse-link uses rate $\frac{1}{3}$ constraint length 9 convolutional coding, with Viterbi decoding.
- *Interleaving*. To protect against possible burst-error patterns, the reverse link interleaves code symbols before transmission, using a 20-ms span.

There is at least one reverse-link access channel for every paging channel on the forward link, with a maximum of 32 access channels per paging channel. The access channels are used for the mobile to initiate a call or respond to a page or information request from the base station. The number of active reverse-link traffic channels is equal to the number of active forward-link traffic channels. Each reverse link channel is distinguished by a distinct phase offset of the same 42-stage long-code PN sequence used on the forward link. The channels as received at the base station are illustrated in Fig. 7, where n (the number of paging channels) and m (the number of traffic channels) are limited by interference.

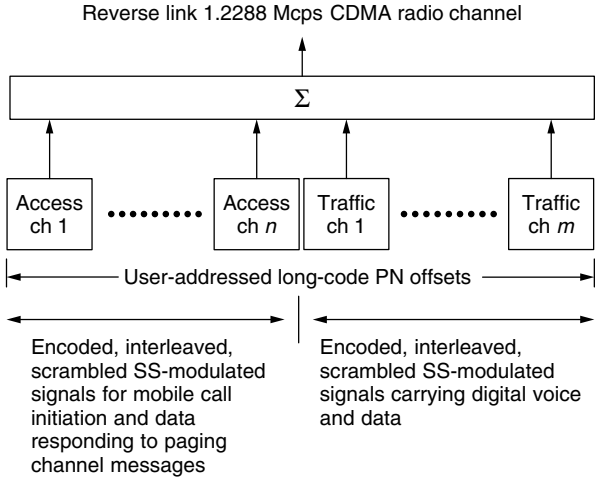


Figure 7. Reverse-link channel assignments at the base station.

The reverse-link transmitter consists of a convolutional encoder and modulator, and a quadrature PN-spreading modulator. The quadrature modulator for the reverse link is different from that used on the forward link in that a half-chip delay is inserted in the quadrature-phase (Q) channel to achieve a form of offset-QPSK (OQPSK) modulation. The one-half chip offset eliminates phase transitions through the origin to provide a modulation scheme that gives a relatively constant envelope. The transmission of the same data by means of a two-quadrature modulation scheme is a form of diversity. Its analytical justification in Refs. 5 and 13 shows that the QPSK CDMA system has a 3-dB advantage over BPSK CDMA system in terms of intersymbol interference performance and also has cochannel interference advantages.

Figure 8 is a block diagram of traffic channel processing. A variable rate vocoder is used to generate a digital voice signal at a rate varying from 0.8 to 8.6 kbps in a given 20-ms traffic channel frame. Depending on the data rate, the data frame is encoded with a CRC block code to enable the base station receiver to determine whether the frame has been received with error. An 8-bit encoder tail is added to the frame to ensure that the convolutional encoder, which follows, is reset to the all-zero state at the end of the frame. These operations result in data rates of 9600 (full rate), 4800 (half rate), 2400 ($\frac{1}{4}$ rate), or 1200 ($\frac{1}{8}$ rate) bps with, respectively, 192, 96, 48, or 24 bits per frame. The frame is then convolutionally encoded at a $\frac{1}{3}$ rate, resulting in $3 \times 192 = 576$ code symbols per frame at full rate, or 28.8 ksps. For other voice data rates, the code symbols are repeated as necessary to cause each rate to input the same number of code symbols to the interleaver in a frame.

Each group of six consecutive encoded symbols out of the interleaver is used to select a 64-chip Walsh sequence for orthogonal modulation, with a chip rate of $28.8 \times 64/6 = 307.2$ kchips/s. Because of the way that the symbols are read out from the interleaver array, these modulation symbols occur in alternating groups of six modulation symbols and $6(n-1)$ repeated modulation symbols, where n is the order of repetition. Altogether in a frame interval there are

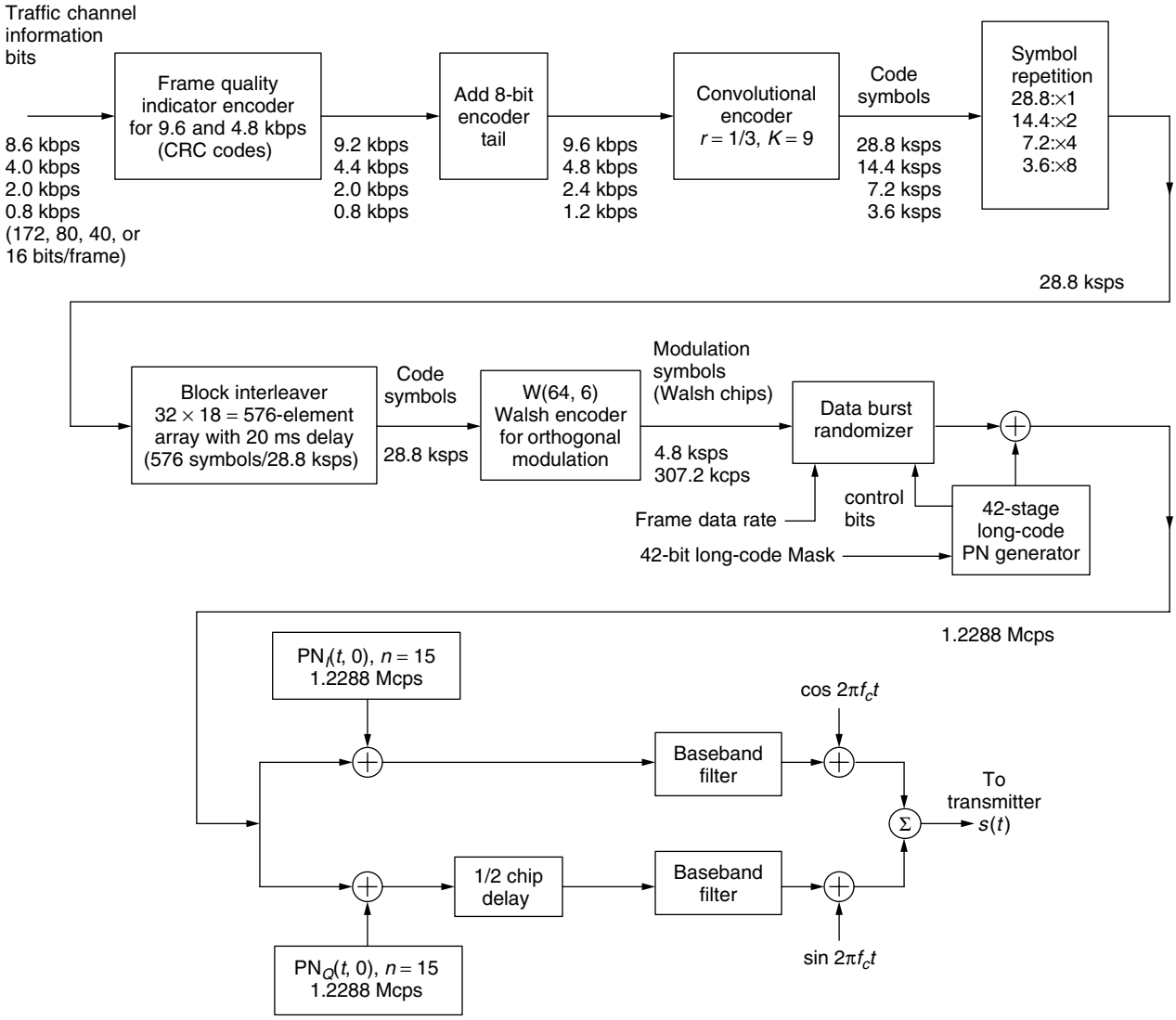


Figure 8. Reverse traffic channel modulation.

$96 \div 6 = 16$ groups of six orthogonal modulation symbols, each composed of $6 \times 64 = 384$ Walsh chips.

To reduce the average amount of reverse link interference and thereby increase user capacity, on reverse link transmissions, repeated symbols are gated off. A "data burst randomizer" is used to select in a pseudorandom manner which of the groups of six symbols are transmitted, based on "control bits" that are values of long PN code sequence bits at a certain time.

The user-distinct offset of the 42-stage long PN code is used to further spread the signal and ensure that the channels can be distinguished. The offset is implemented using a mask that depends on the electronic serial number (ESN) of the mobile. The masks and offsets on both forward and reverse traffic channels are identical for a given mobile user. The modulation parameters of the reverse traffic channel are summarized in Table 2.

Transmissions on the reverse traffic channel begin with a preamble of all-zero data frames to aid the base station in acquiring the signal. Signaling messages from the mobile to the base station may be sent on the reverse traffic channel as well as the access channel. When a message is to be sent, it can be sent in a "dim and burst" mode during periods of active speech, in which a portion of the voice data in a frame is overwritten by the message data, or in a "blank and burst" mode during periods of speech inactivity, in which all the data in the frame are message data.

2.4. Diversity Features of the IS95 System Design

Special features of the design of IS95 and of its common implementations are described in Ref. 5. Here we mention that the digital SS design of the IS95 forward- and reverse-link waveforms permits the use of several forms of diversity in addition to the time diversity inherent in the repetition, encoding, and interleaving of the data symbols. These forms of diversity include multipath diversity and base station diversity; the latter is available with or without the prospect of handing off the call to a different base station.

Table 2. Reverse Traffic Channel Modulation Parameters

Parameter	Value				Units
Data rate	9600	4800	2400	1200	bps
PN chip rate	1.2288				Mchips/s
Code rate	1/3				bits/code symbol
Transmit duty cycle	100	50	25	12.5	%
Code symbol rate	28,800				sps
Modulation rate	6				Code sym/mod sym
Mod symbol rate	4800				sps
Walsh chip rate	307.2				Kchips/s
Mod symbol duration	208.33				μs
PN chips/code symbol	42.67				—
PN chips/modulation symbol	256				—
PN chips/Walsh chip	4				—

Because the IS95 waveform for a particular channel is a dual-quadrature direct-sequence SS signal, it is possible to employ SS correlation techniques to isolate a single multipath component of that channel's signal and to discriminate not only against other channels' signals but also against multipath components of the same channel's received signal. Using the RAKE technique [14], in which the receiver uses several parallel receiver "fingers" to isolate multipath components, on the forward link it is possible to extract several multipath components from the total received signal and to align them for optimal combining.

If, at a particular mobile station, another cell site pilot signal becomes significantly stronger than the current pilot signal, the control processor initiates handoff procedures during which the forward links of both cell sites transmit the same call data to that mobile, which uses different fingers to process the two base station signals. With both sites handling the call, additional space diversity is obtained. When handoff is not contemplated, in a cell site diversity mode the strongest paths from multiple cell sites are determined by a search receiver, and the digital data receivers in the RAKE fingers are assigned to demodulate these paths. The data from multiple digital receivers are combined for improved resistance to fading.

Soft handoff methods have several advantages over conventional hard handoff methods [15]. Contact with the new base station is made before the call is switched, which prevents the mobile from losing contact with the system if the handoff signal is not heard or incorrectly interpreted. Diversity combining is used between multiple cell sites, allowing for additional resistance to fading.

2.5. System Evolution Toward Third Generation

The IS95B standard was published in 1999 [16]. This revision of the IS95 CAI features new multiplexing options that provide for transmission of up to eight simultaneous (parallel) full-rate "code channels" to constitute the forward- or reverse-link traffic channels. On the forward link, active users are assigned one *forward fundamental code channel* (including power control bit subchannel) with variable rate when only one code channel is operative, and the full rate (9600 or 14,400 bps) when multiple code channels are used. For data users needing rates greater than 9600 or 14,400 bps, from 0 to 7 *forward supplemental code channel* at full rate can used to transmit data in parallel on orthogonally multiplexed forward code channels, as illustrated in Fig. 9.

Multiple 64-chip Walsh functions are allocated as needed to implement orthogonal Walsh function multiplexing of the forward-code channels — the designation "code channel" thus refers to Walsh functions on the forward link. When multiple Walsh code channels are used

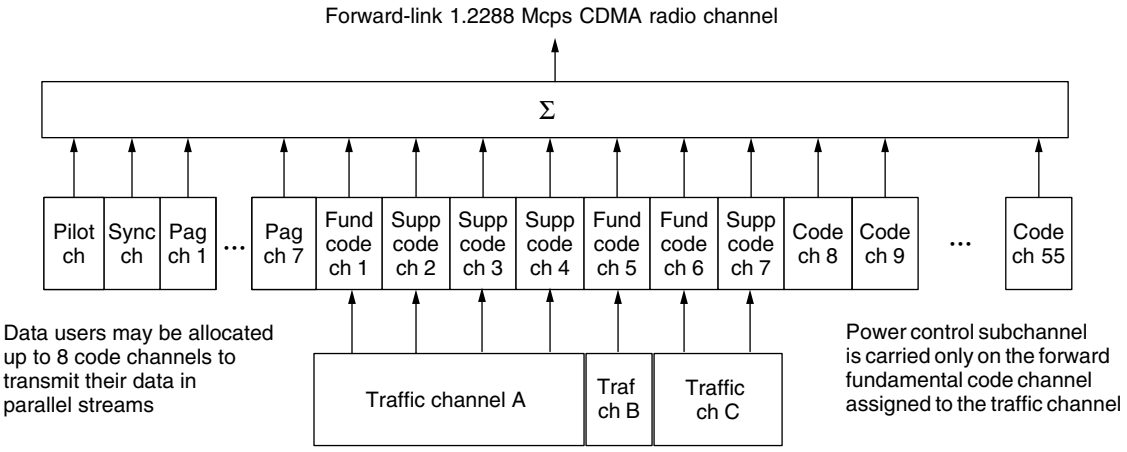


Figure 9. Forward-link channel aggregation under IS95B.

for a traffic channel, the scrambling in each code channel uses the same long code mask. The maximum bit rate for a traffic channel using eight code channels is 76.8 kbps for rate set 1 (1928 bits per 20-ms frame) and 115.2 kbps for rate set 2 (2888 bits per 20-ms frame).

The maximum information bit rate for a forward traffic channel (excluding overhead bits) using eight code channels is 68.8 kbps for rate set 1 (1728 bits per 20-ms frame) and 106.8 kbps for rate set 2 (2678 bits per 20-ms frame). Each mobile receiver "finger" must be able to demodulate up to eight forward code channels (and, by implication, reassemble them into the original single, high-speed bitstream). In 1998, Qualcomm announced the availability of the fifth-generation single-chip Mobile Station Modem (MSM3000) that uses a new "superfinger" demodulator architecture to support simultaneous demodulation in each finger of up to eight 9600-bps forward-link channels (76.8 kbps total) and up to six 14,400-bps channels (86.4 kbps total). In 1999, Qualcomm brought out the sixth generation of the chip, MSM3100. Features include voice recognition, echo cancellation, and GPS processing for position location. However, as of mid-2001, the SuperFinger is still limited to 86.4 kbps for a single user.

In a similar manner, the IS95B reverse link provides each active user with a reverse fundamental code channel with variable rate and random burst transmission when only one code channel is operative, and the full rate (9600 or 14,400 bps) when multiple code channels are used. For data users needing rates greater than 9600 or 14,400 bps, from 0 to 7 reverse supplemental code channels at full rate can be used to transmit data in parallel over PN code-division-multiplexed reverse code channels. When multiple code channels are allocated to a particular mobile user, the long PN code masks of the code channels are indexed to the code channel numbers for code-division multiplexing of the reverse code channels — the mask starts with 11xxx11000, where xxx = 000, 001, and so on. In addition, supplemental code channels, as added, are offset in carrier phase by certain increments of $\pi/4$ ($\pi/2$, $\pi/4$, $3\pi/4$, 0, $\pi/2$, $\pi/4$, $3\pi/4$, in that order) that provide at least partial carrier phase orthogonality to the supplemental code channels. Provision is made for intermittent adding or dropping of supplemental channels, with preambles to aid acquisition.

An advanced version of IS95, known as *cdma2000* and described in the TIA interim standard IS2000 [17], was proposed to the International Telecommunications Union as a candidate for the international third-generation cellular system. It was accepted as one of several technologies for future development and possible deployment.

BIOGRAPHIES

Leonard E. Miller received his B.E.E. degree in 1964 from Rensselaer Polytechnic Institute, Troy, New York; a M.S.E.E. degree in 1966 from Purdue University, Lafayette, Indiana; and a Ph.D. degree in electrical engineering from Catholic University of America, Washington, D.C., in 1973. He joined the Naval Ordnance Laboratory, Silver Spring, Maryland (later called Naval Surface Weapons Center) in 1964 as an electronics engineer. At NOL he designed instrumentation for underwater weapons and systems, and he developed algorithms for processing sonar signals. In 1978, he joined J. S. Lee Associates, Inc., Rockville, Maryland, where he performed R&D related to military surveillance and communication systems and also studied digital cellular telephone technology. Since May 2000, he has been with the Wireless Communication Technologies Group of the Advanced Network Technologies Division at the National Institute of Standards and Technology, Gaithersburg, Maryland. At NIST he is involved in R&D related to wireless ad-hoc networks. Dr. Miller, a senior member of IEEE, is coauthor of CDMA Systems Engineering Handbook (Artech House, 1998) as well as many journal and conference publications.

Jhong Sam Lee received his B.S. degree in electrical engineering in 1959 from the University of Oklahoma and his M.S.E. and D.Sc. degrees in electrical engineering from the George Washington University, Washington, D.C., in 1961 and 1967, respectively.

From 1959 to 1964 he worked in the industry in radar systems and microwave components development. From 1965 to 1968 he was an assistant professor at the George Washington University, Washington, D.C., where he taught courses in digital communication, information, and coding theories. From 1968 to 1969 he was an advisory engineer at IBM Corporation. From 1969 to 1973 he was an associate professor of electrical engineering at the Catholic University of America, Washington, D.C. From 1965 to 1973 he was a consultant at the U.S. Naval Research Laboratory in the areas of underwater signal processing and spread spectrum communication systems. In 1976, he founded J.S. Lee Associates, Inc. (JSLAI), for development of techniques in satellite and electronic warfare systems for DoD (Department of Defense) and its component services. He also founded in 2000 Advanced Technology Systems, Inc., in Seoul, Korea, for development (and for manufacturing) of interference cancellation systems for applications in the CDMA wireless transmission networks. Dr. Lee coauthored (with Dr. L.E. Miller) a book, CDMA Systems Engineering Handbook (1200pp), Artech House (1998), which is also translated into the Chinese language and was published in the People's Republic of China in 2001. He holds several patents in the area of CDMA wireless communications. Dr. Lee is a fellow of the Institute of Electrical and Electronics Engineers, Inc. (IEEE).

BIBLIOGRAPHY

1. *Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System*, TIA/EIA Interim Standard 95 (IS95), Telecommunications Industry Assoc., Washington, DC, July 1993 (amended as IS95A in May 1995).
2. V. H. MacDonald, The cellular concept, *Bell Syst. Tech. J.* **58**(1): 15–41 (Jan. 1979).
3. *Cellular System Dual-Mode Mobile Station–Base Station Compatibility Standard*, TIA/EIA Interim Standard 54 (IS54B), Telecommunications Industry Assoc., Washington, DC, April 1992.

4. U.S. Patent 5,103,459 (April 7, 1992), K. S. Gilhousen et al., System and method for generating signal waveforms in a CDMA cellular telephone system.
5. J. S. Lee and L. E. Miller, *CDMA Systems Engineering Handbook*, Artech House, Boston, 1998.
6. R. Padovani, Reverse link performance of IS95 based cellular systems, *IEEE Pers. Commun. Mag.* 28–34 (3rd quarter 1998).
7. K. I. Kim, CDMA cellular engineering issues, *IEEE Trans. Vehic. Technol.* **42**: 345–350 (Aug. 1993).
8. C. Wheatley, Trading coverage for capacity in cellular systems: A system perspective, *Microwave J.* **38**: 62–79 (July 1995).
9. K. S. Gilhousen et al., On the capacity of a cellular CDMA system, *IEEE Trans. Vehic. Technol.* **40**: 303–312 (May 1991).
10. K. Gilhousen, On the "optimum" bandwidth for spread spectrum, *Proc. 2nd Int. Conf. Personal, Mobile, and Spread Spectrum Communications*, Beijing, 1994, pp. 202–210.
11. H. Harmuth, A generalized concept of frequency and some applications, *IEEE Trans. Inform. Theory* **IT-14**: 375–382 (May 1968).
12. *Personal Station-Base Station Compatibility Requirements for 1.8 to 2.0 GHz Code Division Multiple Access (CDMA) Personal Communications Systems*, ANSI J-STD-008, Telecommunications Industry Assoc., Washington, DC, 1996.
13. A. J. Viterbi, *Principles of Spread Spectrum Multiple Access Communication*, Addison-Wesley, New York, 1995.
14. R. Price and P. E. Green, Jr., A communication technique for multipath channels, *Proc. Inst. Radio Engineers*, March 1958, Vol. 47, pp. 555–570.
15. A. J. Viterbi, A. M. Viterbi, K. S. Gilhousen, and E. Zehavi, Soft handoff extends CDMA cell coverage and increases reverse link capacity, *IEEE J. Select. Areas Commun.* **12**(8): 1281–1288 (Oct. 1994).
16. *Mobile Station-Base Station Compatibility Standard for Wideband Spread Spectrum Cellular Systems* (ANSI/TIA/EIA-95-B-99), Telecommunications Industry Assoc., Washington, DC, Feb. 1999.
17. *Physical Layer Standards for cdma2000 Spread Spectrum Systems*, TIA/EIA/IS-2000-1a, March 2000.

v

cdma2000

GIOVANNI EMANUELE CORAZZA
ALESSANDRO VANELLI-CORALLI
University of Bologna
Bologna, Italy

1. OVERVIEW

cdma2000 is the multicarrier wideband code-division multiple access (CDMA) radio interface included in the International Mobile Telecommunications 2000 (IMT-2000) family of standards. As such, cdma2000 shares with various other standardized radio interfaces many features that are best described with an introduction to IMT-2000. IMT-2000, developed under the auspices of the International Telecommunications Union (ITU), is the third-generation (3G) mobile telecommunications standard system. The objectives pursued by IMT-2000 are global service capability, standardized radio interfaces, flexible/seamless service provision, advanced multimedia services and applications, integrated terrestrial and satellite networks, and finally improved operational efficiency with respect to second-generation (2G) standards. At the present state of the standardization effort, IMT-2000 includes five terrestrial and six satellite radio interfaces [1]. The terrestrial radio interfaces are organized in two groups: CDMA interfaces and TDMA interfaces. The CDMA interfaces are IMT-DS (W-CDMA, FDD), IMT-MC (cdma2000, FDD), and IMT-TC (TD-SCDMA, TDD), whereas the TDMA interfaces are IMT-SC (UWC-136, FDD) and IMT-FT (DECT, TDD). All IMT-2000-compliant systems must support voice and data services. The latter can be symmetric or asymmetric, circuit- or packet-switched, with data rates ranging from 16 kbps to 2 Mbps. The IMT-2000 spectrum allocation was set in the works of three World Radio Conferences (WRC92, WRC95, and WRC00) and includes the following frequency bands: 806–960 MHz, 1710–1885 MHz, 1885–2025 MHz, 2110–2200 MHz, and 2500–2690 MHz.

The cdma2000 radio interface has been developed within the Third-Generation Partnership Project 2 (3GPP2) [2]. 3GPP2 is a partnership of standards development organizations, aiming at defining a 3G system based on the ANSI/TIA/EIA-41 core network. 3GPP2 partners are ARIB (Japan), CWTS (China), TIA (USA), TTA (Korea), and TTC (Japan). 3GPP2 works are organized in four main committees: Organizational Partners Committee, Steering Committee, Technical Specification Groups (TSGs), and ad hoc groups. The technical specification groups are TSG-A (access network interface), TSG-C (cdma2000), TSG-N (Intersystem operations), TSG-P (wireless packet data networking), and TSG-S (services and systems aspects). The cdma2000 standard is presently under development within 3GPP2. To date, releases 0, A, and B of the technical specifications have been published, whereas release C is in progress. To provide the most up-to-date information, this article discusses the release C specifications, which build strongly on releases A and B while adding a further radio configuration for high-speed packet data transmission, identified as 1XEVolved high-speed integrated Data and Voice (1X EV-DV) [3]. The 1X EV-DV radio configuration is also known as the high-data-rate (HDR) system.

A description of a few notable cdma2000 features is in order to conclude this introductory overview. The multicarrier structure pertains strictly to the forward link, where N ($N = 1, 3$, and optionally 6, 9, 12) adjacent direct-sequence spread RF carriers are used, while in the reverse link a single direct-spread RF carrier is adopted, with flexible spreading factor. This multicarrier flexibility is instrumental in guaranteeing backward compatibility with the 2G standard TIA/EIA-IS95B [4], also known as *cdmaOne*, thus easing the transition and coexistence between the two standards. Code-division multiple access is achieved adopting spread-spectrum techniques with long spreading codes. Notably, the time epoch offsets of these codes identify base stations and users in the forward and reverse

links, respectively. This time epoch management is made possible by the fact that all base stations in the network are synchronized to a common time reference. The time reference system takes advantage of the Global Positioning System (GPS) for clock synchronization. In the forward link, the chip rate amounts to 1.2288 Mcps on each carrier, which adds up to 3.6864 Mcps when three carriers are used. In the return link both 1.2288 and 3.6864 Mcps chip rates are imposed on a single carrier by varying the spreading factor. These two spreading rate modes are referred to as *spreading rate 1* (SR1) and *spreading rate 3* (SR3). SR1 and SR3 are commonly referred to as "1X" and "3X," respectively. Finally, the cdma2000 core network specifications are based on the evolved ANSI 41 and IP networks. Additionally, to maximize customer roaming capabilities the cross-mode operation with the GSM-MAP core network, identified as MC-MAP, is also supported [5].

2. THE cdma2000 AIR INTERFACE STANDARD

The cdma2000 core air interface standard is reported in 3GPP2 specifications C.S0001–C.S0006 [6–10]. An additional specification [11] is provided to support analog operations for dual-mode mobile stations (MSs) and base stations (BSs).

The protocol architecture of the air interface has been developed with reference to the ISO/OSI model, and is reported in Fig. 1. The ISO/OSI layer 1, or physical layer [7], provides for transmission and reception of radio signals between the base station and the mobile station. The physical layer services are offered to the upper layers through physical channels (represented in Fig. 1 by dotted lines with uppercase labels), which are the means for information transport over the air. The physical channels are characterized by the coding technique and rate, the spreading factor, and the digital modulation scheme. The precise parameters of each physical channel are defined in a set of radio configurations (RCs). There are 10 RCs for the forward link and 6 RCs for the return link, which collectively form the FDD MC-CDMA 1X/3X air interface. The ISO/OSI layer 2 (data-link layer) provides for delivery of signaling messages generated by layer 3 (network layer), making use of the services provided by layer 1. It has been subdivided into a medium access control (MAC) layer [8], and a link access control (LAC) layer [9]. The MAC layer is further subdivided into the multiplexing and QoS entity, the radio link protocol (RLP) entity, the signaling radio burst protocol (SRBP) entity, and the packet data channel control function (PDCHCF) entity. The MAC services are provided to the upper layers through logical channels (shown by solid lines with lowercase labels in Fig. 1), which are identified by the carried information typology, and are mapped onto physical channels. The LAC layer provides signaling, packet data voice, and data service transportation for the upper layers. Finally, the upper signaling layer [10] makes use of the services provided by layer 2 to support a wide range of radio interface signaling alternatives, namely, the native cdma2000 upper-layer signaling, backward-compatible TIA/EIA-IS95B signaling, and other existing or future upper-layer signaling entities. In layer 3 signaling messages between BS and MS are originated and terminated.

2.1. The Physical Layer

The physical layer (layer 1) offers information transfer between the mobile station and the base station to MAC and higher layers by means of physical channels. Physical layer specifications are defined in Ref. 7. Spreading rates, data rates, modulation parameters, forward error correction (FEC) schemes, puncturing, repetition rates, interleaving, and channel structures for the forward- and reverse-link signals are specified by RC1–RC10 and RC1–RC6, respectively. RC1 and RC2 are backward-compatible with the TIA/EIA-IS95B standard.

2.1.1. Forward Link. In the forward link, RC1–RC5 plus RC10 employ SR1, whereas RC6–RC9 correspond to SR3. Data rates for the different radio configurations are up to 9.6 kbps for RC1, 14.4 kbps for RC2, 153.6 kbps for RC3, 307.2 kbps for RC4, 230.4 kbps for RC5, 307.2 kbps for RC6, 614.4 kbps for RC7, 460.8 kbps for RC8, and 1.0368 Mbps for RC9. In RC10 data rates per subpacket range from 81.6 to 3.0912 Mbps.

The forward-link physical channels are shown in Fig. 2. They are the *pilot channels*, used for channel estimation and power-level measurements; the *common power control channel*, which carries as many power control bits as the number of active *reverse traffic*, *common control*, *acknowledgment*, or *channel quality indicator channels*; the *common assignment channel*, which is used for quick assignment of a *reverse-link channel* for random-access packets; the *common control channel*, which carries MS-specific messages; the *synchronization channel*, used to aid the initial time synchronization procedure; the *broadcast control channel*, used to transmit BS-specific, systemwide information and MS-specific messages; the *paging channel*, used in SR1 to transmit system overhead information and MS specific messages; the *quick paging channel*, used to inform MSs in idle state and slotted mode, as to whether they should receive the *forward common control* or the *paging channels* in the next slot; the *packet data control channels* and the *traffic channels*, which are used to transmit signaling and user information to a specific MS during a call.

The *traffic channel* includes the *fundamental channel*, which carries user and signaling information during a call; the *packet data channel*, which is used to transmit user packet data in RC10 with SR1; the *dedicated control channel*, which is employed to send user and signaling information during a call; the *power control subchannel*, which is used to transmit power control messages in association with a fundamental channel or a forward dedicated channel; the *supplemental* and *supplemental code channels*, which are used to transmit user information to a specific MS during a call in RC3–RC9 and RC1–RC2, respectively. There can be up to two supplemental channels and up to seven supplemental code channels for each traffic channel.

The pilot channels are a set of unmodulated spread-spectrum signals consisting of the *forward pilot channel*, which provides a phase reference for coherent

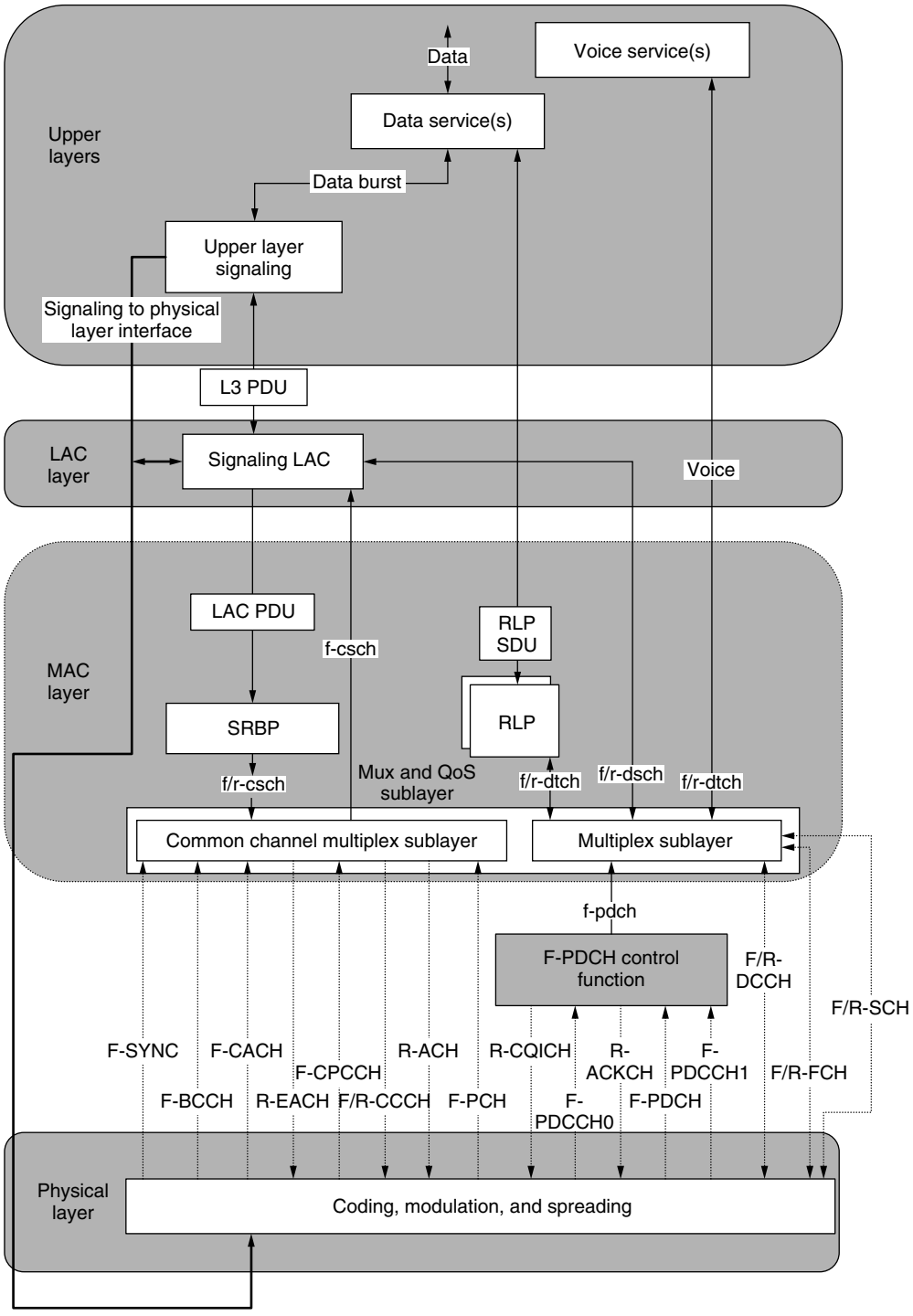


Figure 1. cdma2000 general radio interface protocol architecture [8]. Reprinted with permission.

demodulation and is used for signal strength comparison for the handoff procedure; the *transmit diversity pilot channel*, which is transmitted whenever transmission diversity is applied; the *auxiliary pilot*, and the *auxiliary transmit diversity pilot channels*.

Physical channels are processed as reported in the simplified block diagrams in Figs. 3 and 4. The actual block diagrams are channel- and RC-dependent, and are reported in Ref. 7. Information coming from the higher layers enters the forward error correction (FEC) block, then undergoes repetition and/or puncturing, interleaving, scrambling, complex modulation mapping, and gain control. The encoded and modulated output symbols are then demultiplexed in N ($N = 1$ or 3) I(cosine)/Q(sine) pairs ($I_i, Q_i, i = 1$ or 3) that are fed to the spreading, filtering, and upconversion sections. The RF output is finally transmitted on N adjacent carriers. Normally, a single antenna is used [non–transmit diversity—(NTD)

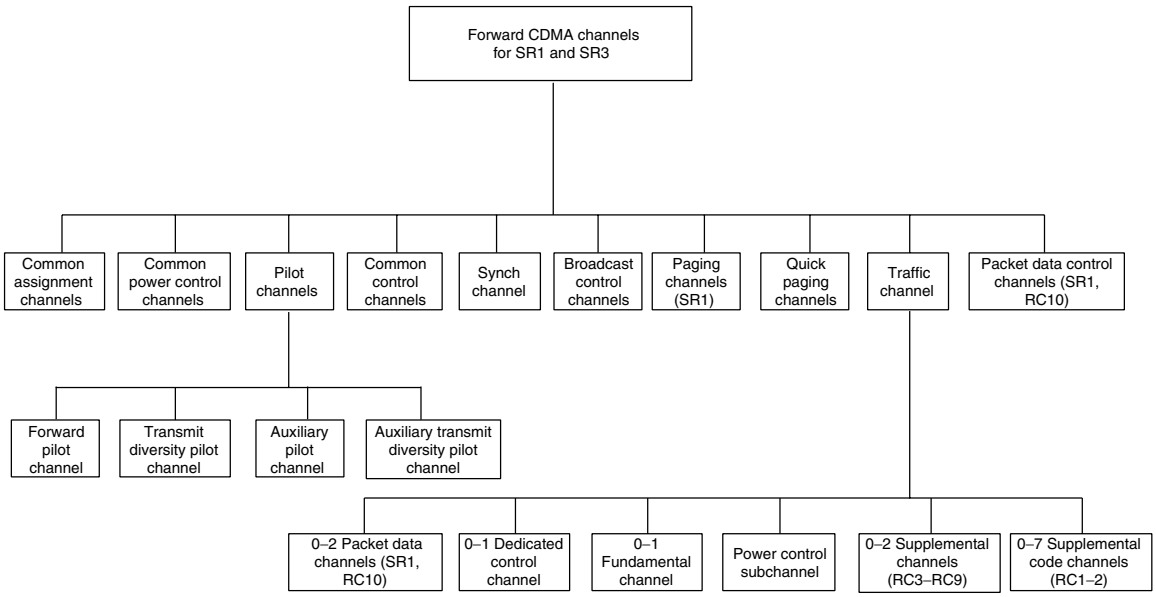


Figure 2. cdma2000 forward channels [7]. Reprinted with permission.

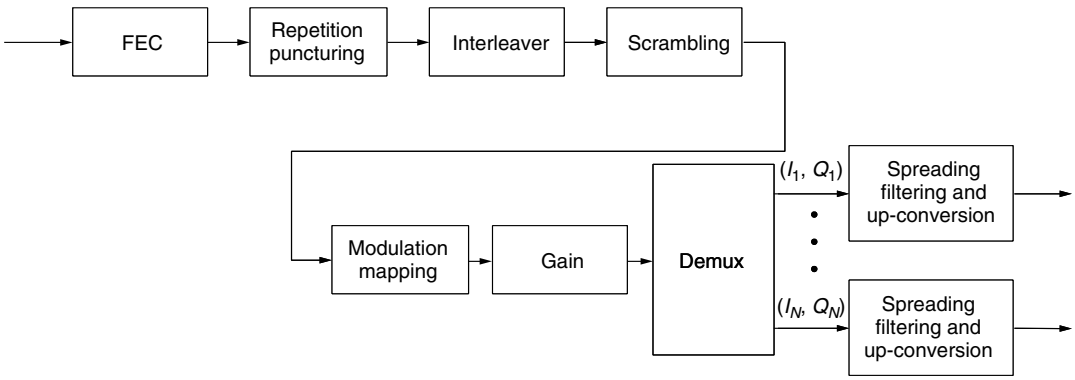


Figure 3. Simplified forward link transmitter block diagram.

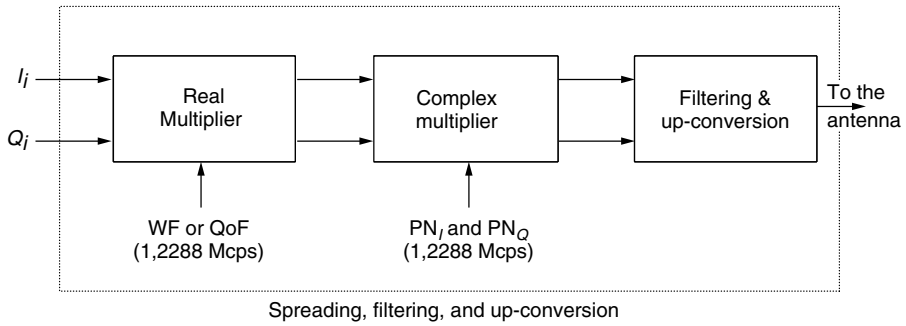


Figure 4. Simplified forward link spreading, filtering, and upconversion.

mode]. As an alternative, orthogonal TD (OTD) can be adopted, whereby two and three transmitting antennas are employed for SR1 and SR3, respectively. In these cases, the *transmit diversity pilot* and the *transmit diversity pilot channels* are also transmitted.

Forward error correction is accomplished through convolutional or turbo encoding. The possible coding rates are $R = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, and $\frac{1}{6}$ for the convolutional encoder, and $R = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, and $\frac{1}{5}$ for the turbo encoder. The coding schemes and rates for each physical channel are RC-dependent. All the convolutional encoders have constraint length $K = 9$. The rate $\frac{1}{2}$ convolutional encoder has generator polynomials $g_0 = 753$ and $g_1 = 561$. The rate $\frac{1}{3}$ convolutional encoder has generator polynomials $g_0 = 577, g_1 = 663$, and $g_2 = 711$. The rate $\frac{1}{4}$ convolutional encoder has generator polynomials $g_0 = 457$, $g_1 = 671$, $g_2 = 513$, and $g_3 = 473$. The rate $\frac{1}{6}$ convolutional encoder has generator polynomials $g_0 = 765$, $g_1 = 755$, $g_2 = 551$, $g_3 = 637$, $g_4 = 625$, and $g_5 = 727$. The turbo encoder consists of two identical recursive convolutional encoders, parallel concatenated,

with a turbo interleaver preceding the second convolutional encoder. The transfer function for the recursive convolutional code used for all coding rates is $G(D) = [1 + n_0(D)/d(D) + n_1(D)/d(D)]$, where $d(D) = 1 + D^2 + D^3$, $n_0(D) = 1 + D + D^3$, and $n_1(D) = 1 + D + D^2 + D^3$. The encoder output is punctured and repeated to obtain the desired coding rate. No coding is applied to the *quick paging* and *common power control channels*.

Data scrambling is obtained by means of a long pseudonoise sequence (PNS), and is applied to all physical channels, with the exception of the common power control, pilot, synchronization, quick paging, and packet data control channels. In RC10, data scrambling on the traffic channel is obtained by means of a different scrambling sequence, produced by a 17-tap linear feedback shift register with generator polynomial $h(D) = D^{17} + D^{14} + 1$.

The modulation schemes are BPSK in RC1–RC2 and QPSK in RC3–RC9. In RC10, QPSK, 8-PSK, or 16-QAM are chosen adaptively depending on the radio propagation channel conditions. Orthogonal spreading is used to ensure separation between channels on each carrier (Fig. 4). The modulated symbols (I_i, Q_i) are spread to a chip rate of 1.2288 Mcps by way of Walsh functions (WFs) in RC1–RC2 and RC10, and by way of WF or alternatively quasiorthogonal functions (QOFs) in RC3–RC9. QOF sequences are obtained using a nonzero sign multiplier and a nonzero rotate enable WF to enlarge the set of orthogonal codes because, depending on the particular deployment and operating environment, the system capacity may result to be limited by the number of Walsh codes. Walsh sequences are indicated as W_n^K, n tagging the nth row of a $K \times K$ Hadamard matrix. A Hadamard matrix is recursively constructed as

$$H_{2K} = \begin{bmatrix} H_K & H_K \\ H_K & H_K^{'} \end{bmatrix}$$

where K is a power of 2, $H_K^{'}$ is the binary complement of H_K, and $H_1 = 0$.

Following orthogonal spreading, the quadrature pairs are chipwise complex-multiplied with an overlay quadrature spreading sequence. The quadrature spreading PNS sequence is formed by a couple of extended maximum-length shift register (MLSR) sequences of length 32768 chips (a 0 is inserted after the run of 14 consecutive zeros) at a chip rate of 1.2288 Mcps. The PNS sequence period is 26.66 ms. Following filtering, upconversion is obtained by way of harmonic in-phase and quadrature modulation.

The MS receiver chain performs complementary operations. Multiple propagation paths can be usefully collected and combined by using a rake receiver with multiple fingers. The rake receiver is also instrumental in implementing the soft handoff procedure.

2.1.2. Reverse Link. In the reverse link, RC1–RC4 use SR1, whereas RC5 and RC6 correspond to SR3. Data rates for RCs are up to 9.6 kbps for RC1, 14.4 kbps for RC2, 307.2 kbps for RC3, 230.4 kbps for RC4, 614.4 kbps for RC5, and 1.0368 Mbps for RC6.

Figure 5 represents the reverse link physical channel organization. The physical channels used by the mobile station to communicate with the base station are the *pilot channel*, used to aid BS operation in detecting a MS transmission and that includes the *reverse power control subchannel* for RC3–RC6; the *access* and the *enhanced access channels*, used to initiate communications or to respond to a message received on the FL; the *reverse common control channel*, employed to transmit user and signaling information when the *reverse traffic channels* are not active; and the *reverse traffic channel*, used to transmit user information and signaling during a call.

The reverse traffic channel includes the *reverse fundamental channel*, aimed at transmitting user and signaling information during a call; the *reverse supplemental code channel*, used in RC1–RC2 to carry user information during a call; the *reverse dedicated control channel*, aimed at transmitting user and signaling information during a call in RC1–RC2; the *reverse supplemental channel*, aimed at transporting user information during a call in RC3–RC6; the *reverse acknowledgment channel*, which provides control for the *forward packet data channel*; and the *reverse quality indicator channel*, which is used to indicate to the BS the channel quality measurements of the serving sector. The reverse acknowledgment and the reverse quality indicator channels are used only in combination with the forward packet data channel (i.e., RC10).

The reverse link physical channels are processed similarly to the forward-link case. Information coming from higher layers undergoes FEC, repetition and/or puncturing, interleaving, and complex modulation mapping. Differently from the forward link, the I and Q channels are used separately to transmit independent information.

FEC is accomplished by means of convolutional, turbo, or block coding. The admissible coding rates are $R = \frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$ for the convolutional code; $R = \frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, and $\frac{1}{5}$ for the turbo code; and $\frac{1}{3}$ for the block code. The convolutional and the turbo encoder schemes and their generator polynomials are identical to those used in the forward link. The block code is (12,4) and is used only for the reverse channel quality indicator channel. As for the forward link, the association between coding schemes and physical channels is specified by RCs. The modulation schemes employed in the reverse link are the backward-compatible TIA/EIA IS95B 64-ary orthogonal modulation in RC1–RC2 and dual-BPSK modulation in RC3–RC6.

Figure 6 shows the simplified spreading, filtering, and upconverting subsystem block diagram for RC3–RC6. Multiple physical channels of a user are separated by spreading with orthogonal Walsh functions, whereas users are separated by means of different PNS sequence offsets. The physical channel symbols are first spread through WF to a chip rate of $N \times 1.2288$ Mcps ($N = 1$ for SR1 and $N = 3$ for SR3), and then complex-multiplied with the long PN sequence at the corresponding rate. Because of the direct spread nature of the reverse link, the RF signal at the output of the upconversion is always transmitted on a single carrier. The block diagram for RC1–RC2 is identical to that of the TIA/EIA-IS95B system [15].

The BS receiver chain performs complementary operations. Multiple propagation paths can be usefully collected and combined by using a rake receiver with multiple fingers.

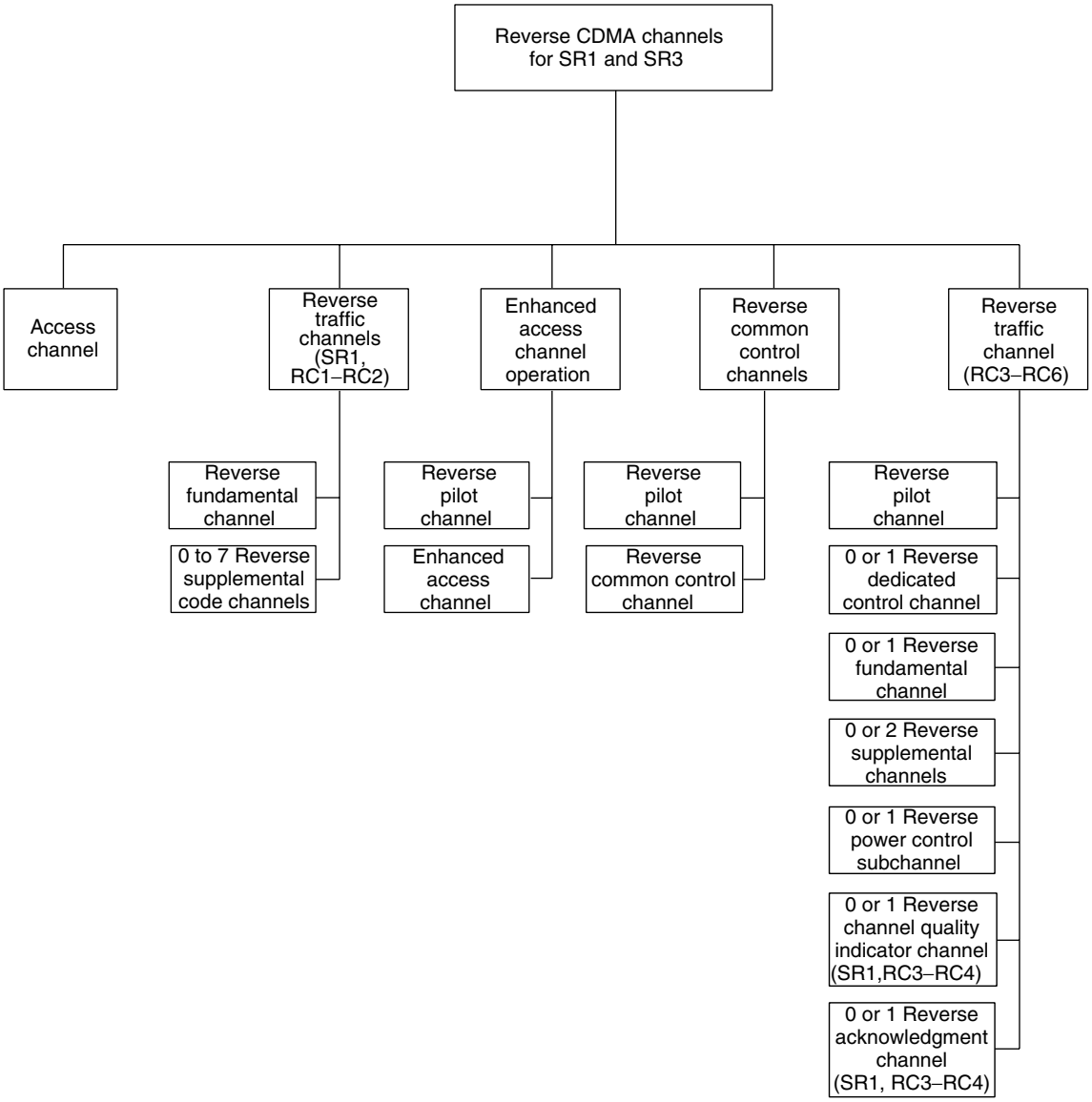


Figure 5. Reverse physical channels [7]. Reprinted with permission.

2.2. The Medium Access Control (MAC) Layer

The MAC layer is the interface toward the physical layer of the ISO/OSI reference model link layer, and it is described in Ref. 8. It provides the two important functions of best-effort delivery and of multiplexing and QoS control. However, when backward compatibility with TIA/EIA-IS95B is adopted, that is, when encoded voice data are transported directly by the physical layer, the MAC services are null.

The MAC services are provided by means of the logical channels and the MAC entities. The logical channels are connections between peer entities and they are defined by the information they carry. Logical channel categories are the *common signaling channel* (csch), *dedicated signaling channel* (dsch), *dedicated traffic channel* (dtch), and, for the forward link only, the *packet data channel* (pdch). Logical channels are associated with physical channels. Association between logical and physical channels can be (1) exclusive and permanent, (2) exclusive but temporary, or (3) shared. Information on how to perform this mapping is contained into the logical-to-physical mapping (LPM) table.

The MAC entities are the RLP, SRBP, multiplexer–QoS delivery, and the PDCHCF entity. The multiplexer–QoS delivery entity has both transmitting and receiving functions. The transmitting function combines information (LAC signaling, data services, voice services) into multiplex protocol data units (MuxPDUs), which in turn are mapped onto physical layer service data units (SDUs) and PDCHCF SDUs. The receiving function separates the physical-layer SDUs and the PDCHCF SDUs and directs information to the appropriate entity. The multiplexer entity may operate in two different modes: mode A for RC1–RC2 and mode B for RC $>$ 2. In mode A, a single MuxPDU is used to form a physical-layer SDU, while in mode B the additional flexibility of mapping one or more

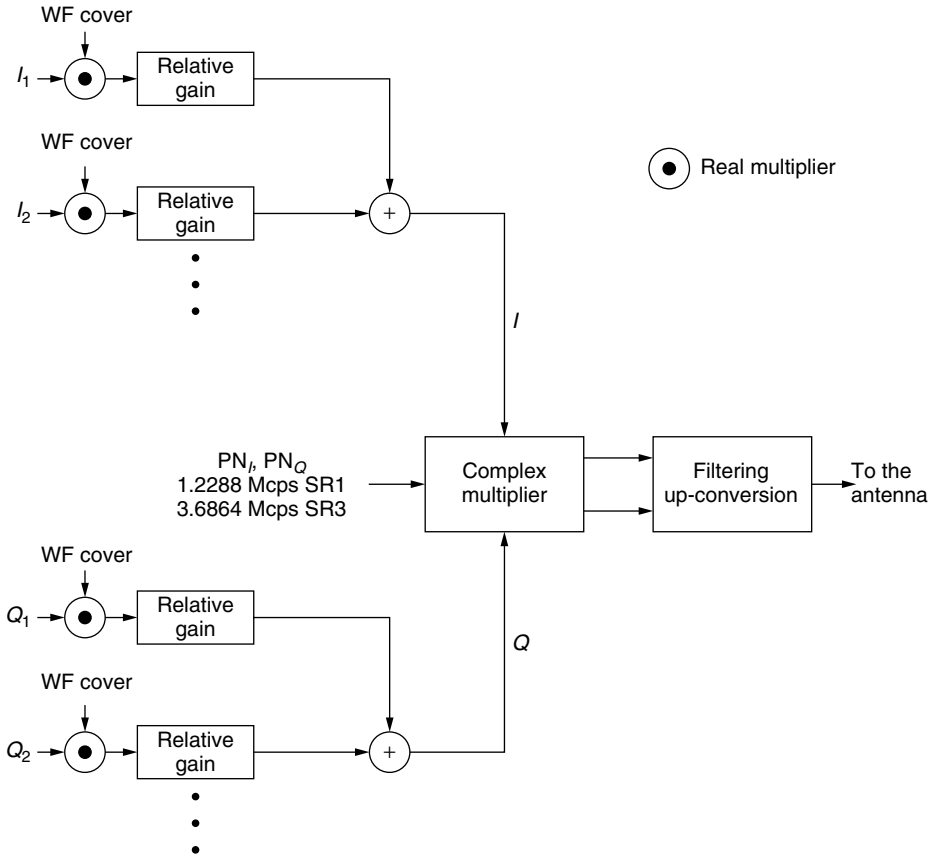


Figure 6. Reverse link I and Q spreading for RC3–RC6.

MuxPDUs into an SDU is provided. The multiplexer–QoS delivery entity determines the relative priority between information coming from higher entities. However, the precise use of priority information to guarantee a required QoS over the air is purposely not specified in the standard.

The SRBP entity manages the synchronization, paging, access, enhanced access, common assignment, broadcast channel, forward common control channel, and the return common control channel procedures. At the base station side the SRBP entity performs also the generation of the channel identifier.

The RLP is described in Ref. 14 and is used with a traffic channel to support connection-oriented negative-acknowledge-based data traffic delivery; that is, the receiver does not acknowledge correct reception and decoding of a data frame, but only requires the retransmission of erroneously received data frames. RLP is used only with RCs > 2. RLP supports both encrypted and nonencrypted data transport modes.

The PDCHCF entity is used in RC10 with the packet data channel to ensure the delivery of encoder data packets from BS to MS. In particular, four independent ARQ channels and code-division multiplexing (CDM) of encoder subpackets are adopted to enhance packet data transmission performance. All physical channels associated with the packet data channel, namely the packet data control channel, the acknowledgment channel, the channel quality indicator channel, and the packet data channel originate and terminate in the PDCHCF entity.

2.3. The Link Access Control (LAC) Layer

The LAC layer corresponds to the "upper" portion (i.e., above the MAC Layer) of the ISO/OSI Reference Model link layer. LAC provides for the correct transport and delivery of layer 3, signaling messages by implementing a data-link protocol. LAC offers the framework and the necessary support for point-to-point transmission over the air for signaling services, circuit data service provision (optionally), and transportation of encoded voice in the form of packet data or circuit data traffic. When backward compatibility with TIA/EIA-IS95B is enforced, then the LAC services are null.

The LAC layer is organized in the interface with the lower (MAC) and upper (L3 signaling) layers, protocol sublayers, and logical channels (Fig. 7). The LAC interfaces are identified as service access points (SAPs). On the L3-LAC SAP, LAC sends and receives SDUs and interface control primitives in the form of message control status blocks (MCSBs). At the LAC-MAC SAP, LAC exchanges LAC protocol data units (PDUs). The received L3-SDUs are serially processed by the protocol sublayer to form LAC PDUs, which in turn are serially processed to reconstruct L3-SDUs. The protocol sublayers are the *authentication-and-message integrity* sublayer, which performs the MS identification; the *ARQ sublayer*, which implements the ARQ protocol; the *addressing sublayer*, which takes care of the PDU addressing; the *utility sublayer*, which assembles and validates well-formed PDUs; and the *segmentation-and-reassembly sublayer*,

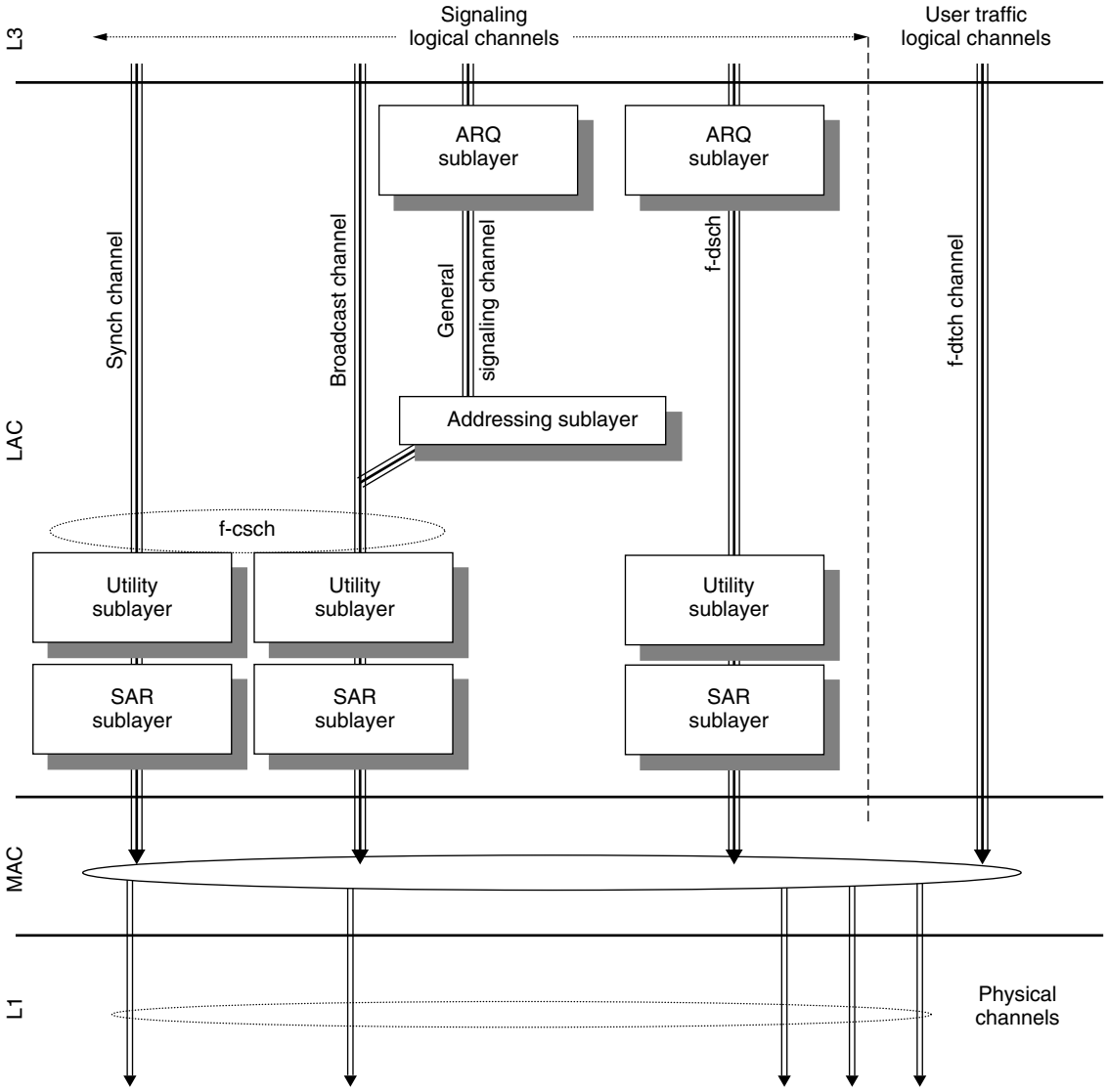


Figure 7. Architecture of the forward logical channels seen by the LAC sublayer [9]. Reprinted with permission.

which segments PDUs into PDU fragments or reassembles PDU fragments into PDUs.

Logical channels are the means for information transport from L3 and LAC to MAC. Logical channels allow the LAC to disregard details of the radio air interface characteristics. Only forward and reverse common signaling channels (f/r-csch), and forward and reverse dedicated signaling channels (f/r-dsch) are employed by LAC for access, synchronization, broadcast, and general or dedicated signaling.

2.4. Upper-Layer (Layer 3) Signaling

Upper-layer (layer 3) signaling corresponds to layer 3 (L3) and above of the ISO/OSI Reference Model. The L3 signaling layer consists of the protocol layer, which sends and receives L3 PDUs to and from lower layers, SAPs, and the communication primitives between L3 and lower layers. The L3 signaling layer is described in Ref. 10.

From the BS point of view, L3 signaling consists of the *pilot* and *synchronization channel processing*, used to transmit the pilot and synchronization channels to the MSs in the initialization state to achieve synchronization; *the common channel processing* used to transmit the paging, forward common control, and broadcast control channels, which are monitored by the MS in either idle or system access state; *the access channel and enhanced access channel processing*, employed to monitor those channels that are used by a MS in the system access state to send messages; and the *traffic channel processing*, used to communicate with a MS in the *control on the traffic channel state*. From the MS point of views, L3 signaling consists of the *initialization*, *idle*, *system access*, and *control on the traffic channel states*.

The *initialization state* is subdivided into the *system determination substate*, in which the MS selects which system to use; the *pilot channel acquisition substate*, in which the MS acquires the *pilot channel* of a CDMA

system; the *synchronization channel acquisition substate*, in which the MS obtains system configuration and timing information; and the *timing change substate*, in which the MS synchronizes its timing to that of the acquired system.

In the *idle state*, the MS monitors the *paging channel*, the *quick paging channel*, the *forward common control channel*, and the *primary broadcast control channel*. MSs in this state can receive messages or incoming calls, can cancel a *priority access and channel assignment* (PACA) call, and can initiate a registration or a message transmission or a call.

In the *system access state*, the MS sends messages to a BS on the r-csch and receives messages from a BS on the f-csch. This state consists of the update overhead information, origination attempt, page response, order/message response, registration access, message transmission, and PACA cancel substates.

In the *control on the traffic channel state*, the MS communicates with a BS using f/r-dsch and f/r-dtch. It consists of the *traffic channel initialization substate*, in which the MS verifies that it can receive the forward traffic channel and begins to transmit on the reverse traffic channel; the *traffic channel substate*, in which the MS exchanges traffic channel frames with a BS in accordance with the current service configuration; and the *release substate*, in which the MS disconnects the calls and the physical channels.

2.5. System Procedures

2.5.1. Power Control Procedures. There are three independent power control methods envisaged in the cdma2000 specifications: open-loop power control, closed-loop power control, and code channel gain adjustment.

In the *open-loop power control*, the MS sets the transmitted power according to the measured ratio of the received energy per chip E_c to the total (i.e., interference plus noise) received power spectral density I_0. The MS measures the E_c value on the received pilot signal and the I_0 level over the entire $N \times 1.25$-MHz bandwidth, considering all rake receiver fingers.

The *closed-loop power control* algorithm consists of two loops: the *outer loop*, which sets the target ratio of the energy per bit E_b to the effective noise power spectral density N_t, according to the desired QoS (i.e., frame error rate) and the *inner loop*, which estimates the received E_b/N_t ratio on the traffic channel, compares it with the target value, and accordingly sets the value of the power control subchannel bit. Power control commands are sent every 1.25 ms performing a fast 800-Hz rate power control. Nominal step values for each power control command are 1, 0.5, or 0.25 dB. Differently from the TIA/EIA-IS95 system, in the cdma2000 forward-link power control, the MS sends directly power control bits to the BS instead of reporting the frame error rate. As an alternative, the *erasure indicator bits* and the *quality indicator bits* can be used to inform the BS about the erroneous reception of a frame.

The *code channel gain adjustment* consists in setting the relative channel gains to maintain the ratio of the mean code channel output power to the mean reverse pilot channel output power within predefined limits [7].

2.5.2. Paging Procedure. The paging procedure is used by a BS to inform a MS of incoming calls or signaling messages, namely, information needed to operate with this BS. The procedure relies on the use of the *paging*, *forward common control*, and *quick paging* channels. These channels are divided into 80-ms slots. When a MS monitors every slot, it is said to operate in nonslotted mode, whereas when it monitors only some preassigned slots, it is said to operate in slotted mode. In the slotted mode, the MS can stop or reduce its processing for power conservation. The nonslotted mode cannot be operated in the idle state.

The *quick paging channel* is used in slotted mode for the transmission of paging, broadcast, and configuration change indicators for a MS. Two paging indicators are reserved for a MS in its preassigned quick paging channel slot. A BS sets the paging indicators for a MS when it needs to start receiving the paging channel or forward common control channel.

2.5.3. Handoff Procedures. The handoff procedure is used to transfer the communication from one BS to another. Depending on the MS state, different handoff procedures are possible. When a MS is in the control on the traffic channel state, the handoff procedures are the *soft handoff*, the *hard handoff*, and the *CDMA-to-analog handoff*. When a MS is in the idle state, only the *idle handoff* is possible.

During a *soft handoff*, the MS starts communications with a new BS without interrupting the communications with the previous one. This procedure applies only to CDMA channels having the same frequency assignments. Notably, soft handoff provides path diversity for the forward traffic and reverse traffic channels at the boundaries between BS coverage. By means of the *hard handoff* procedure, a MS transits between disjoint sets of BSs, band classes, and frequency assignments. The hard handoff is characterized by a temporary disconnection of the traffic channel. The *CDMA-to-analog handoff* procedure is used whenever a MS is directed from a CDMA traffic channel to an analog voice channel. In this case temporary disconnection occurs. The *idle handoff* procedure applies when a MS in the idle state detects a pilot channel signal that is sufficiently stronger than that of the serving BS.

To perform handoff a MS maintains a list of available pilot channels. The pilot channels are grouped into sets describing their status with regard to pilot searching, on the base of their relative offset to the zero-offset pilot PNS sequence. The pilot channel sets are: the *active set*, consisting of the pilot channels corresponding to the paging channel or the forward common control channel currently monitored; the *neighbor set*, which consists of all the pilot channels that are likely candidates for the idle handoff and that are specified by broadcast messages on the paging–broadcast control channel; the *remaining set*, which consists of all possible pilot channels in the current system excluding those already included in the preceding two sets; and the *private neighbor set*, which consists of all the pilot channels available for the private systems and that are specified in a dedicate broadcast list.

2.5.4. Access Procedure. The *access procedure* is a power ramping slotted ALOHA procedure performed by

a MS aiming at sending a message to a BS and receiving an acknowledgment for that message. The access procedure is based on *access attempts*. Each attempt consists of a sequence of one or more *access subattempts*, in turn made up of a sequence of one or more *access probe* sent on the access channels.

The *access probe* consists of a *preamble part* and a *message part*. The preamble part is a sequence of all-zero frames sent at the 4800 bps rate and is the actual instrument for the power ramping procedure. The message part includes the message body, length field, and cyclic redundancy check (CRC).

3. cdma2000 KEY FEATURES

The most notable cdma2000 air interface characteristics are summarized here:

Core Network Compatibility. cdma2000 has been developed with reference to the evolved ANSI-41 and all IP core networks, identified as native MC-41 mode. However, cross-modes, namely, MC-MAP and DS-41, are developed under the auspices of the Operator Harmonization Group (OHG) [5] and supported [12,13] to extend user roaming capability. In particular, in the MC-MAP cross-mode L1, MAC, LAC, and radio resource control of the cdma2000 standard are combined with the connection management and mobility management layers of the UMTS W-CDMA FDD standard.

Backward Compatibility with TIA/EIA-IS95B. Backward compatibility with the TIA/EIA-IS95B system is fully enforced by the cdma2000 standard [6]. Backward compatibility ensures that any TIA/EIA-IS95B MS can place and receive calls in any cdma2000 system, and that any cdma2000 MS can place and receive calls in any TIA/EIA-IS95B system. In the latter case, the cdma2000 MS is limited to the IS95B service capabilities (i.e., only SR1 can be used). The compatibility between cdma2000 and TIA/EIA-IS95B systems involves also the handoff procedures. A cdma2000 system in fact supports handoff of voice, data, and other supported services from and toward a TIA/EIA-IS95B network. In particular, handoffs between cdma2000 and TIA/EIA-IS95B networks can occur at cell boundaries or in the same cell, either in the same or between different frequency bands.

Overlay Capabilities with TIA/EIA-IS95B. cdma2000 supports different channel bandwidths, 1.25 MHz channel bandwidth in SR1, and 3.75 MHz channel bandwidth in SR3. This enables cdma2000 to be deployed as an overlay of a TIA/EIA-IS95B system with many different configurations. For example, combining different forward- and reverse-link RCs the following deployments are possible: 1X forward and reverse links, 3X forward link and 1X reverse link, 3X forward and reverse links, 1X forward link and 3X reverse link. Therefore, a seamless and flexible transition from TIA/EIA-IS95B to cdma2000 networks and exploitation of the existing TIA/EIA-IS95 network coverage during cdma2000 deployment is possible.

Fast Power Control. cdma2000 supports fast closed-loop power control (800 Hz) in both forward and reverse links. In particular, in the reverse link following signal to interference plus noise measurements, the MS itself sends power control commands to the BS.

Forward-Link Transmit Diversity. cdma2000 offers transmission diversity capabilities by means of the OTD mode. In the case OTD is employed, two or three transmitting antennas are used for SR1 and SR3, respectively. In the SR3 case, the multicarrier feature is exploited to transmit a carrier per antenna. This can be shown to provide a very efficient form of diversity. Auxiliary pilot and auxiliary transmit diversity pilot physical channels are supplied for each antenna to ease MS synchronization, channel estimation, and power-level measurements.

Coherent Reverse Link. Carrier phase estimation for coherent reception in the reverse link is made possible by means of the transmission of reverse pilot physical channels whenever a reverse traffic channel is assigned.

Enhanced Channel Structure. cdma2000 employs 5, 10, 20, 40 and 80 ms frame lengths, providing a means for trading off overhead and delay.

Turbo Codes. cdma2000 forward- and reverse-link RCs employ high-performance turbo codes.

Synchronous Base Stations. The cdma2000 system architecture is characterized by synchronous base stations. The system time is synchronous with the Universal Coordinated Time (UTC) and it is derived from the GPS system.

Multiple Access Spreading Codes and Code Planning. cdma2000 uses two levels of spreading. The first level is used to separate different physical channels in a CDM flux transmitted by either a BS or a user. Separation is achieved by means of orthogonal Walsh functions or quasiorthogonal Walsh functions. The second level of spreading aims at separating different CDM fluxes. Separation is achieved by means of complex long PNS sequences. Exploiting the BS synchronism, different CDM fluxes are associated to different code phases of the same PNS sequence. This is a significant difference and advantage with respect to the code planning necessity of the W-CDMA standard. In fact, since different BSs are identified by a code offset, cell planning simplifies to the association between code offsets and BSs. To relax the requirements associated with timing, among all the possible phase-shifted codes, those with minimum phase distance are avoided.

Initial Synchronization. Since a MS needs only to search for different phases of the unique PNS sequence, initial synchronization is significantly simpler in cdma2000 than in the other 3G CDMA air interfaces.

High Data Rate (*HDR*) *Packet Transmission*. cdma-2000 envisages a high-data-rate packet transmission (RC10) using the 1X mode, 1X EV-DV, which employs adaptive coding and modulation, fast retransmission of erroneously received frames, multiple (up to 4) time multiplexed ARQ channels for each MS, best serving BS selection driven by the MS, and megadiversity via sector selection. The MAC layer employs the new PDCHCF entity to support 1X EV-DV.

BIOGRAPHIES

Giovanni Emanuele Corazza received the Dr. Ing. degree (summa cum laude) in Electronic Engineering in 1988 from the University of Bologna (Italy), and a Ph.D. in 1995 from the University of Rome "Tor Vergata" (Italy). He is currently a Full Professor at DEIS, University of Bologna. He holds the chair for Telecommunications inside the Faculty of Engineering, and he is responsible for the area of Wireless Communications inside the Advanced Research Centre for Electronic Systems (ARCES). He is Vice-Chairman of the Advanced Satellite Mobile Systems Task Force (ASMS-TF), a European forum on satellite communications with more than 40 industrial partners. He visited ESA/ESTEC (Noordwjik, NL) as a Research Fellow, the University of Southern California (Los Angles, CA) as a Visiting Professor, and Qualcomm Inc. (San Diego, CA) as a Principal Engineer. He is associate Editor on spread spectrum for the *IEEE Transactions on Communications*. He received the Marconi International Fellowship Young Scientist Award in 1995 and two Best Paper Awards at IEEE Conferences. Professor Corazza has research interests in the areas of communication theory, wireless communications systems (including cellular, satellite, and fixed systems), spread-spectrum techniques, and synchronization. He is author or co-author of more than 70 papers published in international journals and conference proceedings.

Alessandro Vanelli-Coralli received the Dr. Ing. Degree (summa cum laude) in electronics Engineering and the Ph.D. in Electronics and Computer Science from the University of Bologna (Italy) in 1991 and 1996, respectively. Since 1996, he has been with the Department of Electronics, Computer Science and Systems (DEIS) at the University of Bologna where he is currently a Research Associate. Since 1995 he has held courses of Digital Communications at the Faculty of Engineering of the University of Bologna. He has been a research consultant on source coding and audio compression and has been involved in several national and international research projects. Dr. Vanelli-Coralli's research interests are in the area of digital communication systems addressing, in particular, satellite communications, spread-spectrum and CDMA systems, synchronization techniques, and digital signal processing. Dr. Vanelli-Coralli is a reviewer for IEEE journals and conferences, and has chaired sessions at IEEE Conferences. Dr. Vanelli-Coralli is co-recipient of the Best Paper Award at the IEEE ICT 2001 Conference. He is co-author of papers published in national and international journals and conference proceedings.

ACRONYMS

1X	Single carrier (i.e., spreading rate 1)
1X EV-DV	1XEVolved high-speed integrated Data and Voice
3G	Third generation
3GPP	Third-Generation Partnership Project
3X	Three carriers (i.e., spreading rate 3)
ANSI	American National Standard Institute
ARIB	Association of Radio Industries and Businesses
ARQ	Automatic repeat request
CDMA	Code-division multiple access
CRC	Cyclic redundancy check
csch	Common signaling channel
CWTS	China Wireless Telecommunication Standard Group
DECT	Digital enhanced cordless telecommunications
dsch	Dedicated signaling channel
dtch	Dedicated traffic channel
EIA	Electronic Industry Alliance
FDD	Frequency-division duplex
FEC	Forward error correction
GSM-MAP	Global System for Mobile communications — Mobile Application Part
IMT-2000	International Mobile Telecommunication 2000
IMT-DS	IMT direct spread
IMT-FT	IMT frequency time
IMT-MC	IMT multicarrier
IMT-SC	IMT single carrier
IMT-TC	IMT time code
ISO/OSI	International Standard Organization/Open System Interconnection
ITU	International Telecommunication Union
kbps	kilobit per second
LAC	Link access control
LPM	Logical-to-physical mapping
MAC	Medium access control
Mbps	megabits per second
MC	Multicarrier
MC-MAP	Multicarrier using GSM MAP
Mcps	Megachips per second
MCSB	Message control status block
MLSR	Maximum-length shift register
MS	Mobile station
PACA	Priority access and channel assignment
pdch	Packet data channel
PDCHCF	Packet data channel control function
PDU	Protocol data unit
PNS	PN sequence
QOF	Quasiorthogonal function
RC	Radio configuration
RF	Radiofrequency
RLP	Radio Link Protocol
SAP	Service access point
SDO	Standard Development Organization

SDU	Service data unit
SR1	Spreading rate 1, corresponding to 1X
SR3	Spreading rate 3 corresponding to 3X
SRBP	Signaling Radio Burst Protocol
TDD	Time-division multiplex
TD-SCDMA	Time-division single carrier CDMA
TIA	Telecommunication Industries Association
TTA	Telecommunications Technology Association
TTC	Telecommunication Technology Committee
UTC	Universal Coordinated Time
UMTS	Universal Mobile Telecommunication System
UWC-136	Universal Wireless Communication — 136
W-CDMA	Wideband CDMA
WF	Walsh function

BIBLIOGRAPHY

1. ITU-R, M.1457, *Detailed specifications of the radio interfaces of International Mobile Telecommunications-2000* (*IMT-2000*), draft revision, Doc 8/BL/6-E, April 17, 2001.
2. For further information on how to obtain 3GPP2 Technical Specifications and Technical Reports, please visit *http://www.3GPP2.org*
3. 3GPP2, S.R0026, *High-Speed Data Enhancements for cdma2000 1x — Integrated Data and Voice — Stage 1 Requirements*, version 1.0, *http://www.3gpp2.org*, Oct. 2000.
4. TIA/EIA-IS-95-B, *Mobile Station-Base Station Compatibility Standard for Wideband Spread Spectrum Cellular Systems*, Feb. 1999.
5. Operators Harmonization Group (OHG), *Specification framework for ITU IMT-2000 CDMA proposal*, *http://www.itu.int/imt/2-datio-dev/ohg/index-es.html*, Jan. 1999.
6. 3GPP2, C.S0001, *Introduction to cdma2000 Standards for Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
7. 3GPP2, C.S0002-C, *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
8. 3GPP2, C.S0003-C, *Medium Access Control* (*MAC*) *Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
9. 3GPP2, C.S0004-C, *Signaling Link Access Control* (*LAC*) *Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
10. 3GPP2, C.S0005-C, *Signalling Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
11. 3GPP2, C.S0006-C, *Analog Signaling Standard for cdma2000 Spread Spectrum Systems*, release C, Version 1.0, *http://www.3gpp2.org*, May 2002.
12. 3GPP2, C.S0007-0, *Direct Spread Specification for Spread Spectrum Systems on ANSI-41* (DS-41) (Upper Layers Air Interface), *http://www.3gpp2.org*, June 2000.
13. 3GPP2, C.S0008-0, *Multi-carrier Specification for Spread Spectrum Systems on GSM MAP* (MC-MAP) (Lower Layers Air Interface), *http://www.3gpp2.org*, June 2000.
14. 3GPP2, C.S0017-0-2, *Data Service Options for Spread Spectrum Systems*, addendum 2, version 2.0, *http://www.3gpp2.org*, Aug. 2000.
15. V. K. Garg, *IS-95 CDMA and cdma2000 Cellular/PCS Systems Implementation*, Prentice-Hall Communications Engineering and Emerging Technologies Series, Englewood Cliffs, NJ, 2000.

CELL PLANNING IN WIRELESS NETWORKS

KAI ROHRBACHER
Head, Department of Mobile Communication Software
LStelcom Lichtenau, Germany

JÜRGEN KEHRBECK
Head, Division of e-Commerce and Mobile Communications
LStelcom Lichtenau, Germany

WERNER WIESBECK
Director, Institute for High Frequency Technology and Electronics Karlsruhe University, Germany

1. FOCUS OF THIS ARTICLE

This article is intended to give an overview of planning and simulating strategies for wireless networks, from the past to the present. The focus is on the most popular system technologies such as GSM900/1800, GPRS/EDGE, and the upcoming so-called third-generation systems (3G) with IMT-2000 framework like UMTS-FDD in Europe, UMTS-TDD, CDMA2000, and TD-SCDMA. The authors present examples and methodologies based on network technologies used in Europe and also discuss general aspects of network planning and optimization.

While most of the licenses for the 3G mobile systems have been granted and research on the rollout of 3G networks is still under way, research projects for even the next generation (4G) has begun.

2. WIRELESS NETWORKS: FROM THE BEGINNING TO THE FOURTH GENERATION

The development of mobile networks can be traced back to the early 1960s. The main target was to offer voice conversation at any point in the country at any time without being bound to fixed network lines. Widespread commercial deployment of the first analog systems begun in the 1980s [1]. The introduction of second-generation systems (2G), especially the GSM system in the late 1980s, then ignited the ultimate success story of mobile networks. In Europe, the year 2000 marked the milestone where in some countries, the number of mobile subscribers has topped the number of fixed (landline) telephone sets.

Customers increasingly, tend to see mobile networks as a natural mobile "extension" if not even a replacement of their fixed-line networks. This puts additional pressure on mobile network developments. This rising demand in quantity and quality of the networks in parallel with stiffering operator competition led to a dramatic reduction in the rollout time of new networks to a few months instead

of 2 or 3 years as in the early years of GSM. But also existing network operators are facing problems — rapidly growing user figures and demand for new services is increasingly pushing their networks to the limit. However, "bandwidth" is a physical constraint.

All these factors together made it increasingly necessary to have powerful computer systems that support the network rollout and optimization process.

Various tools have been developed since 1990, starting with simple propagation algorithms, which led to highly sophisticated computer and database systems that support the complete planning process from green field layouts to live network optimization.

2.1. 3G: From "Evolution" to "Revolution"

Having said that, the "mobile evolution" now turns into a "mobile revolution" with the upcoming third-generation mobile systems. These systems do not simply extend existing 2G, they are completely new and indeed, represent a paradigm shift in how to plan them. With their inherent vast additional level of complexity, it may become virtually impossible to plan 3G (let alone 4G) networks without massive computational support and guidance. This forces planning tool suppliers as well as network operators to use a whole new set of algorithms and planning processes to create and engineer these systems.

3. PLANNING 1G, 2G, 2.5G, 3G, 4G: A SHORT SURVEY

It is interesting to see how the method of planning cellular networks changed over time and what particular demands influenced (and do influence) that development.

The main driving factor of the development was the increasing demand for capacity on the air interface. This always has been a very challenging topic, because, as seen from a scientific viewpoint, earth's atmosphere is a poor medium for electronic signal transmissions. This led to the effect that increasing effort and complexity went into the development of the air interface whereas the backbone network remained relatively stable over time. Other parameters also affected the evolution path of cellular systems, which are also discussed here.

It is common usage to speak about "generations" of cellular networks. This is due to the fact that progress in cellular networks was a process that "warped" at certain times and then continued in a rather smooth evolution up to the next "warp"; it's these intervals that we call "generations."

3.1. 1G Systems: When It All Started

First-generation systems are characterized mostly by

- Analog transmissions
- Coarse site placements
- High power transmitters
- Simple modulation schemes
- Nationwide, incompatible systems
- Voice-only systems

The cellular systems are surprisingly old; it was only 33 years after German physicist Heinrich Rudolf Hertz had discovered the electromagnetic waves that the first public cellular phone was introduced in Germany by the Deutsche Reichsbahn on the Zossen-Berlin railway in 1918.

AT&T launched a commercial cellular network in 1946 in the city area of St. Louis: It used six fixed FM channels and manual switching by an operator. A year later, the Bell Labs obtained a patent on a frequency re-usage scheme deploying a regular cell grid.

The German "A network" was implemented in 1957 and is a classic example of a 1G system — It did provide nationwide coverage, but was divided into 136 areas. Each area used 37 frequency pairs (for up/downlink) and call connection could be done only by an operator.

Such 1G systems still were very similar to the radiobroadcast systems from which they originated; as with broadcast stations, transmitters for 1G systems were placed on high mountains or hills to cover large areas.

Spectrum efficiency was not a major issue; the "A network," for example, never exceeded 11.000 users, while occupying a frequency range of 156–174 MHz (=18 MHz bandwidth). "Planning" these networks consisted merely in finding a few sites on a hilltop and installing the system technology.

Later systems enabled self-establishment of calls, automatic call routing, and handover of outgoing calls to neighboring cells. The networks became able to track and handle the mobile user's position and establish a call to him/her automatically. This did not become possible before the wider availability of digital computers in the 1980s due to the required process automation.

Also, transmitter density increased, not only for better coverage, but also to decrease transmission power and improve spectrum efficiency.

One example of such late 1G systems was the German "C network," which worked with 287 channels in a frequency range of 450–465.74 MHz.

3.2. 2G Systems: Digital Technology

Key characteristics of 2G systems are

- Digital transmissions
- Dense site placements
- Low power transmitters
- Enhanced modulation schemes
- TDMA/FDMA access structures
- Semicompatible systems
- Voice focus, but first data service support
- Step-by-step planning: coverage before QoS (quality of service)

Typical representatives are GSM, DCS1800, PHS, IS-95, IS-136, PCS, and DECT.

In the early 1980s, it had became obvious that even with reuse patterns and denser networks, analog transmission couldn't cope with the increasing user figures any longer.

The C network reached its theoretical limit at about 1 million users only — not much for a potential 80 million German users.

Fortunately, the advances in computer technology provided a new solution for the capacity demands in the form of digital transmissions.

Once human voice is digitized, it can be sent as *compressed* data over the air interface. Exploiting the fact that the human perception system is quite forgiving in slight deviations of spoken voice, even more effective *lossy* compression algorithms could be applied. Enhanced modulation schemes such as GMSK and BPSK (Gaussian minimum shift keying and binary phase shift keying) added to this development. The density of the networks could increase further as time and frequency diversity were combined (GSM) or CDMA principles were used (IS95). This led to lower battery consumption and — together with advances in miniaturization — gave rise to truly small handsets. All this added to the capacity of the networks so that today, for example, a GSM-based network could serve theoretically about 100 million voice users.

Also, globalization effects started to drive standardization. GSM (standardized in 1982, officially released in 1990) became an especially impressive pan-European success.

Even if 2G systems began to offer some basic data services, they are still networks for mobile *voice* traffic only.

This one-service-only property simplifies the planning process to a large extent, as it does not necessitate consideration of the time-domain parameter. As we will see later in more detail, it suffices without significant loss of accuracy to work on *averaged* data and separate the planning process in distinct phases not impacting each other.

3.3. 2.5G and 3G Systems: Two Steps Instead of One

Key characteristics of 2.5G systems are

- Data and voice services
- Mixed services
- Circuit- packet-switched traffic

Key characteristics of 3G systems are

- Mainly data services
- A broad range of very differing services, multi media support
- Mainly packet-switched data
- Enhanced multiple-access schemes
- International standardization
- All-in-one planning: coverage *and* QoS
- Coexistence with 2G systems

In the late 1990s it was realized once again that the air interface would soon become a bottleneck, this time not because of the number of *voice* users (which could be handled by the existing 2G systems), but because of the growing demand for *data* services. The success of the (fixed-network) Internet is expected to lead to a similar demand for wireless Internet services. However, data services require large bandwidths, and, of course, only lossless compression schemes can be used applied, instead of the (more efficient) lossy algorithms used for voice traffic. For example, state-of-the art voice codecs can transmit human voice at acceptable speech quality with a data rate of less than 4 kbps, whereas an "acceptable" WWW session today seems to be in the range of ~64 kbps with demands rising, as multimedia contents begin to overtake WWW contents.

The required bit rate of a typical multimedia service seems to grow by nearly 50% per year. This puts pressure on the capabilities of the air interface, as customers come to expect the same behavior in their mobile service, that they have in their Fixed-network (landline) connections.

3G systems therefore focus mainly on data transmission. This, however, constitutes a complete paradigm change; even the most sophisticated 2G systems are still *circuit-switched* networks — using exactly the same principle as Philipp Reis did with world's first "fixed-network phone" experiment back in 1861! Data services traffic, on the other hand, is usually very bursty in nature, of a *packet-switched* nature. Assigning a dedicated (high-capacity) transmission channel for the entire connection time, although the channel is used for only a small fraction of time by some data packets, thus constitutes a waste of potential: This can't be afforded for the scarce air interface resource.

Similarly, international harmonization efforts for a truly worldwide 3G standard are ongoing. However, existing but still mutually incompatible 2G systems strongly dominated the market and lobbied local interests to such an extent that the outcome of the standardization process was a very complex and demanding one, and also reflected the political situation. Emphasis in the resulting five official air interfaces was placed on CDMA technology, which again represented is a giant transition for most of the existing 2G networks.

Even for the existing CDMA networks (e.g., IS95), the simultaneous support of the mixed traffic scenarios poses severe problems.

3.4. 4G Systems: True Multimedia

Although 3G networks are not even in commercial operation, R&D work on a successor technology, commonly called "4G," has already started.

Key characteristics seem to be

- Very high bandwidths and data rates
- Very asymmetric uplink/downlink traffic
- Data services clearly predominating
- Spotty coverage
- Adaptive antennas

After the initial hype (hyperbole) about 3G possibilities had cooled down to reality, it became obvious that the theoretical limits of this technology couldn't be realized practically.

The required financial requirements would simply be prohibitive. As of today, more and more incumbent UMTS operators have started to reduce customer expectations from 2 Mbps to 384 kbps and even 144 kbps in most for "typical" urban environments. However, the expected demand for true multimedia services is ~20 Mbps.

Also, the inherently assumed asymmetry in the network load of 2–1 for the downlink–uplink ratio does not fit the reality of multimedia applications; a ratio of 5–1 or even 10–1 seems more realistic. These discrepancies indicate the necessity for a new technology.

4. PLANNING OF WIRELESS NETWORKS

The process of planning wireless networks always entails a set of parameters that must be optimized simultaneously. Not surprisingly, some of these "global" parameters interfere each other. For example, good network coverage can be achieved not only with a large number of base stations and low network interference but also with a smaller number of sites — for the price of a resulting higher average interference. These "global" goals must thus be guided by some "local" planning goals.

As this brief example demonstrates, there is no single "optimal" planning solution but a *set* of "equally optimal" solutions. Seen from a scientific point of view, the task of planning a mobile network is a so-called multidimensional optimization problem with Pareto optimality criteria [19].

It can be shown that even subproblems of this task are computationally NP-complete. Without discussing of theoretical computer sciences in detail, this means that the effort to find such optimal network plans explodes exponentially in the size of the network and thus can be solved only *approximately* in reality and/or by simplifying the planning process partially.

Fortunately, the global conditions could be divided into several independent steps, forming the conventional-approach of network planning for 2G and 2.5G networks as was done with manual and PC-based planning systems:

The first classical step of radio network planning is to achieve the necessary coverage area of the desired area (and/or customers). In 2G and 2.5G systems site location and placement planning could be optimized in terms of coverage analysis alone and can be separated from the further QoS (quality-of-service) and GoS (grade-of-service) evaluations. In 3G, however, this will substantially change. Even in the first step of planning, the effects of the CDMA physics on coverage issues (e.g., cell breathing) will dramatically affect the planning strategy.

Basically the traffic, which represents a mix of services (bit rate, whether packet- or circuit-switched, etc.), will result in additional noise in a (W)CDMA cell and thus change the cell's size. This dependency between load in a cell and cell size will cause the cell edges to "float" dynamically and thereby lead to dropped users, if the network is designed poorly. This dependency leads to a situation in which, the well-known separation between coverage and QoS planning as used in 2G and 2.5G will no longer work in 3G systems, which is a challenge for network operators and planning system suppliers. The following sections discuss the conventional 2G and 2.5G planning approaches. We will also outline a more advanced planning algorithm that leads the way to 3G planning. In addition, we sketch how the previously mainly manually performed network planning and optimization task can be automated.

We then address true 3G systems to in a separate section. The results, illustrations, and other aspects are derived using state-of-the-art computer systems and are already used by network operators [2–4].

5. SITE LOCATION AND PLACEMENT

A conventional site is a base station with one or more antennas, where each antenna normally, corresponds to one cell of the network. Several antenna configurations have shown up in the past, while antenna space diversity is a popular configuration in GSM and also in 3G CDMA networks. However, for the planning task itself the antenna configuration is not that interesting, while the antenna characteristics (antenna patterns, azimuth, mechanico-electrical downtilt and gain) are important parameters for controlling coverage and interference [5].

In the 1990s the first planning systems involving a Geographic Information System (GIS) came into professional use. The amount of memory, hard-disk capacity, and processor speed as well as the availability and price of digital terrain models (DTMs) so far prevented an intensive use of digital maps in combination with PC based systems. Nowadays, powerful radio network planning (RNP) tools in combination with GIS and relational database systems are common.

Network planning starts by defining the sites and sectors and defining them in the tool. Multitechnology, multiband planning systems require combining different system technologies (e.g., GSM900, GSM1800, TETRA450) into one database–simulation scenario.

5.1. The Conventional Approach

Sites are still placed manually one, based on a high level of expertise and skill by the planner. A "green field" layout starts by using a regular hexagonal grid based on flat-earth propagation assumptions. This can be done by using a reference site and distributing sites of this type in a user-defined area with a specified distance between the site locations. Even in that simplified approach, defining a cellular site involves at least the following:

- Minimum site parameters
  - Site coordinates
  - Site identifier
  - Number of sectors (BTS)
- Minimum sector parameters
- Antenna parameters
  - Type
  - Azimuth
  - Mechanicoelectron downtilt
  - Antenna height
- Other parameters
  - System technology
  - Radiated power.

Modern tools [2] support the user in doing this with high-efficiency user interfaces of RNP tools and automatically factoring in terrain and other environmental information. Graphical interaction within the GIS system and the database enables the user by easy click and drop mechanisms to define the sites. The user can select from

a template of sites and place them on the geographic location, move, copy, re-arrange, create, or delete sites and/or cells and quickly evaluate these scenarios.

Figure 1 shows an example of a German city (Munich) after automatically placing a regular hexagonal site grid without optimization as the first network scenario. The next step of the radio network planner is then a manual refinement of the grid in order to account for additional conditions such as street orientations and high traffic areas, sectorization). Such a manual refinement of the network elements, taking clutter structure into account, leads to the layout depicted in Fig. 2.

In the scene shown in Fig. 2, the responsible engineer has optimized the grid layout, changing site positions, the number of sectors, antenna azimuth and distance between sites to accommodate the topology and terrain types within metropolitan Munich. The tools support this task by quick analysis of expected cell field strengths, line-of-sight (LoS) checks, and other methods see Fig. 3. To check the overall improvements resulting from his site placement, the planner will recalculate the combined coverage of the cells he/she created. Modern planning tools have different algorithms for wave propagation analysis, which is accompanied by calibration features from the network planning tool. The different propagation models used for the different cell types are discussed in Section 6.

Figure 4 shows the coverage of the Munich network (75 regular cells) based on a regular grid without optimization. One can see that in the city center the coverage level is poor because of the high degree of attenuation by building

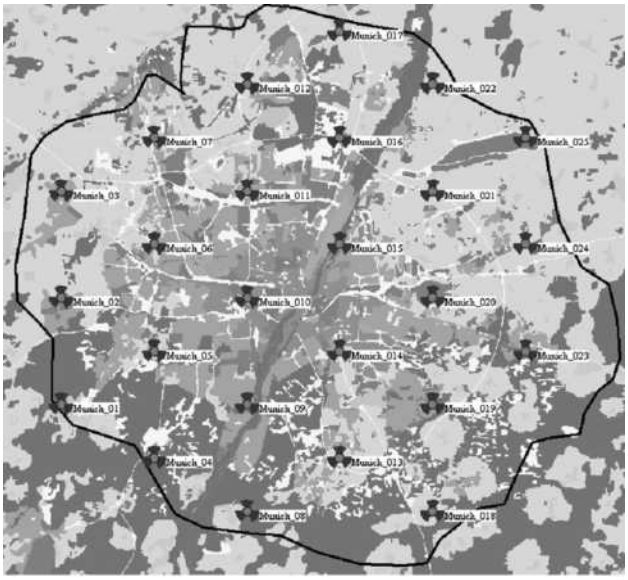


Figure 1. Regular grid layout (Region of Munich, Germany) with network area borders.

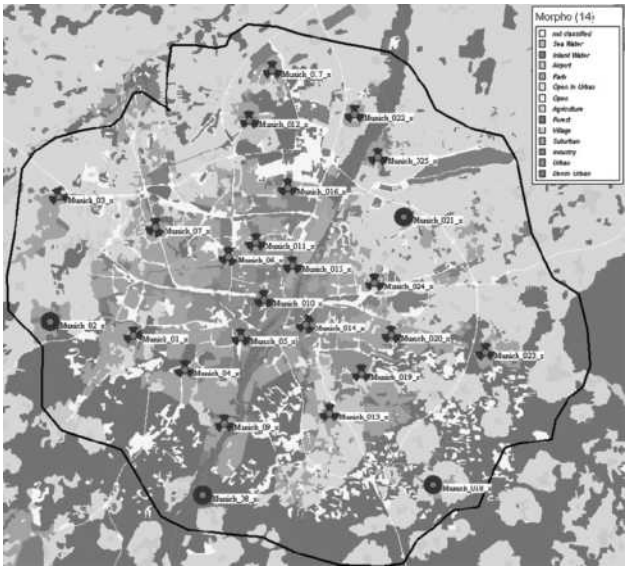


Figure 2. Manually optimized site locations taking into account morphology structure to accommodate possible high traffic areas.

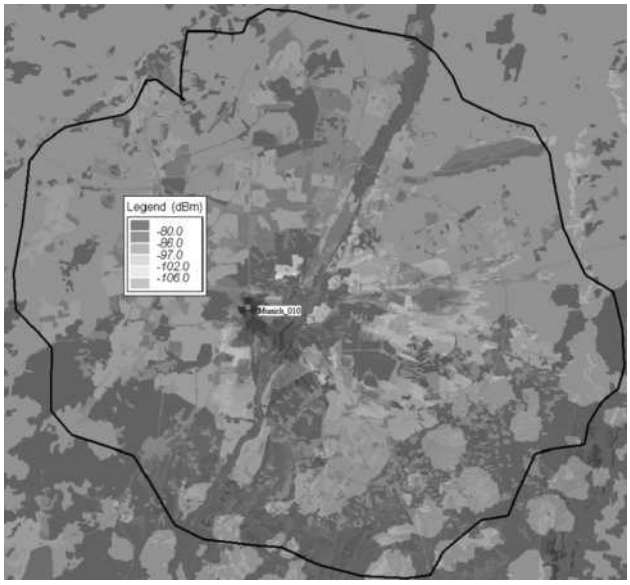


Figure 3. Single-cell field strength transparently overlaid on clutter background map.

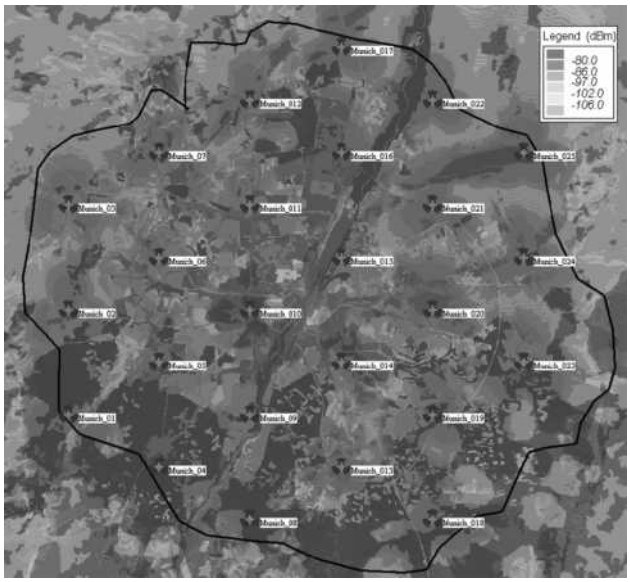


Figure 4. Networkwide coverage for regular grid layout.

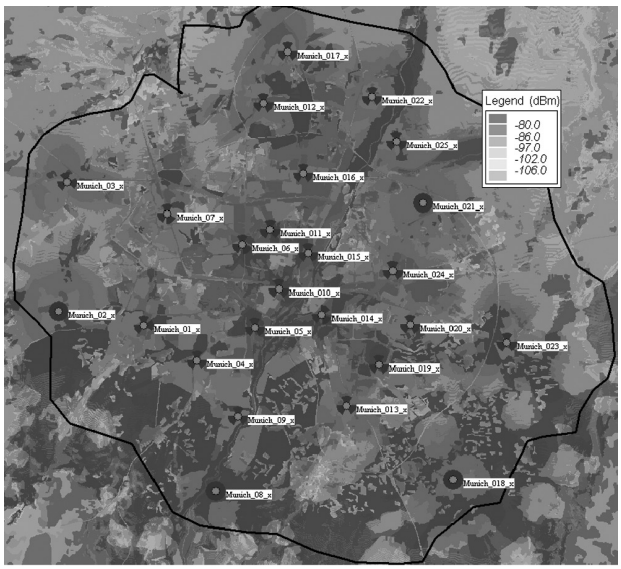


Figure 5. Coverage areas optimized for expected cell traffic especially in the city center.

structures. In Fig. 5 the coverage is enhanced according to the clutter structure and expected service areas with a lower number of cells (66 instead of 75) achieved by manual optimization of the site locations as well as changing of the site configuration (OMNI sites in lower traffic areas). Obviously, the expertise of the experienced planner paid off by a higher coverage, although nine sites less are required than in the first hexagonal grid scenario.

5.2. Automatic Site Placement

From the above paragraph it is obvious that the radio network planner's expertise and experience can be used to optimize the problem. Although the outcome seems to be fine and the number of base stations may appear to have reached a minimum, there is still plenty of room for optimization. This is because, as mentioned above, this is a multidimensional optimization problem with Pareto optimality criteria and NP-complete subproblems. Translated into simple words, this means that it is impossible for a human planner to actually find an optimal network solution. Watching how experienced network planners tackle the planning job, it seems that most of them aggressively optimize for coverage by using "promising" cell candidates first — and sticking with these, even if giving up on one cell for one or two other candidates nearby leads to better network solutions. None of them was able to plan for coverage and interference *in parallel*. The multidimensioning problem was "solved" by assuming that a low number of cells also automatically corresponds to a low interference (a heuristic that may not be the case, however). A new method has therefore been developed to overcome the time-consuming, yet suboptimal manual cell selection (choosing optimal sites from a given set of candidates), cell placement (finding the set of optimal new candidates in a "green field" situation), and cell dimensioning (finding the set of optimal parameters for a given set of cells) problem [6].

An initial network situation is iteratively improved. Several steps involving placement, selection, and parameter dimensioning are alternated and repeated. All of this is steered by a special "genetic algorithm" to break down the huge search domain into important, but practically searchable subspaces. An algorithm based on the technique of cell splitting is also used. This is intended to accommodate for traffic issues.

The boundary conditions for the local and global conditions for the objective function to be minimized are *area coverage rate*, *traffic coverage rate*, *spectral costs*, *geographic functions*, and *financial cost functions*. Here, the operator can guide the tool by defining which of the (Pareto-) equivalent solutions is preferred.

Again, the regular grid layout (Fig. 6) is used as a starting point for cell placement, while a mode exists to use the algorithm for cell selection as well (finding the optimum cells of a given set of cells) (see also Fig. 7).

For example, the coverage rate is calculated using a propagation model based on DEM data (topography and clutter) and used to calculate the influence of neighboring cells using the assignment probability of each cell in the selected network DEM (Digital Elevation Model)

$$A_{\text{cov}}(S) = \sum_{x=0}^{n-1}\sum_{y=0}^{m-1} \delta \cdot P_{\text{cov}}(x, y, S)$$

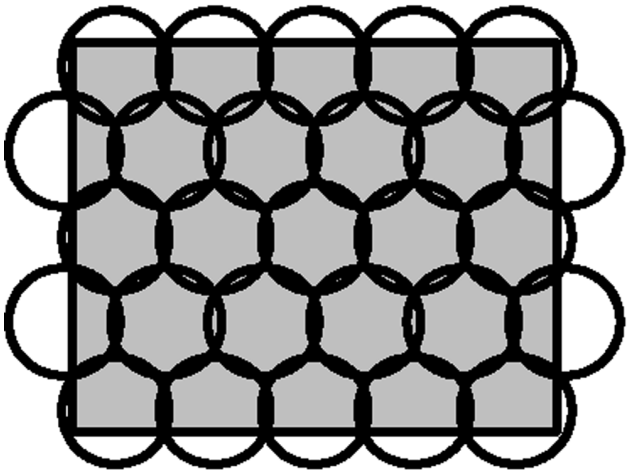


Figure 6. Regular cell grid.

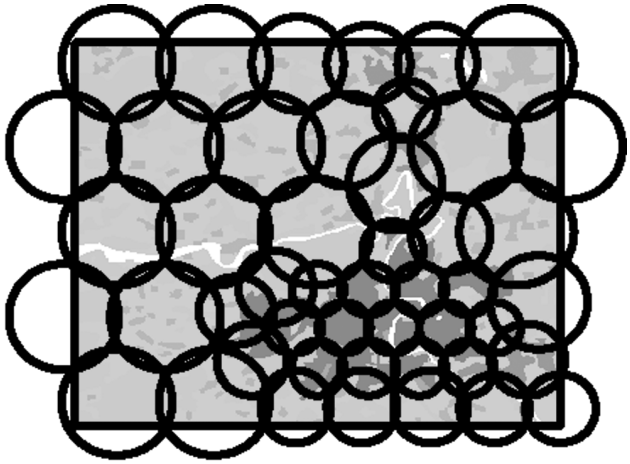


Figure 7. Cell refinement.

where δ is the pixel size and $p_{\text{cov}}(x, y, S)$ is the probability of a mobile at (x, y) being served by *any* cell in the admissible cell set S. This takes into account the handover between cells. The area coverage rate is then calculated as

$$f_1(S) = \frac{A_{\text{cov}}(S)}{A_s}$$

where A_s is the service area size. For the other costs, similar objective functions can be defined and used in a multiobjective optimization algorithm. A hierarchical workflow of optimization is shown in Fig. 8 (see also Fig. 9).

Until now one cell per site is considered and the cell size is as large as possible. What happens if the traffic demand increases? What has to be done if the existing cells cannot handle the scaled-up traffic? The method commonly used in practice is cell splitting, that is artificially reducing the existing cell sizes and adding new cells in between. The problem is then to find optimal base station (BS) sites for new cells and optimize dimensions of both original cells and additional cells so that the growing demand can be met effectively, while offering minimum disturbance to the existing network structure. Nominal cell splitting is illustrated in Fig. 10. Initially, the largest possible cell sizes are used, one cell per site. In the next step, a cell is divided into a number of sectors. Here only the three-sector case is considered. Each sector is served by a different set of channels and illuminated by a directional antenna. The sector can therefore be considered as a new cell. As a consequence, there are three cells per site using the original BS sites. BSs are located at the corner of cells, as shown in Fig. 10b. Now the number of sites is still the same, but the number of cells is 3 times higher than before. The following step is to do further cell splitting, specifically, reducing the size

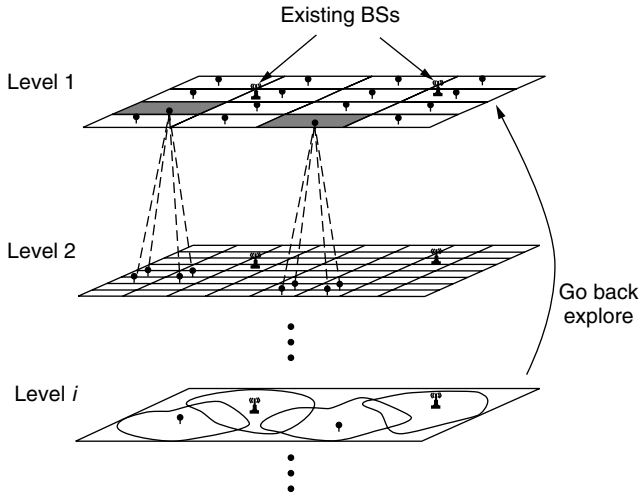


Figure 9. Structure of the hierarchical approach.

of existing cells and adding new cells. As can be seen in Fig. 10c, the former sites are still used in the new cell plan, but additional sites are now required for serving new cells.

Such algorithms will be integrated into modern network planning tools in order to accelerate network optimization. These are now the edge of modern research and will influence the future planning process.

6. PROPAGATION MODELING

Propagation modeling is the key issue for proper network planning. All further results such as coverage probability, interference, and frequency assignment depend directly on the quality of propagation prediction. Various models

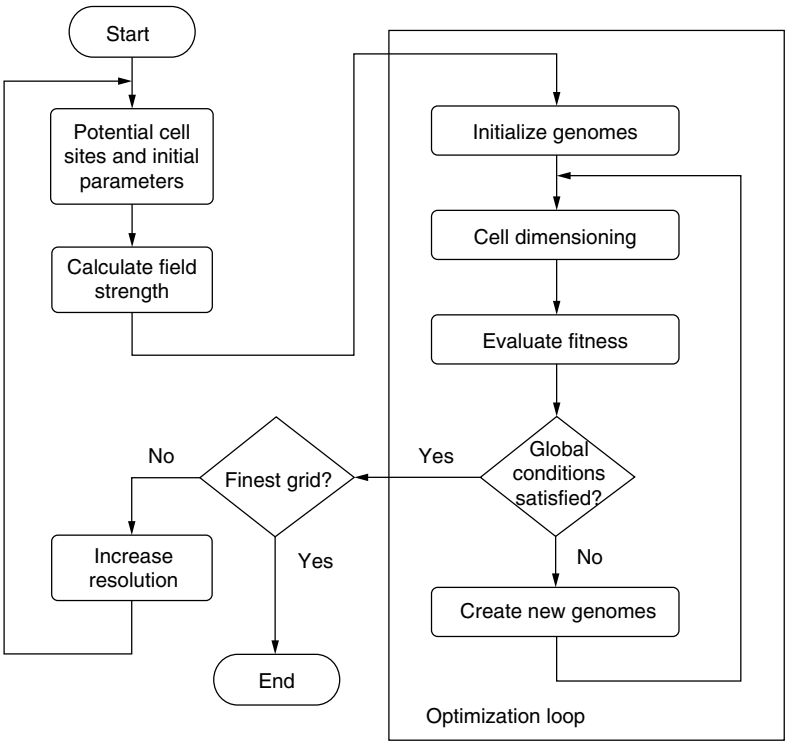


Figure 8. Hierarchical optimization process.

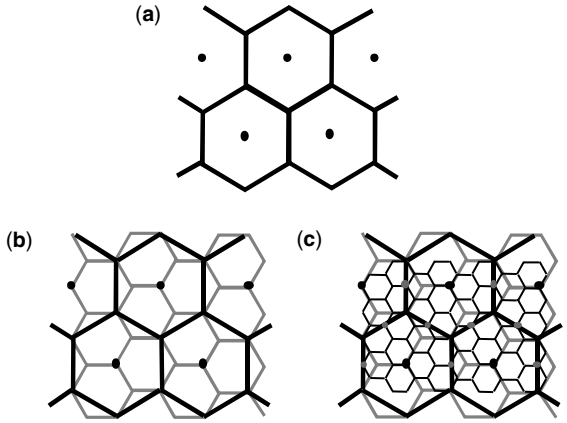


Figure 10. Nominal cell splitting: (**a**) initial cell plan; (**b**) phase 1 — each cell is divided into three cells, using original sites; (**c**) phase 2 — old cells are reduced, new cells are added, requiring additional sites.

have been developed for different frequency ranges and services. In the area of cellular (mobile) communications, the best known models are the Okumura–Hata and Walfish–Ikegami models, which have both predominated in the technology of 2G planning tools since 1980 or so.

6.1. Macrocell–Minicell–Microcell

Modern RNP tools can accommodate the various types of models for different cell layers. The macrocell layer still represents about 90% of the whole network planning; the rest is divided into microcells (cell ranges of ~100 m) and minicells (cell ranges ~500 m–2 km). The most important propagation calculation to be discussed in this section is the macrocell and its calibration, as this model is still the most widely used one in order to reduce computation time and cost for complex building data.

6.2. Model Calibration

The most widely used propagation model for macrocells is a parametric one such as the Okumura–Hata (OH) model. Many investigations and measurement campaigns have been performed in different frequency ranges and environments in order to adapt those simple models to the desired services. Because of the parametric nature of these models, calibration is quite simple and serves as the starting point for the network planning. Model calibration always compares measured drives with the predicted values. Manual, semiautomatic as well as fully automatic calibration techniques are used. The quality of calibration depends on the measurements taken for specific cells and on the quality and resolution of the topological and clutter data used in the tool used for calibration. Care must also be taken to avoid using only the transmitter point only, but the whole profile path to the receiver for conducting the calibration. The parameters of the OH model are well known from various study groups [14] for the frequency ranges around 900 and 1800 MHz (see also Fig. 11). The parameters to be optimized are mainly the clutter correction data, the gain, and the height of different land-use classes (see Fig. 12), and, to some extent, the other

Parameter	Nom. Value	min	max
a1	46.30	40	50
a2	33.90	28	40
a3	–13.80	–20	–8
b1	44.90	35	55
b2	–6.50	fix	fix

Figure 11. OH parameters and their nominal ranges for calibration (1800 MHz).

Clutter Class	Gain dB	Range (+/–)	Height (m)	Range (+/–)
not classified	0	0	0	0
Sea Water	28	4	0	0
Inland Water	27	4	0	0
Airport	6	4	5	2
Park	6	3	7	3
Open In Urban	8	8	0	0
Open	23	2	0	0
Agriculture	8	6	3	2
Forest	20	6	12	6
Village	15	8	7	3
Suburban	12	4	10	5
Industry	2	2	25	10
Urban	8	4	14	7
Dense Urban	6	3	17	9

Figure 12. Clutter parameters and their nominal ranges for calibration in Europe (1800 MHz).

parameters of the Hata equation. Typical parameters to be calibrated for a macrocell model such as OH for one frequency range (e.g., GSM1800):

$$L = a_1 + a_2 * \log_{10}(f) - a_3 * \log(h_{\text{eff}}) + (b_1 - b_2) * \log_{10}(h_{\text{eff}})(\log_{10} d)$$

where f is the frequency of operation and h_{eff} is the effective height. Several methods exist for the determination of effective height between BS (base station), and MS (mobile station):

- Height of BS antenna above ground
- Height of BS antenna above mean sea level
- Height of BS antenna in relation to the effective terrain height according to ITU
- Height of BS antenna in relation to the effective terrain height over the entire profile
- Height of BS antenna in relation to the MS antenna
- Effective height/distance ratios by rotation of the terrain against its ascent

Measurement drives are imported into modern RNP tools [2] where multiscreen, coupled cursor optimization features are used to calibrate the propagation model/clutter parameters for the averaged measured values (average using time or space windows smoothing the measured drive) (see Figs. 13–14).

The typical outcome for calibrating macrocell models are a mean value of around 0 dB of the difference and a standard deviation of 5–10 dB.

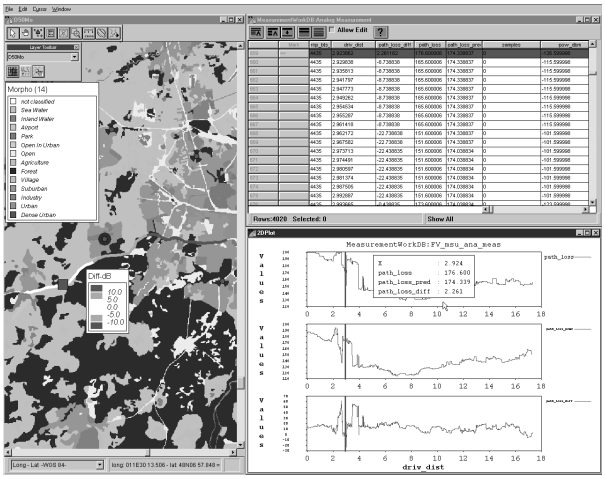


Figure 13. Measurement evaluation screens in modern RNP tools.

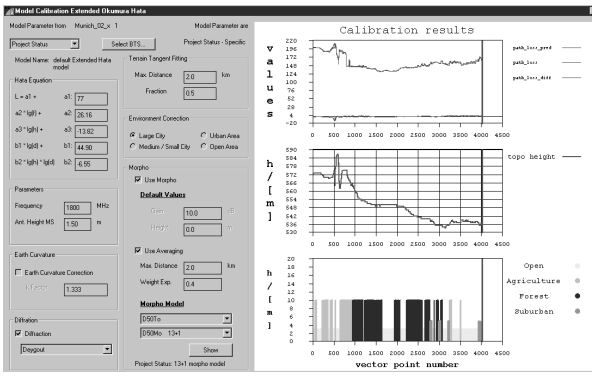


Figure 14. Calibration engine GUI, (graphical user interface), example.

7. 2G NETWORKS (GSM)

Let us summarize the strategies for 2G network planning first. The planning process, conducted either manually or automatically, consists of a number of *separate* planning steps:

- Determine a (an initial) cell selection
- Issue a coverage predictions (physical propagation modeling)
- Derive the "best server" areas
- Determine the traffic load per cell
- Calculate the number of frequencies needed
- Do a frequency plan assignment
- Analyze the resulting interference situation
- Repeat the loop for network optimization

We address some of the more important results required for this process in the next section. Please note the critical prerequisite that these steps are assumed to be *independent* of each other to a large extent; that is, if the traffic load exceeds a cell's capacity, it can be extended by adding another transceiver to the cell. If the required frequency is chosen from a previously calculated "candidate" set, the whole network can be assumed to not have changed significantly. This assumption will not hold true for 3G networks (or 2G, CDMA-based networks) and thus, require additional effort.

The strategy of applying planning tools to solve the different tasks can be generally divided into two basic approaches, a so-called "deterministic approach" used by most RNP tools and a more enhanced "probability based" approach. The following sections describe the different strategies.

7.1. Deterministic Approach

Traditional RNP tools deal with coverage aspects only, where the propagation model calculates for each possible mobile station (MS) location the power level according to the appropriate model. In the deterministic approach (see Fig. 15), only these cell specific coverage results are used to evaluate networkwide coverage and interference. An advantage of the deterministic approach is the simple calculation dependencies as the only inputs for networkwide evaluations are the cell-specific power files. Another advantage is the rapid computation time even for networks with thousands of base stations.

The input files (the cell power results) are combined with a network engine to achieve the following results described.

7.1.1. Coverage-Based Results. The following results are typically used to plan or to optimize the coverage of the network:

Input parameters — network access level [e.g., −98 dBm and maximum allowed timing advance (TA) for TDMA systems].

Generated outputs — only pixels that have a power level higher than the network access level and are inside the TA of the corresponding TDMA system (GSM: TA ∼35 km) are considered.

7.1.1.1. Maximum Server. For each pixel in the calculation area shown in Fig. 16, the transmitter produces the strongest power level compared to all others at that pixel.

7.1.1.2. Networkwide Coverage. Figure 17, shows the strongest power level in dBm for each pixel in the calculation area, indicating the maximum value of all contributing signal sources.

7.1.1.3. Strongest Interferer at Strongest Server. Figure 18 shows for each pixel in the calculation area the transmitter that causes the highest interference relative to the serving cell.

7.1.2. Interference-Based Results. The following results are typically used to plan or optimize the interference (service) of the network: *input parameters* — network access level (e.g., −98 dBm), maximum TA (e.g., 35 km in standard GSM; i.e., maximum signal delay that can be compensated by the system to remain synchronized) and

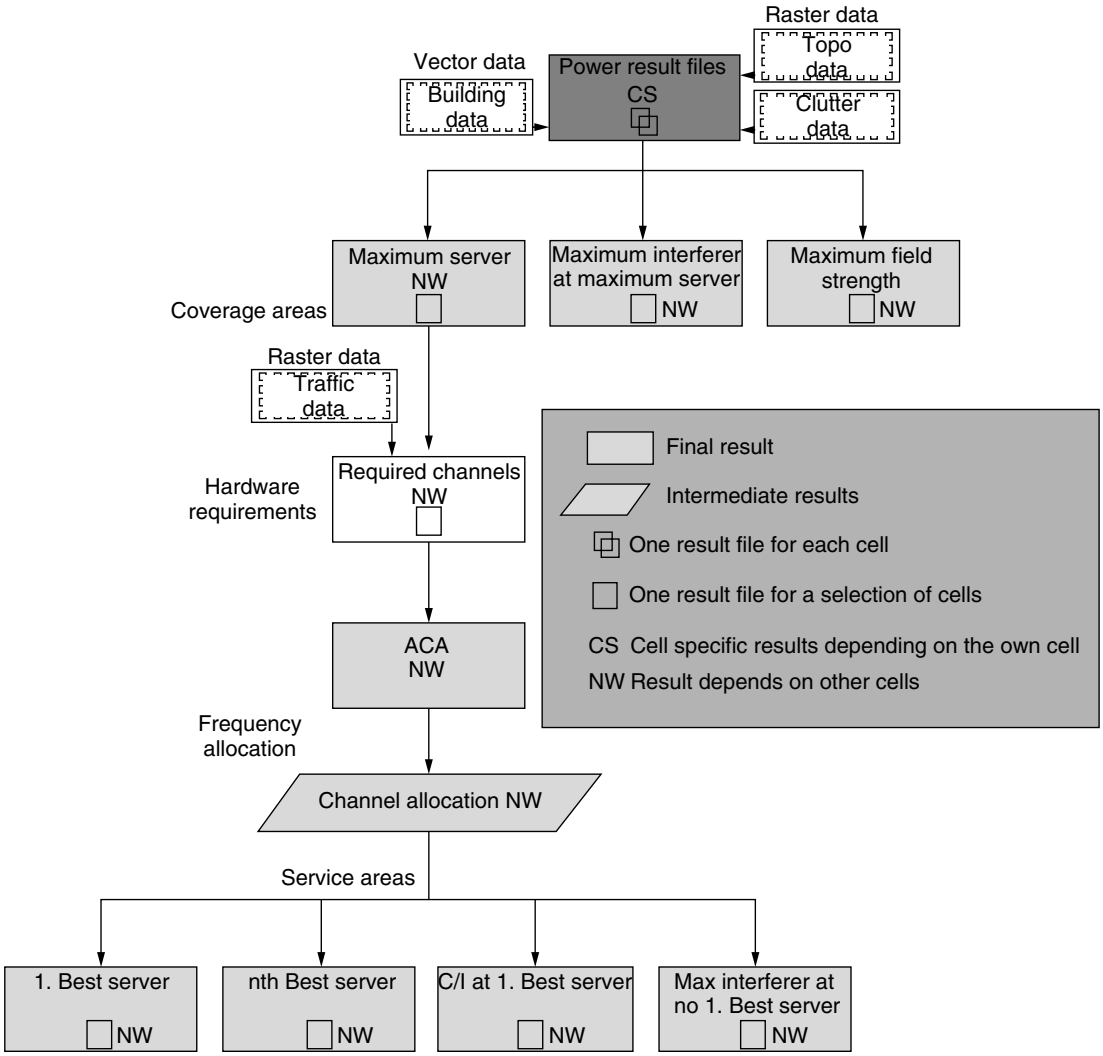


Figure 15. Calculation workflow in the deterministic approach.

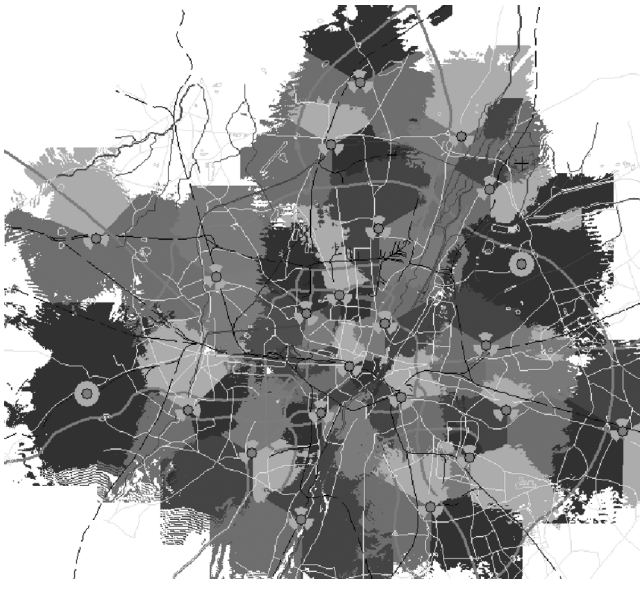


Figure 16. Maximum server result.

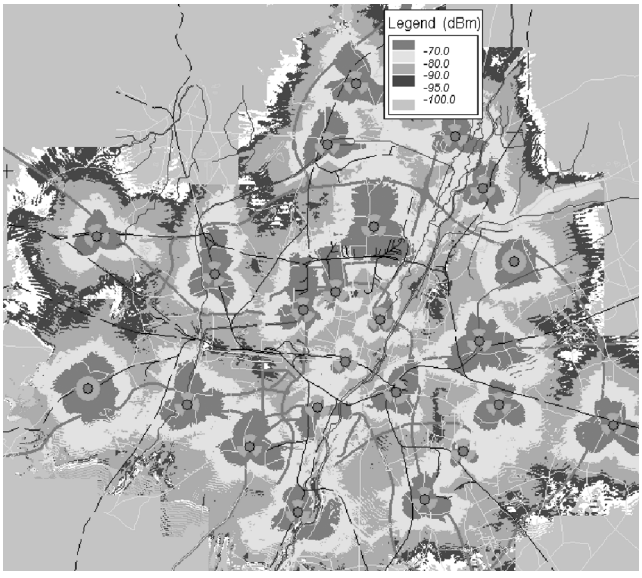


Figure 17. Networkwide coverage result.

	Co	1. Adjacent	2. Adjacent	3. Adjacent
Channel distance	0	1	2	3
Ratio/dB	9	–9	–41	–49

Figure 18. Protection ratios for GSM system.

protection ratios for co- and adjacent channel interference. For the GSM system, the values are listed in Fig. 18.

The table in Fig. 18 should be read as follows. In case the server produces at a pixel in the calculation area a power level of −65 dBm (say), then this server is assumed to be interference-free and can provide service if a potential interferer at the same channel ("co channel") is not stronger than −74 dBm. A potential interferer having two channels' distance ("2 adjacent") is allowed to produce a level of −14 dBm. Those protection ratios are also used as a quality criterion for the frequency planning.

*7.1.2.1. **Best Server/Best Server.*** This type of result shows for each pixel in the calculation area the transmitter that causes the first or nth strongest field strength at that pixel *and* is not disturbed by interference (according to the protection ratios). For this, only pixels with a serving field strength above a certain threshold *and* lying inside the timing advance range (relative to the serving transmitter) are considered. For example, if $n = 3$, the user will get three results. The first is the first best server, second best server, and so on.

*7.1.2.2. **Carrier-to-Interference Ratio.*** Figure 19 shows a QoS (quality-of-service) type of result. A high carrier-to-interferer ratio (C/I in decibels) ensures a high quality of the connection. The C/I at best server determines the carrier-to-interferer ratio C/I (in dB) for each pixel that fulfills the best server criterion in the calculation area. Therefore, in the result window all pixels are colored (served) according to the C/I ratio in dB. QoS is an important criterion for network operators as high-quality voice and data connection is a key issue to attract potential customers to a specific operator.

*7.1.2.3. **GoS (Grade of Service).*** Another important point is the so-called grade of service (GoS) of the network. Even if the network is excellent in terms of interference, a shortfall of system equipment can dramatically reduce the performance of the cells. The expected number of users producing the air traffic is a further key input to network planning. From marketing surveys an estimated load is extracted and a traffic map can be generated. Load of a (circuit-switched) telephone network is described using Erlang's [10,11] formulation for blocking and queuing systems (Fig. 20). GSM is a typical blocking system, where TETRA, for instance, is a queuing system [15–17]. The dependency of Erlang B (blocking) and Erlang C (queuing) and the equations can be found, for example, in Ref. 11.

The Erlang B formula expresses the relation between the expected traffic in a cell and its hardware, and in a GSM system, the number of time slots necessary to carry the traffic in the cell.

$$P_{\text{block}} = \frac{\dfrac{A^n}{n!}}{\displaystyle\sum_{i=0}^{n} \frac{A^i}{i!}}$$

This equation describes the relationship between three variables: the blocking probability P_{block}, the traffic load A (in erlangs), and the number of channels n. Obviously, the blocking probability increases with the traffic load and decreases with the number of channels, but *not* linearly, as shown graphically in Fig. 21. Simply stated, this nonlinearity means that an operator runs out of the last 10% (say) much network capacity a much more quickly than he/she would for any previous 10%, making capacity

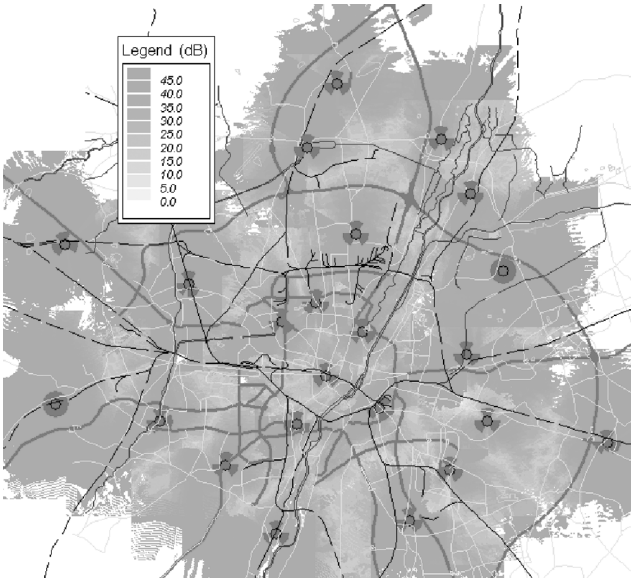


Figure 19. C/I result for QoS optimization.

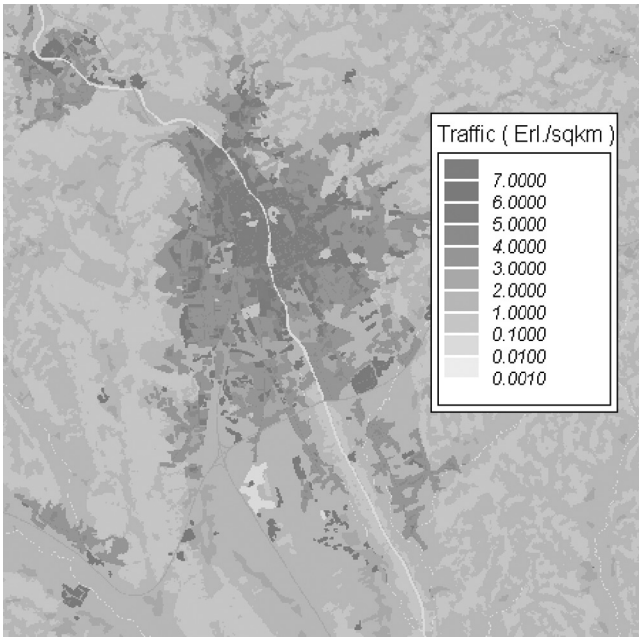


Figure 20. Traffic layer in Erlang Formula (in km^2).

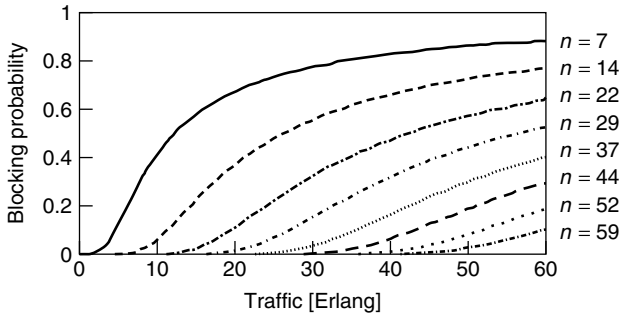


Figure 21. Example of Erlang B curves.

planning a crucial aspect of network planning. Because of its importance, the formula has been widely tabulated. Readers are referred to Lee [7] for a complete Erlang B table. By knowing any two of the three variables (A, n, and P_{block}), one can derive the third.

Assuming that the average conversation time is T seconds and the number of calls per subscriber at the busy hour is λ, the traffic produced by an average subscriber is

$$a = \frac{\lambda \cdot T}{3600}$$

and can be interpreted as the fraction of time that each user occupies a channel. A typical mobile European customer is loading a cell with about 30 millierlangs, where this figure strongly depends on the region and habit of the subscribers. The load of a subscriber varies between roughly 10 and 80 millierlangs throughout the world.

The number of time slots in a TDMA system is correlated with the number of transmitters (TRXs) for a given sector.

Typically network operators will equip their networks mostly with predefined site configurations to minimize the materials management and to overcome the problem of getting suitable marketing — and therefore traffic information. Typical site configurations for an initial GSM network (eight time slots per physical channel) are listed in Fig. 22.

7.2. Probability Approach

The convention approach described in the last section is a bit simplistic — a pixel is said to be either served (completely) by a cell or not be served by a cell (at all). This assumption does not hold true in reality, especially for pixels at the very edge of its best server area. Replacing the digital yes/no "being served" information by a *probability* of being served translates the whole planning process into the fields of probability theory and allows more detailed insight into the planning process. The major difference compared to the deterministic approach is the conversion of deterministic cell power results into assignment probabilities. Assignment probability results are still cell-specific (with one result file per cell) but dependent on the power level of each cell *and* other cells *and* the neighborhood relations and parameters between cells. One new important point addressed is the handover simulation (which has been completely neglected in the old deterministic approach). Cellular systems make it necessary to hand users from one cell to the next cell by handover (handoff) strategies in order to continuously serve their moving users. The probability approach uses the same strategy a mobile station would use, measuring and reporting all received field strength from the neighboring cells and the base station, deciding whether the mobile is handed to one cell or to the other. The dependencies of calculation for the probability approach are shown in Fig. 23.

The calculation of the assignment probability uses the following equations to convert the cell power results to an assignment probability. The assignment probability of a cell b_i at location (x, y), $p_{\text{ass}}(b_i, x, y)$, is defined as the probability of a MS at (x, y) being served by the cell b_i. It is obvious if the MS is served only by one cell that the probability will then be 100%:

$$p_{\text{ass}}(b_i, x, y) = \frac{F(b_i, x, y) - F(b_1, x, y) + DEF\_HO\_MARGIN}{\sum_{j=1}^{N} (F(b_j, x, y) - F(b_1, x, y) + DEF\_HO\_MARGIN)} \cdot P_{\text{tot}}$$

where
$p_{\text{ass}}(b_i, x, y)$ = assignment probability of cell b_i at coordinates x, y
$F_{\text{ass}}(b_i, x, y)$ = Power level of cell b_i at coordinates x, y
p_{tot} = Sum of all assignment probabilities of all cells in the selection
DEF_HO_MARGIN = handover margin for neighborhood relations

This is the basic difference between the deterministic and the probability approach. The cell-specific assignment takes into account the handover margins and therefore simulates a more realistic behavior of the mobile in the network, especially at cell edges. The following figures show examples for the cell-specific assignment. In Fig. 24 a single cell is calculated using an omnidirectional antenna on a flat terrain showing 100% probability that a mobile inside the red area is served by this cell (because no other cell is serving that area). The edge of the red area is in this case determined by the minimum access level of the MS which was set to −95 dBm.

Figures 25–27 show the assignment of each cell in a small network. These more detailed results are used for

	Sectors	Antenna	TRXs/Cell	TRXs(total)	Capacity	Subscribers/Site
Rural 1	1	1 × 360 deg	2	2	8 Erl.	250
Rural 2	3	3 × 120 deg	2	6	35 Erl.	1100
Urban	3	3 × 90 deg	3	9	55 Erl.	1800
Dense Urban	4	4 × 60 deg	3	12	80 Erl.	2600

Figure 22. GSM site configurations.

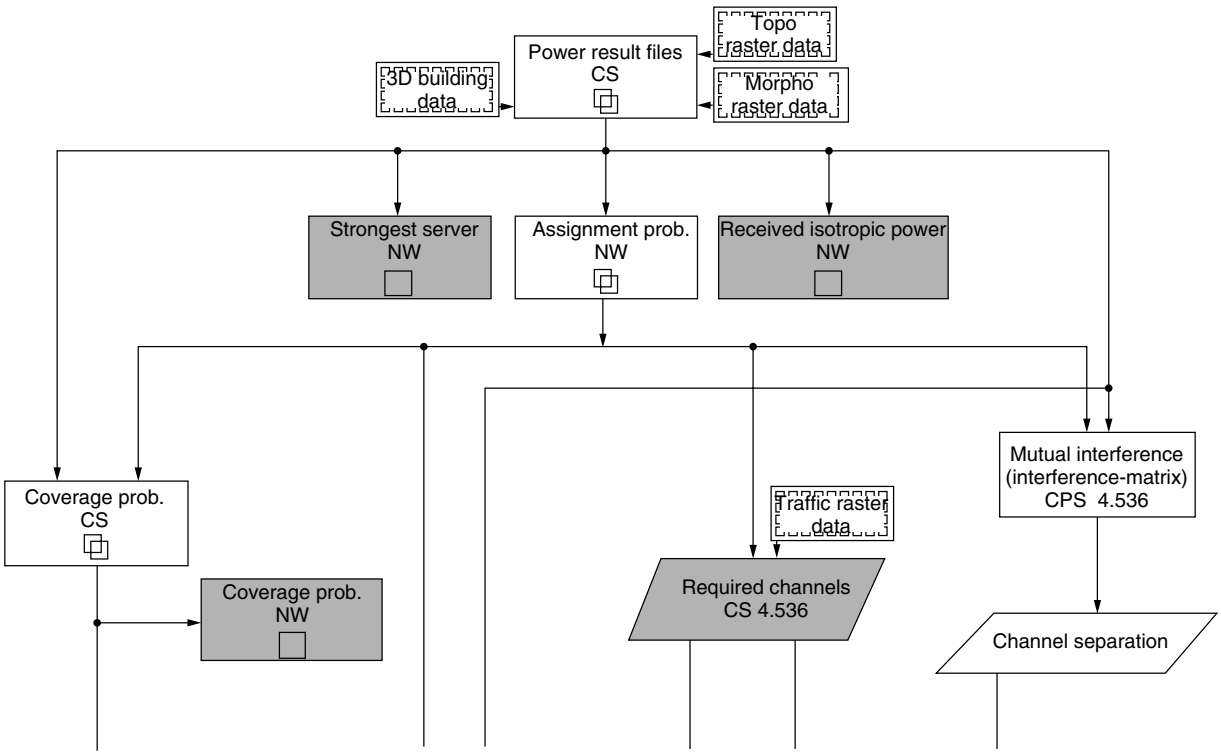


Figure 23. Dependencies of calculation; major differences to deterministic approach.

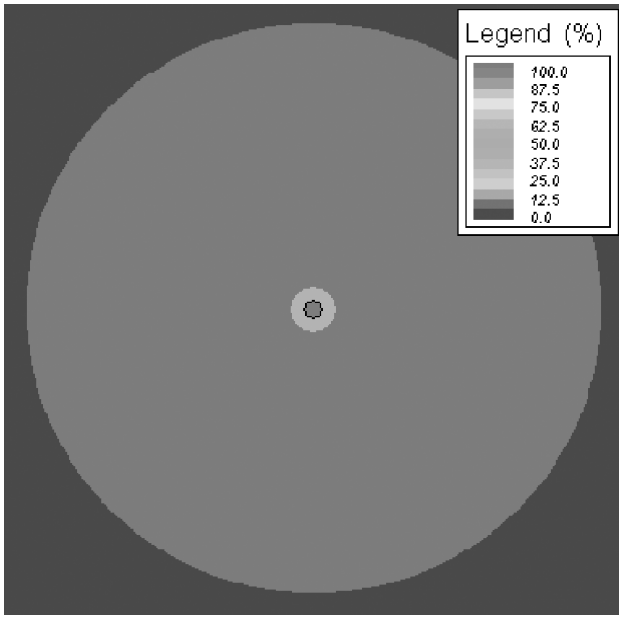


Figure 24. Assignment of a single cell.

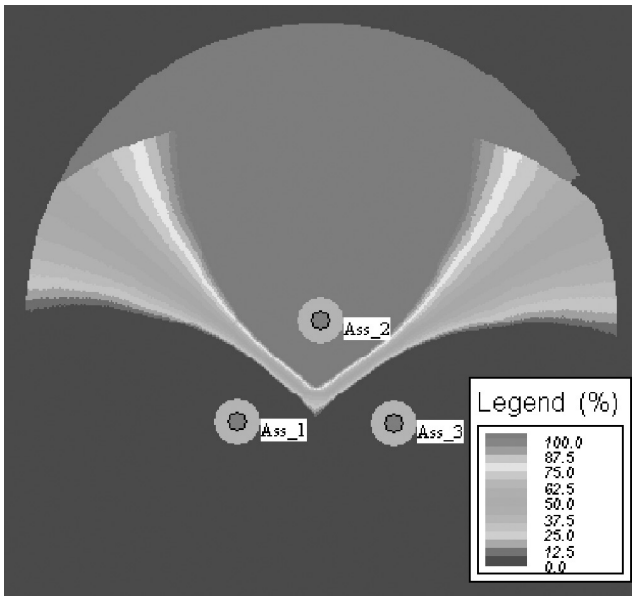


Figure 25. Assignment of cell Ass_2 while Ass_1 and Ass_3 are also serving.

further evaluations-such as coverage probability, mutual interference, and channel separation.

The coverage probability is determined by

$$P_{\text{cov}}(x,y) = \text{erf}\left(\frac{F(x,y) - F_{\text{thr}}}{\sigma(x,y)}\right)$$

where $P_{\text{cov}}(x,y)$ = coverage probability
erf = error function
$\sigma(x,y)$ = standard deviation
F_{thr} = power threshold
$F(x,y)$ = power level at x, y

Figure 28 shows the networkwide coverage probability of this simple network. The coverage probability is defined as the probability that the field strength of the signal from the BTS is greater than a given threshold.

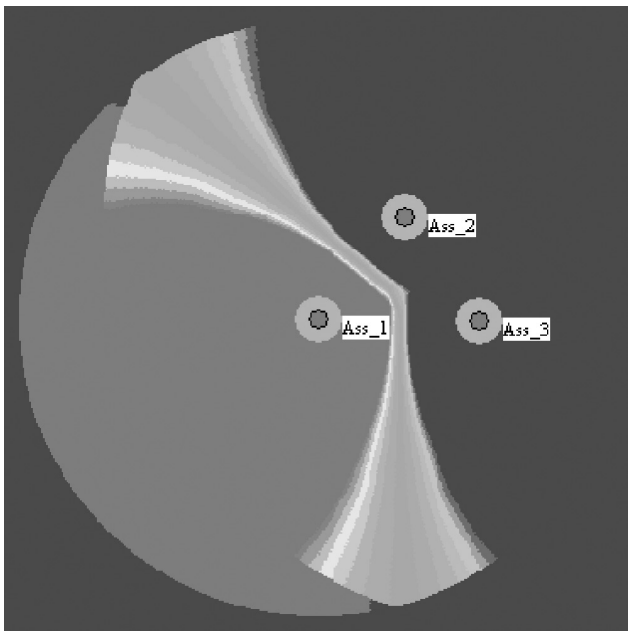


Figure 26. Assignment of cell *Ass_1* while *Ass_2* and *Ass_3* are also serving.

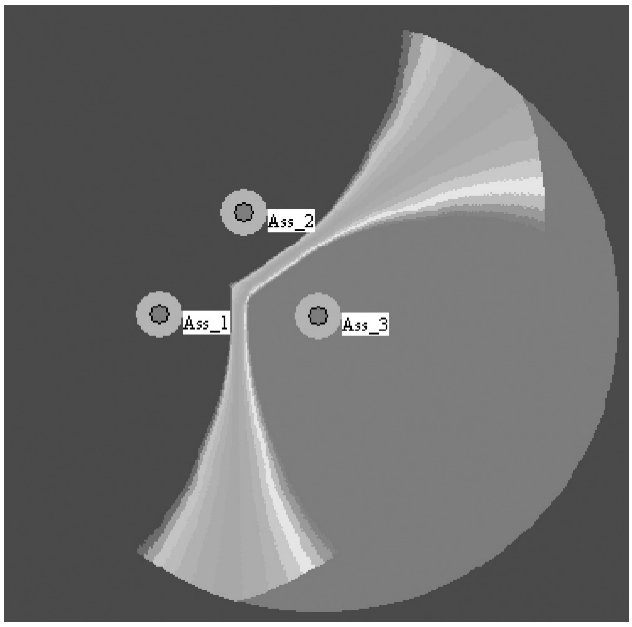


Figure 27. Assignment of cell *Ass_3* while *Ass_1* and *Ass_2* are also serving.

In a realistic network the coverage probability will resemble Fig. 29 [where F_{thr} was set to -95 dBm, $\sigma(x, y)$ to 6 dB].

7.3. Frequency Assignment and Optimization

Worldwide research activities have been done on the channel assignment problem (CAP). It is also called frequency assignment problem (FAP) in some literature [8] and has been shown to be NP-complete for subproblems. Once the cell sites and dimensions are determined, there is a lower-bound, which the minimum frequency spectrum

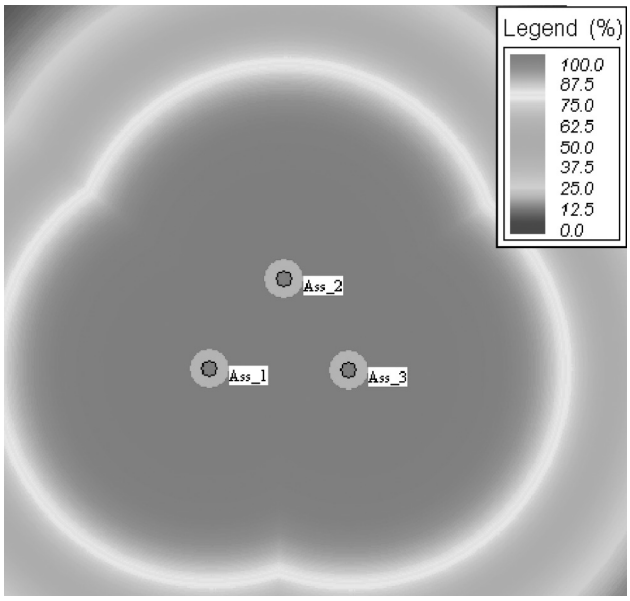


Figure 28. Coverage probability.

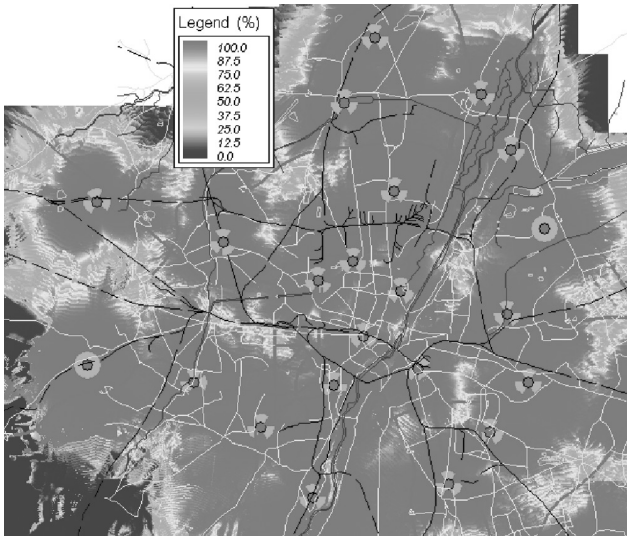


Figure 29. Coverage probability of Munich network.

required for assigning each cell a sufficient number of channels [9]. The base for solving the CAP is the so-called channel separation matrix, where the minimum channel separation between all cells in the network under evaluation are stored. This is a pair-to-pair relationship as each cell is considered as a server and all remaining cells considered as potential interferers. The boundary conditions for a channel assignment problem are: (see also Fig. 32).

7.3.1. Global Conditions. These are as follows:

Allowed frequency band (channels) — for example, GSM 900 (1-124), GSM1800 (512-885).

Channel types — some operators additionally divide their assigned spectrum by channel types (traffic or control channel to further enhance the quality

as control channels are separated from the traffic channels), for instance, to protect the more "valuable" control channels (BCCH) compared to the regular traffic channels.

Figure 30 shows a typical division of the frequency band in the GSM 900 range for one of two GSM 900 operators in a country.

7.3.2. Network-Specific Conditions

Channel separation between all cells can either be calculated or set manually.

Neighborhood relations — cells in the handover list typically will have additional channel separation requirements

Number of required TRXs: these are either calculated (by Erlang equations and assumed spatial traffic load) or manually set; see Section 7.1.2.3.

Neighboring countries — coordination issues, frequencies that are not allowed to be used at country borders, international or mutual agreements between foreign regulators or operators (Vienna agreement [18]). In a planning tool coordination issues can be set up by forbidden channels for cells radiating toward the border.

The outcome of a successful channel assignment is a channel/frequency plan that fulfills the boundary conditions or minimizes cost functions. The channel plans are stored in the RNP tools database as shown in Fig. 33. The allowed frequency spectrum for this GSM 1800 example network was from 600 to 725 and from 800 to 860 as defined in the RNP tool as global conditions (see Figs. 30 and 31 as cell-specific conditions).

The results in Figs. 34 and 35 show a test case of a C/I calculation for the network of Munich in the theoretical case that all TRXs would be operating on the same frequency compared to the C/I after an automatical frequency assignment.

8. 2.5G NETWORK EXTENSIONS (HSCSD/GPRS/EDGE)

GPRS (general packet radio service), EDGE (enhanced data rates for GSM evolution) and HSCSD (high-speed circuit-switched data) have been designed primarily as upgrades to the well-known and heavily deployed GSM standard. The same applies to IS136+ and IS136 HS in the case of the IS136 standard. In the starting phase of GSM and IS136 systems, data transmission issues were of minor importance compared to voice transmission. Beside this fact, the maximum transmission speed of 9.6 kbps that plain GSM and IS136 offered, appeared to be sufficient and was comparable with analogue wireline modem speed at the time when the white papers were drafted. In the 1990s, in particular with the increasing usage of the internet, higher data rates were provided on the fixed modem lines while GSM and IS136 still stuck with the 9.6–14.4 kbps.

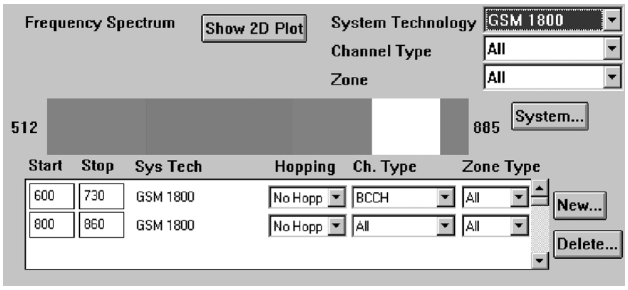


Figure 30. Global conditions for CAP.

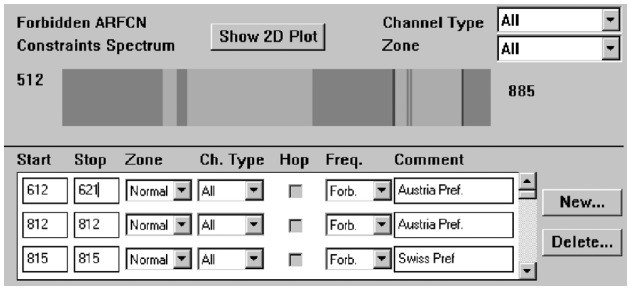


Figure 31. Cell-specific boundary conditions for channel assignment.

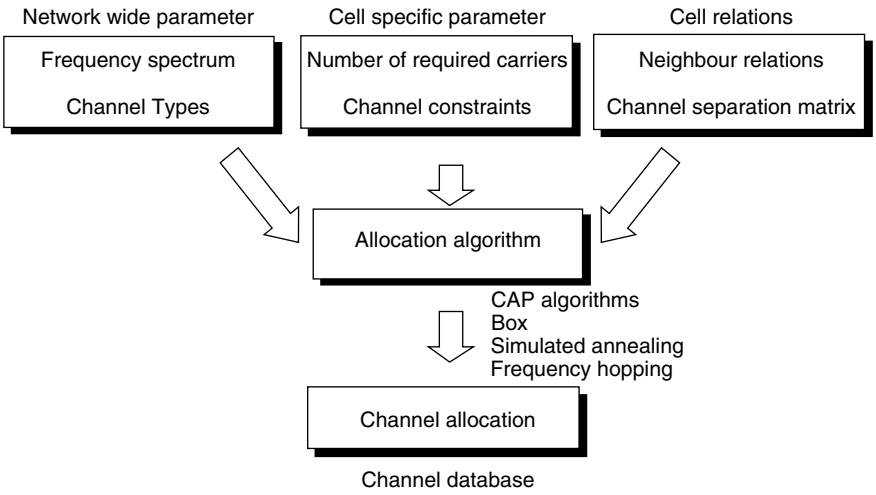


Figure 32. CAP inputs.

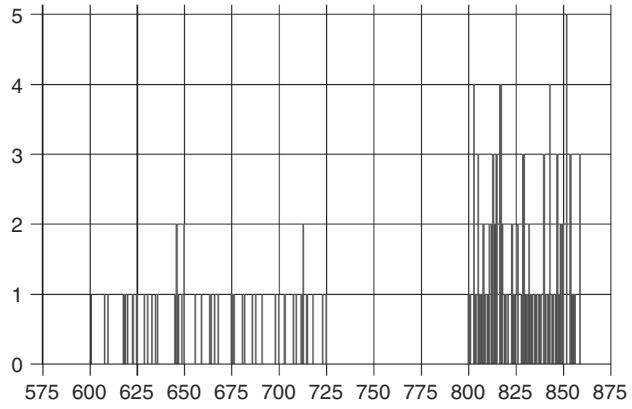


Figure 33. After successful channel assignment for a GSM 1800 system.

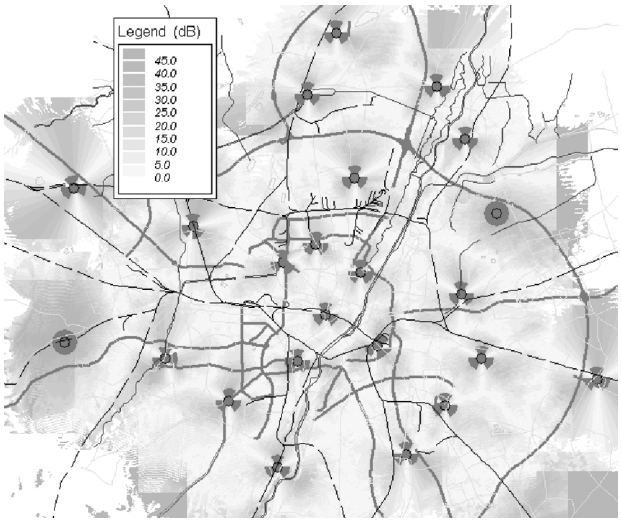


Figure 34. C/I for Munich network before channel assignment.

An optimistic timeline for deployment of data services is shown in Fig. 36.

Because of the complex change of migrating 2G networks to the 3G world, there were strong demands from the market to enable higher data rates in the already existing 2G infrastructure. Pressure increased not to wait for the "mobile Internet" up to the year 2005, where UMTS services are currently expected to be really deployed and ready for widespread use, including handset availability. The success of the possible 2.5G technologies depends largely on the support from system and handset manufacturers and the real-time schedule for 3G availability. Commercial amounts of the first 2.5G handsets

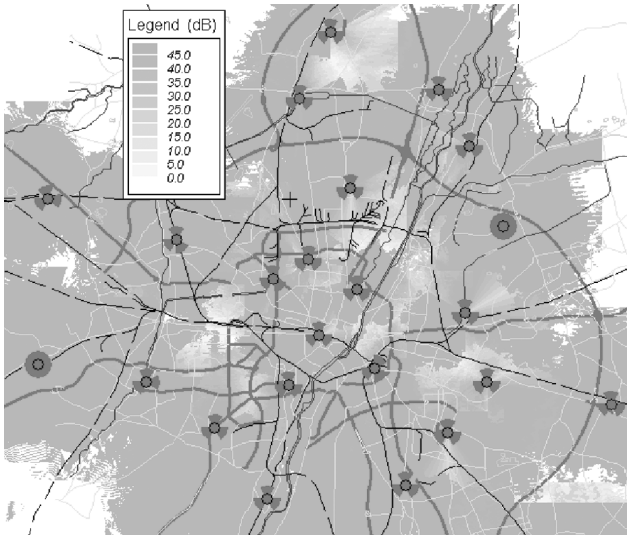


Figure 35. C/I for Munich area after channel assignment.

are currently (summer 2001) sold on the market. The first network operator in Germany (D1) has officially launched GPRS services, but the data rate offered is still about only half the speed of an analog modem of the fixed network side and available only in specific areas. EDGE in this respect is a consequent step further using existing GPRS equipment for higher data rates (up to theoretically 384 kbit/s). Simply put, GPRS is a key enabler technology to use the voice and circuit-switched GSM infrastructure for packet-switched data services. The major point for enabling this data service is an adaptation of the GSM backbone network for the coexistence of both technologies. The major advantage for all network operators is that the main parts of the existing infrastructure, especially site installations, can be used further on and reduce investment costs to a minimum. Possible data rates offered depend on the number of packet-switched users in the cell and on different modulation types, so-called coding schemes (CSs). Depending on the hardware suppliers, different QoS ratings are necessary to enable a specific data rate. However, the limits (maximum possible data rates of GPRS and EDGE [also referred as enhanced GPRS (EGPRS)] concerning the data rates are shown in Figs. 37–39.

EGPRS is expected to introduce nine new coding schemes, where the higher coding schemes use different modulation in the same timeslots. The applied modulation for higher EGPRS data rates is 8-PSK (phase shift keying) and contains phase and amplitude modulation. Thus, 8-PSK-modulated signals in EGPRS need to be transmitted with a smaller power than in GMSK

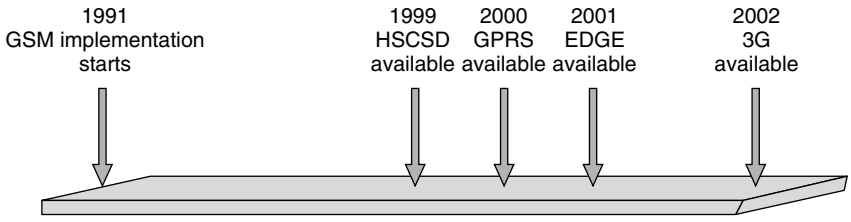


Figure 36. Optimistic view of enhanced data service deployment.

	1 Timeslot	2 Timeslot	8 Timeslots
CS-1	9.2 kBit/s	18.4 kBit/s	73.6 kBit/s
CS-2	13.55 kBit/s	27.1 kBit/s	108.4 kBit/s
CS-3	15.75 kBit/s	31.5 kBit/s	126 kBit/s
CS-4	21.55 kBit/s	43.1 kBit/s	172.4 kBit/s

Figure 37. Possible coding schemes and data rates for GPRS.

Coding Scheme	Modulation	Throughput/Timeslot
MCS-1	GMSK	8.8 kbit/s
MCS-2	GMSK	11.2 kbit/s
MCS-3	GMSK	14.8 kbit/s
MCS-4	GMSK	17.6 kbit/s
MCS-5	8-PSK	22.4 kbit/s
MCS-6	8-PSK	29.6 kbit/s
MCS-7	8-PSK	44.8 kbit/s
MCS-8	8-PSK	59.2 kbit/s
MCS-9	8-PSK	59.2 kbit/s

Figure 38. Possible data rates for EDGE/EGPRS.

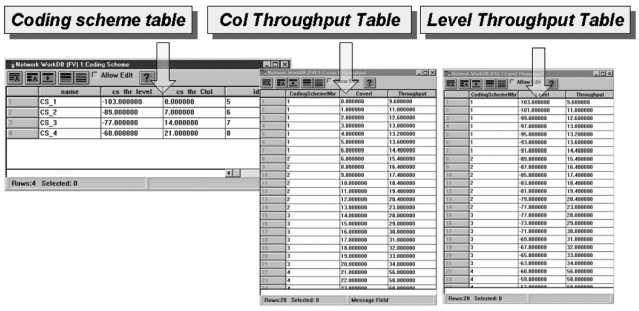


Figure 39. Hardware-dependent GPRS parameters.

(Gaussian minimum shift keying) as is currently used for GPRS and GSM. Otherwise, the respective output power amplifier would be driven to nonlinear operation, which would then result in a garbled signal. Also, 8-PSK is more error-prone than is GMSK. Therefore, the coverage of a cell running EGPRS will shrink and this fact must be accounted for in the network design.

What do the tables in Figs. 37–39 mean? For reasonable data rates for a single user, it is necessary to combine (reserve) at least two time slots in each cell used for GPRS. The coding scheme that will be used during the connection is adapted according to the QoS currently measured for this connection. Therefore a high C/I enables a higher coding scheme and thus a considerably higher data rate. However, even for high-quality networks this will increase efforts in deploying new TRXs on existing sites in order to assure a data rate at least comparable to analog fixed-line modems. As mentioned at the beginning of this section, the coding scheme that can be applied strongly depends on the hardware performance. So, for a RNP tool, it is mandatory to scope for this flexibility in the database. A possible solution is to have a database allowing the user to input lookup tables for different hardware vendors.

The possible data rates for the different coding schemes (throughput in kbps) are stored as a function of power level and C/I for each hardware unit used in the network. The calculation is divided into two streams: a maximum possible throughput depending on the power level of each cell and a best possible based on C/I calculations of the current fully loaded network.

Note, however, that the theoretical bit rate limits of GPRS, EDGE, and EGPRS are far from reality because of the transmission errors, necessary amount of error correction overhead, handset restrictions, and sharing the capacity between all users. In reality, roughly 10–20% of these maximum values can be expected.

Results of both streams are the coding scheme and throughput possible in the network deployed (see Figs. 40–42).

9. 3G NETWORKS (UMTS)

The following sections cover the most important 3G standards.

9.1. FDD/TDD/CDMA2000-TDSCDMA

The so-called "3G" standard has become a mixture of several allowed air interfaces that do not harmonize with

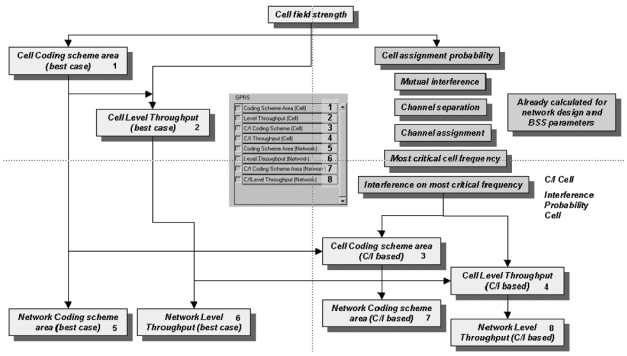


Figure 40. GPRS calculation implementation in a RNP tool.

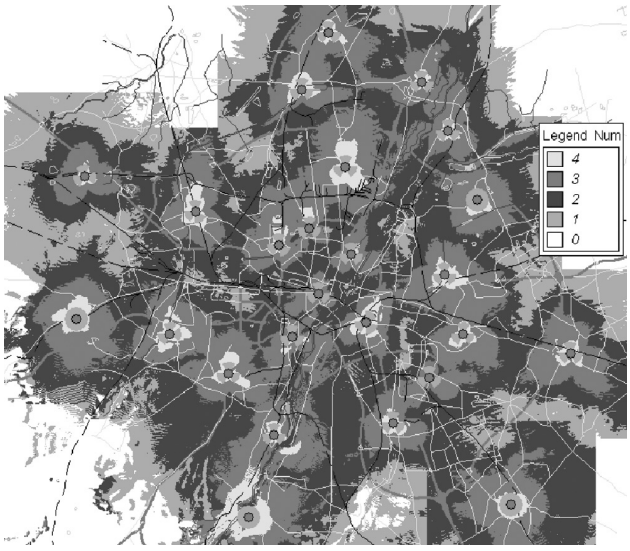


Figure 41. Possible coding schemes in the Munich network.

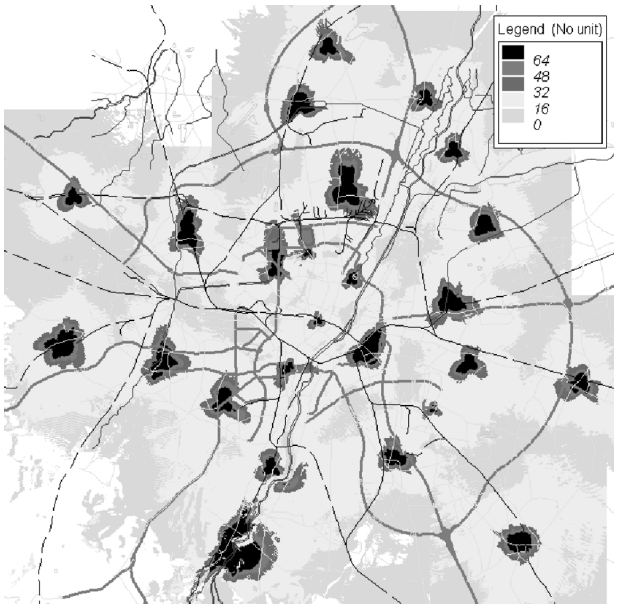


Figure 42. Throughput in kilobits per second.

each other. The framework of this 3G standard(s) is IMT-2000 and has several family members (see Fig. 43).

The process of obtaining a 3G standard was driven by political and commercial interests more than of inventing only one real worldwide standard. Each part of the world has different 2G systems running, which determined the development of a next generation into different directions [12,13]. Therefore, the official 3G standard now consists of several different technologies, all more or less incompatible to each other. "3G" in Europe means for operators and network engineers mostly "UMTS" (wideband CDMA, FDD mode, and TDD mode), where during the first rollout phase operators will stick to the FDD mode to be deployed for macro cellular approach (see Fig. 44). In China (already today world's biggest cellular market), TDSCDMA might dominate the market: TDSCDMA is a special development for the Chinese market and driven by mostly Siemens and Chinese companies, but looks promising for urban situations in general. The following sections will deal with the most popular 3G technologies in Europe: the UMTS-WCDMA FDD and TDD mode used for urban and suburban environments.

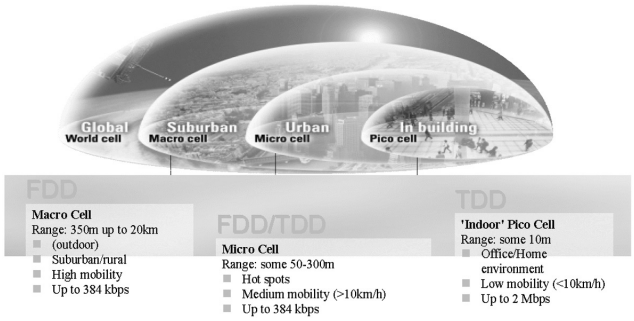


Figure 44. Hierarchical rollout strategy.

9.2. WCDMA Planning Aspects

Especially in Europe, with its dominance of GSM, CDMA is a completely new technology. Perhaps the reader of this article are familiar with UMTS planning, which is the one that most people are familiar with today. It is a wideband CDMA technology, but still, it is a CDMA technology as we have encountered already in IS95 networks. There seem to be some differences compared to IS95 (as in the mixed-traffic scenarios) and also compared to GSM (e.g., the cell breathing effects), but all in all, it seems not that terribly different from planning a 2G/2.5G network. But this is not true. . . .

9.3. UMTS: The Impact of the Service Mixes

One key feature of the UMTS system is its inherent flexibility regarding data rates and service types. The network attempts to satisfy the service requirements with a mixture of

- Spreading factors
- Forward error correction (FEC) coding types
- Signal-to-interference ratio (SIR) targets
- Numbers of spreading sequences
- Code puncturing rates

These all interact, which makes an "all in one" planning process necessary.

A given UMTS "service" offered to any user may be achieved by the allocation of a spreading sequence or a

Standard	UTRA-FDD	UTRA-TDD	TD-SCDMA	CDMA2000	UWCC136	DECT
Freq. band	Paired	Unpaired	Unpaired	Paired	Paired	Unpaired
IMT-2000	IMT-DS	IMT-TD	IMT-TD	IMT-MC	IMT-SC	IMT-FT
	IMT-2000 CDMA DS (direct spread)	Other mode of IMT-2000 CDMA TDD (⇨TD-SCDMA)	One mode of IMT-2000 CDMA TDD (⇨UTRA-TDD)	IMT-2000 CDMA MC (multi carrier)	IMT-2000 TDMA SC (single carrier)	IMT-2000 FDMa/TDMA
Core network compatibility	GSM MAP	GSM MAP	GSM MAP	ANSI-41	ANSI-41	ISDN
Primary standardisation bodies	3GPP	3GPP	CWTS 3GPP	3GPP2	TIA (U.S.)	ETSI

Figure 43. IMT-2000 family members.

number of sequences depending on orthogonal variable spreading factor (OVSF) sequence availability. The network may achieve the target bit error rate (BER) through higher transmit power or higher SIR target setting, or through use of more powerful forward error correction (FEC) coding depending on the delay constraints requested for the service.

Unfortunately, these settings cannot be optimized *per service* but only *per service mix* as a UMTS base station ("node *B*" for historical reasons[1]) must serve all users and their specific services within its cell simultaneously. But as it has only one hardware transceiver unit, the mixture of services offered by a UMTS network means any and all resources and mixes of parameters must be employed to satisfy competing user requirements.

The complexity of the air interface thereby rules out many of the conventional "set and forget" approaches used in planning systems such as GSM or IS95.

For example, assume that you have a stable UMTS network in a city area and suddenly, a bus of tourists arrives, step out and switch on their multimedia 3G handsets to send live movies of their voyage back home. What will be the impact on your network? Is a planning tool capable of simulating that situation?

In GSM, network planning assumed *static* cell boundaries (so-called best server areas), more or less given by power constraints only; traffic issues were addressed by additional TRXs. In 2G CDMA networks (and GSM), there was only *one* service, so cointerference caused by other services simply did not exist. For example, although the famous "cell breathing" already did exist in IS95, it wasn't such a big issue, as it affected only the one and only "voice" service. Given a maximum traffic load for the service, the resulting interference, and thus the amount of cell breathing could be estimated.

In UMTS, however, not only are there a vast number of services, but these services are *very* different from each other: Not only in their data rates, but also their traffic types and QoS demands. So what impact does that one new 2-Mbps packet-switched data user in a cell have on the 20 (circuit-switched) voice users and the one 384-kbps packet-switched data user currently logged on? Must you drop the latter and/or some of your voice users?

Given the mix of services and parameters, the fact is simply that the old-fashioned "static" or even empirical models of the GSM and IS95 world won't work anymore and completely new simulation and analysis strategies must be implemented in an "UMTS RNP tool."

9.4. UMTS Planning Strategy

Another bad habit is to assume the use of conventional RAKE receiver technology from narrowband CDMA networks such as IS95 for UMTS rollout.

Some of these problematic assumptions are

- Gaussian nature of interference
- A small number of multipath components
- Long spreading code analysis of RAKE receiver performance
- Homogeneous network traffic (low data rate or voice)

Given these assumptions, many of the physical-layer performance issues can be "characterized" and abstracted to higher planning levels; this makes IS95 planning simpler to some extent, but this can't be assumed for UMTS any longer. The mixed-traffic, mixed-quality, wideband nature of UMTS invalidate these assumptions. It is well known that RAKE analysis assumptions have significant weaknesses when

- Mobile signals are not tightly power-controlled
- Short spreading sequences are employed
- The number of RAKE fingers is finite and must be shared between cells in the users active set

Nevertheless, many "UMTS planning" tools still use the same old propagation-based network analysis concept.

This approach has significant disadvantages, and many generalizing assumptions must then be made, such as

- The performance of the receiver employed
- The nature of multipath channel
- The nature of the interference produced by the competing users
- The impact of active sets
- The nature and performance of advanced FEC coding schemes

While these assumptions allow a simple analysis of the network, they are so wide-sweeping that they render the results highly inaccurate. Even worse, they give *unduly optimistic* planning results and are far off from reality, the more traffic mixes occur and the more high-data-rate services are involved.

Research results show that

- The conventional "propagation only" approach may serve only as a quick first glance at the network situation. It is fast but way too optimistic for realistic network planning.
- The "static Monte Carlo" method is quite popular today, but again falls short for detailed analysis, especially for mixed-traffic scenarios and involvement of high-bit-rate services.
- An enhanced dynamic method is needed to account for the impact of user mobility and the dynamics that occur with high-data-rate users.

It seems feasible to use at least a static Monte Carlo (MC) analysis for the macrocell environments and address high-traffic environments (e.g., microcells, urban areas, hotspots) and high-mobility environments (e.g., streets and railways) with dynamic analysis. Even the static MC should be replaced by a "quasidynamic" approach, as described in one of the next sections.

[1] The strange term "node *B*" was initially used as a temporary working item during the standardization process but then never changed.

9.5. UMTS Network Design Process

Extensive R&D in WCDMA hardware and software has shown that it is necessary to use a UMTS planning tool that addresses all the issues mentioned before.

A major strength of a strong UMTS planning tool [3] is its ability to model users, service characteristics, and network features in great detail. A powerful network simulation engine provides the designer with the flexibility to explore true dynamic UMTS *real-time* scenarios in detail or examine the "achieved versus designed" network performance.

The design process can be extended beyond the approach used for purely propagation prediction based planning by allowing users and services to be modeled at either an SINR level or even a chip level. This allows the operator to examine

- Link-level performance (dropping, blocking, achieved QoS, BER)
- Mixed-traffic types and their impact on each other
- Handover (active set size change) regions to be examined in detail
- Realistic interference rather than doing simplistic interference modeling

While there are many features and parameters in the proposed UMTS standard, it is likely that only a subset of these parameters will be of interest to the network planner.

A state-of-the-art UMTS planning tool offers three modes of simulation and analysis to the user to tackle the different UMTS planning problems:

- *Static-mode analysis*, which requires only propagation prediction results and network configuration information. The results are service independent. Analysis results are based on link level calculations only. The results available include
  - Least path loss
  - Best server
  - Delay spread
  - RMS delay spread
  - Received CPiCH power
  - Handover regions (active set size changes)
- *Quasidynamic mode*, which requires propagation prediction results plus additional information about the services being offered and the average traffic load in the network. Analysis results are based on link-level calculations and iterative attempts to solve power control equations. The results partly consider the time domain by showing averaged network behavior and should at least include
  - Service coverage [uplink and downlink (UL/DL)] for each service
  - Number of served, blocked, or stolen mobiles for each cell
  - Power used (UDL) for each service
  - OVSF tree utilization
- *Dynamic mode*, which allows the user to examine the link-level performance of the real-time network behavior that includes all the dynamic characteristics of the UMTS air interface. It also allows many of the assumptions associated with ideal RAKE performance to be removed. Results produced by the dynamic simulation mode include
  - Dropping and blocking rates (hard and soft blocking)
  - BER calculations
  - Achieved SIR and SIR target setting performance
  - Dynamic active set utilization
  - Dynamic transmit and receive power requirements

The usual way to plan a network consists of

- RF coverage planning
- Capacity and quality maximization
- Network optimization and maturation

Initially, the designer's task is to provide a rapid RF rollout solution that satisfies the RF coverage design rules and provides acceptable levels of capacity in the planned regions. In 2G, this is done by static analysis of link budget calculations for received pilot signal power, BCCH, and so on. As outlined, this can be done for a first, optimistic indication about coverage, but it is not a suitable way to determine the network capacity for UMTS since there are potentially many different services with different quality requirements competing for the same radio resources. The related issue of capacity and coverage in UMTS can only be addressed by some form of dynamic or quasidynamic analysis that considers the solution to the limited power budget, intercell interference and the competing quality requirements. This task is made more difficult with microcell or "hotspot" environments. In particular, many of the assumptions that are required to be made to determine capacity and coverage begin to break down in heavily loaded regions or when high data rates are considered.

The nature of mixed-traffic-type CDMA also means that the quality of the network is of a highly variable nature. While quasidynamic simulation and analysis techniques will report average performance characteristics, many of the users in the network will experience significantly worse conditions for a significant proportion of the time leading to unacceptable levels of dropping and quality of service.

The final significant task area is to optimize and improve the quality of network. This includes determining the impact of new network equipment, considering the impact of particular dynamic traffic scenarios (e.g., at train stations, highways, or ferry terminals), and improving quality of service through operator-controlled parameters (e.g., SIR target settings). The nature of mixed-traffic CDMA means that many of the physical layer abstractions which are possible in 2G can't be done for UMTS any more; without link-level simulation taking into account the many features of UMTS, many of the optimization tasks are just not possible.

Figure 45 lists some typical tasks for the network planner and the type of analysis tool that is suitable

Design Task	Relative Complexity	Static	Quasi-Dynamic	Dynamic
RF Coverage	Low	✓		
Calculating Service Regions	Medium	✗	✓	✓
"Hot Spot" Region Planning	Medium / High	✗	✓	✓
"What IF" Traffic Profile Scenario	High	✗	✗	✓
Evaluation of New Equipment Features	High	✗	✗	✓
Calculating Blocking / Dropping Rates	High	✗	✗	✓
Determining Network Capacity	High	✗	✓	✓
Determining Network Link Quality	High	✗	✗	✓

Figure 45. Network planning and design tasks.

for carrying out these tasks as well as some relative level of complexity of the task.

For many tasks, quasidynamic analysis is suitable. With all the implicit assumptions that form part of this approach, however, the quasidynamic mode (let alone the simple Monte Carlo approaches) is only able to indicate where support of the desired services is *at all* possible in the region of interest *in the average*. But it will not allow the planner to determine whether the desired service can be supported with the required level of quality at a specific moment of time.

9.6. UMTS Network Planning Examples

The following section shows examples of a fictitious UMTS network in the southeast Sydney, Australia. It consists of 30 sites with 90 sectors.

9.6.1. Propagation-Based Results: Static Approach. The results of importance and achievable with the *static* approach based on path-loss calculations and traffic and interference assumptions are

- Least path loss (Fig. 46)
- Best server (Fig. 47)
- RX CPiCH power level (pilot received Power level)
- Handover regions (active set size changes) (Fig. 48)

The results of the static approach are easily derived and can be adapted from 2G calculations, as all influences from traffic mixes, link level, and OVSF utilization are considered by fixed margin during calculation. The propagation-based approach deals with the parameters listed in Fig. 49.

"Static" in this context means that the time domain is completely ignored. Traffic is considered only indirect by one single margin to be added as "noise." No service mixes can be considered by this approach, let alone specific service requirements.

9.6.2. Propagation-Based Results: Quasidynamic Approach. This approach adds static traffic definitions and

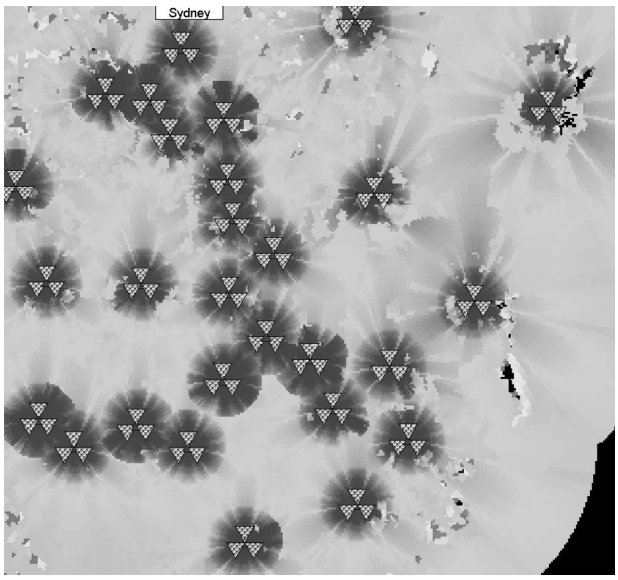


Figure 46. Least path loss.

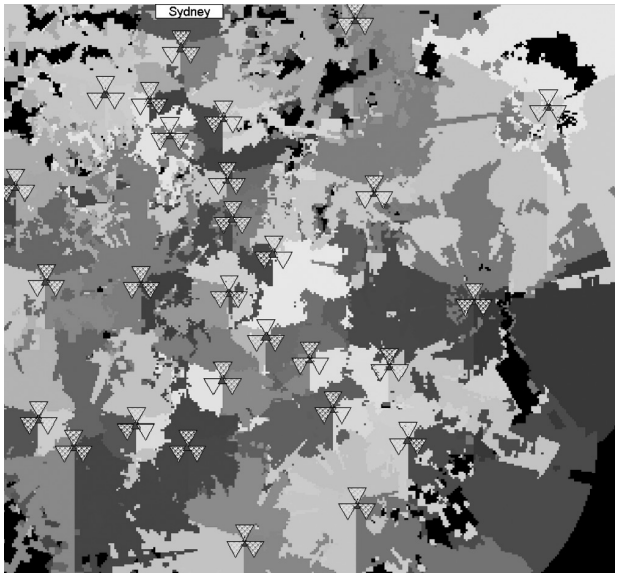


Figure 47. Best server.

traffic mixes to the analysis. The traffic load can be defined on a per cell basis as shown in Fig. 50.

The next result types are examples of the output of the quasidynamic network simulation illustrating the influence of load in a network. The calculation is done by solving the power control equations with a Monte Carlo simulation until the network is in its equilibrium. Then the last mobile unit served in a specific service of the network is analyzed. This shows the cell breathing effect, which depends directly on the load added to the network.

- Test mobile connected at voice (Fig. 51)
- Test mobile connected at 384 kbps (Fig. 52)

Comparing Figs. 51 and 52, it is obvious that the presence of a served user at 384 kbps will dramatically

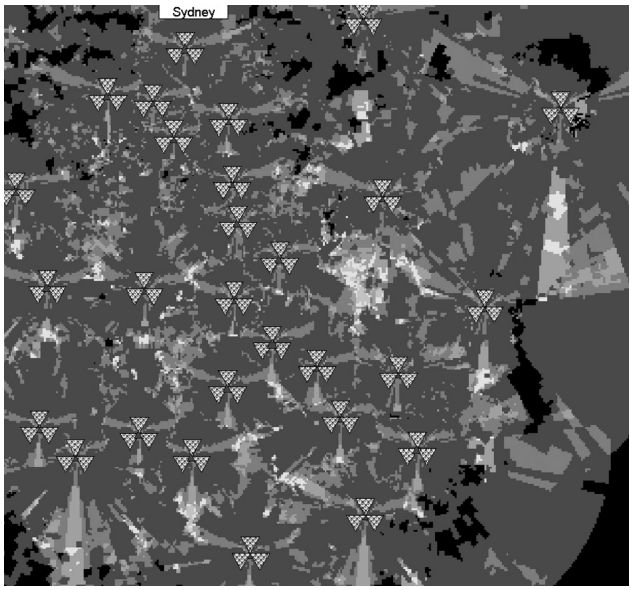


Figure 48. Handover regions.

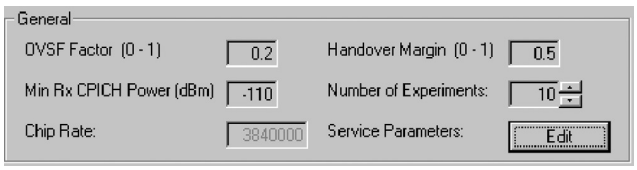


Figure 49. Parameters for static analysis.

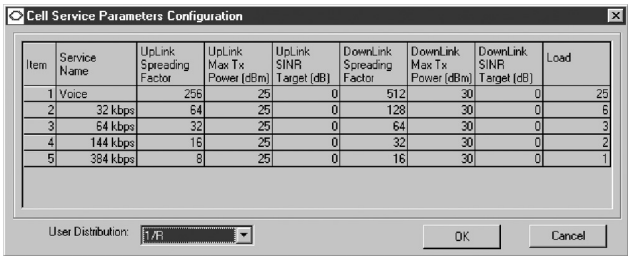


Figure 50. Additional service parameters for quasidynamic simulation.

raise the interference and force all mobiles in the network to raise their transmit power. This is exactly the cell breathing effect. Depending on the maximum TX power level of the mobiles, the "black holes" in the network become larger and larger relative to the load. The term *quasidynamic* is used here because the time domain is taken into consideration by taking several "snapshots" of the network dynamics and building an average out of this. The hope is that if enough snapshots are taken, statistical confidence in the "average" network behavior is reached. Note that this is true only for circuitswitched services and that the number of snapshots (i.e., number of iterations) depends mainly on the services taking part in the traffic mix and the network size itself.

9.6.3. Dynamic Simulation. This approach enables the user to examine in great detail the relations in the network. Instead of traffic definitions on a per cell basis, the users are defined by a statistical function or along deterministic paths. The probability density function is to describe the "move and turn" behavior of the mobile users (see Figs. 53 and 54). Also the number of users and their specific traffic model can be defined on a per user group basis.

The full dynamic approach is to solve detailed problems in the network and can be used to simulate algorithms for

- any new network features / new hardware
- competing / mixed vendor equipment
- advanced UMTS features (smart antennas, transmit diversity, non-RAKE receivers to name but a few)
- operator configurable parameters (SIR target setting, call admission)

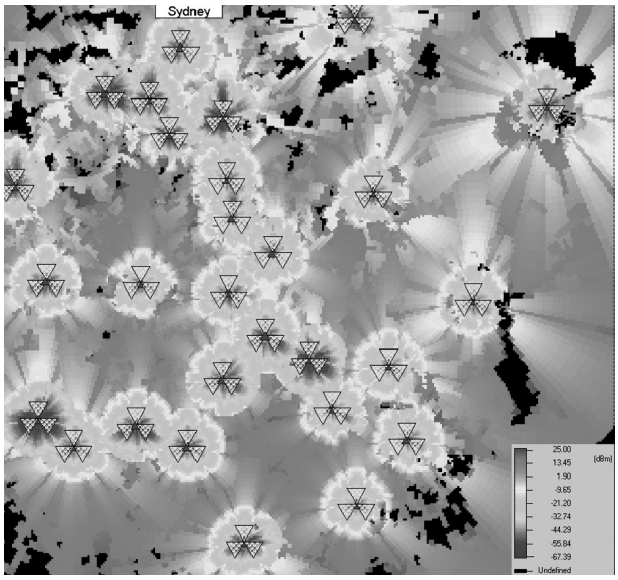


Figure 51. MS TX power at voice connection.

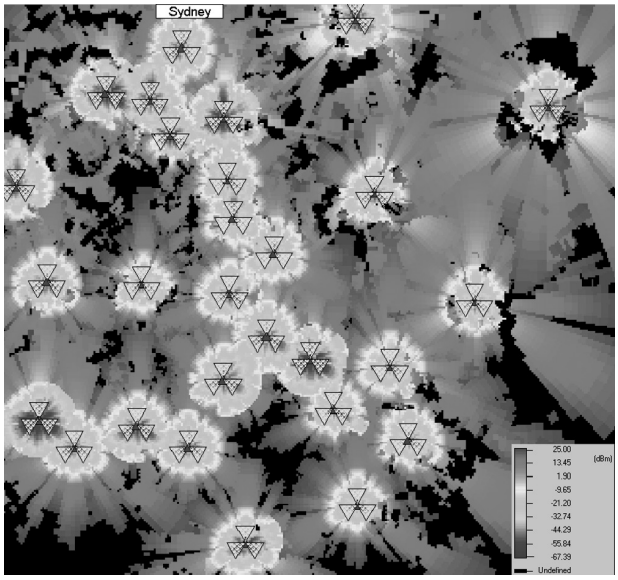


Figure 52. MS TX power for test mobile with 384 kbps.

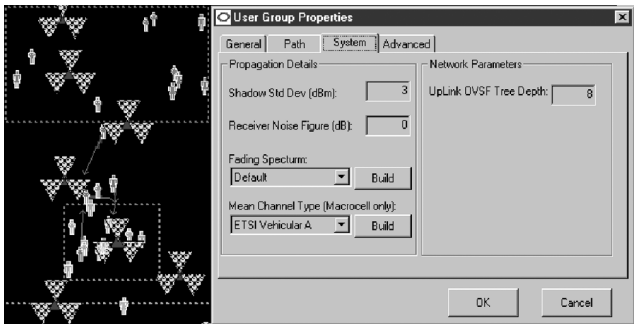


Figure 53. User group properties.

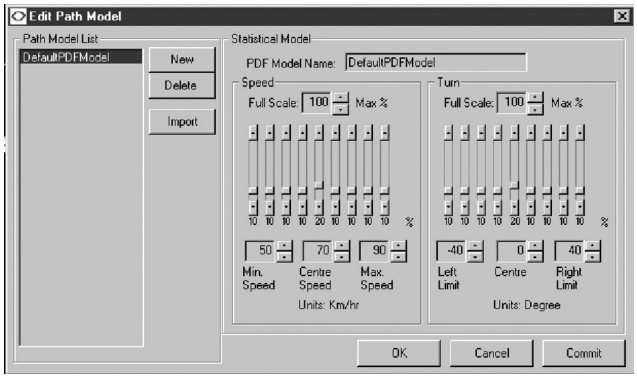


Figure 54. PDF model for a statistical user group.

As it is able to actually *simulate* the UMTS system on a chip level (so each chip which is transmitted by each mobile in up and downlink is considered) dynamic results as the SINR versus time for the different sent and received chips are calculated (see Figs. 55 and 56).

In particular, this truly dynamic simulation approach allows one to actually trace the behavior of 3G systems in dynamic situations, for example, how a UMTS system in equilibrium will behave when suddenly, a high-bit-rate user demands service.

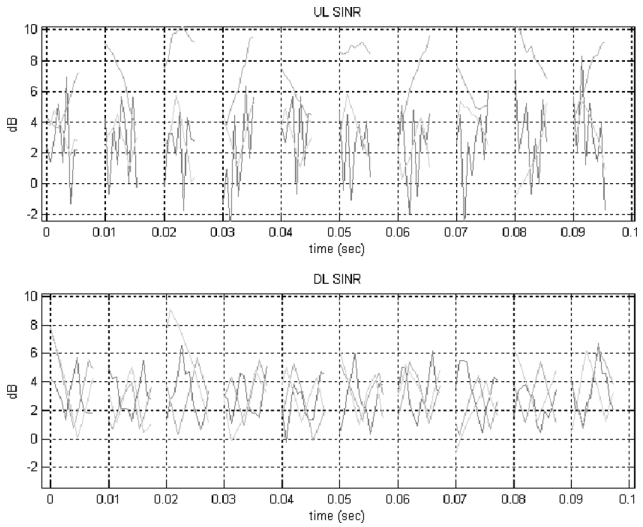


Figure 55. Signal-to-noise ratio versus time for uplink and downlink for different user groups.

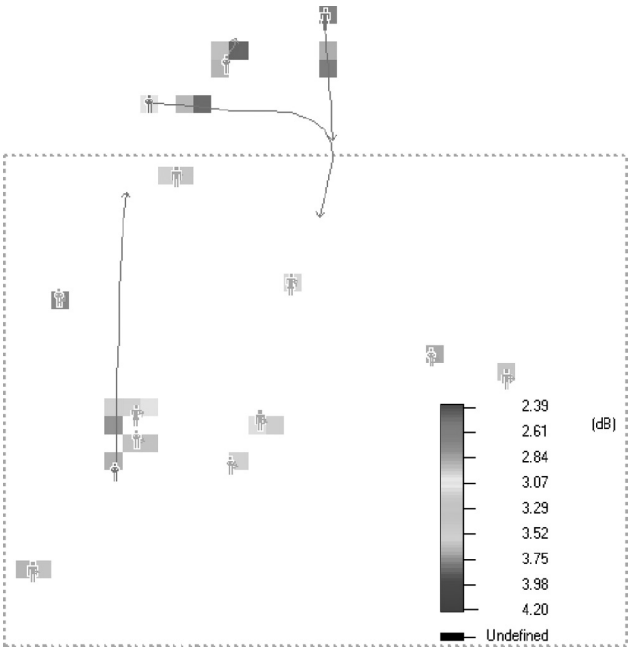


Figure 56. Spatial SINR result for the analyzed user groups. The red lines are the result of mobility model applied.

This method is truly "dynamic" as it completely considers the time domain. As there is no more averaging effect, very detailed simulations are possible. However, the price for accuracy is a lot of computational effort, which currently restricts this method to model evaluation, hotspot analysis, and planning.

9.7. UMTS — Is It Good or Bad?

This question is hard to answer. UMTS is a powerful, flexible standard that allows many new services and offers great promise for the future of mobile communications. The planning and optimization of a UMTS network is made more difficult by this same flexibility. Although it is clear that simple static simulations may be able to predict RF coverage based on pilot powers, in order to tackle the closely coupled problem of coverage and capacity analysis in a mixed-service environment, quasidynamic or dynamic analysis of the network must be performed.

Many planning jobs from 2G networks reappear in UMTS networks also, but often more delicate (e.g., neighborhood planning) or simultaneously (e.g., coverage and capacity analysis). New problems add to this, rendering the network design a very challenging process.

With the fierce competition for subscribers and the high prices paid for UMTS licenses, squeezing the best out of the network is critical. Operators who can support the required services, with the demanded quality of service, will have a clear advantage in the UMTS race.

10. WHERE DO WE GO WITH 4G?

As depicted in the beginning, 4G research and development has just begun. This should not surprise, as for 2G and 3G, the time from first R&D to commercial availability also has been around 10 years. 4G is thus to be expected around 2010.

The necessity for a new mobile generation arises from the progress in fixed network services: Customers increasingly demand the same services and accompanying bandwidths that are available at home (with landline service) when they are using their mobiles.

So, things that we foresee in fixed networks will be driving for development for mobile communication as well. As stated in this article, 20 Mbps in the downlink and roughly 2 Mbps in uplink are to be expected. Having said that, again the question arises how to realize networks that are able to deliver such data rates. If we can't improve propagation predictions, we have to concentrate the energy to the communication path between transmitter and receiver. Logically, we will encounter very small cells, enhanced by adaptive antenna arrays and beamforming. As this can't be deployed nationwide, it is also clear that 4G will live in close internetwork roaming conditions with its predecessor technologies; this can also be a path for a truly global 4G standard.

The backbone network will surely be IPv6 to cater the high demand of IP addresses and bring the mobile world in line with the fixed network world. Higher bit rates also mean higher bandwidth demand in higher frequency ranges: Japan's DoCoMo (which claims to dedicate currently about 80% of its R&D already to 4G) sees 4G in the frequency range of 3–8 GHz. This will need massive worldwide coordination effort, but fortunately, ITU started discussion about such 4G systems that are beyond the current scope of IMT-2000 in the WP-8F initiative in November 1999 and WRC2000 framework. These high-frequency ranges mean worse propagation conditions and higher power demands. Taking the power restrictions of handsets into account, this is also a clear indication for very small cells. There are also ideas of "pumping" networks, such as the example of a highway-bridge-mounted transceiver that covers only a few meters of the highway, but "pumps" that many data with 150 Mbps++ into a car, and that this will yield enough data received until they reach the next serving cell.

Dr. Nobuo Nakajima, former Senior Vice President of DoCoMo Wireless Laboratories, sees an increase in total mobile traffic of 2200% by 2015 compared to that of 1999. He assumes a required bandwidth for 4G of ~1350 MHz.

This all depicts that with 4G, again a "revolution" in system technology will occur, again requiring a whole new planning approach.

BIOGRAPHIES

Jürgen Kehrbeck was born in Karlsruhe, Germany, in 1961. He received the Dr. Ing degree in high-frequency electronics engineering at the University of Karlsruhe, Germany in 1993. His main research activities from 1989 to 1995 were small-vehicle radar sensors in the high gigahertz range. Since 1995 he has been responsible for the development of mobile network planning tools at LS telcom. Currently he is head of the Division of e-Commerce and Mobile Communications at LS telcom.

Kai Rohrbacher was born in Karlsruhe, Germany, in 1966. He received Dipl.-Inform. degree in computer sciences at the University of Karlsruhe, Germany in 1993. He worked for major German computer and telecommunications magazines (and still does so as a sideline). After 4 years of work for a German GSM operator, he joined LS telcom in 1998. Currently, he is head of department mobile communication software at LS telcom.

Werner Wiesbeck (SM'87, F'94) received the Dipl.-Ing. (M.S.E.E.) and the Dr.-Ing. (Ph.D.E.E.) degrees from the Technical University Munich, Germany in 1969 and 1972, respectively. From 1972 to 1983 he was with AEG-Telefunken in various positions, including that of head of R&D of marketing director Receiver and Direction Finder Division, Ulm. During this period he had product responsibility for mm-wave radars, receivers, direction finders, and electronic warfare systems. Since 1983 he has been director of the Institut für Höchstfrequenztechnik und Elektronik (IHE) at the Universität Karlsruhe (TH), Germany.

His research topics include radar, remote sensing, wave propagation, and antennas. In 1989 and 1994, respectively, he spent a 6-month sabbatical at the Jet Propulsion Laboratory, Pasadena. He is a member of the IEEE GRS-S AdCom (1992–2000), chairman of the GRS-S Awards Committee (1994–1998), executive vice president IEEE GRS-S (1998–1999), president IEEE GRS-S (2000–2001), associate editor *IEEE-AP Transactions* (1996–1999), and past treasurer of the IEEE German Section.

He has been general chairman of the 1988 Heinrich Hertz Centennial Symposium, the 1993 Conference on Microwaves and Optics (MIOP'93), and he has been a member of scientific committees of many conferences. For the Carl Cranz Series for Scientific Education, he serves as a permanent lecturer in radar system engineering and for wave propagation. He is a member of an Advisory Committee of the EU–Joint Research Centre (Ispra/Italy), and he is an advisor to the German Research Council (DFG), to the Federal German Ministry for Research (BMBF), and to industry in Germany.

BIBLIOGRAPHY

1. I. S. Redl, M. Weber, and O. Malcolm, *An Introduction to GSM*, Artech House, Boston, 1995.
2. CHIRplus_M, *Computer Based Planning System for 2G and 2.5G Cellular Networks*, LS telcom AG, Lichtenau, Germany, 1999.
3. UTRAplan, *Computer Based Planning System 3G WCDMA FDD/TDD Cellular Networks*, LS telcom AG, 77839 Lichtenau, Germany, 1899.
4. www.placeAbase.com, *Web Based Site Marketplace*, LS telcom AG, 77839 Lichtenau, Germany, 2000.
5. Bureau de Développement des Télécommuni-cations, *Manual on Mobile Communication Development*, ITU-Geneva, 1997.
6. X. Huang, *Automatic Cell Planning for Mobile Network Design: Optimization Models and Algorithms*, Ph.D. thesis, Faculty of Electrical Engineering, Univ. of Karlsruhe, Germany, May, 2001.
7. W. C. Y. Lee, Mobile Cellular Telecommunications Systems, McGraw-Hill, New York, 1998.
8. A. M. C. A. Koster, *Frequency Assignment—Models and Algorithms*, Ph.D. thesis, Maastricht Univ., The Netherlands, 1999.

9. A. Gamst, Some lower bounds for a class of frequency assignment problems, *IEEE Trans. Vehic. Technol.* **35**(1): 8–14.
10. D. Minoli, Broadband Network Analysis and Design, Artech House, Boston.
11. J. S. Lee and L. E. Miller, *CDMA Systems Engineering Handbook*, Artech House, Boston.
12. T. Ojanperä and R. Prasad, *Wideband CDMA for Third Generation Communications*, Artech House, Boston.
13. H. Holma and A. Toskala, *WCDMA for UMTS*, Artech House, Boston.
14. COST 231, *Digital Mobile Radio Towards Future Generation Systems* (1989–1996), EUR 18957.
15. ETSI Technical Report, *Terrestrial Trunked Radio*, ETR 086-1, Jan. 1994.
16. ETSI Technical Report, *Terrestrial Trunked Radio*, ETR 300-1, May 1997.
17. ETSI Technical Report, *Terrestrial Trunked Radio*, ETR 300-2, May 1997.
18. Agreement between the telecommunications authorities of Austria, Belgium, the Czech Republic, Germany, France, Hungary, the Netherlands, Croatia, Italy, Lithuania, Luxembourg, Poland, Romania, the Slovak Republic, Slovenia and Switzerland, on the coordination of frequencies between 29.7 MHz and 43.5 GHz for fixed services and land mobile services, Vienna, June 30, 2000.
19. E. Zitzler and L. Thiele, Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach, *IEEE Trans. Evol. Comput.* **3**(4): 257–271.

CELLULAR COMMUNICATIONS CHANNELS

Aarne Mämmelä
Pertti Järvensivu
VTT Electronics
Oulu, Finland

1. INTRODUCTION

This article gives an overview of cellular radio or wireless channels for mobile digital communications [1]. The main emphasis is in channel models for radiowave propagation in terrestrial outdoor mobile cellular systems between a base station and a mobile station in the downlink and uplink, in either microcells or macrocells. Such systems typically work in the frequency range from about 1 to 2 GHz with the corresponding wavelengths between 0.3 and 0.15 m. The bandwidth of the transmitted signals is in the order of 100 kHz–1 MHz.

The location of the base station antenna has a significant effect on channel modeling. In microcells the cell radius is about 0.1–1 km, and the base station antenna is below the rooftop level of the surrounding buildings. On the other hand, in macrocells the base station antenna is above the rooftop level and the cell radius is about 1–30 km. The area types are usually divided into urban, suburban, and rural, each of which may be nonhilly or hilly.

A radio channel is almost always linear. Because of its mobility, the channel is also time-variant. It is thus fully described by its impulse response $h(\tau, t)$, where τ is the delay parameter and t is the time. The complex impulse response $h(\tau, t)$ is a lowpass equivalent model of the actual real bandpass impulse response. Equivalently, the channel is characterized by its transfer function $H(f, t) = \int_{-\infty}^{\infty} h(\tau, t) \exp(-j2\pi f\tau) d\tau$, which is the Fourier transform of the impulse response with respect to the delay parameter.

The magnitude $|H(f, t)|$ of the transfer function at a given frequency f is changing randomly in time, and we say that the mobile radio channel is a fading channel. The phase $\arg H(f, t)$ is also a random function of time. Fading is caused mainly by multipath propagation due to reflection, scattering, and diffraction of the radiowaves from nearby objects such as buildings, vehicles, hills, and mountains. With respect to a stationary base station, multipath propagation creates a stochastic standing-wave pattern, through which the mobile station moves. Additional fading is caused by shadowing when the propagation environment is changing significantly, but this fading is typically much slower than the multipath fading. Modem design is affected mainly by the faster multipath fading, which can be normally assumed to be locally wide-sense stationary (WSS). Early important work on WSS fading multipath models in a more general framework is due to Turin, Kailath, and Bello in the 1950s and 1960s [2,3].

Some modern systems use directive antennas to amplify the desired signal and to reject the interfering signals. Conventionally only horizontal directions are taken into account. In such systems the direction of arrival of the received signals as well as the azimuthal power gain of the antenna are important issues, and the models are two-dimensional (2D). The two dimensions are the delay and the azimuth whereas in one-dimensional (1D) models the only dimension (in addition to time) is the delay. Important early work on 2D models was done by Clarke in the 1960s [4].

The channel models are used for performance analysis and simulations of mobile systems. The models can also be used for measurements in a controlled environment, to guarantee repeatability and to avoid the expensive measurements in the field. However, any model is only an approximation of the actual propagation in the field. For measurements, the average received signal-to-noise ratio must be defined. It is estimated by making a link power budget, which includes the transmitter power, distance-dependent attenuation of the channel, antenna gains in the transmitter and receiver, and various loss factors and margins. It depends on the system designer whether a margin for fading is taken into account or whether the performance simulations or measurements with the channel model will include fading. The power of additive noise is also estimated for modeling purposes. Usually only the thermal noise with a certain noise figure in the receiver is considered. The additive noise is assumed to be white Gaussian noise (WGN) within the signal bandwidth. Unless otherwise stated, we will exclude any other noise sources.

The organization of this chapter is as follows. In Section 2 we give the statistical description of one-dimensional and two-dimensional channel models. In Section 3 we summarize the methods by which the channel is measured. In Section 4 widely available simulation models are described. Finally, in Section 5 some more recent trends are noticed. More extensive reviews are included for the one-dimensional models [5,6] and two-dimensional models [4]. Models for indoor communications are summarized [7]. Much of the theory is valid also for outdoor communications. Our list of references is not exhaustive, and some very important work has been left out because of space limitations. Additional references can be found from the papers cited.

2. STATISTICS OF THE TIME-VARIANT IMPULSE RESPONSE

The most important propagation phenomena to be included in a channel model are shadowing and multipath fading, and the model is either one-dimensional or two-dimensional. We will first consider 1D models, which are characterized by the time-variant impulse response and transfer function. If the transmitted signal is denoted by $z(t)$, the received signal $w(t)$, without noise, is given by $w(t) = \int_{-\infty}^{\infty} z(t-\tau)h(t,\tau)d\tau$. Alternatively, the received signal is $w(t) = \int_{-\infty}^{\infty} Z(f)H(f,t)df$, where $Z(f)$ is the Fourier transform of $z(t)$.

If the transmitted signal has a bandwidth of W, the delay resolution of the measurement is approximately $1/W$, which means that the receiver cannot resolve delay differences smaller than $1/W$. We define such unresolved multipath components as clusters on the delay axis [3]. The receiver can resolve multipath components whose delay differences are larger than $1/W$. We will apply the central-limit theorem for the clusters. The impulse response has the general form $h(\tau,t) = \sum_{l=0}^{L-1} h_l(t)\delta(\tau-\tau_l)$, where L is the number of resolvable clusters whose amplitudes and delays are $h_l(t)$ and τ_l, respectively. Since the channel is random, we need a stochastic description for it. The delays of the clusters are usually assumed to be constant in channel models, but it must be noted that fading is caused mainly by the randomly changing delays, which change the relative phase shift between the multipath components within the clusters.

If several multipath signals due to scattering with approximately equal amplitudes, or alternatively with random amplitudes, are added at random phases, the resultant has a complex Gaussian distribution with a zero mean. The amplitude of such a cluster is Rayleigh distributed and the phase is uniformly distributed. The channel is then said to be a Rayleigh fading channel. Alternatively, if in addition to the scattered components, the received signal includes a strong component, which is a line-of-sight (LOS) signal coming either directly from the transmitter or from a specular reflection, the impulse response at that delay will have a Gaussian distribution with a nonzero mean and the amplitude will be Rice distributed. The channel is in this case a Rice fading channel. In both Rayleigh and Rice fading channels, only the first- and second-order statistics, including the mean and autocorrelation functions, are needed to fully describe them. A more general description is the covariance matrix of a discretized impulse response.

For multipath fading, a widely used model is a wide-sense stationary uncorrelated scattering (WSSUS) model. It is WSS with respect to the time variable. Uncorrelated scattering (US) means that the autocorrelation function of the WSS Rayleigh fading impulse response has the form $E\{h^*(\tau,t)h(\tau+\Delta\tau,t+\Delta t)\} = P_h(\tau,\Delta t)\delta(\Delta\tau)$, or there is no correlation on the τ axis, but some correlation may exist on the time axis. The function $P_h(\tau,\Delta t)$ is the autocorrelation of the impulse response at the delay τ with the time difference Δt. The impulse response is nonstationary white noise in the delay variable. It can be shown that in a WSSUS channel the transfer function is WSS also with respect to the frequency variable [2].

The Fourier transform of $P_h(\tau,\Delta t)$ with respect of the time difference Δt is the scattering function $S(\tau,\lambda)$ of the channel, or $S(\tau,\lambda) = \int_{-\infty}^{\infty} P_h(\tau,\Delta t)\exp(-j2\pi\lambda\Delta t)d\Delta t$, where λ is the Doppler shift variable. The scattering function is a measure of the average power output as a function of the time delay τ and the Doppler shift variable λ. The delay power spectrum is $P_h(\tau) = \int_{-\infty}^{\infty} S(\tau,\lambda)d\lambda$. Equivalently, the delay power spectrum is $P_h(\tau) = E\{|h(\tau,t)|^2\}$. The Doppler power spectrum is $S_H(\lambda) = \int_{-\infty}^{\infty} S(\tau,\lambda)d\tau$.

The width of the delay power spectrum is referred to as the *delay spread*, and the width of the Doppler power spectrum is referred to as the *Doppler spread*. A suitable engineering definition is used for the width. The channel is frequency-nonselective or flat fading if the signal bandwidth W is smaller than the inverse of the delay spread, or the coherence bandwidth; otherwise the channel is frequency-selective. The Doppler spread and its inverse, or the coherence time, are measures of the rapidity of fading.

A typical approximation for the delay power spectrum is exponential. A typical approximation for the Doppler power spectrum is $S_H(\lambda) = (1/\pi f_m)[1-(\lambda/f_m)^2]^{1/2}$, where $f_m = (v/c)f_0$ is the Doppler frequency, v is the velocity of the mobile station, c is the velocity of the radiowaves, and f_0 is the carrier frequency. This Doppler power spectrum is based on the assumption that the multipath components arrive the omnidirectional antenna uniformly from all horizontal directions. It is often referred to as Jakes's Doppler power spectrum [1] even though it was derived earlier by Clarke [5].

In addition to Rayleigh and Rice distributions, a useful amplitude distribution for the multipath fading is Nakagami m distribution, which is in fact a form of the generalized Rayleigh distribution. When selecting a suitable distribution, one should note that for system performance, the most notable effect has the distribution at small amplitudes [8].

Shadowing is essentially frequency-nonselective fading, much slower than multipath fading, and it is usually described by the lognormal distribution; thus, the received

power in decibels is normally distributed. The lognormal distribution is also based on the central-limit theorem [9]. The product of several random variables may be approximated as being lognormally distributed. The product comes from the various attenuation factors due to the obstacles between the transmitter and the receiver.

The WSSUS model described above can be generalized to the 2D case as follows. The azimuth or the angle of arrival relative to the velocity vector of the mobile station is denoted by α. As previously, the multipath components are combined into clusters in space with a delay resolution of $1/W$ and with an angular resolution $\Delta\alpha$ of the receiver antenna. Each cluster has a Rayleigh or Rice fading amplitude. Each complex gain of the impulse response has now the general form $h_l(t) = \sum_{n=0}^{N-1} a_{ln} \exp[j\varphi_{ln} + 2\pi f_m t \cos(\alpha_{ln})]$ where α_{ln} is the azimuth and φ_{ln} is the phase of the nth component in the lth delay and N is the number of components in the model at the lth delay. As a generalization of the delay power spectrum, we can define an azimuthal delay power spectrum that shows the distribution of the received power versus azimuth and delay. For a given velocity and direction of the mobile station, and for a given azimuthal power gain of the antenna, the azimuthal delay power spectrum corresponds to a certain scattering function of the WSSUS channel (see Fig. 1). The scattering function is an aliased form of the azimuthal delay power spectrum since two different azimuths α_{ln} and $-\alpha_{ln}$ of arriving clusters create the same Doppler shift due to the cosine function in $h_l(t)$. Given a uniform distribution $p(\alpha)$ for the received power, we obtain Clarke's Doppler power spectrum given earlier.

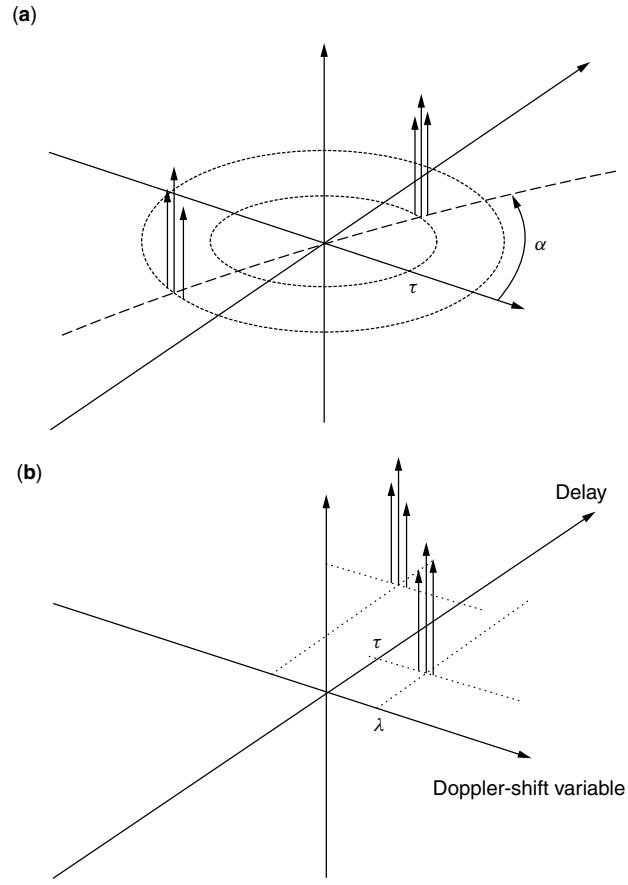


Figure 1. Azimuthal delay power spectrum (**a**) and scattering function (**b**) of the channel.

3. MEASUREMENT OF LINEAR TIME-VARIANT CHANNELS

Channel measurements can be divided into narrowband and wideband measurements. Wideband measurements use measurement signals, which have about the same bandwidth as the intended information signal. Unlike narrowband measurements, which use a single unmodulated carrier as a measurement signal, wideband measurements provide information on the multipath propagation as well as frequency selectivity of the channel. Therefore, only wideband measurements are discussed here. The measurements can also be divided into measurement of [1] instantaneous values of the impulse response and [2] the average parameters of the channel. The average parameters include first/second-order statistics, or the mean and the autocorrelation function of the impulse response, and the scattering function of the channel. Several ways to perform systematic measurements have been listed in Ref. 6.

The problem of the measurement of system functions of time-variant channels differs from that of the time-invariant case. Even in the absence of noise the system function of a time-variant channel may be unmeasurable. The condition on the measurability of a linear time-varying WSSUS channel was first presented by Kailath and later extended by Bello [10]. It turns out that the channel measurability depends on the value of the area spread factor, which in effect is the area of region where the scattering function is effectively nonzero. If the area spread factor is less than or equal to a threshold, the channel impulse response could be measured unambiguously [10]. The value of the threshold is of the order of unity. The channels for which the area spread factor fulfills the criterion mentioned above are called *underspread*; otherwise they are called *overspread*. If the channel is overspread, it is not possible to measure the instantaneous values of the impulse response, even in the absence of the noise. Fortunately, most physical channels are underspread. The average parameters can be determined either from the instantaneous values of the channel impulse response or by cross-correlation methods. Since the statistical averages contain much less information than the instantaneous values, the channel need not always be underspread before the average parameters could be measured.

Various channel measurement techniques have been proposed [11]. However, two practical methods of measuring the impulse response of the underspread cellular channel can be identified. One method is to transmit a very short impulselike pulse to the channel and observe the multiple pulses received. In order to follow the time variation of the channel, the pulses need to be transmitted periodically. The short pulses result in a high ratio of peak to average transmission power, which could be undesirable. Another, more efficient, method to measure the impulse

response is the use of direct sequence spread-spectrum signals [6]. A pseudonoise (PN) sequence is used to modulate the carrier. Maximal-length sequences (m sequences) are widely used because of their excellent periodic autocorrelation properties. The receiver is based on a correlation principle. It can be implemented by a matched filter or a sliding correlator [5]. Nowadays, high-speed digital signal processing techniques can be employed to implement real-time matched filter channel sounders. In a WSSUS channel, time averaging can be used to obtain the autocorrelation function of the measured impulse response. An estimate of the scattering function can then be calculated by the Fourier transform [12,13]. Angle-of-arrival measurements can be conducted by using directional or array antennas at the receiver [4]. To some extent, angle of arrival can be deduced from the measured scattering function [12].

One of the earliest measurement results of the impulse response of the cellular channel in urban and suburban areas by using short pulses was reported by Turin in 1972 [3]. The delay, amplitude, and phase of multipath components were measured at three different frequencies simultaneously. It was found that spatial correlation distances of these variables at neighboring geographic points vary considerably. They ranged from less than a wavelength for the phases, through tens of wavelengths for the amplitudes and delays, to hundreds of wavelengths for the means and variances, or powers, of the amplitudes.

More recent measurement studies have used almost exclusively the direct-sequence spread-spectrum signals with m sequences to measure the channel. Wideband macrocell measurements conducted at 1 GHz show that in typical urban areas the delay spread from 1 to 2 μs is characteristic. In suburban areas delay spreads from 10 to 20 μs are typical. The longest delay spreads occur in mountainous environments, where delay spreads from 100 to 150 μs have been encountered. In open areas the delay spread is practically nonexistent and the received signal consists of the directly propagated component only [12]. Wideband measurements in urban microcell environments at 2 GHz have been reported [13]. The delay power spectrum, average normalized correlation functions, and scattering functions have been calculated from the measured impulse response. Under LOS conditions, the direct component with unresolvable specular reflections dominated the propagation. Some resolvable specular reflections existed at delays of up to 1.5 μs. In non-LOS situations, the powers of the strongest received signals were more than 15 dB below the LOS components that resulted in nearby locations. The propagation process was found to be dominated by multiple reflections and scattering along the streets, and not by diffraction (13).

Relatively few results on spatial channel measurements have been published. Ertel et al. have summarized some of the results [4]. Measurements conducted at 2 GHz with a 10-MHz bandwidth, by using a rotating azimuth beam directional antenna at the receiver, have shown that delay and angle-of-arrival spreads are small in rural, suburban, and even many urban environments. Measurements in urban areas have shown that most of the major features of the delay angle of arrival spectra can be accounted for by considering the large buildings in the environment. Finally, variations in the spatial characteristics with both time and frequency have been measured. The results indicate that the uplink spatial characteristics cannot be directly applied for downlink beamforming in most of the present cellular and personal communication systems that have 45-MHz and 80-MHz separations between the uplink and downlink frequencies, respectively [4].

4. SIMULATION MODELS

The evolution of channel simulation models has been parallel to that of cellular systems. The early models considered only the signal amplitude-level distributions and Doppler spreads of the received signals. A delay spread information was later added to the channel models. In addition to those, modern channel models also include such concepts as angle-of-arrival and adaptive array antenna geometries. The signal parameters that need to be simulated in these models for each multipath component are the amplitude, carrier phase shift, time delay, angle of arrival, and Doppler shift. All of these parameters are in general time-varying, causing Doppler spread in addition to delay spread [4].

The channel simulators can be categorized into three classes according to the way the channel impulse response is modeled: stored channel impulse response, ray tracing models, and stochastic parametric models for the channel impulse response. The stored channel impulse responses are based on selected measurements, which are then stored for later use. Although this method provides actual information from the channel, the proper selection criterion for the measurements may be difficult to identify. Also, the large amount of data needed to store the measurement results could be difficult to handle. However, some models have been proposed [4,14]. The ray tracing models are deterministic. They are based on geometric propagation theory and reflection, diffraction, and scattering models. Accurate channel models are possible using this method. However, the high computational burden and lack of detailed terrain and building databases make these models difficult to use [4]. By far the most popular channel simulation models are stochastic parametric models. In this approach, the channel impulse response is characterized by a set of deterministic and random parameters. The values of the parameters and the probability distributions governing their behavior are selected according to measurements. The remaining challenge is to develop models, that exhaustively reproduce the propagation scenarios accounted in reality [14]. A recommended summary of stochastic channel models can be found in Ref. 14. Spatial, or 2D, stochastic channel models are summarized in Ref. 4.

Usually, discrete-time channel impulse responses in the form of transversal filters are used in the stochastic channel simulators. The transversal filter model allows the simulators to be implemented either by software or hardware. The time-varying complex coefficients and delays of the transversal filter are generated according to the statistics associated with the different parameters. The amplitudes are usually assumed to be Rayleigh or Rice distributed as a result of multipath fading. A uniform and

Poisson distribution is usually assumed for the phases and delays, respectively. However, in a Rice fading channel the phase is concentrated around the phase of the strong component unless there is a Doppler shift in it. As mentioned earlier, the delay power spectrum is typically approximated by an exponential function. There are several methods to simulate the angles of arrival. For example, in Ref. 15 they are modeled as normally distributed random variables. Other methods have been summarized in Ref. 4. Rayleigh and Rice processes needed to simulate the amplitudes can be generated by using colored Gaussian noise processes. A well-known method to produce colored Gaussian noise processes is to filter WGN with a filter having a transfer function of the square root of the Doppler power spectrum. Typically, the Doppler power spectrum by Clarke is used. Another method is based on Rice's sum of sinusoids. In this case, a colored Gaussian noise process is approximated by a finite sum of weighted and properly designed sinusoids [16]. The long-term variations in the channel impulse response can be modeled by making the delays drift with time and by using an attenuation filter to model the lognormal fading caused by shadowing or transitions between different environments [14]. An exponential function is used to approximate the autocorrelation of the shadowing as a function of distance. A correlation distance of 20 m is typically used for urban environments. For suburban environments, much larger correlation distances should be used. In a hardware implementation, digitally controllable attenuators can be used to simulate the attenuation caused by the shadowing [15].

Different standardization organizations are actively defining channel models as a part of specification of new mobile cellular systems. Their motivation is to specify the operational environment of the system and to provide test parameters for manufacturers. The channel models for second-generation digital advanced mobile phone system (DAMPS) and global system for mobile communications (GSM) mobile cellular systems were specified by the Telecommunications Industry Association (TIA) in the United States and the European Telecommunications Standards Institute (ETSI) in Europe. For the third-generation cellular systems a global standard has been defined as the International Mobile Telecommunications (IMT-2000) proposal by the International Telecommunication Union (ITU). For a more thorough discussion, see Ref. 6. The standardization work of third-generation systems has now shifted to the international 3rd Generation Partnership Project (3GPP).

The channel models have also been developed by different research consortia in international research programs. In Europe, particularly Cooperation in the Field of Scientific and Technical Research (COST) projects have been extremely influential when GSM and digital communication system at 1800 MHz (DCS 1800) systems were developed. The achievements in COST partly stimulated the Universal Mobile Telecommunications System (UMTS) Code-Division Testbed (CODIT) project within the Research and Development in Advanced Communication Technologies in Europe (RACE-II) program. The CODIT 2D channel models seem to be state-of-the-art. A summary of European research programs can be found in Ref. 14.

5. MORE RECENT TRENDS

Channel modeling for cellular communications is a rapidly changing area, and the models are becoming increasingly accurate. Some of the more recent trends are summarized here. Higher frequencies of up to ~60 GHz will be used in the future. The cell size is made smaller since the channel attenuation is larger at higher frequencies. Frequencies above ~10 GHz are also affected more by air molecules and rain. Consequently, the highest frequencies can be used only in indoor environments. Also, with the increasing data rates, the bandwidths are becoming larger, approaching 10–100 MHz. 3D models are important in macrocells, for example, in urban and mountainous areas where the base station antenna is much higher than the mobile station antenna. The 3D models take into account the elevation angle of the arriving waves, in addition to the azimuthal angle and the delay. Various nonstationary models are often used. In addition to the lognormal distribution, shadowing effects are modeled with birth–death processes, where some delays suddenly appear and disappear, simulating rapid changes as in street corners and tunnels. In some models the delays are time-variant to test the delay tracking ability of the receiver. Furthermore, the models are becoming more comprehensive in the sense that they will have multiple inputs and outputs. In this way the models can be used to simulate diversity systems with many users. Even handoffs between base stations should be simulated. Correlation and crosstalk between the multiple channels are important effects in such systems. An example of crosstalk is the cochannel and adjacent-channel interference between the various users of the same frequency band in the same geographic region.

BIOGRAPHIES

Aarne Mämmelä was born in Vihanti, Finland, in 1957. He received the degrees of M.Sc. (Eng), Lic.Tech., and Ph.D. (all with distinction) from the Department of Electrical Engineering, University of Oulu, Finland, in 1983, 1988, and 1996, respectively. His doctoral thesis was on diversity receivers in fast fading multipath channels. From 1982 to 1993 he was with the Telecommunication Laboratory at the University of Oulu and researched adaptive algorithms in spread-spectrum systems. In 1990–1991 he visited the University of Kaiserslautern, Germany. In 1993 he joined VTT, Computer Technology Laboratory, which was merged to VTT Electronics in 1994. In 1996–1997 he was on leave as a postdoctoral fellow at the University of Canterbury, Christchurch, New Zealand. Since 1996 he has been a research professor of digital signal processing at VTT Electronics. His research area is the design of digital transmitter-receivers in wireless communications. In addition, since 2000 he has been a docent or adjunct professor of receiver signal processing at the Helsinki University of Technology, Espoo, Finland. He

is especially interested in synchronization and estimation problems in wireless digital communications, both in single- and multi-carrier systems.

Pertti Järvensivu was born in Laukaa, Finland, in 1966. He received a M.Sc. degree in electrical engineering from the University of Oulu, Finland, in 1992. From June 1988 to December 1999 he was in various teaching and research positions at the University of Oulu. Since January 2000 he has been with VTT Electronics as research scientist in digital signal processing. His current research interests are channel estimation and wireless adaptive radio systems.

BIBLIOGRAPHY

1. W. C. Jakes, ed., *Microwave Mobile Communications*, Wiley, New York, 1974.
2. P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* **CS-11**: 360–393 (1963).
3. G. L. Turin, Introduction to spread-spectrum antimultipath techniques and their application to urban digital radio, *Proc. IEEE* **68**: 328–353 (1980).
4. R. B. Ertel et al., Overview of spatial channel models for antenna array communication systems, *IEEE Pers. Commun.* **5**: 10–22 (1998).
5. D. Parsons, *The Mobile Radio Propagation Channel*, Pentech Press, London, 1992.
6. K. Pahlavan and A. H. Levesque, *Wireless Information Networks*, Wiley, New York, 1995.
7. H. Hashemi, The indoor radio propagation channel, *Proc. IEEE* **81**: 943–968 (1993).
8. S. Stein, Fading channel issues in system engineering, *IEEE J. Select. Areas Commun.* **SAC-5**: 68–89 (1987).
9. A. J. Coulson, A. G. Williamson, and R. G. Vaughan, A statistical basis for lognormal shadowing effects in multipath fading channels, *IEEE Trans. Commun.* **46**: 494–502 (1998).
10. P. A. Bello, Measurement of random time-variant linear channels, *IEEE Trans. Inform. Theory* **IT-15**: 469–475 (1969).
11. A. Hewitt and E. Vilar, Selective fading on LOS microwave links: Classical and spread-spectrum measurement techniques, *IEEE Trans. Commun.* **36**: 789–796 (1988).
12. W. R. Braun and U. Dersch, A physical mobile radio channel model, *IEEE Trans. Vehic. Technol.* **40**: 427–482 (1991).
13. U. Dersch and E. Zollinger, Physical characteristics of urban micro-cellular propagation, *IEEE Trans. Antennas Propag.* **42**: 1528–1539 (1994).
14. B. H. Fleury and P. E. Leuthold, Radiowave propagation in mobile communications: An overview of European research, *IEEE Commun. Mag.* **34**: 70–81 (1996).
15. J. J. Olmos, A. Gelonch, F. J. Casadevall, and G. Femenias, Design and implementation of a wide-band real-time mobile channel emulator, *IEEE Trans. Vehic. Technol.* **48**: 746–764 (1999).
16. M. Pätzold, U. Killat, F. Laue, and Y. Li, On the statistical properties of deterministic simulation models for mobile fading channels, *IEEE Trans. Vehic. Technol.* **47**: 254–269 (1998).

CHANNEL MODELING AND ESTIMATION

LANG TONG
Cornell University
Ithaca, New York

1. INTRODUCTION

One of the objectives of receiver design for digital communications is to minimize the probability of detection error. In general, the design of the optimal detector requires certain prior knowledge of the channel characteristics, which are usually estimated through the use of pilot symbols. A typical example is the voiceband modem where, on establishing the initial connection, a signal known to the receiver is transmitted through the telephone channel. The receiver is then tuned to compensate for the distortions caused by the channel. The process of using pilot symbols to estimate the channel, or directly, the receiver coefficients is referred to as *training*.

If there is a sufficient amount of training time, and the channel does not vary significantly, the problem of channel estimation can be formulated either as the classical point estimation or as the Bayesian estimation. Techniques such as maximum-likelihood estimation and methods of least squares are readily applicable, and they generally offer good performance. For these applications, the choice of algorithm is often determined by the complexity of implementation.

The explosive growth in wireless communications and the increasing emphasis on packet-switched transmissions present a new set of challenges in channel estimation and receiver design. Although critical to reliable communications, channel acquisition and tracking are difficult because of the rapid variations caused by multipath fading and the mobility of users. The use of training is no longer straightforward, as the receiver must be trained repeatedly for time-varying channels. Since the time used for training is the time lost for transmitting information, there is a tradeoff between the quality of channel estimation and the efficiency of channel utilization.

This article presents an overview of the modeling and estimation of channels for digital transmission of single carrier linearly modulated signals. In Section 2, we present the complex baseband representation of intersymbol interference channels. A discrete-time linear model is obtained that relates the received data samples with the channel coefficients and the transmitted pilot and data symbols. Within the framework of parametric estimation, we present techniques and performance analysis of various channel estimation problems in Sections 3 and 4. A brief bibliography note is provided at the end of this article.

2. THE BASEBAND MODEL OF BAND-LIMITED CHANNELS

In this section, we consider the baseband model, in both continuous and discrete time, for linearly modulated signals transmitted over band-limited passband channels.

2.1. The Continuous-Time Model

Figure 1 illustrates the passband transmission of linearly modulated baseband signals. To transmit a sequence of

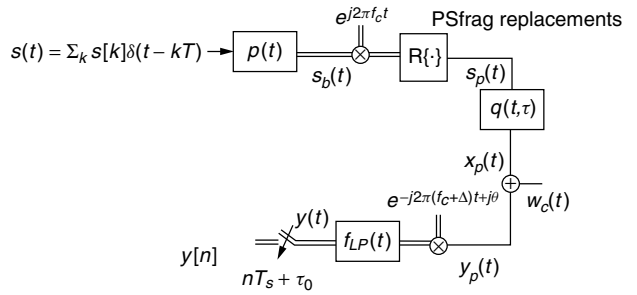


Figure 1. The transmission of linearly modulated baseband signal through a band-limited passband channel. Double lines are paths for complex signals and single lines for real signals. The operator $\Re\{\cdot\}$ takes the real part of its argument. The filter $f_{LP}(t)$ is the impulse response of an ideal lowpass filter.

information-carrying symbols $\{s[k]\}$, the baseband signal $s_b(t)$ is formed as

$$s_b(t) = \sum_k s[k]p(t - kT) \tag{1}$$

where $p(t)$ is the baseband pulse, T the symbol interval, and $1/T$ the symbol rate. For passband transmissions, where the transmitted signals do not contain any DC component, the symbol sequence $\{s[k]\}$ may be complex in general. If the transmission is at the baseband, then $s[k]$ is real. In this article, we will assume that $\{s[k]\}$ is a complex sequence and the results are also valid when $\{s[k]\}$ is a real.

If $s_b(t)$ is transmitted over a band-limited channel, the Nyquist criterion for choosing $p(t)$ needs to be satisfied. In particular, $p(t)$ should be such that

$$\frac{1}{T}\sum_{i=-\infty}^{\infty}\left|P\left(f + \frac{i}{T}\right)\right|^2 = 1 \tag{2}$$

where $P(f)$ is the Fourier transform of $p(t)$. A usual choice is from the class of square-root raised-cosine pulses. The minimum bandwidth pulse is the ideal lowpass filter with bandwidth $1/2T$. In practical implementations, the actual bandwidth is between $1/2T$ and $1/T$ for narrowband transmissions and much greater than $1/T$ for spread spectrum transmissions.

To transmit $s_b(t)$ through a particular frequency band, the baseband signal is converted to the passband signal $s_p(t)$ by (quadrature) amplitude modulation. Hence, the transmitted signal is represented as

$$\begin{aligned} s_p(t) = \Re\{s_b(t)e^{j2\pi f_c t}\} &= \Re\{s_b(t)\}\cos(2\pi f_c t) \\ &\quad - \Im\{s_b(t)\}\sin(2\pi f_c t), \end{aligned} \tag{3}$$

where the operator $\Re\{\cdot\}$ takes the real part of its argument and $\Im\{\cdot\}$ the complex part of its argument. The (real) passband signal is transmitted through a linear, possibly time-varying, propagation channel $q(t, \tau)$ whose output $x_p(t)$ is given by

$$\begin{aligned} x_p(t) &= \int q(t,\tau)s_p(t-\tau)\,d\tau \\ &= \Re\left\{e^{j2\pi f_c t}\int q(t,\tau)e^{-j2\pi f_c \tau}s_b(t-\tau)\,d\tau\right\} \\ &= \Re\left\{e^{j2\pi f_c t}\int q_b(t,\tau)s_b(t-\tau)\,d\tau\right\} \end{aligned} \tag{4}$$

where we note that the passband propagation channel $q(t, \tau)$ is always real and denote the baseband propagation channel as

$$q_b(t,\tau) \triangleq q(t,\tau)e^{-j2\pi f_c \tau} \tag{5}$$

The received *passband* signal $y_p(t)$ is corrupted by noise $w_c(t)$ assumed to be zero mean, white, and Gaussian. The passband signal is then converted back to the baseband signal by frequency downshifting and lowpass filtering:

$$\begin{aligned} y(t) &= f_{LP}(t) * [y_p(t)e^{-j2\pi(f_c+\Delta)t+j\theta}] \\ &= e^{j(2\pi\Delta t+\theta)}\int q_b(t,\tau)s_b(t-\tau)d_\tau + w(t) \end{aligned} \tag{6}$$

where Δ is the frequency offset and θ the phase offset, and $w(t)$ is zero mean complex Gaussian with constant power spectrum density within the spectral range of the signal. Substituting (1) into Eq. (6), we obtain

$$y(t) = e^{j2\pi\Delta t}\sum_k h_k(t)s[k] + w(t) \tag{7}$$

where

$$h_k(t) = e^{j\theta}\int q_b(t,\tau)p(t - kT - \tau)d\tau \tag{8}$$

Note that the transmitted signal is distorted by two major factors: the propagation channel $q_b(t, \tau)$ and carrier-phase synchronization errors Δ and θ. Typically, carrier synchronization is performed separately from channel estimation. Assuming that the frequency offset Δ has been corrected before channel estimation, we can let $\Delta = 0$ and combine the phase error θ with baseband channel $h_k(t)$.

The problem of channel estimation can then be formulated as estimating $h_k(t)$, which combines the propagation channel $q(t, \tau)$, the signal waveform $p(t)$, and the phase error θ. Notice that $p(t)$ is known in general, and that it can be exploited to improve channel estimation.

If the channel can be modeled as time-invariant within the interval that channel estimation is performed, we then have $q_b(t, \tau) = q_b(\tau)$, and

$$y(t) = \sum_k s[k]h(t - kT) \tag{9}$$

where

$$h(t) = e^{j\theta}\int q_b(\tau)p(t-\tau)d_\tau \tag{10}$$

is called the *composite baseband channel*, which is, in general, complex.

2.2. The Discrete-Time Model

For band-limited transmissions, the baseband signal $y(t)$ can be sampled without loss of information if the sampling rate $f_s = 1/T_s$ exceeds the Nyquist rate. This implies that the sampling rate should be at least the symbol rate. If the transmitted pulse $p(t)$ has the bandwidth that exceeds the minimum bandwidth of $1/2T$, the sampling rate should be higher than the symbol rate. We will assume that the received signal is "over" sampled at a rate G times the symbol rate, namely, $T = GT_s$. For narrowband

transmissions, $G = 2$ satisfies the Nyquist rate, whereas, for spread-spectrum communications, G should be greater than or equal to the spreading gain of the system.

Denote the sampled discrete-time baseband signals as

$$y[n] \triangleq y(nT_s + \tau_0), \quad w[n] \triangleq w(nT_s + \tau_0) \tag{11}$$

where τ_0 is the sampling phase. For time-invariant channels with synchronized carrier as defined in (9), the received data samples satisfy

$$y[n] = \sum_k h(nT_s - kGT_s + \tau_0)s[k] + w[n] \tag{12}$$

$$= \sum_k h[n - kG]s[k] + w[n], \quad h[n] \triangleq h(nT_s + \tau_0) \tag{13}$$

Note that, when the input sequence $s[n]$ is stationary, $y[n]$ is not stationary unless $G = 1$. In general, the oversampled signal $y[n]$ is cyclostationary.

2.2.1. The SIMO Model. A convenient model for the oversampled discrete-time channel is the vectorized single-input multioutput (SIMO) model shown in Fig. 2. This model is obtained by noting that, if the received signal is sampled G times faster than the symbol rate $1/T$, there are G samples per symbol period, and $y[n]$ can be split into G subsequences. Specifically, for $i = 1, \ldots, G$, denote

$$y_i[n] \triangleq y[nG + i - 1], \mathbf{y}[n] = [y_1[n], \ldots, y_G[n]]^T \tag{14}$$

$$w_i[n] \triangleq w[nG + i - 1], \mathbf{w}[n] = [w_1[n], \ldots, w_G[n]]^T \tag{15}$$

$$h_i[n] \triangleq h[nG + i - 1], \mathbf{h}[n] = [h_1[n], \ldots, h_G[n]]^T \tag{16}$$

We then have the SIMO model given by

$$\mathbf{y}[n] = \sum_{k=0}^{L} \mathbf{h}[k]s[n - k] + \mathbf{w}[n], \tag{17}$$

where L, referred to as the channel order, is such that $k > L$ and $k < 0$. We assume that L is finite.[1]

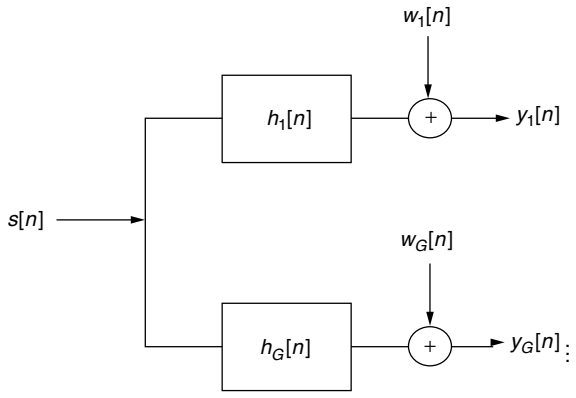


Figure 2. The SIMO vector channel model.

[1] For strictly band-limited signals, $L = \infty$. On the other hand, for strictly time-limited signals, L is always finite but, unfortunately, the Nyquist sampling frequency is ∞. In practice, L can be chosen large enough so that the model is sufficiently accurate.

If we collect all the transmitted symbols in a vector $\mathbf{s}$, all the received data samples in $\mathbf{y}$, and all channel parameters in $\mathbf{h}$

$$\mathbf{y} \triangleq \begin{pmatrix} \mathbf{y}[N-1] \\ \vdots \\ \mathbf{y}[0] \end{pmatrix}, \mathbf{h} \triangleq \begin{pmatrix} \mathbf{h}[0] \\ \vdots \\ \mathbf{h}[L] \end{pmatrix}, \mathbf{s} \triangleq \begin{pmatrix} s[N-1] \\ \vdots \\ s[-L] \end{pmatrix}$$

with noise vector $\mathbf{w}$ similarly defined, we have the following model equations:

$$\mathbf{y} = \mathcal{H}(\mathbf{h})\mathbf{s} + \mathbf{w} = \mathcal{F}(\mathbf{s})\mathbf{h} + \mathbf{w} \tag{18}$$

where $\mathcal{H}(\mathbf{h})$ is a block Toeplitz matrix generated from the channel $\mathbf{h}$ and $\mathcal{F}(\mathbf{s})$ a block Hankel matrix generated from the input $\mathbf{s}$ with dimensions matched to those of $\mathbf{y}$ and $\mathbf{s}$

$$\mathcal{H}(\mathbf{h}) = \begin{pmatrix} \mathbf{h}[0] & \cdots & \mathbf{h}[L] & & \\ & \ddots & & \ddots & \\ & & \mathbf{h}[0] & \cdots & \mathbf{h}[L] \end{pmatrix} \tag{19}$$

$$\mathcal{F}(\mathbf{s}) = \begin{pmatrix} s[N-1]\mathbf{I}_G & \cdots & s[N-L-1]\mathbf{I}_G \\ \vdots & \text{Block Hankel} & \vdots \\ s[0]\mathbf{I}_G & \cdots & s[-L]\mathbf{I}_G \end{pmatrix} \tag{20}$$

$$= \begin{pmatrix} s[N-1] & \cdots & s[N-L-1] \\ \vdots & \text{Hankel} & \vdots \\ s[0] & \cdots & s[-L] \end{pmatrix} \otimes \mathbf{I}_G, \tag{21}$$

where the operator $\otimes$ is the Kronecker product and $\mathbf{I}_G$ is the $G \times G$ identity matrix.

2.2.2. The MIMO Model. The SIMO model can be easily extended to incorporate systems that involve multiple users and multiple transmitting and receiving antennas. A general schematic is shown in Fig. 3, where there are K users, each transmitting a sequence of symbols $s_i[n]$ using a particular waveform. The signals $\{y_i(t)\}$ received by M receivers are distorted by noise, their corresponding propagation channels, and cross-interference. Let $\mathbf{y}_j[k]$ and $\mathbf{w}_j[n]$ be the received signal vector and the additive noise at the jth antenna, respectively, and $\mathbf{h}_{ji}[k]$ be the channel

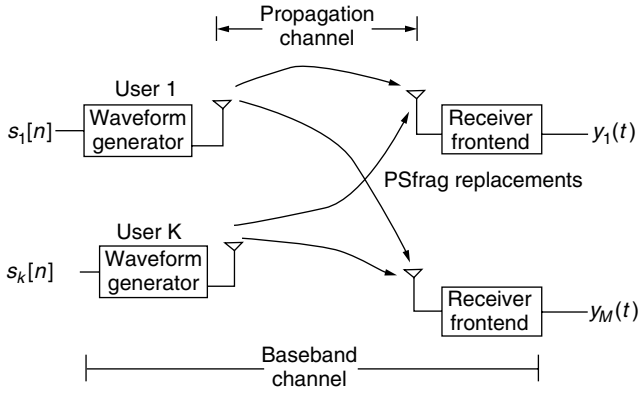


Figure 3. A general multiuser communication system.

between the ith user and the jth antenna. We then have the multiinput multioutput (MIMO) channel model

$$\mathbf{y}_j[n] = \sum_{i=1}^{K} \sum_{k} s_i[k]\mathbf{h}_{ij}[n-k] + \mathbf{w}_j[n], \quad j = 1, \ldots, M$$

Stacking data from all antennas as

$$\mathbf{y}[k] \triangleq \begin{pmatrix} \mathbf{y}_1[k] \\ \vdots \\ \mathbf{y}_M[k] \end{pmatrix}, \mathbf{h}_j[k] \triangleq \begin{pmatrix} \mathbf{h}_{1j}[k] \\ \vdots \\ \mathbf{h}_{Mj}[k] \end{pmatrix}, \quad j = 1, \ldots, M$$

one obtains the MIMO model

$$\mathbf{y}[k] = \sum_{i=1}^{M} \sum_{k} s_i[k]\mathbf{h}_i[n-k] + \mathbf{w}[n]$$

where $\mathbf{w}[n]$ is the noise vector similarly defined, and $\mathbf{h}_j[k]$ is the (vector) channel impulse response from the jth user to all receiving antennas. Again collecting all received data in a single vector $\mathbf{y}$ and transmitted symbols in $\mathbf{s} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T]^T$, we obtain the (batch) MIMO equation

$$\mathbf{y} = [\mathcal{H}(\mathbf{h}_1), \ldots, \mathcal{H}(\mathbf{h}_K)]\mathbf{s} + \mathbf{w} = [\mathcal{F}(\mathbf{s}_1), \ldots, \mathcal{F}(\mathbf{s}_K)]\mathbf{h} + \mathbf{w} \tag{22}$$

While we shall restrict our discussion to the single-user case multiple-antenna systems, many results apply directly to the estimation of MIMO channels described in Eq. (22).

3. CHANNEL ESTIMATION: GENERAL CONCEPTS

The objective of channel estimation is to infer channel parameters from the received signal. The function that maps the received signal and prior knowledge about the channel and pilot symbols is called the *estimator*. In this section, we discuss the formulation of the estimation problem. The development of specific estimators will be considered in Section 4.

3.1. Channel Estimation Techniques

3.1.1. Transmissions with Embedded Pilot Symbols. The development of channel estimation algorithms depends on the format of the transmitted symbols. Traditionally, the transmission is divided into two phases: the training phase and the transmission phase. In the training phase, pilot symbols known to the receiver are transmitted so that channel parameters can be estimated. The estimated channel is then used in the design of the receiver. When a feedback channel is available, the estimated channel can also be utilized to design an optimal transmitter.

For packet transmissions, especially in a wireless environment, it may be necessary that the channel be estimated separately for each packet. For example, the base station in a cellular system receives packets from different users, each with a different propagation channel. Therefore, pilot symbols need to be inserted into every data packet. The presence (or the absence) of pilot symbols, the number of pilot symbols, and the placement of pilot symbols all affect the parametric model from which channel estimators are derived. Figure 4 shows a typical packet format for the transmitted symbol $s[n]$ and its corresponding received signal $\mathbf{y}[n]$. The shaded area in the transmitted data frame is where pilot symbols are located. In general, there may be multiple pilot clusters $\{\mathbf{s}_{pi}\}$ whose location and values are known to the receiver.

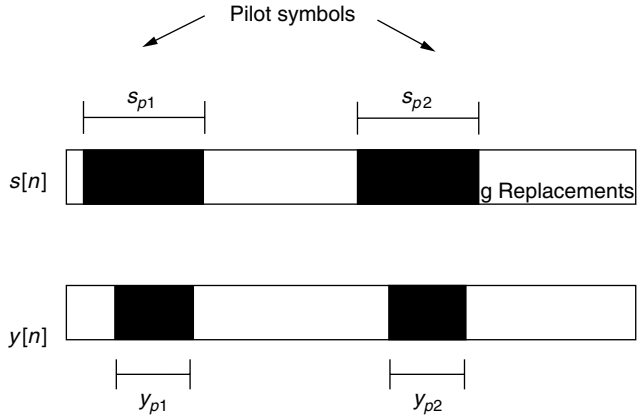


Figure 4. Signal frame structure. The shaded areas in the transmitted data frame are where pilot symbols are located. The shaded areas in the received signal frame are where samples corresponding to only pilot symbols are located.

3.1.2. Training-Based, Semiblind, and Blind Estimators. If the channel is memoryless, that is, if $L = 0$ in (17), then any received sample, say, $y[k]$, is a function of either a data symbol or a pilot symbol. However, if the channel has memory, $L > 0$, a received sample $\mathbf{y}[k]$ may be a function of (1) the (unknown) data symbols only, (2) the pilot symbols only, or (3) both data and pilot symbols. As illustrated by the shaded parts in Fig. 4, let $\mathbf{y}_{pi}$ be the cluster of received samples corresponding only to pilot cluster $\mathbf{s}_{pi}$. In other words, every sample in $\mathbf{y}_{pi}$ is a function of the pilot symbols in $\mathbf{s}_{pi}$ only. With this in mind, we consider three types of channel estimators, depending on the information that is used by the estimator.

- *Training-Based Channel Estimators*. A training-based channel estimator only uses data that correspond to the pilot symbols. In the case illustrated in Fig. 4, the estimator takes $\{\mathbf{y}_{p1}, \mathbf{s}_{p1}\}$ and $\{\mathbf{y}_{p2}, \mathbf{s}_{p2}\}$ as its input and produces an estimate of the channel. If $\mathbf{h}$ is the vector containing all channel parameters, $\mathbf{s}_p$ the vector containing all pilot clusters, and $\mathbf{y}_p$ the vector containing all received data corresponding to $\mathbf{s}_p$, a training-based channel estimator can be written as

$$\hat{\mathbf{h}} = G_T(\mathbf{s}_p, \mathbf{y}_p). \tag{23}$$

  Notice that although other received data also contain information about the channel, they are not utilized by the estimator. Training-based channel estimators are commonly used in practice. These estimators are easy to derive and analyze because the unknown data are not part of the observation. On the other hand, there needs to be a sufficient number of pilot

symbols present for good performance, and there are restrictions on how they should be placed in the data packet. For example, the size of pilot clusters must be at least $L+1$ in order to have one received data sample that is related to pilot symbols only.

- *Blind Channel Estimator*. When there are no pilot symbols available, the channel estimator is called "blind." In this case, the channel estimator uses the received signal $\mathbf{y}$ to estimate the channel

$$\hat{\mathbf{h}} = G_B(\mathbf{y}) \tag{24}$$

where the estimator $G_B(\cdot)$ is derived on the basis of certain qualitative information about the model. For example, although the input data are not known to the receiver, their statistical properties may be known. Other techniques include the exploitation of the finite-alphabet property of the source and certain parametric models of the channel. It is not obvious that blind channel estimation is even possible as neither the input nor the channel is known to the receiver. Indeed, such estimation is possible only under certain identifiability conditions, and the identification can be achieved only up to a scaling factor. In some applications such as terrestrial broadcasting of high-definition television (HDTV) where the requirement of efficient bandwidth utilization is stringent, pilot symbols may be so scarce that the receiver must acquire the channel without training. In such cases, blind channel estimation is necessary.

- *Semiblind Channel Estimator*. Between training-based and blind estimators is the class of semiblind channel estimators that utilize not only that part of signal corresponding to the training symbols but also the part corresponding to data symbols. In particular, a semiblind channel estimator takes $\{\mathbf{s}_{p1}, \mathbf{s}_{p2}, \mathbf{y}\}$ to generate a channel estimate. A semiblind channel estimator can be expressed as

$$\hat{\mathbf{h}} = G_{SB}(\mathbf{s}_p, \mathbf{y}) \tag{25}$$

By fully exploiting the information about the channel contained in the entire data record, semiblind channel estimation may provide considerable gain over training-based algorithms as shown in Section 4.2.

3.2. Performance Measure and Performance Bound

In estimating the channel, we can model $\mathbf{h}$ as a deterministic vector or a random vector with a certain probability distribution. If $\mathbf{h}$ is deterministic, we have the problem of point estimation whereas, when $\mathbf{h}$ is random, the estimation problem is Bayesian. In the following discussion, we will restrict ourselves to the case when $\mathbf{h}$ is deterministic but unknown.

The problem of channel estimation is to find an estimator $\hat{\mathbf{h}}$ that is close to the true channel under a certain performance measure. Typically, we will be concerned about the bias and covariance of the estimator defined by

$$\mathcal{B}(\hat{\mathbf{h}}) \triangleq E\{\hat{\mathbf{h}} - \mathbf{h}\}, \mathcal{V}(\hat{\mathbf{h}}) \triangleq E\{[\hat{\mathbf{h}} - E(\hat{\mathbf{h}})][\hat{\mathbf{h}} - E(\hat{\mathbf{h}})]^H\} \tag{26}$$

where both, in general, are functions of $\mathbf{h}$. If $\mathcal{B}(\mathbf{h}) = 0$ for all possible $\mathbf{h}$, then the estimator is unbiased. The performance of an estimator can also be measured by the covariance matrix of the estimation error

$$\mathcal{M}(\hat{\mathbf{h}}) \triangleq E(\hat{\mathbf{h}} - \mathbf{h})(\hat{\mathbf{h}} - \mathbf{h})^H \tag{27}$$

from which we obtain the *mean-square error*

$$E(\|\hat{\mathbf{h}} - \mathbf{h}\|^2) = tr\{\mathcal{M}(\hat{\mathbf{h}})\} \tag{28}$$

These definitions also apply to random channels.

In assessing the performance of the estimator, it is often useful to compare the covariance of the estimation error with the *Cramér–Rao bound* (CRB), which is a lower bound on the MSE of any unbiased estimator. Given a deterministic channel $\mathbf{h}$, assume that we have a well-defined probability density function $\mathbf{f}(\mathbf{y}; \mathbf{h})$. Viewed as a function of $\mathbf{h}$, $\boldsymbol{f}(\mathbf{y}; \mathbf{h})$ is the likelihood function of the channel parameter $\mathbf{h}$. The complex Fisher information matrix (FIM) $\boldsymbol{I}(\mathbf{h})$ is defined by

$$\mathbf{I}(\mathbf{h}) \triangleq E\{[\nabla_{\mathbf{h}^*} \ln f(\mathbf{y}; \mathbf{h})][\nabla_{\mathbf{h}^*}^H \ln f(\mathbf{y}; \mathbf{h})]\} \tag{29}$$

where the complex gradient operator applied to a real function $g(\mathbf{x})$ with complex argument $\mathbf{x} \in C^K$ is defined by

$$\nabla_{\mathbf{x}^*} g(\mathbf{y}) \triangleq \frac{1}{2}\begin{pmatrix} \dfrac{\partial g(\mathbf{x})}{\partial \Re\{x_1\}} \\ \vdots \\ \dfrac{\partial g(\mathbf{x})}{\partial \Re\{x_K\}} \end{pmatrix} + \frac{j}{2}\begin{pmatrix} \dfrac{\partial g(\mathbf{x})}{\partial \Im\{x_1\}} \\ \vdots \\ \dfrac{\partial g(\mathbf{x})}{\partial \Im\{x_K\}} \end{pmatrix} \tag{30}$$

Under regularity conditions [19,34], the MSE of any unbiased estimator $\hat{\mathbf{h}}$ is lower-bounded by $\mathbf{I}^{-1}$ $(\mathbf{h})$ and MSE by $tr\{\mathbf{I}^{-1}(\mathbf{h})\}$:

$$E(\hat{\mathbf{h}} - \mathbf{h})(\hat{\mathbf{h}} - \mathbf{h})^H \geq \mathbf{I}^{-1}(\mathbf{h}), \tag{31}$$

$$E(\|\hat{\mathbf{h}} - \mathbf{h}\|^2) \geq tr\{\mathbf{I}^{-1}(\mathbf{h})\} \tag{32}$$

An unbiased estimator that achieves the CRB is called *efficient*. The same expression also holds when $\mathbf{h}$ is random except that the expectation in (29) is taken over $\mathbf{y}$ and the channel vector $\mathbf{h}$.

3.3. Estimation Techniques

3.3.1. The Maximum-Likelihood Methods. One of the most popular parameter estimation algorithms is the maximum-likelihood (ML) method. The ML estimator can usually be derived in a systematic way by maximizing the likelihood function

$$\hat{\mathbf{h}}_{ML} = \arg\max_{\mathbf{h}\in\Theta} \boldsymbol{f}(\mathbf{y}; \mathbf{h}) \tag{33}$$

where Θ is the set of channels that satisfy certain constraints.

The ML estimator has a number of attractive properties. If an efficient estimator exists, it must be an ML estimator; it can be shown that the class of maximum-likelihood estimators are asymptotically efficient [23],

although examples exist that the ML estimator may perform poorly when the data size is small.

While the ML estimator is conceptually simple, the implementation of the ML estimator is sometimes computationally intensive. Furthermore, the optimization of the likelihood function in (33) is often hampered by the existence of local maxima. Therefore, it is desirable that effective initialization techniques are used in conjunction with ML estimation.

3.3.2. The Moment Methods. For some applications, the knowledge of the model is incomplete and the likelihood function cannot be specified explicitly. In such cases, the method of moments may be applied. Suppose that we know the explicit form that the ith moments $\mathbf{M}_i(\mathbf{h})$ of $\mathbf{y}$ relate to the channel parameter. The moment estimator is then given by matching the moment functions $\mathbf{M}_i(\mathbf{h})$ with moments estimated from the data. Often, simple estimators can be obtained from solving for $\hat{\mathbf{h}}$ directly from

$$\mathbf{M}_i(\mathbf{h}) = \hat{\mathbf{M}}_i \tag{34}$$

The matching of moments can also be done using least-squares techniques.

4. ESTIMATION OF SIMO CHANNELS

We now apply the performance bound and general estimation techniques to the estimation of the SIMO channel model (17) developed in Section 2. Training-based estimation is presented first followed by blind and semiblind estimation.

4.1. Training-Based Channel Estimation Algorithms

Training-based estimators use only those parts of the received signal corresponding to pilot symbols. If there is a single cluster of training symbols, we can use the model given in (18) assuming that all symbols $s[n]$ are known to the receiver, and

$$\mathbf{y} = \mathcal{F}(\mathbf{s})\mathbf{h} + \mathbf{w}, \mathcal{F}(\mathbf{s}) = \mathcal{F}_1(\mathbf{s}) \otimes \mathbf{I}_G \tag{35}$$

where the noise vector $\mathbf{w}$ is zero mean, Gaussian with covariance $\sigma^2\mathbf{I}$, and

$$\mathcal{F}_1(\mathbf{s}) \triangleq \begin{pmatrix} s[N-1] & \cdots & s[N-L-1] \\ \vdots & \text{Hankel} & \vdots \\ s[0] & \cdots & s[-L] \end{pmatrix}$$

If there are multiple clusters, the preceding equations remain valid with $\mathcal{F}(\mathbf{s})$ replaced by a stack of $\mathcal{F}(\mathbf{S}_{pi})$, each corresponding to one cluster of pilot symbols.

4.1.1. Performance Bound and Identifiability. The likelihood function $f(\mathbf{y}, \theta)$ for the parameter $\boldsymbol{\theta} = \begin{pmatrix} \mathbf{h} \\ \sigma^2 \end{pmatrix}$ is given by

$$f(\mathbf{y};\theta) = \frac{1}{(\pi\sigma^2)^N} \exp\left\{-\frac{1}{\sigma^2}\|\mathbf{y} - \mathcal{F}(\mathbf{s})\mathbf{h}\|^2\right\} \tag{36}$$

The (complex) Fisher information matrix is given by

$$I(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} \mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}) & \mathbf{0} \\ \mathbf{0} & \dfrac{N}{\sigma^2} \end{pmatrix} \tag{37}$$

and the CRB for the training-based channel estimators is

$$CRB_T(\mathbf{h}) = \sigma^2[\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})]^{-1} \tag{38}$$

assuming the inverse exists.

The assumption that $\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})$ is invertible is significant. If this condition is not satisfied, the channel is not identifiable. Specifically, if $\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})$ is not invertible, columns of $\mathcal{F}(\mathbf{s})$ are linearly dependent. Hence, there exists a vector $\Delta\mathbf{h}$ such that

$$\mathcal{F}(\mathbf{s})\Delta\mathbf{h} = \mathbf{0}$$

which implies that

$$\mathbf{y} = \mathcal{F}(\mathbf{s})(\mathbf{h} + \gamma\,\Delta\mathbf{h}) + \mathbf{w}$$

for any γ. In other words, the estimation error can be arbitrarily large.

4.1.2. Design of Pilot Sequence. The condition of identifiability imposes certain constraints on the training sequence. To ensure that $\mathcal{F}(\mathbf{s})$ has full column rank, it is necessary and sufficient that $\mathcal{F}_1(\mathbf{s})$ have full column rank. An equivalent condition is that the *linear complexity*[2] [5] of the pilot sequence $s[n]$ should be greater than L. This implies that, to estimate a set of parallel channels of order L, the minimum number of training symbols must be greater than $2L$.

Note that the CRB in (38) is not a function of the channel parameter. It is, however, a function of the transmitted (pilot) symbol vector $\mathbf{s}$. Since the CRB can be achieved by the ML estimator described below, the training sequence should be designed to minimize the CRB. Specifically, we may choose the training sequence with constant amplitude σ_s according to the following optimization:

$$\min_{\mathbf{s}} tr\{(\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}))^{-1}\}$$

It can be shown that among all possible transmitted symbols with constant amplitude σ_s, the one that minimizes the CRB satisfies the orthogonality condition

$$\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}) = N\sigma_s^2\mathbf{I}$$

The sequence that gives the minimum CRB can be chosen from points on the circle with radius σ_s in the complex plane [6].

[2] The linear complexity of a sequence $s[n]$ is defined as the smallest number c such that there exist α_i such that $s[n] = \sum_{i=1}^{c} \alpha_i s[n-i]$ for all $n \geq c$.

4.1.3. THE ML ESTIMATOR

The maximum-likelihood (ML) estimator of the channel is given by

$$\mathbf{h}_{ML} = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathcal{F}(\mathbf{s})\mathbf{h}\|^2 \tag{39}$$

$$= [\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})]^{-1}\mathcal{F}^H(\mathbf{s})\mathbf{y}$$

$$= (\{[\mathcal{F}_1^H(\mathbf{s})\mathcal{F}_1(\mathbf{s})]^{-1}\mathcal{F}_1^H(\mathbf{s})\} \otimes \mathbf{I}_G)\mathbf{y} \tag{40}$$

assuming again the inverse exists. It is easily verified that

$$E(\hat{\mathbf{h}}_{ML}) = \mathbf{h},\ \mathcal{V}(\hat{\mathbf{h}}_{ML}) = \sigma^2(\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}))^{-1} \tag{41}$$

In other words, the ML channel estimator is efficient. When the optimal training sequence is used, from (41), we obtain

$$\mathcal{V}(\hat{\mathbf{h}}_{ML}) = \frac{1}{N} \xrightarrow{N\to\infty} 0 \tag{42}$$

Hence, the estimator is consistent and the estimation error decreases to zero at the rate of $1/N$.

The implementation of the ML estimator can be simplified by treating the G subchannels in Fig. 2 separately. Let $\mathbf{h}^{(i)}$ be the channel vector containing the impulse response of the ith subchannel, and $\mathbf{y}^{(i)}$ be the observation corresponding to the ith subchannel. Because the assumption that the noise samples in $\mathbf{w}$ are independent, we have

$$\hat{\mathbf{h}}^{(i)} = [\mathcal{F}_1^H(\mathbf{s})\mathcal{F}_1(\mathbf{s})]^{-1}\mathcal{F}_1^H(\mathbf{s})\mathbf{y}^{(i)} \tag{43}$$

which involves the inversion of a $(L_i+1)\times(L_i+1)$ Hermitian matrix where L_i is the order of the ith channel. In contrast, the inversion of a $G(L+1)\times G(L+1)$ matrix is involved in (40).

4.1.4. Recursive Least Squares. The ML channel estimator, under the assumption that $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is zero mean, Gaussian with covariance $\sigma^2\mathbf{I}$, is also the least-squares estimator defined by (40). In addition to reducing computation complexity by avoiding direct matrix inverse, the recursive least-squares (RLS) algorithm computes the channel estimate recursively, allowing updates as more data become available.

Since all sub-channels can be estimated independently, without loss of generality, we assume that $G = 1$. Suppose that we have already obtained the LS estimate $\hat{\mathbf{h}}_n$ using all observation $\mathbf{y}_n$ up to time n and their corresponding input symbols $\mathbf{s}_n$ defined by

$$\mathbf{y}_n = \begin{pmatrix} y[n] \\ \vdots \\ y[0] \end{pmatrix},\ \mathbf{s}_n = \begin{pmatrix} s[n] \\ \vdots \\ s[-L] \end{pmatrix} \tag{44}$$

From (40), we have

$$\hat{\mathbf{h}}_n \triangleq \underbrace{[\mathcal{F}^H(\mathbf{s}_n)\mathcal{F}(\mathbf{s}_n)]}_{\mathbf{R}_n}{}^{-1}\ \underbrace{\mathcal{F}^H(\mathbf{s}_n)\mathbf{y}_n}_{\mathbf{r}_n}$$

$$= \mathbf{R}_n^{-1}\mathbf{r}_n = \mathbf{P}_n\mathbf{r}_n \tag{45}$$

where $\mathbf{P}_n \triangleq \mathbf{R}_n^{-1}$. Letting $\mathbf{s}_{L+1}[n] \triangleq [s[n], \ldots, s[n-L]]^T$, we note that $\mathbf{R}_n$ can be computed recursively

$$\mathbf{R}_n = \mathbf{R}_{n-1} + \mathbf{s}_{L+1}^*[n]\mathbf{s}_{L+1}^T[n]$$

Recursive relations also hold for $\mathbf{P}_n$ and $\mathbf{r}_n$

$$\mathbf{P}_n \triangleq \mathbf{R}_n^{-1} = \mathbf{P}_{n-1} - \frac{\mathbf{P}_{n-1}\mathbf{s}_{L+1}^*[n]\mathbf{s}_{L+1}^T[n]\mathbf{P}_{n-1}^H}{1 + \mathbf{s}_{L+1}^T[n]\mathbf{P}_{n-1}\mathbf{s}_{L+1}^*[n]} \tag{46}$$

$$\mathbf{r}_n \triangleq \mathcal{F}^H(\mathbf{s}_n)\mathbf{y}_n = \mathbf{r}[n-1] + y[n]\mathbf{s}_{L+1}^*[n] \tag{47}$$

Suppose now that we are made available the next pilot $s[n+1]$ and the corresponding observation $y[n+1]$. Then, the ML estimator using data up to time $n+1$ can be updated from the previous estimate by

$$\hat{\mathbf{h}}_{n+1} \triangleq \hat{\mathbf{h}}_n + \mathbf{g}_{n+1}\varepsilon[n+1] \tag{48}$$

where $\varepsilon[n+1]$ is the error of the predicted observation using $\hat{\mathbf{h}}[n]$

$$\varepsilon[n+1] = y[n+1] - \mathbf{s}_{L+1}^T[n+1]\hat{\mathbf{h}}_n, \tag{49}$$

and $\mathbf{g}_{n+1}$ is the gain vector

$$\mathbf{g}_{n+1} = \frac{\mathbf{P}_n\mathbf{s}_{L+1}^*[n+1]}{1 + \mathbf{s}_{L+1}^T[n+1]\mathbf{P}_n\mathbf{s}_{L+1}^*[n+1]} \tag{50}$$

It is interesting to note that the amount of update in the channel estimate depends on the prediction error by the channel estimate. The RLS algorithm reduces the computation of the batch ML estimation to $\mathcal{O}((L+1)^2)$.

4.1.5. The LMS Algorithm. For its simplicity, the LMS algorithm, originally proposed by Widrow and Hoff [36], is perhaps the most widely used adaptive estimation algorithm. It resembles the RLS update and may be expressed as follows

$$\hat{\mathbf{h}}_{n+1} \triangleq \hat{\mathbf{h}}_n + \mu\mathbf{s}_{L+1}^*[n+1]\varepsilon[n+1] \tag{51}$$

where the computationally more expensive gain vector $\mathbf{g}_{n+1}$ in RLS is replaced by the readily available input vector and a constant-step-size μ for the update. The derivation of the LMS algorithm does not have a direct connection to the ML estimation. Under the assumption that the input sequence $s[n]$ is random, the LMS can be viewed as the stochastic gradient implementation of the minimization of the average prediction error

$$\min_{\mathbf{h}} E\left\{|y[n] - \sum_i h[i]s[n-i]|^2\right\} \tag{52}$$

where the expectation is taken over the noise and the input process.

4.2. Blind and Semiblind Channel Estimation Algorithms

We now consider the semiblind and blind channel estimation problem where the channel estimator uses the entire data record.

The input sequence can be partitioned into two parts: the pilot symbols $\mathbf{s}_p$ and data symbols $\mathbf{s}_d$. With the corresponding partition in the channel matrix, we have

$$\mathbf{y} = \mathcal{H}_p(\mathbf{h})\mathbf{s}_p + \mathcal{H}_d(\mathbf{h})\mathbf{s}_d + \mathbf{w} \tag{53}$$

where $\mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is again assumed to be zero mean Gaussian with covariance $\sigma^2\mathbf{I}$. The estimation problem can now be formulated on the basis of two models of the input data vector $\mathbf{s}_d$. The *deterministic model* assumes that $\mathbf{s}_d$ is a deterministic, unknown nuisance parameter in the channel estimation. The *stochastic model*, on the other hand, assumes that $\mathbf{s}_d$ is random with certain distribution. These two models lead to different likelihood functions and therefore, different performance bounds and estimation algorithms. The choice of modeling depends naturally on the application. Typically, if the channel is to be estimated using a small number of samples, the deterministic model is more effective.

4.2.1. The Performance Bound. Under the deterministic model, the likelihood function of the unknown parameter $\boldsymbol{\theta} = [\mathbf{h}^T \mathbf{s}_d^T \sigma^2]^T$ is given by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \frac{1}{\pi^N \sigma^{2N}} \exp\left\{-\frac{1}{\sigma^2}\|\mathbf{y} - \mathcal{H}_p(\mathbf{h})\mathbf{s}_p - \mathcal{H}_d(\mathbf{h})\mathbf{s}_d\|^2\right\} \tag{54}$$

The complex Fisher information matrix is given by

$$I(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} \mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}) & \mathcal{F}^H(\mathbf{s})\mathcal{H}_d(\mathbf{h}) & \mathbf{0} \\ \mathcal{H}_d^H(\mathbf{h})\mathcal{F}(\mathbf{s}) & \mathcal{H}_d^H(\mathbf{h})\mathcal{H}_d(\mathbf{h}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dfrac{N}{\sigma^2} \end{pmatrix} \tag{55}$$

Using the block matrix inversion formula, we obtain the CRB for any unbiased semiblind channel estimator $\hat{\mathbf{h}}$

$$\mathrm{CRB}_{SB}(\mathbf{h}) = \sigma^2[\mathcal{F}^H(\mathbf{s})\mathcal{P}^{\perp}_{\mathcal{H}_d(\mathbf{h})}\mathcal{F}(\mathbf{s})]^{-1} \tag{56}$$

where

$$\mathcal{P}^{\perp}_{\mathcal{H}_d(\mathbf{h})} \triangleq \mathbf{I} - \mathcal{H}_d(\mathbf{h})[\mathcal{H}_d^H(\mathbf{h})\mathcal{H}_d(\mathbf{h})]^{-1}\mathcal{H}_d(\mathbf{h}) \tag{57}$$

is the projection matrix onto the null space of $\mathcal{H}_d(\mathbf{h})$. It is clear from the above expressions and (38) that

$$\mathrm{CRB}_{SB}(\mathbf{h}) \le \mathrm{CRB}_T(\mathbf{h})$$

with equality when all symbols are known.

The derivation for the CRB under the stochastic model is more complicated unless one assumes that the input sequence is Gaussian. Details can be found in Ref. 7.

4.2.2. The Design of Pilot Symbols and Their Placement. The design of training for semiblind channel estimation involves the joint design of the number of pilot symbols, the pilot symbols, and their placements. The CRB can be used as the performance measure for this design.

It is natural to expect that more pilot symbols lead to better performance. On the other hand, increasing the number of pilot symbols reduces the number of data symbols transmitted in a packet. The gain in performance can be evaluated against the percentage of known symbols in a packet using (56). Figure 5 shows an example of the relation between the MSE and the percentage η of known symbols in the data packet for a multipath channel of order $L = 5$ at 10 dB SNR. Notice that $\eta = 100\%$ corresponds to the performance of the training-based ML algorithm. It can be seen that the gain of using all pilot symbols ($\eta = 100\%$) over that of using 1% pilot symbols is about 45 dB, and about 40 dB of gain can already be achieved at $\eta = 20\%$.

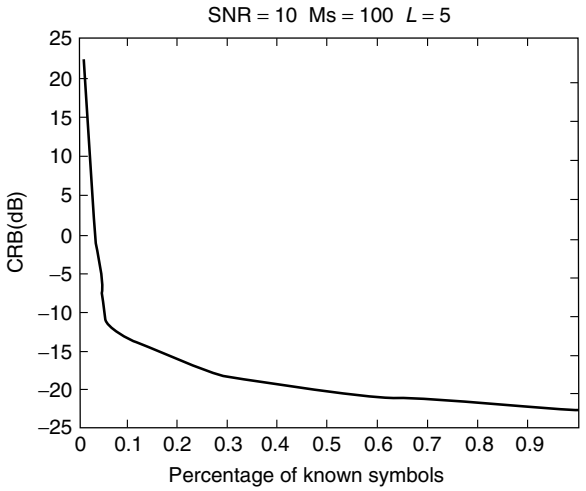


Figure 5. CRB for a multipath channel with $L = 5$; SNR = 10 dB.

Given the percentage of pilot symbols in a data packet, the problem of pilot design can be formulated as minimizing $tr\{\mathrm{CRB}_{SB}(\mathbf{h})\}$ among choices of pilot symbols and their placement. This optimization, unfortunately, depends on the channel coefficients. For random channels, however, the minimization of CRB does lead to the optimal design of pilot symbols and their placement, which are independent of the channel [9].

4.2.3. The ML Estimation. When some or all of the input symbols are unknown, the likelihood function of the channel parameters and the unknown symbols depends on the model assumed for the unknown symbols. We then have the so-called deterministic maximum likelihood (DML) where the unknown input symbols are deterministic, and the stochastic maximum likelihood (SML), where the unknown symbols are assumed to be random with some known distribution.

4.2.3.1. The SML Estimation. While the input vector $\mathbf{s}$ is unknown, it may be modeled as a random vector with known distribution. In such a case, the likelihood function of the channel parameter $\mathbf{h}$ can be obtained by

$$f(\mathbf{y}; \mathbf{h}) = \int f(\mathbf{y}|\mathbf{s}_d; \mathbf{h}) f(\mathbf{s}_d) d\mathbf{s}_d \tag{58}$$

where $f(\mathbf{s}_d)$ is the marginal pdf of the unknown data vector and $f(\mathbf{y}|\mathbf{s}_d; \mathbf{h})$ is the likelihood function of $\mathbf{h}$ for a particular choice of $\mathbf{s}_d$. If the input symbols $\{s[k]\}$ take, with equal probability, a finite number of values, the data vector $\mathbf{s}_d$ also takes values from the signal set $\{\mathbf{v}_1, \ldots, \mathbf{v}_Q\}$ with equal probability.

The likelihood function of the channel parameter is then given by

$$f(\mathbf{y};\mathbf{h}) = \sum_{i=1}^{Q} f(\mathbf{y}|\mathbf{s}_d = \mathbf{v}_i;\mathbf{h})\Pr(\mathbf{s}_d = \mathbf{v}_i)$$
$$= C\sum_{i=1}^{Q} \exp\left\{-\frac{\|\mathbf{y} - \mathcal{F}(\mathbf{v}_i)\mathbf{h} - \mathcal{F}(\mathbf{s}_p)\mathbf{h}\|^2}{\sigma^2}\right\} \quad (59)$$

where C is a constant. The stochastic maximum-likelihood estimator is given by

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \sum_{i=1}^{Q} \exp\left\{-\frac{\|\mathbf{y} - \mathcal{F}(\mathbf{v}_i)\mathbf{h} - \mathcal{F}(\mathbf{s}_p)\|^2}{\sigma^2}\right\} \quad (60)$$

The maximization of the likelihood function defined in (58) is in general difficult. The expectation–maximization (EM) algorithm [2,8] can be applied to transform the complicated optimization to a sequence of quadratic optimizations. Kaleh and Vallet [18] first applied the EM algorithm to the equalization of communication channels with the input sequence having the finite-alphabet property. By using a *hidden Markov model* (HMM), the authors of Ref. 18 developed a batch (offline) procedure that includes the so-called forward and backward recursions [27]. Unfortunately, the complexity of this algorithm increases exponentially with the channel memory. To relax the memory requirements and facilitate channel tracking, "online" sequential approaches have been proposed [30,31,35] for a general input, and for an input with finite alphabet properties under a HMM formulation [21]. Given the appropriate regularity conditions [30] and a good initialization, it can be shown that these algorithms converge (almost surely and in the mean square sense) to the true channel value.

4.2.3.2. DML Estimation. When the noise is zero-mean Gaussian with covariance $\sigma^2 I$, the DML estimator can be obtained by the nonlinear least-squares optimization

$$\{\hat{\mathbf{h}}, \hat{\mathbf{s}}_d\} = \arg\min_{\mathbf{h},\mathbf{s}_d} \|\mathbf{y} - \mathcal{H}_p(\mathbf{h})\mathbf{s}_p - \mathcal{H}_d(\mathbf{h})\mathbf{s}_d\|^2 \quad (61)$$

The joint minimization of the likelihood function with respect to both $\mathbf{h}$ and $\mathbf{s}_d$ is also difficult in general. However, for the general estimation model (18) considered here, the observation vector $\mathbf{y}$ is linear in both the channel and the input parameters individually. We therefore have a separable nonlinear least-squares problem that can be solved sequentially:

$$\{\hat{\mathbf{h}}, \hat{\mathbf{s}}_d\} = \arg\min_{\mathbf{s}_d}\{\min_{\mathbf{h}} \|\mathbf{y} - \mathcal{F}(\mathbf{s})\mathbf{h}\|^2\} \quad (62)$$
$$= \arg\min_{\mathbf{h}}\{\min_{\mathbf{s}_d} \|\mathbf{y} - \mathcal{H}(\mathbf{h})\mathbf{s}\|^2\} \quad (63)$$

If we are interested only in estimating the channel, the preceding minimization can be rewritten as

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \| \underbrace{(\mathbf{I} - \mathcal{H}(\mathbf{h})\mathcal{H}^{\dagger}(\mathbf{h}))}_{\mathcal{P}(\mathbf{h})} \mathbf{y}\|^2 = \arg\min_{\mathbf{h}} \|\mathcal{P}(\mathbf{h})\mathbf{y}\|^2 \quad (64)$$

where $\mathcal{P}(\mathbf{h})$ is a projection transform of $\mathbf{y}$ into the orthogonal complement of the range space of $\mathcal{H}(\mathbf{h})$, or the noise subspace of the observation. Discussions of algorithms of this type can be found in an earlier study [32].

Similar to the hidden Markov model (HMM) for the statistical maximum-likelihood approach, the finite-alphabet properties of the input sequence can also be incorporated into the deterministic maximum-likelihood methods. These algorithms, first proposed by Seshadri [28] and Ghosh and Weber [12], iterate between estimates of the channel and the input. At iteration k, with an initial guess of the channel $\mathbf{h}^{(k)}$, the algorithm estimates the input sequence $\mathbf{s}_d^{(k)}$ and the channel $\mathbf{h}^{(k+1)}$ for the next iteration by

$$\mathbf{s}_d^{(k)} = \arg\min_{\mathbf{s}_d \in S} \|\mathbf{y} - \mathcal{H}(\mathbf{h}^{(k)})\mathbf{s}\|^2 \quad (65)$$

$$\mathbf{h}^{(k+1)} = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathcal{H}(\mathbf{s}^{(k)})\mathbf{h}\|^2 \quad (66)$$

where $\mathcal{S}$ is the (discrete) domain of $\mathbf{s}_d$. The optimization in (66) is a linear least-squares problem whereas the optimization in (65) can be achieved by using the Viterbi algorithm [10]. The convergence of such approaches is not guaranteed in general.

Although the ML channel estimator usually provides better performance, the computation complexity and the existence of local optima are the two major impediments. Next we present two classes of suboptimal techniques that avoid the problem of local optima with significantly reduced computation complexity.

4.2.4. Moment Techniques: The Subspace Algorithms. Subspace techniques convert the problem of blind or semiblind channel estimation to the identification of a one-dimensional subspace that contains the channel vector. By exploiting the multichannel aspects of the channel, many of these techniques lead to a constrained quadratic optimization

$$\hat{\mathbf{h}} = \arg\min_{\|\mathbf{h}\|=1} \mathbf{h}^H Q(\mathbf{y}, \mathbf{s}_p)\mathbf{h} \quad (67)$$

where $Q(\mathbf{y}, \mathbf{s}_p)$ is a positive-definite matrix constructed from the observation and pilot symbols. The solution to the preceding optimization is then given by the eigenvector of $Q(\mathbf{y}, \mathbf{s}_p)$ associated with the minimum eigenvalue.

A simple yet informative approach [37] illustrates the basic idea in a noiseless two-channel scenario. From Fig. 2, if there is no noise, the received signals from the two channels satisfy the relation

$$y_1[n] = h_1[n] * s[n], y_2[n] = h_2[n] * s[n] \quad (68)$$

where $*$ is the linear convolution. Consequently, we have

$$y_1[n] * h_2[n] = y_2[n] * h_1[n] \quad (69)$$

Since the convolution operation is linear with respect to the channel and $y_i[n]$ is available, the above equation is equivalent to solving a homogeneous linear equation

$$\mathbf{Rh} = \mathbf{0} \quad (70)$$

where $\mathbf{R}$ is a matrix made of observations from the two channels. It can be shown that under certain identifiability conditions [32], the null space of $\mathbf{R}$ has dimension 1, which means that the channel can be identified up to a constant. When there is noise, the channel estimator can be obtained from a constrained quadratic optimization

$$\hat{\mathbf{h}} = \arg \min_{\|\mathbf{h}\|=1} \mathbf{h}^H \mathbf{R}^H \mathbf{R} \mathbf{h}, \tag{71}$$

which implies that $\hat{\mathbf{h}}$ is the eigenvector corresponds to the smallest eigenvalue of $\mathbf{Q} = \mathbf{R}^t \mathbf{R}$.

Some insight into the identifiability condition can be gained in the frequency domain. Equation (68) implies that

$$\frac{y_1(z)}{y_2(z)} = \frac{h_1(z)}{h_2(z)}$$

It is clear that if the two subchannels have common zeros, it is not possible to obtain all the zeros from the observation $\{y_1(z), y_2(z)\}$.

4.2.5. Projection Algorithms. The subspace algorithms are batch algorithms, and they are not easily amenable to adaptive forms. The projection based techniques [1,11,29,33,39], on the other hand, convert the problem of channel estimation to the classic problem of linear prediction or smoothing. As a result, these estimators can be implemented adaptively in time and recursively with respect to the delay spread of the channel. The first projection-based algorithm was proposed by Slock [29], where linear prediction is used to obtain the subspace of the channel matrix. Subsequent development [1,11] based on linear predictions assumed that the input sequence is a white sequence. Under the deterministic model, a least-squares smoothing (LSS) technique was developed [33,39] that offers finite sample convergence property in the absence of noise. It also allows a lattice filter implementation that is recursive both in time and the delay spread of the channel. In fact, both the channel and the input sequence can be obtained simultaneously by using oblique projections [38].

1.5. BIBLIOGRAPHY NOTES

The modeling of linearly modulated signals can be found in standard textbooks [22,26]. For fading dispersive channels encountered in wireless communications, earlier treatments can be found in Refs. 3, 16, and 20 and more recent developments in Ref. 4. The general theory of parameter estimation, including the Cramér-Rao bound, the maximum-likelihood estimation, and the moment methods, are presented in many books. See, for example, Lehmann [23] for the mathematical treatment of the subject and Refs. 19, 24, 25, and 34 from engineering application perspectives. The estimation of complex parameters is discussed by Kay [19].

The training-based channel estimation under additive white Gaussian noise is a form of a linear least-squares problem, which is discussed extensively in Ref. 17. See also Ref. 15 for various adaptive implementations. A survey of blind channel estimation algorithms can be found in Ref. 32. The problem of semiblind channel estimation is discussed in detail in Ref. 7. Articles about more recent trends in channel estimation and equalization can be found in Refs. 13 and 14.

BIOGRAPHY

Lang Tong received his B.E. degree from Tsinghua University, Beijing, China, in 1985, and M.S. and Ph.D. degrees in electrical engineering in 1987 and 1990, respectively, from the University of Notre Dame, Indiana. He was a postdoctoral research affiliate at the Information Systems Laboratory, Stanford University, in 1991. Currently, he is an associate professor in the School of Electrical and Computer Engineering, Cornell University, Ithaca, New York.

Dr. Tong received Young Investigator Award from the Office of Naval Research in 1996, and the Outstanding Young Author Award from the IEEE Circuits and Systems Society. His areas of interest include statistical signal processing, adaptive receiver design for communication systems, signal processing for communication networks, and information theory.

BIBLIOGRAPHY

1. K. Abed-Meraim, E. Moulines, and P. Loubaton, Prediction error method for second-order blind identification, *IEEE Trans. Signal Process.* **SP-45**(3): 694–705 (March 1997).
2. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.* **41**: 164–171 (1970).
3. P. A. Bellow, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* 360–393 (Dec. 1963).
4. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: Information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* **44**(4): (Oct. 1998).
5. R. E. Blahut, *Algebraic Methods for Signal Processing and Communications Coding*, Springer-Verlag, New York, 1992.
6. D. C. Chu, Polyphase codes with good periodic correlation properties, *IEEE Trans. Inform. Theory* **3**(4): 531–532 (July 1972).
7. E. de Carvalho and D. T. M. Slock, Semi-blind Methods for FIR multichannel estimation, G. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless & Mobile Communications: Trends in Channel Estimation and Equalization*, Prentice-Hall, Englewood Cliffs, NJ, 2001.
8. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Stat. Soc.* **39**(Ser. B): (1977).
9. M. Dong and L. Tong, Optimal design and placement of pilot symbols for channel estimation, *Proc. ICASSP'2001*, 2001 (an extended journal submission to the *IEEE Trans. Signal Process.* is available from *http://www.ece.cornell.edu/~ltong/pubj.html*).
10. G. D. Forney, The Viterbi algorithm, *IEEE Proc.* **61**: 268–278 (March 1972).

11. D. Gesbert and P. Duhamel, Robust blind identification and equalization based on multi-step predictor, *Proc. IEEE Int. Conf. Acoustics. Speech Signal Processing*, Munich, Germany, April 1997, Vol. 5, pp. 2621–2624.

12. M. Ghosh and C. L. Weber, Maximum-likelihood blind equalization, *Opt. Eng.* **31**(6): 1224–1228 (June 1992).

13. G. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal Processing Advances in Wireless Communications: Trends in Channel Estimation and Equalization*, PTR Prentice-Hall, Englewood Cliffs, NJ, 2001.

14. G. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal Processing Advances in Wireless Communications: Trends in Single- and Multi-User Systems*, PTR Prentice-Hall, Englewood Cliffs, NJ, 2001.

15. S. Haykin, *Array Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

16. T. Kailath, Channel characterization: Time-variant dispersive channels, in E. Baghdadi ed., *Lectures on Communication Theory*, McGraw-Hill, New York, Chap. 6.

17. T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 2000.

18. G. K. Kaleh and R. Vallet, Joint parameter estimation and symbol detection for linear or non linear unknown dispersive channels, *IEEE Trans. Commun.* **42**(7): 2406–2413 (July 1994).

19. S. Kay, *Modern Spectral Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

20. R. S. Kennedy, *Fading Dispersive Communication Channels*, Wiley-Interscience, New York, 1969.

21. V. Krishnamurthy and J. B. Moore, On-line estimation of hidden Markov model parameters based on Kullback-Leibler information measure, *IEEE Trans. Signal Process.* **41**(8): 2557–2573 (Aug. 1993).

22. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer, Norwell, MA, 1988.

23. E. L. Lehmann, *Theory of Point Estimation*, Chapman & Hall, New York, 1991.

24. H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, 1994.

25. B. Porat, *Digital Processing of Random Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

26. J. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, 2001.

27. L. Rabiner, A tutorial on hidden Markov Models and selected applications in speech recognition, *IEEE Proc.* **77**(2): 257–285 (Feb. 1989).

28. N. Seshadri, Joint data and channel estimation using fast blind trellis search techniques, *Proc. Globecom'90*, 1991, pp. 1659–1663.

29. D. Slock, Blind fractionally-spaced equalization, perfect reconstruction filterbanks, and multilinear prediction, In *Proc. ICASSP'94 Conf.*, Adelaide, Australia, April 1994.

30. D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley Series in Probability and Mathematical Statistics, New York, 1985.

31. D. M. Titterington, Recursive parameter estimation using incomplete data, *J. Roy Stat. Soc. B* **46**(2): 257–267 (1984).

32. L. Tong and S. Perreau, Multichannel blind channel estimation: From subspace to maximum likelihood methods, *IEEE Proc.* **86**(10): 1951–1968 (Oct. 1998).

33. L. Tong and Q. Zhao, Joint order detection and blind channel estimation by least squares smoothing, *IEEE Trans. Signal Process.* **47**(9): (Sept. 1999).

34. H. L. Van Trees, *Detection, Estimation and Modulation Theory*, Vol. 1, Wiley, New York, 1968.

35. E. Weinstein, M. Feder, and A. Oppenheim, Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure, *IEEE Trans. Signal Process.* **SP-38**(9): 1652–1654 (Sept. 1990).

36. B. Widrow and Jr. M. E. Hoff, Adaptive switching circuits, *IRE WESCON Conf. Rec.*, 1960, Vol. 4, pp. 96–104.

37. G. Xu, H. Liu, L. Tong, and T. Kailath, A Least-squares approach to blind channel identification, *IEEE Trans. Signal Process.* **SP-43**(12): 2982–2993 (Dec. 1995).

38. Z. Yu and L. Tong, Joint channel and symbol estimation by oblique projections, *IEEE Trans. Signal Process.* **49**(12) (Dec. 2001).

39. Q. Zhao and L. Tong, Adaptive blind channel estimation by least squares smoothing, *IEEE Trans. Signal Process.* **47**(11) (Nov. 1999).

CHANNEL TRACKING IN WIRELESS COMMUNICATION SYSTEMS

GREGORY E. BOTTOMLEY
HÜSEYIN ARSLAN
Ericsson Inc.
Research Triangle Park, North Carolina

1. INTRODUCTION

In digital wireless communication systems, information is transmitted to a receiver, as illustrated in Fig. 1. The transmitted information reaches the receiver after passing through a radio channel, which can be represented as an unknown, time-varying filter. For conventional, coherent receivers, the effect of the channel on the transmitted signal must be estimated to recover the transmitted information. For example, with binary phase shift keying (BPSK), binary information is represented as $+1$ and -1 symbol values. The radio channel can apply a phase shift to the transmitted symbols, possibly inverting the symbol values. As long as the receiver estimates what the channel did to the transmitted signal, it can accurately recover the information sent.

Channel estimation is a challenging problem in wireless communications. Transmitted signals are typically reflected and scattered, arriving at the receiver along multiple paths. How these signals interact depends on

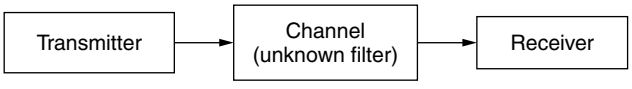


Figure 1. A wireless communication system.

their relative delays. When the relative delays are small compared to the transmitted symbol period, then different "images" of the *same* symbol arrive at the same time, adding either constructively or destructively. The overall effect is a random, fading channel response. When the relative path delays are on the order of a symbol period or more, then images of *different symbols* arrive at the same time. For example, when a particular symbol arrives at the receiver along one path, the previous symbol is arriving along another, delayed path. This is analogous to an acoustic echo and results in a more complicated channel response. Finally, because of the mobility of the transmitter, the receiver, or the scattering objects, the channel response can change rapidly with time.

This article provides an overview of channel tracking approaches commonly applied to digital cellular communication systems. Related work can be found in the study of system identification [1] and in high-frequency (HF) modem design [2].

As shown in Fig. 1, the channel impulse response is modeled as an unknown filter. Specifically, a finite-impulse-response (FIR) filter with discrete filter delays and coefficients is used. We focus on estimation of the channel coefficients, given a set of delays. Also, we focus on conventional demodulation approaches and single-antenna receivers.

In typical digital cellular systems, some part of the transmitted signal is known. In one approach, the transmitter periodically provides known *pilot symbols*, as illustrated in Fig. 2a, which can be used for channel estimation [3,4]. This approach is used in one of the downlink slot structures of the Telecommunications Industry Association/Electronics Industry Association/Interim Standard 136 (TIA/EIA/IS-136 or simply IS-136) system. In this time-division multiple-access (TDMA) system, information is transmitted in time slots. Within each time slot, clusters of known pilot symbols are provided to assist in channel estimation.

The pilot symbol approach has also been used in direct-sequence code-division multiple-access (DS-CDMA) systems. With these systems, each information symbol is represented by a sequence of "chip" symbols. This results in a spreading of the bandwidth (spread-spectrum), allowing multiple information signals to be transmitted in parallel at the same time. For convenience, we will refer to DS-CDMA systems as *wideband* systems and TDMA systems as *narrowband* systems. Pilot symbols are available in the following DS-CDMA systems: the IS-2000 system (uplink, mobile to base station) and the wideband CDMA (WCDMA) system (uplink and downlink).

In a second approach, a *pilot channel* is provided for channel estimation [5], as illustrated in Fig. 2b. This approach is related to the pilot tone approach, developed for narrowband systems [6]. The pilot channel approach is used for the downlink in the IS-95, IS-2000, and WCDMA systems. Usually the pilot channel is shared by many users and is stronger in power than an information channel.

A third approach is to provide a *training sequence* during part of the transmission, which can be used to provide an initial channel estimate. In this case, the channel must be tracked over the data portion of the signal using this signal in some way. This approach, illustrated in Fig. 2c, is used in one of the slot formats of the IS-136 system.

A training sequence is also used in the global system for mobile communications (GSM). This is a TDMA system with time slots that have short duration relative to the maximum rate of channel variation. As a result, the initial channel estimate obtained from the training sequence can be used to demodulate the data in the slot, without having to track the channel. Approaches for channel estimation in this situation are given in a separate article in this encyclopedia and elsewhere [7].

The article is organized as follows. In Section 2, a baseband-equivalent system model is given, including a model for the time-varying channel. Approximate channel models commonly used to develop tracking approaches are given in Section 3. In Section 4, filtering approaches to channel tracking are presented, based on either periodic pilot symbols or a pilot channel. Recursive approaches are presented in Section 5. Data-directed channel tracking is considered in Sections 6 and 7. Section 8 concludes the article.

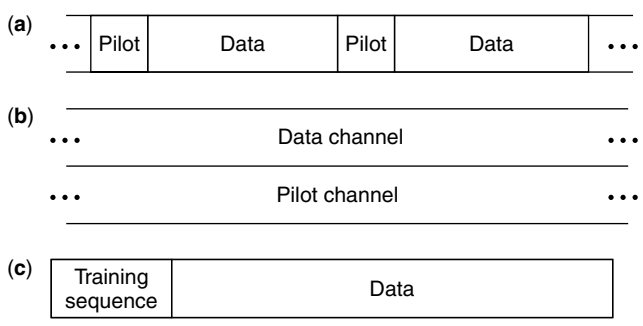


Figure 2. Systems with pilot information: (**a**) pilot symbols, (**b**) pilot channel, and (**c**) training sequence.

2. SYSTEM MODEL

A narrowband system model is presented and then extended to a wideband system model.

2.1. Narrowband System

A complex, baseband-equivalent system model is given in Fig. 3. The complex values correspond to in-phase (cosine) and quadrature (sine) components of the radio signal. At the transmitter, a sequence of digital symbols are transmitted using pulse shaping, giving

$$x(t) = \sum_k b_k f(t - kT) \tag{1}$$

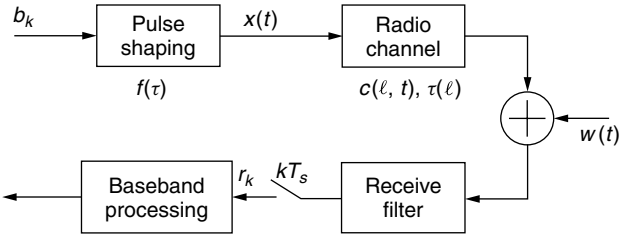


Figure 3. System model.

where b_k corresponds to the sequence of symbols, $f(\tau)$ is the pulse shape as a function of delay τ, and T is the symbol period. We assume either BPSK or quadrature phase shift keying (QPSK) modulation, so that all possible symbol values have the same amplitude ($|b_k|^2 = 1$).

The transmitted signal passes through a radio channel that can be modeled as a linear FIR filter. The resulting signal is received in the presence of noise, giving

$$y(t) = \sum_{\ell=0}^{L-1} c(\ell, t)x(t - \tau(\ell)) + w(t) \tag{2}$$

where $c(\ell, t)$ and $\tau(\ell)$ are the ℓth complex, time-varying channel coefficient and ℓth delay, respectively, and L is the number of channel taps. The noise term $w(t)$ is assumed to be white, complex (circular) Gaussian noise.

The delays are often assumed to be equally spaced; specifically, $\tau(\ell) = \ell T/M$, where M is an integer, and the spacing (T/M) for accurate modeling depends on the bandwidth of the system [8]. Typically M is 1 (symbol-spaced channel modeling) or 2 (fractionally-spaced channel modeling). For simplicity, symbol-spaced channel modeling is assumed, although extension to fractionally-spaced channel modeling is possible.

The coefficients represent the result of different multipath signal images adding together, constructively or destructively. They are well modeled as random variables. Specifically, they are modeled as uncorrelated, zero-mean complex Gaussian random variables [9]. This corresponds to "Rayleigh" fading, in that channel tap magnitudes (amplitudes) are Rayleigh-distributed. Also, the phases of the channel taps are uniformly distributed.

As the mobile transmitter or receiver moves, the phases of all multipath signal images change. This changes how the multipath images add together, so that the channel coefficient varies with time. This time variation is characterized by an autocorrelation function. The Jakes model [10], which is commonly used, assumes that the Gaussian channel coefficients have the following autocorrelation function:

$$R_\ell(\tau) = \mathrm{E}\{c^*(\ell, t)c(\ell, t + \tau)\} = \sigma_\ell^2 J_0(2\pi f_D \tau) \tag{3}$$

where the asterisk superscript denotes complex conjugation, index ℓ denotes the ℓth channel coefficient, τ is the autocorrelation delay, σ_ℓ^2 is the mean-square value of the channel coefficient, f_D is the Doppler spread, and $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind. The corresponding power spectrum of the fading process is shown in Fig. 4. This spectrum shows that the fading process has different frequency components, corresponding to different rates of change. Most of the energy is near the maximum frequency component, the Doppler spread (f_D).

The Doppler spread is proportional to the radio carrier frequency and the speed of the transmitter or receiver. For the examples given, the carrier frequency is either ~900 MHz or ~2 GHz. At a high vehicle speed of 100 km/h, these carrier frequencies correspond to Doppler spread values of 83 and 185 Hz, respectively. The ability to track channel variation depends on how rapidly the channel changes from symbol to symbol. In all the examples given, the symbol rate is much higher than the Doppler spread, so that the channel coefficient value is highly correlated

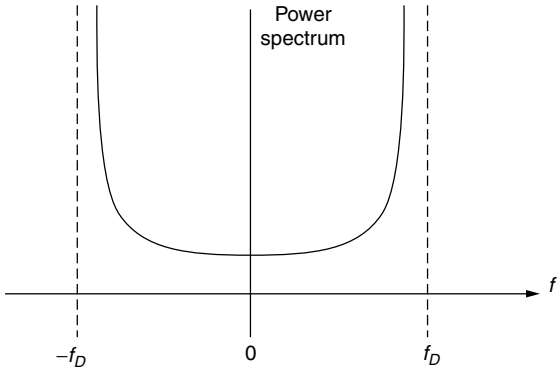


Figure 4. Spectrum of fading process.

from symbol to symbol, making channel tracking possible. Also, the Doppler spread is considered constant, because the speed of the transmitter or receiver changes slowly relative to the transmission rate. However, it is possible to model the time variation of the Doppler spread [11].

At the receiver, $y(t)$ is passed through a filter matched to the pulse shape and sampled, giving received samples

$$r_k = \int f^*(\tau)y(\tau + kT_s)\, d\tau, \quad k = 0, 1, \ldots \tag{4}$$

where $T_s = T/M = T$ is the sampling period. Because the fading varies slowly from symbol to symbol, it can be approximated as constant over the pulse shape. With this approximation, substituting (2) into (4) gives

$$\boxed{r_k = \sum_{j=0}^{J-1} h_k(j)b_{k-j} + z_k} \tag{5}$$

where $h_k(j)$ is the jth composite channel coefficient, reflecting the influence of the transmit filter, the radio channel, and the receive filter

$$h_k(j) = \sum_{\ell=0}^{L-1} c(\ell, kT)R_{ff}(jT - \ell T) \tag{6}$$

where $R_{ff}(\tau)$ is the pulse shape autocorrelation function. Note that z_k corresponds to a sequence of complex Gaussian noise samples, which may be correlated depending on the pulse shape autocorrelation function. Observe that at the receiver, there is intersymbol interference (ISI), as the received samples contain the current symbol b_k as well as interference from previous symbols. The number (J) of composite coefficients needed to accurately model r_k depends on L and on the shape of $R_{ff}(\tau)$. For the special case of root-Nyquist pulse shaping, $J = L$, $h_k(j) = c(j, kT)$, and the noise samples z_k are uncorrelated.

When developing certain channel tracking approaches, it is convenient to formulate the tracker in terms of the conjugate of $h_k(j)$, i.e., $g_k(j) = h_k^*(j)$. This gives the alternative formulation

$$\boxed{r_k = \mathbf{g}_k^H \mathbf{b}_k + z_k} \tag{7}$$

where superscript H denotes Hermitian (conjugate) transpose, $\mathbf{g}_k = [g_k(0)\ g_k(1)\ \cdots\ g_k(J-1)]^T$ is a vector of channel coefficients, $\mathbf{b}_k = [b_k, \ldots, b_{k-J+1}]^T$ is a vector of symbols, and superscript T denotes transpose.

2.2. Simple Narrowband Demodulator Example

Here, a simple coherent receiver for the case of a one-tap channel is given. Using the formulation in (5), we obtain

$$r_k = h_k(0)b_k + z_k \tag{8}$$

Assuming that the information symbol b_k is either $+1$ or -1 (BPSK), we can recover the information using

$$\hat{b}_k = \text{sign}(\text{Re}\{\hat{h}_k^*(0)r_k\}) \tag{9}$$

where $\hat{h}_k(0)$ is an estimate of the channel coefficient and $\text{Re}\{\cdot\}$ denotes the real part of a complex number. When quadrature phase shift keying (QPSK) is used, each symbol represents two bit values. One of the bit values is recovered using (9), and the other bit value is recovered using a similar expression in which the imaginary part is taken instead of the real part.

Multiplying the received value by the conjugate of the channel coefficient estimate removes the phase rotation introduced by the channel and weights the value proportional to how strong the channel coefficient is, which is important when soft information [sign operation omitted in (9)] is used in subsequent forward error correction (FEC) decoding.

2.3. Wideband System

A similar system model can be used for DS-CDMA systems. For these systems, the pulse shape $f(\tau)$ is replaced with the convolution of the chip sequence and a chip pulse shape. Basically, each symbol is represented by a sequence of N_c chips, so that $T = N_c T_c$, where T_c is the chip period. We refer to N_c as the *spreading factor*, which is typically a large integer (e.g., 64, 128) for speech applications. In the DS-CDMA examples given, the chip sequence changes each symbol period, so that the overall symbol pulse shape is time-dependent [$f(\tau)$ is replaced by $f_k(\tau)$]. Also, channel tap delays are typically modeled on the order of the chip period T_c, not the symbol period T. Thus, $\tau(\ell) = \ell T_c/M$. As in the narrowband case, we assume $M = 1$.

As in the narrowband case, the receiver correlates to the symbol pulse shape. For each symbol period k, it produces a "despread" value for each of the L channel taps:

$$r_{k,\ell} = \int f_k^*(\tau)y(\tau + \tau(\ell))\,d\tau \quad \ell = 1, \ldots, L \tag{10}$$

In practice, this involves filtering matched to the chip pulse shape followed by correlation (despreading) using the chip sequence for symbol k. Assuming the spreading factor is large enough, the contribution from adjacent symbols (ISI) can be ignored. Thus, for symbol period k, the despread value for the ℓth channel tap can be modeled as

$$r_{k,\ell} \approx b_k \sum_{\ell=0}^{L-1} c(\ell, kT)R_{f_k f_k}(kT_c - \ell T_c) + z_{k,\ell} \tag{11}$$

Typically $R_{f_k f_k}(iT_c) \approx \delta(i)$ (e.g., when N_c is large), so that

$$r_{k,\ell} \approx c(\ell, kT)b_k + z_{k,\ell} = h_k(\ell)b_k + z_{k,\ell} \tag{12}$$

Comparing (12) to (5), we see that the DS-CDMA case can be treated as L separate, one-tap channels.

The RAKE receiver [9] combines signal energy from each signal image to form a decision variable. For BPSK modulation, this gives the detected bit value

$$\hat{b}_k = \text{sign}\left(\text{Re}\left\{\sum_{\ell=1}^{L} \hat{h}_k^*(\ell, kT)r_{k,\ell}\right\}\right) \tag{13}$$

Observe that the despread values are weighted by the conjugates of the channel coefficient estimates, then added together. Thus, channel estimates are needed to combine the signal images properly.

3. MODELS FOR CHANNEL TRACKING

Channel tracking approaches are often developed from a model of the channel, such as the Rayleigh fading model described in Section 2. However, to obtain reasonable complexity, the model used can be an approximate, simpler model. In this section, such models are described.

There are basically two types of models used to develop channel trackers. The type that is used most often is a *stochastic model*, in which the channel coefficient is modeled as a random process. The Jakes model is an example of a stochastic model. These models are fairly robust, as they allow for random fluctuations in the channel coefficient.

The second type of model is a *deterministic model*, in which the channel coefficient variation in time is represented by a function with parameters. This model is useful when representing the channel variation over a limited period of time, for which the channel variation fits well to a particular functional form. It is particularly useful in predicting future coefficient values when the channel is varying rapidly but the model parameters are varying slowly.

3.1. Stochastic Models

With stochastic models, the channel coefficients are modeled as stochastic random processes. The most commonly used models can be described using the Kalman state-space model [12] given in Fig. 5. This model is fairly general and can well approximate the Jakes model given in Section 2.

With the Kalman model, the $N_s \times 1$ state vector $\mathbf{s}_k$ includes the channel coefficients. The updated state value $\mathbf{s}_{k+1}$ depends on the previous value $\mathbf{s}_k$ through a state transition matrix $\mathbf{F}$ as well as the plant noise $\mathbf{u}_k$ through a gain matrix $\mathbf{G}$. The plant noise is assumed to be a zero-mean, complex white Gaussian process with covariance $\mathbf{I}$ (the identity matrix). This leads to Gaussian channel coefficients, which is consistent with the Rayleigh fading assumption.

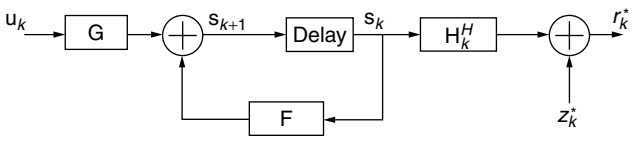


Figure 5. Kalman signal model.

The state is mapped to the output (observation) through a measurement matrix $\mathbf{H}_k$, which includes the symbol values. The output is observed in the presence of zero-mean, complex white Gaussian measurement noise z_k with mean-square value σ_z^2. The observation is defined as the conjugate of the received samples (r_k^*) so that standard expressions can be used. Mathematically, the system is described by the following *process* and *measurement* equations:

$$\mathbf{s}_{k+1} = \mathbf{F}\mathbf{s}_k + \mathbf{G}\mathbf{u}_k \quad \text{(process)} \tag{14}$$

$$r_k^* = \mathbf{H}_k^H \mathbf{s}_k + z_k^* \quad \text{(measurement)} \tag{15}$$

In general, these equations can model an autoregressive moving-average (ARMA) process [13].

One of the simplest channel models is the random-walk model [12]. With this model, the state vector is the conjugate channel coefficient vector ($\mathbf{s}_k = \mathbf{g}_k$), the measurement matrix is the symbol vector ($\mathbf{H}_k = \mathbf{b}_k$), $\mathbf{F} = \mathbf{I}$, and $\mathbf{G} = \sigma_g \mathbf{I}$, so that

$$\mathbf{g}_{k+1} = \mathbf{g}_k + \mathbf{u}_k \tag{16}$$

$$r_k^* = \mathbf{b}_k^H \mathbf{g}_k + z_k^* \tag{17}$$

Note that σ_g is a parameter related to tap strength.

A slightly more general channel model is a first-order autoregressive (AR) process (AR1) [14,15]. This process is similar to the random-walk process, except that $\mathbf{F} = \beta \mathbf{I}$, where β is a parameter between 0 and 1. The choice of β can be related to Doppler spread [14,16]. The *process* equation becomes

$$\mathbf{g}_{k+1} = \beta \mathbf{g}_k + \mathbf{u}_k \tag{18}$$

The random-walk and AR1 models are fairly simple, first-order models that are useful when the channel variation is relatively slow. Note that with these models, as with many other models, the channel coefficients corresponding to different taps are assumed to be uncorrelated (the off-diagonal elements of $\mathbf{F}$ and $\mathbf{G}$ are zero).

When channel variation is relatively fast, a second-order model is used, such as the second-order autoregressive (AR2) channel model [17,18]. Intuitively, such an approach can more accurately model the two spectral peaks that occur at plus and minus the Doppler spread (see Fig. 4). For the simple case of a one-tap channel model ($J = 1$), the model quantities become [17]

$$\mathbf{s}_k = \begin{pmatrix} g_k \\ -a_2 g_{k-1} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} -a_1 & 1 \\ -a_2 & 0 \end{pmatrix},$$
$$\mathbf{G} = \begin{pmatrix} G_1 \\ 0 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} b_k \\ 0 \end{pmatrix} \tag{19}$$

where G_1 is a model parameter related to tap strength. In Refs. 19 and 20, higher-order AR modeling is used, including adaptive estimation of the AR parameters.

Higher-order modeling can also be obtained by tracking different order derivatives of the channel coefficients. For example, the channel tap and its derivative are tracked in the second-order integrated random-walk (IRW) channel model [17,21]. For the simple case of a one-tap channel model ($J = 1$), the model quantities become [17]

$$\mathbf{s}_k = \begin{pmatrix} g_k \\ \dot{g}_k \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 0 \\ G_2 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} b_k \\ 0 \end{pmatrix} \tag{20}$$

where $\dot{g}_k$ denotes the time derivative of g_k and G_2 is a model parameter. This approach can be generalized to include higher-order derivatives [22].

3.2. Deterministic Models

The time evolution of channel coefficients can also be expressed as a deterministic function of time with parameters. Once the function's parameters have been estimated, the channel coefficient as a function of time is determined. The parameters are assumed to vary more slowly than the channel coefficients, which is helpful when the channel varies rapidly.

One model is a polynomial function of time [23,24]. For example, quadratic variation of the channel can be represented as

$$g_k = d_0 + d_1 k + d_2 k^2 \tag{21}$$

The model parameters can be estimated and tracked using a least-squares approach. Model order can be determined by the fading rate, signal-to-noise ratio (SNR), and the duration of the received data. Linear and quadratic functions are most commonly used.

Another deterministic model is the complex sinusoidal model [25]. With this model, a tap is modeled as

$$g_k = \sum_{n=1}^{N_e} A_n e^{j(2\pi f_n k T_s + \phi_n)} \tag{22}$$

where A_n, f_n, and ϕ_n are the amplitude, frequency, and phase of the nth exponential, respectively, and N_e is the number of exponentials (typically <9).

4. FILTERING APPROACHES FOR CHANNEL TRACKING

When there is a sufficiently strong pilot channel or sufficient pilot symbols, the channel can be tracked by filtering channel measurements obtained from the pilot information. The filter smooths the noisy measurements over time and works best when the channel estimate is based on future as well as past channel measurements. Specifically, for a given filter, channel estimation performance depends on the pilot information, fading channel characteristics, and noise level. Pilot information, in terms of how much energy and how often it is available, is a tradeoff between minimizing overhead and optimizing channel estimation performance. For example, with pilot symbols, how often symbols must be sent depends on how rapidly the channel is changing.

In this section, we first examine filtering approaches based on continuous measurements of the channel, which can be obtained from a pilot channel. The simple moving-average filter is examined, as well as the more advanced Wiener filter. Learning the filter weights adaptively,

using adaptive filter techniques, is also discussed. We then examine filtering approaches based on discontinuous measurements of the channel, which can be obtained from pilot symbols. Simple linear interpolation is discussed, as well as Wiener interpolation. To simplify the discussion, the example of a one-tap channel model ($J = 1$) is used throughout this section.

4.1. Estimation Using a Pilot Channel

When a continuous pilot channel is available, a sequence of *channel measurements* can be obtained, as illustrated in Fig. 6. From (8) or (12), measurements of $h_k(0)$ can be obtained by multiplying r_k by b_k^* (assuming $|b_k|^2 = 1$). To estimate the conjugate of $h_k(0)$, which is $g_k(0)$ or simply g_k, we use $r_k^* b_k$ for the channel measurements. These measurements are filtered, giving

$$\hat{g}_k = \sum_{n=N_1}^{N_2} w_{n,k}^* [r^*(k-n)b(k-n)] = \mathbf{w}_k^H \tilde{\mathbf{g}} \tag{23}$$

where $\hat{g}_k$ is the channel estimate at the kth sample position, $\mathbf{w}_k$ represents the vector filter weights as a function of time k, and $\tilde{\mathbf{g}}$ is a vector of measurements. Typically the filter "slides" across the measurements, providing a continuous sequence of channel estimates. The filter weights change slowly in time (k), due to changes in the radio environment.

The parameters N_1 and N_2 are integers ($N_2 \geq N_1$). If only the past samples are used for estimation ($N_1 > 0$), then this operation is called "prediction." Often one-step prediction is used ($N_1 = 1$). If future samples are used ($N_1 < 0$), then this is called "smoothing." In the subsequent subsections, different choices for the filter weights are given.

4.1.1. Moving-Average Filter. The moving-average or sliding-rectangular-window approach is commonly used [26–28]. The channel estimate at time k is obtained by averaging measurements from time $N_1 = k - N$ through $N_2 = k + N$ [i.e., $w_{k,n} = 1/(2N+1)$ in (23)]. The sliding-window approach implies a channel model in which the channel is constant over the averaging window period. The choice of window size is a tradeoff between tracking ability and noise suppression [26]. Thus, it depends on Doppler spread and signal-to-noise ratio (SNR) [28].

4.1.2. Wiener Filtering. With Wiener filtering, the filter weights are designed to minimize the mean-square error (MSE) between the channel estimate and the true channel coefficient [29,30]. The vector of filter weights is given by solving the Wiener–Hopf equation [12]

$$\mathbf{w}_k = \mathbf{R}_{\tilde{g}}^{-1} \mathbf{p} \tag{24}$$

where $\mathbf{R}_{\tilde{g}}$ is the correlation matrix of the measurement vector ($\mathbf{R}_{\tilde{g}} = \mathrm{E}\{\tilde{\mathbf{g}}\tilde{\mathbf{g}}^H\}$) and $\mathbf{p}$ is the correlation vector between the measurement vector and the true channel coefficient ($\mathbf{p} = \mathrm{E}\{\tilde{\mathbf{g}} g_k^*\}$). These quantities typically change slowly in time.

The expressions for $\mathbf{R}_{\tilde{g}}$ and $\mathbf{p}$ depend on the model for the channel coefficient and which channel measurements are used. Consider a simple example, in which the channel is to be estimated at time kT using measurements at times $(k-1)T$, kT, and $(k+1)T$. Also, assume that the channel follows the model given in (3) and has a mean-square value of 1. Then, the Wiener filter quantities are given by

$$\mathbf{R}_{\tilde{g}} = \begin{bmatrix} 1+\sigma_z^2 & J_0(2\pi f_D T) & J_0(2\pi f_D 2T) \\ J_0(2\pi f_D T) & 1+\sigma_z^2 & J_0(2\pi f_D T) \\ J_0(2\pi f_D 2T) & J_0(2\pi f_D T) & 1+\sigma_z^2 \end{bmatrix},$$

$$\mathbf{p} = \begin{bmatrix} J_0(2\pi f_D T) \\ 1 \\ J_0(2\pi f_D T) \end{bmatrix} \tag{25}$$

where σ_z^2 is the noise power. In this example, the Wiener filter design requires knowledge of Doppler spread (f_D) and noise power (σ_z^2). The filter design can be obtained on the basis of the worst-case expected Doppler spread value [31]. Alternatively, the Wiener quantities $\mathbf{R}_{\tilde{g}}$ and $\mathbf{p}$ can be estimated [29].

Approximations to the Wiener filter can be used. A simple approximation is to use a lowpass filter with a cutoff frequency greater than or equal to the maximum expected Doppler frequency [30,32]. In Ref. 33, the channel is approximately modeled as having constant amplitude and linearly varying phase, and other approximations are also made.

4.1.3. Adaptive Filter. Adaptive filtering approaches can be used to "learn" the filter weights [34,35]. The filter output for time k is compared to the channel measurement at time k to generate an error signal, which is used to update the filter weights. Updating approaches are given in Section 5. This approach is often used for prediction.

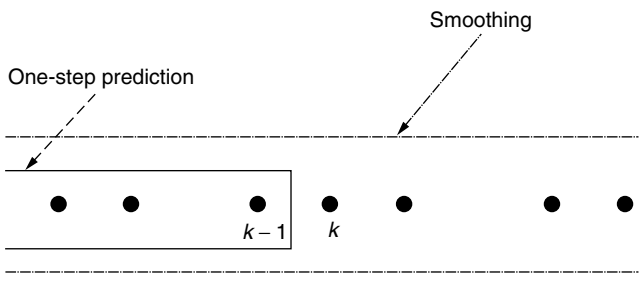


Figure 6. Prediction and smoothing of channel measurements.

4.2. Estimation Using Pilot Symbol Clusters

The channel can be estimated by periodically inserting one or more pilot symbols into the stream of data. Like the pilot channel case, channel measurements can be obtained at the pilot symbol locations. When there is a cluster of pilot symbols, it is often assumed that the channel is approximately constant over the cluster, so that these measurements can be added to give one measurement. In the case of very long pilot clusters, the cluster can be divided into smaller segments [36].

For the case of a multitap channel model, a time-invariant approach [7] can be used to obtain a channel measurement using a cluster of pilot symbols [37]. Another

approach is to run a recursive channel tracker over the pilot cluster [38].

4.2.1. Linear Interpolation. One of the simplest forms of channel estimation using pilot symbols is linear interpolation [4]. With linear interpolation, the channel estimate at a certain time period is a linear combination of the two "nearest" channel measurements. For example, suppose that there are measurements from pilot symbols at times $k = 0$ and $k = M$, denoted $\tilde{g}_0$ and $\tilde{g}_M$, respectively. Then, the channel estimate at time k, $0 < k < M$, is given by

$$\hat{g}_k = w_{0,k}\tilde{g}_0 + w_{M,k}\tilde{g}_M \tag{26}$$

where $w_{0,k}$ and $w_{M,k}$ are given as

$$w_{0,k} = \frac{M-k}{M}, \quad w_{M,k} = \frac{k}{M} \tag{27}$$

Linear interpolation can be viewed as applying a filter with symbol-spaced taps to the channel measurements, which contain zeros at the unknown data symbol points, as illustrated in Fig. 7.

Other simple interpolation filters include lowpass filters [3,38], Gaussian filters [4], and truncated Nyquist interpolation [37], as well as other filter forms [39]. Sometimes there is a tradeoff between using a simpler filter but requiring more closely spaced pilot symbols [4].

4.2.2. Wiener Interpolation. The Wiener filter described previously can also be applied to discontinuous channel measurements [36,40–42]. Continuing with the example given previously, suppose that the channel is to be estimated at time kT using pilot symbol measurements at times $(k-2)T$ and $(k+5)T$. Then, the Wiener filter quantities would be given by

$$\mathbf{R}_{\tilde{g}} = \begin{bmatrix} 1+\sigma_z^2 & J_0(2\pi f_D 7T) \\ J_0(2\pi f_D 7T) & 1+\sigma_z^2 \end{bmatrix}, \quad \mathbf{p} = \begin{bmatrix} J_0(2\pi f_D 2T) \\ J_0(2\pi f_D 5T) \end{bmatrix} \tag{28}$$

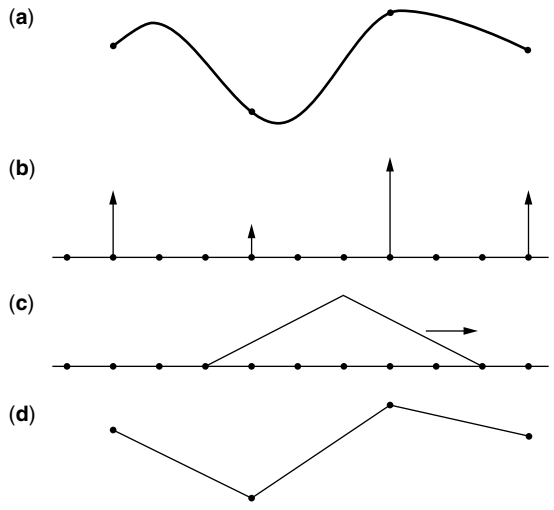


Figure 7. Linear interpolation filtering using pilot symbols: (**a**) fading channel, (**b**) measurements at pilot locations, (**c**) interpolation filter, (**d**) estimated channel.

Observe that the Wiener filter quantities will be different for time $(k+1)T$, requiring the filter weights to change. In general, Wiener interpolation is more complex than the previously described filtering approaches, but its performance is usually better, depending on the accuracy of the Wiener filter quantities.

4.3. Summary

Filtering approaches track the channel by filtering (smoothing) measurements of the channel. Simple filters, with fixed filter coefficients, work well when the channel variation is slow. When the channel variation is rapid, Wiener filtering works better. However, Wiener filtering requires knowledge of the statistics of the fading process and the statistics of the measurement noise process. The fading process statistics can be related to parameters of a channel model, such as Doppler spread and average channel coefficient power. The measurement noise process statistics are usually represented as a noise power or signal-to-noise ratio (SNR). When such information is unavailable, it is possible to "learn" good filter weights using adaptive filter techniques.

5. RECURSIVE APPROACHES FOR CHANNEL TRACKING

With recursive approaches, channel measurements are used one at a time to update channel estimates. For example, recursive approaches are often used in conjunction with data-directed tracking, in which channel estimates are used to detect a data symbol. The detected symbol value is then used to form a channel measurement for updating the channel estimates before detecting the next symbol. Data-directed tracking will be discussed in detail in Sections 6 and 7. Recursive approaches can also be used with pilot channel measurements as an alternative to filtering approaches.

To get started, an initial channel estimate value is needed. The simplest method is to set the initial value to zero, which means that there is a delay before the channel estimate becomes reliable. Another common method is to initially estimate the channel from a training sequence.

In this section, recursive channel tracking approaches are developed assuming a continuous sequence of known symbols (e.g., pilot channel). We start with two simple, related approaches, the least-mean-square (LMS) algorithm and the exponential filter. We then continue with more complex approaches, the recursive-least-squares (RLS) and Kalman filtering approaches. An approximation to the Kalman filter, the Kalman LMS approach, is then discussed.

5.1. Least-Mean-Square (LMS) Algorithm

The least-mean-square (LMS) algorithm [12,43] is one of the simplest approaches to channel tracking. LMS channel tracking is performed according to

$$\hat{\mathbf{g}}_{k+1} = \hat{\mathbf{g}}_k + \mu \mathbf{b}_k e_k^* \tag{29}$$

$$e_k = r_k - \hat{\mathbf{g}}_k^H \mathbf{b}_k \tag{30}$$

In essence, at symbol period k, an error e_k between what is received and what is modeled is formed. This error is used to update the channel coefficient estimate for the next symbol period. The step size μ is a parameter.

For a time-invariant channel, LMS channel tracking can be interpreted as an iterative, stochastic gradient approach for finding the channel coefficients that minimize the mean-square error (MSE) between the *received samples* and the model of the received samples [43]. For two noncomplex (real) channel coefficients, the MSE as a function for the two channel coefficients has a bowl shape [43]. At symbol period k, the LMS algorithm forms a noisy estimate of the slope or gradient at the place on the bowl corresponding to the current channel coefficient estimates. It then updates the taps in the direction of the negative gradient, so as to find the bottom of the bowl. Selection of the step size μ is a tradeoff between rate of convergence and how noisy the model is at convergence (misadjustment noise). For a time-varying channel, the step size μ trades tracking ability for misadjustment noise.

The LMS algorithm is a popular approach for channel tracking [44–46]. It can be derived from the Kalman filter (see Section 5.4), assuming a random-walk model for the channel coefficients [1]. The "leaky" LMS algorithm [12] [obtained by multiplying the first term on the right-hand side of (29) by leakage factor β] can be derived from the Kalman filter assuming an AR1 model [16]. The LMS structure has been further generalized [47], introducing additional parameters whose values are based on first-order or higher-order models of the channel coefficients. Conventional LMS and leaky LMS are special cases of this general, "Wiener LMS" approach.

Different step sizes can be used for different taps depending on the average tap power [16,48], which can also be derived from a Kalman filter formulation [16]. Note that the step size can be adaptive in time [49,50].

5.2. Exponential Filtering (Alpha Tracker)

Exponential filtering, also referred to as an *alpha tracker*, is commonly used with one-tap channels [51,52]. The channel estimate is updated using

$$\hat{g}_{k+1} = \alpha \hat{g}_k + (1-\alpha)\left\{r_k^* b_k\right\} \tag{31}$$

This approach can be derived from the LMS tracker by assuming that the symbols have constant magnitude ($|b_k| = 1$) [51]. It can also be interpreted as a filtering approach employing a first-order infinite-impulse-response (IIR) filter.

5.3. Recursive Least-Squares (RLS)

Recursive least-squares (RLS) [12,43] has also been applied to channel tracking, due to its rapid convergence properties. Conventional RLS channel tracking is performed according to

$$\hat{\mathbf{g}}_{k+1} = \hat{\mathbf{g}}_k + \left(\frac{\mathbf{A}_k}{\lambda + \mathbf{b}_k^H \mathbf{A}_k \mathbf{b}_k}\right) \mathbf{b}_k e_k^* \tag{32}$$

$$\mathbf{A}_{k+1} = \frac{1}{\lambda}\left(\mathbf{A}_k - \frac{\mathbf{A}_k \mathbf{b}_k \mathbf{b}_k^H \mathbf{A}_k}{\lambda + \mathbf{b}_k^H \mathbf{A}_k \mathbf{b}_k}\right) \tag{33}$$

where e_k is given in (30). Compared to LMS, RLS updates a second quantity, the matrix $\mathbf{A}_k$. This quantity is typically initialized to a diagonal matrix with large diagonal entries. The term λ is the "forgetting factor" and, like the LMS step size, determines convergence/tracking rate and misadjustment noise properties. Typically, λ is chosen to be slightly less than 1.

RLS channel tracking can be viewed as finding the set of channel coefficients that minimizes a deterministic weighted squared error between the received samples and the modeled samples [43]:

$$E = |r_k - \mathbf{g}_{k+1}^H \mathbf{b}_k|^2 + \lambda |r_{k-1} - \mathbf{g}_{k+1}^H \mathbf{b}_{k-1}|^2 + \lambda^2 |r_{k-2} - \mathbf{g}_{k+1}^H \mathbf{b}_{k-2}|^2 + \cdots \tag{34}$$

As errors in modeling past received samples are weighted less, the RLS approach tracks the channel by trying to accurately model the most recent data. Because of the exponential weighting, this form of RLS algorithm is also referred to as *exponentially windowed RLS* (EW-RLS) [53,54]. Conventional EW-RLS has been used to track wireless channels [55,56]. It has also been extended to track the channel tap and its derivative [57,58]. Alternatively, a sliding-window RLS (SW-RLS) approach can be used [54], in which the window can be tapered based on statistical knowledge of the fading channel and SNR.

RLS channel tracking can be related to Kalman filtering, based on a first-order AR process with zero plant noise [12,59]. Improved tracking approaches have been developed by considering nonzero plant noise and a time-varying state transition matrix [60].

5.4. Kalman Filtering

Kalman filtering [12,13] provides a recursive form of MMSE filtering. For the Kalman signal model given previously, the corresponding one-step prediction Kalman filter is given by

$$\hat{\mathbf{s}}_{k+1} = \mathbf{F}\hat{\mathbf{s}}_k + \mathbf{K}_k e_k^* \tag{35}$$

$$\mathbf{P}_{k+1} = \mathbf{F}\left(\mathbf{P}_k - \frac{\mathbf{P}_k \mathbf{H}_k \mathbf{H}_k^H \mathbf{P}_k}{\mathbf{H}_k^H \mathbf{P}_k \mathbf{H}_k + \sigma_z^2}\right)\mathbf{F}^H + \mathbf{G}\mathbf{G}^H \tag{36}$$

where

$$\mathbf{K}_k = \frac{\mathbf{F}\mathbf{P}_k \mathbf{H}_k}{\mathbf{H}_k^H \mathbf{P}_k \mathbf{H}_k + \sigma_z^2} \tag{37}$$

and e_k is given by (30). Note that $\mathbf{K}_k$ is a $N_s \times 1$ vector and $\mathbf{P}_k$ is a $N_s \times N_s$ matrix. Also, the Kalman filter requires knowledge of $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}_k$.

Kalman filtering has been applied to channel tracking using a variety of channel models, such as the random-walk model [1,61]. Using the random-walk expressions (16) and (17), Eqs. (35) through (37) simplify to

$$\hat{\mathbf{g}}_{k+1} = \hat{\mathbf{g}}_k + \mathbf{K}_k \mathbf{e}_k^* \tag{38}$$

$$\mathbf{P}_{k+1} = \left(\mathbf{P}_k - \frac{\mathbf{P}_k \mathbf{b}_k \mathbf{b}_k^H \mathbf{P}_k}{\mathbf{b}_k^H \mathbf{P}_k \mathbf{b}_k + \sigma_z^2}\right) + \sigma_g^2 \mathbf{I} \tag{39}$$

$$\mathbf{K}_k = \frac{\mathbf{P}_k \mathbf{b}_k}{\mathbf{b}_k^H \mathbf{P}_k \mathbf{b}_k + \sigma_z^2}. \tag{40}$$

Kalman filtering has been used with the AR1 model [14], higher-order AR models [19], and ARMA models [62]. When model parameters are unknown, extended Kalman filtering can be used to estimate the channel and the unknown parameters [11,63].

To reduce complexity, an approximate form of the Kalman filter can be used, which decouples the tracking of the different channel taps [15]. Another form of approximation is given in Section 5.5, as follows.

5.5. Kalman LMS

In Ref. 17, a series of approximations are applied to the Kalman filter to obtain a lower complexity tracking approach similar to LMS. The key approximation is to average out the effect of the time-varying symbol vector $\mathbf{b}_k$ (part of $\mathbf{H}_k$), which causes the Kalman gain $\mathbf{K}_k$ to vary with time. The resulting "Kalman LMS" (KLMS) approach was applied to two second order models: the IRW model and the AR2 model.

For a single channel tap, the KLMS tracking expressions for these two models are given by

$$\hat{\mathbf{s}}_{k+1} = \mathbf{F}\hat{\mathbf{s}}_k + \boldsymbol{\mu} b_k e_k^* \tag{41}$$

where $\boldsymbol{\mu}$ is a 2×1 vector of step sizes, and $\hat{\mathbf{s}}_k$ and $\mathbf{F}$ are defined earlier for the two models. The two elements in $\boldsymbol{\mu}$ are related by design formulas [17]. The KLMS AR2 form is particularly useful for rapid fading channels [18].

The IRW Kalman LMS form, which tracks the channel coefficient and its derivative, can alternatively be developed from a form of least-squares prediction [22]. This approach was extended to track acceleration (second derivative) as well.

5.6. Summary

When the channel varies slowly, simple approaches such as the LMS algorithm and the exponential filter work well. If rapid convergence at initialization is a concern, the RLS approach can be used. However, these three approaches are based on a first-order model of the channel coefficient. For rapid fading, a higher-order model is needed. In this case, the Kalman filter can be used, as it allows for higher-order models. To reduce complexity, the Kalman LMS approach is useful.

6. DATA-DIRECTED TRACKING, SEPARATE TRACKING

Data-directed tracking is used when there is insufficient pilot information. For systems with only an initial training sequence, the channel may vary significantly over the data portion of the received signal. For other systems, which employ a pilot channel or pilot symbols, data-directed tracking may be beneficial when the power allocated to the pilot information is low, so the channel estimates are very noisy.

In this section, we consider examples in which each channel coefficient is tracked separately. This includes narrowband systems in which there is only one channel tap as well as wideband systems. Joint tracking of multiple coefficients is addressed in Section 7.

6.1. Decision Feedback

With decision feedback, previously detected symbols are used to update the channel estimate before detecting the next symbol [51,52,64]. The detected symbol values fed back can be tentative decisions, with final decisions made using the updated channel estimates [11,22]. Decision feedback has been used with exponential filtering [52], polynomial modeling [65], linear prediction [33,66], and Kalman-based prediction [11].

Before making final symbol decisions, the channel can be estimated again, so that both channel estimation and symbol detection end up being performed twice [66,67]. In the first stage, conventional decision feedback is applied, in which channel estimates are based only on past decisions. In the second stage, channel estimates are based on first-stage decisions of future symbols as well as second-stage decisions of past symbols. The two-stage approach can be extended to multiple stages, further refining the channel estimate and data decisions [67,68]. Depending on the modulation, the first stage may not require channel estimation, as symbols can be detected noncoherently [68].

One problem with decision feedback is that decisions errors can cause the channel estimate to be phase-rotated from the true channel. For example, with BPSK modulation, decision errors correspond to a 180° rotation in the symbol values. To compensate for this, the channel estimates become rotated 180° with respect to the true channel coefficients. This is because the received signal can be equally modeled by an inverted symbol sequence and rotated channel coefficients, sometimes referred to as the *phase ambiguity problem*. Differential modulation can be used to mitigate this problem [51].

6.2. Per-Survivor Processing

Decision errors can also be mitigated by keeping multiple channel estimates. Ideally, a channel estimate should be formed for each possible sequence of symbols [64]. In practice, multiple channel estimates can be formed, corresponding to all possible values for the previous K symbols [29,69–71]. A cost function can be used to decide which channel estimate to keep. This approach can be interpreted as a form of per-survivor processing (PSP), which will be discussed in Section 7.

6.3. Data-Directed Tracking with Pilot Information

When the pilot information is weak, a combination of reference-assisted and data-directed channel tracking can be used. Similar to the case without pilot information, a multistage approach can be used [72,73]. In the first stage, a channel estimate is obtained from the pilot information and used to make tentative symbol decisions. In the second stage, these tentative decisions provide more channel measurements, which are used with the pilot information to produce a refined channel estimate and new symbol values. Note that in the first stage, it is possible to use a mixture of pilot information and decision feedback [34] or a PSP-based approach [29].

A PSP-based approach can be used with pilot information in a single-stage approach as well. Channel measurements from a weak pilot channel can be combined with

PSP-based data-directed channel measurements to form improved channel estimates for demodulation [70,71]. Alternatively, periodic pilot symbols can be used to simply constrain the symbol hypotheses of the PSP process [29,69]. This helps resolve phase ambiguity problem [69] as well as prevent the channel estimator from breaking down because of a high level of decision errors [29].

7. DATA-DIRECTED TRACKING, JOINT TRACKING

Data-directed tracking can also be used when multiple channel coefficients must be estimated together. In this situation, data symbols interfere with one another (ISI) and must be detected together using some form of equalization. Here we consider only equalization approaches that rely on channel estimates, focusing mostly on decision feedback equalization (DFE) and maximum-likelihood sequence estimation (MLSE). There is also the situation in which there may be only one channel coefficient, but the transmitter causes ISI due to the modulation (e.g., partial response pulse shaping).

First, the basic principles for DFE and MLSE are reviewed. Initial channel estimation using a training sequence is then discussed. Various data-directed tracking approaches are presented, showing the tradeoffs between tracking delay, symbol value accuracy, and complexity.

7.1. Equalization

Adaptive equalization has an extensive history [74,75]. In narrowband wireless communication systems, DFE and MLSE are two commonly used forms of equalization. To understand the basic principles of these approaches, consider the simple case of BPSK symbols ($b_k = \pm 1$) and a two-tap channel model, so that

$$r_k = g_k^*(0)b_k + g_k^*(1)b_{k-1} + z_k \tag{42}$$

The DFE, shown in Fig. 8, consists of a feedforward filter, a feedback filter, and a decision device [9]. Conceptually, the feedforward filter tries to collect all signal energy for b_k (which appears in both r_k and r_{k+1}) while suppressing intersymbol interference (ISI) from subsequent symbols (e.g., b_{k+1}). The feedback filter removes ISI from previous symbols (e.g., b_{k-1}). Notice that the feedforward filter introduces delay between when the first image of a symbol arrives and when that symbol is decided.

With MLSE, the likelihood of the received data samples, conditioned on the symbol values, is maximized [9]. The conditional loglikelihood of the kth data sample, assuming that z_k is Gaussian and uncorrelated in time, is related to the following metric or cost function:

$$M(\hat{b}_k, \hat{b}_{k-1}) = |r_k - (g_k^*(0)\hat{b}_k + g_k^*(1)\hat{b}_{k-1})|^2 \tag{43}$$

For different symbol sequence hypotheses ("paths"), this metric is accumulated, generating a path metric. Once all the data samples have been processed, the sequence corresponding to the smallest path metric determines the detected sequence. Intuitively, the metric indicates how well the model of the received data fits the actual received data.

A brute-force search of all possible sequences would require high computational complexity. However, it is possible to determine the smallest metric sequence through a process of path pruning known as the Viterbi algorithm [76]. This involves defining a set of "states" corresponding to a set of paths. Tentative decisions can be made after some delay D. The path "history" (sequence of symbol values) corresponding to the state with the best path metric after processing the kth sample can be used to determine the $(k - D)$th symbol value.

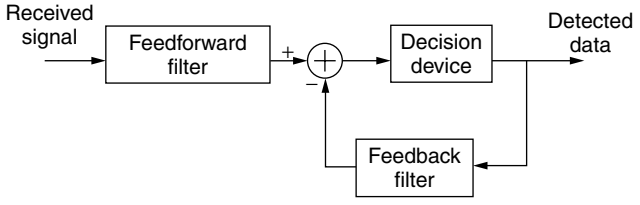


Figure 8. DFE receiver.

7.2. Initialization

A time-invariant approach [7] can be used to initially estimate the channel with a training sequence [77]. Recursive channel tracking can also be used to obtain an initial estimate [78] or to refine an estimate obtained by a time-invariant approach [18].

7.3. Decision Feedback

As in Section 6, decision feedback can be used with MLSE [44]. To obtain reliable tentative symbol decisions, a certain delay (D) is needed. To compensate for this delay, D-step channel prediction can be used, such as linear prediction [78] or a Kalman-based approach [18]. However, channel prediction becomes less reliable with larger D values. Thus, there is still a tradeoff between accuracy of symbol values and tracking delay.

There are several other issues related to decision feedback tracking and MLSE:

1. The "best" path can suddenly change, so that the detected symbol sequence does not correspond to any one path history. Regularization can be used to improve channel tracking in this situation [79].
2. As discussed in Section 6, the channel estimate can be phase rotated [80]. Bidirectional channel tracking can be used to resolve this problem [81].
3. Overmodeling the finite-impulse-response (FIR) channel can cause equalizer timing divergence after fading dips in fast-fading channels [18]. Figure 9 shows an example where a two-tap model is used when only one tap is needed. During recovery from a deep fade, the position of the nonzero tap might change (false lock or time slip). When this happens, the channel is, in fact, being tracked, but with the wrong delay. This problem can be mitigated by using

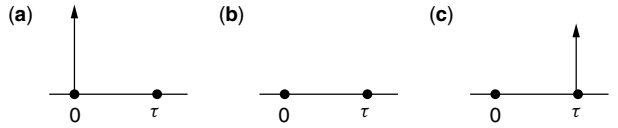


Figure 9. Time-slip problem, channel estimates (**a**) before, (**b**) during, and (**c**) after fade.

different recursive tracking step sizes for different taps [18] or by employing bidirectional tracking [82].

4. Finally, as in Section 6, a multistage approach can be used [29,83].

Decision feedback for channel estimation can also be used with DFE [14,45,48]. Similar to MLSE, delay can be addressed with prediction [14,48]. Decision feedback has also been applied to a form of linear equalization that employs channel estimates [84].

7.4. Per-Survivor Processing

For MLSE, tracking delay can be eliminated by keeping a different channel model for each state [85,86]. In the pruning process, the path that is kept also determines which channel model is updated for the new state. Such an approach is a form of *per-survivor processing* (PSP) [86]. While the channel estimate is usually updated after pruning, it is possible to update the channel estimate before pruning [87].

To reduce complexity, the number of channel models can be less than the number of states [88,89]. Conversely, to improve performance, the size of the state space can be increased [90] to create more channel models.

When the channel is estimated, Viterbi pruning does not necessarily lead to the smallest metric sequence [91]. This leads to keeping channel models for each sequence hypothesis (path) [90]. The M-algorithm [92] can be used to prune paths, keeping the M best paths [93].

8. CONCLUSION

Channel tracking is an important part of receiver design in digital wireless communication systems. Tracking approaches are based on a model of how the channel changes in time. First-order models are often used for slowly varying channels, whereas higher-order models are used for rapidly varying channels.

Reference-assisted channel estimation, using either pilot symbols or a pilot channel, is commonly used to estimate the time-varying channel impulse response. When reference information is unavailable or insufficient, data-directed tracking can be used. Obtaining reliable symbol decisions and reducing tracking delay are the main challenges. The tracking delay can be minimized by keeping multiple channel models, corresponding to different hypotheses of the data symbol values.

Currently, third-generation digital wireless communication systems are being deployed. As modulation formats for these systems require coherent reception, channel estimation will continue to be a key element in receiver design.

Acknowledgment

The authors would like to thank A. Khayrallah, K. Molnar, J. G. Proakis, and Y.-P. E. Wang for reviewing a draft of this article. The authors gratefully acknowledge the help of others in identifying references. Finally, the authors wish to thank their colleagues for many helpful discussions on channel estimation.

BIOGRAPHIES

Gregory E. Bottomley received his B.S. and his M.S. degrees from Virginia Polytechnic Institute and State University, Blacksburg, Virginia, in 1983 and 1985, respectively, and his Ph.D. degree from North Carolina State University, Raleigh, in 1989, all in electrical engineering.

From 1985 to 1987 he was with AT&T Bell Laboratories, Whippany, NJ, working in the area of sonar signal processing. In 1990, he was a visiting lecturer at North Carolina State University, Raleigh. Since 1991, he has been with Ericsson Inc., Research Triangle Park, NC, where he is currently a member of the Advanced Development and Research Department. He is listed as an inventor on over 30 patents in wireless communications and was a recipient of Ericsson's Inventor of the Year Award in 1997.

Dr. Bottomley is an associate member of Sigma Xi and a senior member of The Institute of Electrical and Electronics Engineers, Inc. (IEEE). In 1998, he was a recipient of the IEEE Eastern North Carolina Section Outstanding Engineer Award. He served as associate editor (1997–2000) and currently serves as editor for the *IEEE Transactions on Vehicular Technology*. His research interests are in baseband signal processing for wireless communications, including equalization, RAKE reception, and interference suppression.

Huseyin Arslan (eushura@rtp.ericsson.se) was born in Nazilli, Turkey, in 1968. He received a B.S. degree from Middle East Technical University, Ankara, Turkey, and M.S. and Ph.D. degrees from Southern Methodist University, Dallas, Texas, in 1992, 1994, and 1998, respectively, all in electrical engineering. Since January 1998, he has been at Ericsson research at RTP, North Carolina. His research interests are in baseband signal processing for mobile communications, including interference cancellation, channel estimation, modulation, demodulation, and equalization.

BIBLIOGRAPHY

1. L. Ljung and S. Gunnarsson, Adaptation and tracking in system identification: A survey, *Automatica* **26**: 7–21 (1990).
2. A. P. Clark, *Adaptive Detectors for Digital Modems*, Pentech, London, 1989.
3. M. L. Moher and J. H. Lodge, TCMP—a modulation and coding strategy for Rician fading channels, *IEEE J. Select. Areas Commun.* **7**: 1347–1355 (Dec. 1989).
4. S. Sampei and T. Sunaga, Rayleigh fading compensation for QAM in land mobile radio communications, *IEEE Trans. Vehic. Technol.* **42**: (May 1993).

5. K. S. Gilhousen et al., On the capacity of a cellular CDMA system, *IEEE Trans. Vehic. Technol.* **40**: 303–312 (May 1991).
6. H. W. Li and J. K. Cavers, An adaptive filtering technique for pilot-aided transmission systems, *IEEE Trans. Vehic. Technol.* **40**: 532–545 (Aug. 1991).
7. H. Arslan and G. E. Bottomley, Channel estimation in narrowband wireless communication systems, *Wireless Commun. Mobile Comput. J.* **1**: 201–219 (April/June 2001).
8. H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Radar-Sonar Signal Processing and Gaussian Signals in Noise*, Krieger, Malabar, FL, 1992.
9. J. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
10. W. C. Jakes, ed., *Microwave Mobile Communications*, IEEE Press, Piscataway, NJ, 1993.
11. A. Aghamohammadi, H. Meyr, and G. Ascheid, Adaptive synchronization and channel parameter estimation using an extended Kalman filter, *IEEE Trans. Commun.* **37**: 1212–1218 (Nov. 1989).
12. S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.
13. B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.
14. M. Stojanovic, J. G. Proakis, and J. A. Catipovic, Analysis of the impact of channel estimation errors on the performance of a decision-feedback equalizer in fading multipath channels, *IEEE Trans. Commun.* **43**: 877–885 (Feb.–April 1995).
15. M. E. Rollins and S. J. Simmons, Simplified per-survivor Kalman processing in fast frequency-selective fading channels, *IEEE Trans. Commun.* **45**: 544–552 (May 1997).
16. W. Liu, Performance of joint data and channel estimation using tap variable step size LMS for multipath fast fading channel, *Proc. IEEE Globecom Conf.*, San Francisco, CA, 1994, pp. 973–978.
17. L. Lindbom, Simplified Kalman estimation of fading mobile radio channels: High performance at LMS computational load, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Minneapolis, MN, 1993, pp. 352–355.
18. K. Jamal, G. Brismark, and B. Gudmundson, Adaptive MLSE performance on the D-AMPS 1900 channel, *IEEE Trans. Vehic. Technol.* **46**: 634–641 (Aug. 1997).
19. H. Zamiri-Jafarian and S. Pasupathy, Adaptive MLSD receiver with identification of flat fading channels, *Proc. IEEE Vehicular Technology Conf.*, Phoenix, AZ, May 4–7, 1997, pp. 695–699.
20. L. M. Davis, I. B. Collings, and R. J. Evans, Coupled estimators for equalization of fast fading mobile channels, *IEEE Trans. Commun.* **46**: 1262–1265 (Oct. 1998).
21. S. Gazor, Prediction in LMS-type adaptive algorithms for smoothly time varying environments, *IEEE Trans. Signal Process.* **47**: 1735–1739 (June 1999).
22. A. P. Clark, Adaptive channel estimator for an HF radio link, *IEEE Trans. Commun.* **37**: 918–926 (Sept. 1989).
23. W. D. Rummler, R. P. Coutts, and M. Liniger, Multipath fading channel models for microwave digital radio, *IEEE Commun. Mag.* **24**: 30–42 (Nov. 1986).
24. D. K. Borah and B. D. Hart, Frequency-selective fading channel estimation with a polynomial time-varying channel model, *IEEE Trans. Commun.* **47**: 862–873 (June 1999).
25. A. Duel-Hallen, S. Hu, and H. Hallen, Long-range prediction of fading signals, *IEEE Signal Process. Mag.* **17**: 62–75 (May 2000).
26. V.-P. Kaasila and A. Mämmelä, The adaptive rake matched filter in a time-variant two-path channel, *Proc. IEEE Int. Symposium on Personal Indoor and Mobile Radio Communications* (*PIMRC*), Boston, MA, Oct. 19–21, 1992, pp. 441–445.
27. U. Fawer, A coherent spread-spectrum diversity-receiver with AFC for multipath fading channels, *IEEE Trans. Commun.* **42**: 1300–1311 (Feb.–April 1994).
28. M. Benthin and K.-D. Kammeyer, Influence of channel estimation on the performance of a coherent DS-CDMA system, *IEEE Trans. Vehic. Technol.* **46**: 262–268 (May 1997).
29. A. N. D'Andrea, A. Diglio, and U. Mengali, Symbol-aided channel estimation with non-selective Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **44**: 41–49 (Feb. 1995).
30. F. Ling, Optimal reception, performance bound, and cut-off rate analysis of reference-assisted coherent CDMA communications with applications, *IEEE Trans. Commun.* **47**: 1583–1592 (Oct. 1999).
31. P. Schramm, Differentially coherent demodulation for differential BPSK in spread spectrum systems, *IEEE Trans. Vehic. Technol.* **48**: 1650–1656 (Sept. 1999).
32. G. Chen, X.-H. Yu, and J. Wang, Adaptive channel estimation and dedicated pilot power adjustment based on the fading-rate measurement for a pilot-aided CDMA systems, *IEEE J. Select. Areas Commun.* **19**: 132–139 (Jan. 2001).
33. L. Bin and P. Ho, Data-aided linear prediction receiver for coherent DPSK and CPM transmitted over Rayleigh flat-fading channels, *IEEE Trans. Vehic. Technol.* **48**: 1229–1236 (July 1999).
34. Y. Liu and S. D. Blostein, Identification of frequency non-selective fading channels using decision feedback and adaptive linear prediction, *IEEE Trans. Commun.* **43**: 1484–1492 (Feb.–April 1995).
35. R. J. Young and J. H. Lodge, Detection of CPM signals in fast Rayleigh flat-fading using adaptive channel estimation, *IEEE Trans. Vehic. Technol.* **44**: 338–347 (May 1995).
36. S. A. Fechtel and H. Meyr, Optimal parametric feedforward estimation of frequency-selective fading radio channels, *IEEE Trans. Commun.* **42**: 1639–1650 (Feb.–April 1994).
37. N. W. K. Lo, D. D. Falconer, and A. U. H. Sheikh, Adaptive equalization and diversity combining for mobile radio using interpolated channel estimates, *IEEE Trans. Vehic. Technol.* **40**: 636–645 (Aug. 1991).
38. A. Aghamohammadi, H. Meyr, and G. Ascheid, A new method for phase synchronization and automatic gain control of linearly modulated signals on frequency-flat fading channels, *IEEE Trans. Commun.* **39**: 25–29 (Jan. 1991).
39. H. Andoh, M. Sawahashi, and F. Adachi, Channel estimation using time multiplexed pilot symbols for coherent rake combining for DS-CDMA mobile radio, *Proc. IEEE Int. Symposium on Personal Indoor and Mobile Radio Communications* (*PIMRC*), Helsinki, Finland, Sept. 1–4 1997, pp. 954–958.
40. J. K. Cavers, An analysis of pilot symbol assisted modulation for Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **40**: 686–693 (Nov. 1991).

41. W.-Y. Kuo and M. P. Fitz, Designs for pilot-symbol-assisted burst-mode communications with fading and frequency uncertainty, *Int. J. Wireless Inform. Networks* **1**: 239–252 (1994).
42. C. D'Amours, M. Moher, and A. Yongaçoğlu, Comparison of pilot symbol-assisted and differentially detected BPSK for DS-CDMA systems employing RAKE receivers in Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **47**: 1258–1267 (Nov. 1998).
43. S. T. Alexander, *Adaptive Signal Processing*, Springer-Verlang, New York, 1986.
44. F. R. Magee and J. G. Proakis, Adaptive maximum-likelihood sequence estimation for digital signaling in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 120–124 (Jan. 1973).
45. P. K. Shukla and L. F. Turner, Channel-estimation-based adaptive DFE for fading multipath radio channels, *IEE Proc.-I Communications, Speech and Vision* **138**: 525–543 (1991).
46. M.-C. Chiu and C.-C. Chao, Analysis of LMS-adaptive MLSE equalization on multipath fading channels, *IEEE Trans. Commun.* **44**: 1684–1692 (Dec. 1996).
47. L. Lindbom, M. Sternad, and A. Ahlén, Tracking of time-varying mobile radio channels—Part I: The Wiener LMS algorithm, *IEEE Trans. Commun.* **49**: 2207–2217 (Dec. 2001).
48. S. A. Fechtel and H. Meyr, An investigation of channel estimation and equalization techniques for moderately rapid HF-channels, *Proc. IEEE Int. Conf. Communications*, Denver, CO, June 23–26 1991, pp. 768–772.
49. H. Shiino, N. Yamaguchi, and Y. Shoji, Performance of an adaptive maximum-likelihood receiver for fast fading multipath channel, *Proc. IEEE Vehicular Technology Conf.*, Denver, CO, May 1992, pp. 380–383.
50. S. Denno and Y. Saito, Orthogonal-transformed variable-gain least mean squares (OVLMS) algorithm for fractional tap-spaced adaptive MLSE equalizers, *IEEE Trans. Commun.* **47**: 1151–1160 (Aug. 1999).
51. K. Pahlavan and J. W. Matthews, Performance of adaptive matched filter receivers over fading multipath channels, *IEEE Trans. Commun.* **38**: 2106–2113 (Dec. 1990).
52. G. J. R. Povey, P. M. Grant, and R. D. Pringle, A decision-directed spread-spectrum rake receiver for fast-fading mobile channels, *IEEE Trans. Vehic. Technol.* **45**: 491–502 (Aug. 1996).
53. E. Eleftheriou and D. D. Falconer, Tracking properties and steady state performance of RLS adaptive filter algorithms, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**: 1097–1110 (Oct. 1986).
54. J. Lin, J. G. Proakis, F. Ling, and H. Lev-Ari, Optimal tracking of time-varying channels: A frequency domain approach for known and new algorithms, *IEEE J. Select. Areas Commun.* **13**: 142–154 (Jan. 1995).
55. P. Newson and B. Mulgrew, Adaptive channel identification and equalization for GSM European digital mobile radio, *Proc. IEEE Int. Conf. Communications*, Denver, CO, June 23–26 1991, pp. 23–27.
56. H.-N. Lee and G. J. Pottie, Fast adaptive equalization/diversity combining for time-varying dispersive channels, *IEEE Trans. Commun.* **46**: 1146–1162 (Sept. 1998).
57. A P. Clark and S. Hariharan, Efficient estimators for an HF radio link, *IEEE Trans. Commun.* **38**: 1173–1180 (Aug. 1990).
58. N. Zhou and N. Holte, Least squares channel estimation for a channel with fast time variations, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1992, pp. 165–168.
59. A. H. Sayed and T. Kailath, A state-space approach to adaptive RLS filtering, *IEEE Signal Process. Mag.* **11**: 18–60 (July 1994).
60. S. Haykin et al., Adaptive tracking of linear time-variant systems by extended RLS algorithms, *IEEE Trans. Signal Process.* **45**: 1118–1128 (May 1997).
61. G. E. Bottomley and K. J. Molnar, Adaptive channel estimation for multichannel MLSE receivers, *IEEE Commun. Lett.* **3**: 40–42 (Feb. 1999).
62. Q. Dai and E. Shwedyk, Detection of bandlimited signals over frequency selective Rayleigh fading channel, *IEEE Trans. Commun.* **42**: 941–950 (Feb.–April 1994).
63. R. A. Iltis and A. W. Fuxjaeger, A digital DS spread-spectrum receiver with joint channel and Doppler shift estimation, *IEEE Trans. Commun.* **39**: 1255–1267 (Aug. 1991).
64. R. Haeb and H. Meyr, A systematic approach to carrier recovery and detection of digitally phase modulated signals on fading channels, *IEEE Trans. Commun.* **37**: 748–754 (July 1989).
65. Y. Sanada, A. Kajiwara, and M. Nakagawa, Adaptive rake system for mobile communications, *Proc. IEEE Int. Conf. Selected Topics in Wireless Communication* (*ICWC'92*), Vancouver, BC, Canada, 1992, pp. 227–230.
66. G. Auer, G. J. R. Povey, and D. I. Laurenson, Mobile channel estimation for decision directed RAKE receivers operating in fast fading radio channels, *Proc. IEEE Int. Symp. Spread Spectrum Techniques and Applications* (*ISSSTA*), Sun City, South Africa, Sept. 2–4, 1998, pp. 576–579.
67. P. Y. Kam, Optimal detection of digital data over the non-selective Rayleigh fading channel with diversity reception, *IEEE Trans. Commun.* **39**: 214–219 (Feb. 1991).
68. B. H. Park, K. J. Kim, S.-Y. Kwon, and K. C. Whang, Multi-stage decision-directed channel estimation scheme for DS-CDMA system with M-ary orthogonal signaling, *IEEE Trans. Vehic. Technol.* **49**: 43–49 (Jan. 2000).
69. G. M. Vitetta and D. P. Taylor, Maximum likelihood decoding of uncoded and coded PSK signal sequences transmitted over Rayleigh flat-fading channels, *IEEE Trans. Commun.* **43**: 2750–2758 (Nov. 1995).
70. J. Choi, Multipath CDMA channel estimation by jointly utilising pilot and traffic channels, *IEE Proc. Commun.* **146**: 312–318 (Oct. 1999).
71. S.-C. Hong, J.-S. Joo, and Y. H. Lee, Per-survivor processing sequence detection for DS/CDMA systems with pilot and traffic channels, *IEEE Commun. Lett.* **5**: 346–348 (Aug. 2001).
72. G. T. Irvine and P. J. McLane, Symbol-aided plus decision-directed reception for PSK/TCM modulation on shadowed mobile satellite fading channels, *IEEE J. Select. Areas Commun.* **10**: 1289–1299 (Oct. 1992).
73. S. Min and K. B. Lee, Pilot and traffic based channel estimation for DS/CDMA systems, *IEE Electron. Lett.* **34**: 1070–1071 (May 1998).
74. R. W. Lucky, A survey of the communication theory literature: 1968–1973, *IEEE Trans. Inform. Theory* **19**: 725–739 (Nov. 1973).
75. J. G. Proakis, Adaptive equalization for TDMA digital mobile radio, *IEEE Trans. Vehic. Technol.* **40**: 333–341 (May 1991).

76. G. D. Forney, The Viterbi algorithm, *Proc. IEEE* **61**: 268–277 (March 1973).

77. R. D'Avella, L. Moreno, and M. Sant'Agostino, An adaptive MLSE receiver for TDMA digital mobile radio, *IEEE J. Select. Areas Commun.* **7**: 122–129 (Jan. 1989).

78. E. Dahlman, New adaptive Viterbi detector for fast-fading mobile-radio channels, *IEE Electron. Lett.* **26**: (Sept. 1990).

79. M. Martone, Optimally regularized channel tracking techniques for sequence estimation based on cross-validated subspace signal processing, *IEEE Trans. Commun.* **48**: 95–105 (Jan. 2000).

80. P. K. Shukla and L. F. Turner, Examination of an adaptive DFE and MLSE/near-MLSE for fading multipath radio channels, *IEE Proc.-I Communications, Speech and Vision* **139**: 418–428 (Aug. 1992).

81. H. Arslan, R. Ramesh, and A. Mostafa, Interpolation and channel tracking based receivers for coherent Mary-PSK modulations, *Proc. IEEE Vehicular Technology Conf.*, Houston, Tex, May 1999, pp. 2194–2199.

82. Y.-J. Liu, M. Wallace, and J. W. Ketchum, A soft-output bidirectional decision feedback equalization technique for TDMA cellular radio, *IEEE J. Select. Areas Commun.* **11**: 1034–1045 (Sept. 1993).

83. D. K. Borah and B. D. Hart, Receiver structures for time-varying frequency-selective fading channels, *IEEE J. Select. Areas Commun.* **17**: 1863–1875 (Nov. 1999).

84. P. Butler and A. Cantoni, Noniterative automatic equalization, *IEEE Trans. Commun.* **COM-23**: 621–633 (June 1975).

85. H. Kubo, K. Murakami, and T. Fujino, An adaptive maximum-likelihood sequence estimator for fast time-varying intersymbol interference channels, *IEEE Trans. Commun.* **42**: 1872–1880 (Feb.–April 1994).

86. R. Raheli, A. Polydoros, and C.-K. Tzou, Per-survivor processing: A general approach to MLSE in uncertain environments, *IEEE Trans. Commun.* **43**: 354–364 (Feb.–April 1995).

87. H. Zamiri-Jafarian and S. Pasupathy, Adaptive MLSDE using the EM algorithm, *IEEE Trans. Commun.* **47**: 1181–1193 (Aug. 1999).

88. R. Raheli, G. Marino, and P. Castoldi, Per-survivor processing and tentative decisions: What is in between?, *IEEE Trans. Commun.* **44**: 127–129 (Feb. 1996).

89. M. J. Bradley and P. Mars, Application of multiple channel estimators in MLSE detectors for fast time-varying and frequency selective channels, *IEE Electron. Lett.* **32**: 620–621 (March 1996).

90. N. Seshadri, Joint data and channel estimation using blind trellis search techniques, *IEEE Trans. Commun.* **42**: 1000–1011 (Feb.–April 1994).

91. K. M. Chugg, The condition for the applicability of the Viterbi algorithm with implications for fading channel MLSD, *IEEE Trans. Commun.* **46**: 1112–1116 (Sept. 1998).

92. J. B. Anderson and S. Mohan, Sequential coding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**: 169–176 (Feb. 1984).

93. P. Castoldi, R. Raheli, and G. Marino, Efficient trellis search algorithms for adaptive MLSE on fast Rayleigh fading channels, *Proc. IEEE Globecom Conf.*, San Francisco, CA, Nov. 1994, pp. 196–200.

CHAOS IN COMMUNICATIONS

KUNG YAO
CHI-CHUNG CHEN
University of California
Los Angeles, California

1. INTRODUCTION

In 1963, Lorenz used a digital computer to study the numerical solutions of nonlinear dynamical systems modeling convection in the atmosphere. He found that even a very small difference in the initial conditions can lead to solutions that can grow rapidly apart with time [1]. The deterministic but unpredictable behaviors of certain classes of nonlinear dynamical systems have been called *chaos*. Later, Lorenz presented a talk with the title "Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" Thus, the essence of the sensitivity of initial conditions to the long-term solutions of these chaotic systems has been colorfully called the "butterfly effect." Since that time, chaos has become a well-developed branch of mathematics, with applications to physics, biology, medicine, engineering, economics, and other fields. Chaos was also shown to be related to the work of fractal by Mandelbrot [2]. Researchers in nonlinear circuits and systems have found large numbers of fairly simple nonlinearities that can induce quite complicated chaotic solutions [3]. Since the early 1960s, thousands of papers and hundreds of books have been written on various aspects of chaos. Since then, a small-scale intellectual industry has been formed in exploiting chaos. Gleick's book, *Chaos: The Amazing Science of the Unpredictable,* which introduced chaos to the general public, is extremely readable, and was a bestseller when it was published [4]. Quoting Gleick's book (pp. 5–6), "The most passionate advocates of this new science go so far as to say that twentieth-century science will be remembered for just three things: relativity, quantum mechanics, and chaos." Chaos, they contend, has become the century's third great revolution in the physical sciences. Like the first two revolutions, chaos cuts away at the tenets of Newton's physics. As one physicist put it: "Relativity eliminated the Newtonian illusion of absolute space and time; quantum mechanics eliminated the Newtonian dread of a controllable measurement process; and chaos eliminates the Laplacian fantasy of "deterministic predictability." Of the three, the revolution in chaos applies to the universe as we see and touch, to objects at human scale. It remains to be seen whether the long-range impacts of chaos both to theory and applications are as profound as some advocates have claimed. In any case, for communication engineers not familiar with chaos, some useful books include: a very readable account of basic chaotic concepts by Williams [5], a medium-level graduate text on chaos by Hilborn [6], and an advanced treatise on stochastic aspects of dynamics and chaos by Lasota and Mackey [7]. Various introductory surveys, overviews, and special issue papers on chaotic communications have appeared elsewhere in the literature [3,8–14].

In the early 1990s, Pecora and Carroll [15] showed that despite the broadband nature of the spectrum, noiselike

behavior, and sensitivity to initial conditions inherented in chaotic systems, two chaotic systems can be synchronized. This rather remarkable discovery has caused much interest and created the research field of "chaotic communication." Many papers motivated by this work have been published since the early 1990s. The drive–response synchronization configuration that Pecora and Carroll proposed is shown in Fig. 1. We will use the Lorenz system example unless mentioned otherwise. If a chaotic system can be decomposed into two subsystems, a drive system x and a conditionally stable response system $\{y, z\}$ as shown in this example, then the identical chaotic system at the receiver can be synchronized when driven with a common signal. The reason for this is simple. In the absence of noise, the output signals y_r and z_r will follow the signals y and z since it is a stable subsystem. For more discussion on chaotic synchronization, see the paper by Pecora et al. [16].

On the basis of the self-synchronization properties of chaotic systems, various chaotic communication systems using chaos as carrier have been proposed. We summarize many of these systems in Section 2. Specifically, in Section 2.1, we deal with the chaotic masking modulation/demodulation scheme of Cuomo et al. [17], which was one of the earliest proposed methods to perform chaotic communication. In order to circumvent an inadequacy of chaotic masking modulation/demodulation due to the small amplitude of the information signal, the dynamical feedback modulation (DFM) scheme of Milanovic and Zaghloul [18] was proposed and treated in Section 2.2. In Section 2.3, we consider the chaotic switching modulation (CSM) scheme of Kocarev et al. [19]. In all three of these schemes, since the communication system performances strongly depend on the synchronization capability of the chaotic systems, the robustness ability of self-synchronization to the white noise needs to be explored [17,20]. However, it is generally difficult to obtain the analytic solutions to the nonlinear stochastic systems, in particular, the chaotic systems. The use of Monte Carlo (MC) simulation is a standard method for the evaluation of complicated communication systems over noisy channels. However, because of the nonlinear stochastic dynamical system modeling of these chaotic communication systems, standard deterministic MC simulation method turns out to be inadequate. In Section 2.4, we show the need to use Ito–Stratonovich stochastic integrations for the performance evaluation of continuous-time chaotic communication systems. Specifically, we show that the use of conventional Runge–Kutta numerical integration method can yield incorrect results. Furthermore, the precise evaluation of the performance of many of these chaotic communication systems showed the required SNRs (signal-to-noise ratios) to achieve nominal acceptable error probabilities are significantly higher than those conventional (nonchaotic) communication systems.

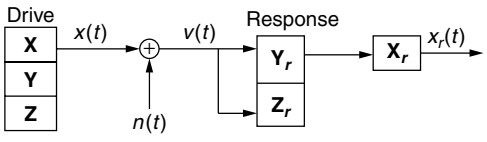


Figure 1. Drive–response self-synchronization scheme.

In order to mitigate the poor system performance of the chaotic modulation schemes due to the ill effect of the self-synchronizing aspects of these schemes, Dedieu et al. [21] first proposed the chaos shift-keyed modulation(CSK), then Kennedy and Kolumban [22–24] proposed the differential shift-keyed modulation (DCSK), and later the frequency modulation DCSK (FM-DCSK) schemes [25,26]. In Section 2.5, we discuss these different forms of CSK schemes and show in particular the FM-DCSK scheme to be competitive to conventional digital modulation schemes in performance. This is to be contrasted to earlier chaotic modulation schemes, which are intellectually interesting but not practical for implementation on a real communication system.

The symbolic dynamic model is one tool that can be used to analyze a complex chaotic dynamical system. Based on the symbolic dynamics of a chaotic system and using chaos control technique, Hayes et al. [27,28] proposed that the information message can be embedded into the chaotic dynamics for transmission to the receiver. This approach is considered in Section 3. Since it is usually difficult to estimate a chaotic signal, Papadopoulos and Wornell [29] developed a maximum-likelihood estimator for the tent map. By setting the information data into the initial condition, the information data can be encoded using a chaotic dynamical system, and can be retrieved by estimating the initial condition by Chen and Wornell [30]. These issues are considered in Section 4.

Instead of transmitting the chaotic waveform, a chaotic pulse position modulation (CPPM) scheme was proposed [31]. This proposed system is similar to an ultra-wide-bandwidth impulse radio [32] that offers a very promising communication platform, especially in a severe multipath environment. A pulse position method is used to modulate the binary information onto the carrier. The separation between the adjacent pulses is chaotic because of the dynamical system with irregular behavior. These methods are discussed in Section 5.

Since the 1960s, first for military, then later for commercial applications, code-division multiple access (CDMA) communication techniques have been used extensively. Early special issues on CDMA appeared [33,34], but extensive paper publications and books appeared later. More recently, the use of CDMA for cellular telephone communication has provided explosive worldwide interests in CDMA systems. System performance of a direct-sequence CDMA (DSCDMA) system critically depends on the auto- and cross-correlation properties of the *spreading sequences*. The chaotic signals have low non-zero-shift autocorrelation and all cross-correlation properties due to the intrinsic broad spectrum and sensitivity to initial conditions. The sequences generated by a logistic map can also be used for a DSCDMA communication system as first proposed [49]. The use of the correlation function characteristics of some specific chaotic sequences and its comparison to m sequences/Gold sequences have been discussed [50,51,53,54,56,57,59–61]. Such optimum sequences derived and generated on chaos-based concepts can support about 15% more users than previously known sequences generated by deterministic methods. These issues are discussed in Section 6. Finally, two

distinct applications of chaos-related stochastic processes to optical communication based on chaos in semiconductor lasers [65] and modeling of radar and radio propagation effects [66,74] are considered in Section 7.

2. COMMUNICATIONS USING CHAOS AS CARRIER

The use of modulating an aperiodic chaotic waveform, instead of a periodic sinusoidal signal, for carrying information messages has been proposed, in particular, chaotic masking, dynamical feedback, chaotic switching, chaos shift keying, and inverse approach modulations [10,17–19,21,22,35,36]. In this section, we summarize several of these chaotic modulation schemes and the impact of self-synchronization in the demodulation process.

2.1. Chaotic Masking Modulation

The basic idea of a chaotic masking modulation (CMM) scheme is based on chaotic signal masking and recovery. As shown in Fig. 2, we add a noiselike chaotic signal to the information signal at the transmitter, and at the receiver the masking signal is removed. The received signal, consisting of masking and information signals, is used to regenerate the masking signal at the receiver. The masking signal is then subtracted from the received signal to retrieve the information signal. The regeneration of the masking signal can be done by synchronizing the receiver chaotic system with the transmitter chaotic system. This communication system could be used for analog and digital information data. Cuomo et al. [17] have built a Lorenz system circuit and demonstrated the performance of a CMM system with a segment of speech from the sentence "He has the bluest eyes." The communication system performance truly relies on the synchronization ability of chaotic system. The masking properties of this scheme work only when the amplitudes of the information signals are much smaller than the masking chaotic signals.

2.2. Dynamical Feedback Modulation

To avoid the small-amplitude restriction of the information signal, another modulation scheme, called *dynamical feedback modulation* (DFM), has been proposed by Milanovic and Zaghloul [18] and is shown in Fig. 3. The basic idea is to feed back the information signal into the chaotic transmitter in order to ensure the identical input signal for the chaotic transmitter and the receiver. Specifically, the transmitted signal $v(t) = x(t) + m(t)$, consisting of the information signal $m(t)$ and the chaotic signal $x(t)$, is communicated to the receiver which is identical to the chaotic

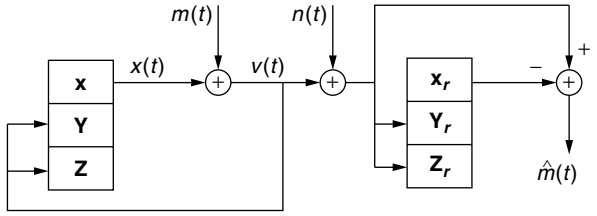


Figure 3. Dynamical feedback modulation.

transmitter. Since the reconstructed signal $x_r(t)$ will be identical to $x(t)$ in the absence of noise $n(t)$, the information signal $m(t)$ can be decoded from the received signal by using $\hat{m}(t) = x(t) + m(t) - x_r(t)$.

This *analog* communication technique can be also applied to binary data communication by setting $m(t) = A$ if binary information data are one, and $m(t) = -A$ if binary data are zero. The sufficient statistic η_d of detection is the average of the error signal at the receiver after discarding a transient period before synchronization, and is given by

$$\eta_d = \frac{1}{T}\int_{t_0}^{t_0+T} e_d(t)\,dt \tag{1}$$

where the error signal $e_d(t)$ is defined by

$$e_d(t) = v(t) + n(t) - x_r(t) \tag{2}$$

Since the feedback information will affect the chaotic property, the information level A should be scaled carefully to make the transmitter still chaotic to maintain the desired communication security.

2.3. Chaotic Switching Modulation

In contrast to DFM, the chaotic switching modulation (CSM) communication system, as illustrated in Fig. 4, does not suffer the above scaling problem as discussed by Kocarev et al. [19]. The basic idea is to encode the binary data $m(t)$ with different chaotic attractors by modulating the transmitter parameters and then transmitting the chaotic drive signal $x_m(t)$. At the receiver, the parameter modulation will produce a synchronization error between the received drive signal and the regenerated drive signal with an error amplitude that depends on the modulation. Using the synchronization error, the binary data can be detected.

In Fig. 4, the parameter b of the transmitter is modulated by the binary data $m(t)$. Assume that the communication link is an AWGN channel; the synchronization

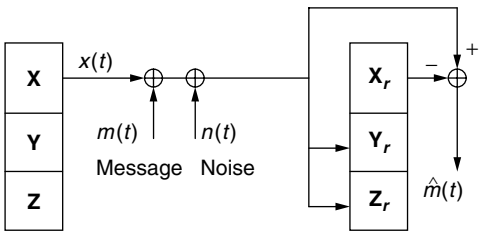


Figure 2. Chaotic masking modulation.

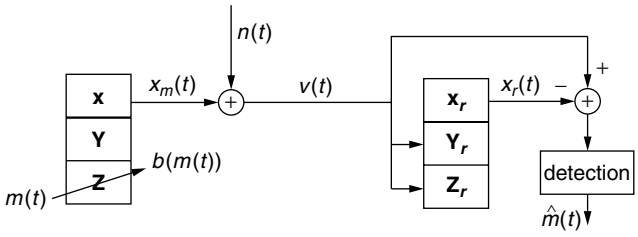


Figure 4. Chaotic switching modulation.

error $e_s(t)$ is defined in Eq. (3) as the difference between the noisy received signal $r(t) = x_m(t) + n(t)$ and the regenerated signal $x_r(t)$ at the receiver, and is given by

$$e_s(t) = r(t) - x_r(t) \tag{3}$$

For this binary hypothesis problem, the sufficient statistic η_s is the squared mean of the synchronization error after discarding some transient data and is defined by

$$\eta_s = \frac{1}{T}\int_{t_0}^{t_0+T} e_s^2(t)\,dt \tag{4}$$

Synchronization of the chaos signal is a common feature of the above three chaotic modulation–demodulation schemes, and system performance depends crucially on this synchronization capability. When the channel condition is so poor that it is impossible to achieve chaotic synchronization, different chaotic modulation techniques for digital communication, based on variations of chaos shift keying modulation (CSK) (see Fig. 5, have been introduced. The details of this approach will be discussed in Section 2.5.

2.4. Numerical Algorithm and Performance Evaluation of Chaotic Communications Based on Stochastic Calculus

Because of inherent nonlinearity of chaotic systems, the analytic performance evaluation of a chaotic communication system using chaos as a carrier is in general very difficult, and thus a numerical simulation approach is needed as shown by Chen and Yao [20]. It is known that commercial numerical computational packages using the standard Euler or Runge–Kutta (RK) algorithm designed for a deterministic differential equation to approximate the solution of a nonlinear stochastic differential equation (SDE) will incur significant errors [38]. This is particularly true for a nonlinear chaotic dynamical system modeling the transmitter inputting into an AWGN channel and followed by another nonlinear chaotic dynamical system.

We use the stochastic calculus approach to perform the integration algorithm for the sample functions of nonlinear dynamic systems excited by the stochastic white noise. Depending on the precise interpretation of the white noise, there are two different solutions to the SDE based on the Stratonovich or Ito integral [40–42]. Using the conversion between them, a correct numerical integration algorithm in the Ito sense is introduced. With this algorithm, the correct error probability of the robust self-synchronization

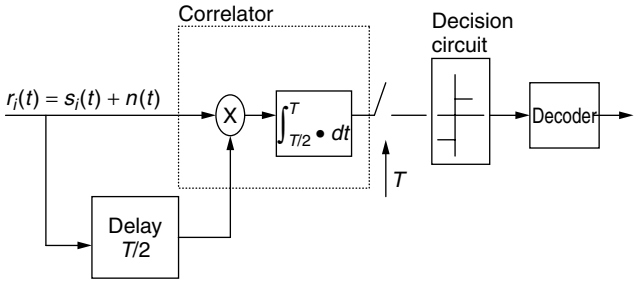


Figure 5. Differential chaos shift keying receiver.

Lorenz communication system with AWGN perturbation can be evaluated numerically. In the following, we present the numerical problem when evaluating the synchronization ability of the Lorenz system. Details regarding the error probabilities of CSM and DFM systems are given in Ref. 20.

2.4.1. Problem Description. A modified Lorenz system [17] is given by

$$\begin{aligned}\frac{dx}{d\tau} &= \sigma(y(t) - x(t))\\ \frac{dy}{d\tau} &= rx(t) - y(t) - 20x(t)z(t)\\ \frac{dz}{d\tau} &= 5x(t)y(t) - bz(t)\end{aligned} \tag{5}$$

where σ, r, and b are system parameters, and $\tau = t/K$, in which K is a timescaling factor. To characterize the robust ability of synchronization to white noise, the modified Lorenz system can be interpreted as the drive system, the signal $v(t)$ is the received waveform at the response system as defined by

$$\begin{aligned}\frac{dx_r}{d\tau} &= \sigma(y_r(t) - x_r(t))\\ \frac{dy_r}{d\tau} &= rv(t) - y_r(t) - 20v(t)z_r(t)\\ \frac{dz_r}{d\tau} &= 5v(t)y_r(t) - bz_r(t)\end{aligned} \tag{6}$$

where $v(t) = x(t) + n(t)$, and $n(t)$ is white Gaussian noise with zero mean and power spectrum density σ_n^2. The chosen coefficients are $\sigma = 16$, $r = 45.6$, and $b = 4$.

We define the vector $\mathbf{X} = [x, y, z, x_r, y_r, z_r]^T$. The entire system composed of the drive subsystem and the response subsystem can be viewed as a nonlinear system with an external white-noise input, and has the following standard form

$$\dot{\mathbf{X}}_i = f_i(\mathbf{X}) + g_i(\mathbf{X})n(t),\ i = 1, 2, \ldots, 6 \tag{7}$$

where f_1, f_2, f_3 are the modified Lorenz system equations, and

$$\begin{aligned}g_1 &= g_2 = g_3 = g_4 = 0\\ f_4 &= \sigma(y_r - x_r)\\ g_5 &= r - 20z_r\\ f_5 &= rx - y_r - 20xz_r\\ g_6 &= 5y_r\\ f_6 &= 5xy_r - bz_r\end{aligned} \tag{8}$$

and $n(t)$ is the WGN with zero mean and unity variance. The evolution of (7) is then given by

$$\mathbf{X}_i(t_0 + h) = \mathbf{X}_i(t_0) + \int_{t_0}^{t_0+h} f_i(\mathbf{X})\,dt + \int_{t_0}^{t_0+h} g_i(\mathbf{X})n(t)\,dt \tag{9}$$

A commonly made mistake is to treat the third term of (9) using the deterministic ordinary calculus method and apply the standard Euler or RK integration algorithm. Thus, assuming $g_i(\mathbf{X})$ is a smooth function, the integration result can be approximated by

$$\int_{t_0}^{t_0+h} g_i(\mathbf{X})n(t)\,dt \approx g_i(\mathbf{X}(t_0))Y_1 h \tag{10}$$

where Y_1 is a Gaussian random variable with zero mean and unity variance.

In order to illustrate this issue clearly, we use the preceding algorithm to characterize the robust self-synchronization ability of a Lorenz system by numerical computation as considered in Ref. 17. The simulation results are shown in Fig. 6 with the middle dashed curve, which is consistent with that in Ref. 17. The definitions of the input SNR and output SNR quantities in Fig. 6 are defined by input $\text{SNR} = 10\log_{10}(\sigma_x^2/\sigma_n^2)$, and output $\text{SNR} = 10\log_{10}(\sigma_x^2/\sigma_e^2)$, where σ_x^2 is the power of transmitted signal $x(t)$, and σ_e^2 is the power of the synchronization error $e(t) = x(t) - x_r(t)$. Clearly, we note the output SNR varies with the integration step size using the standard Euler/RK integration algorithm. Furthermore, the output SNR decreases by 3 dB as the step size is doubled. This is not a reasonable consequence for a given system which is excited by a stationary external white noise.

2.4.2. Numerical Algorithm for SDE. For the above-described chaotic system, the corresponding SDE in the sense of Ito is

$$dX_i = f_i(\mathbf{X})\,dt + g_i(\mathbf{X})\,dw(t), \mathbf{X}(t_0) = \mathbf{X}_0 \tag{11}$$

where $w(t)$ is the one-dimensional Wiener process or Brownian motion, and $\mathbf{X}_0$ is the initial conditions.

Mannella and Palleschi [39] have derived an accurate integration algorithm in the sense of Stratonovich [40], which treats the white noise as the limiting behavior of band-limited white noise, and the approximate results are summarized by the following algorithm

$$X_i(h) - X_i(0) = \delta X_i^{1/2} + \delta X_i^1 + \delta X_i^{3/2} + \delta X_i^2 + \cdots \tag{12}$$

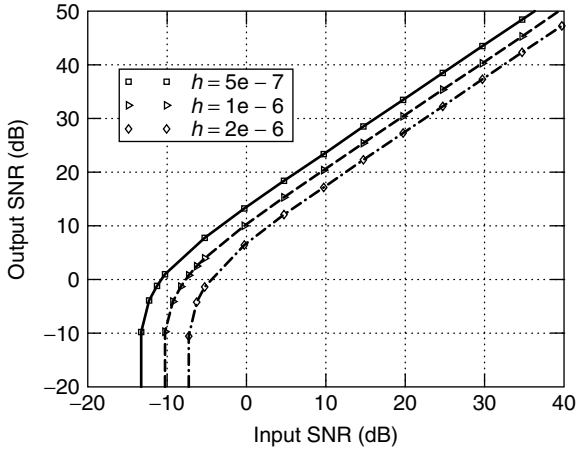


Figure 6. Robust ability for incorrect algorithm simulation with $K = 2505$.

where h is the integration step size, and $\delta X_i^{1/2} = g_i \int_0^h dw(t) = \sqrt{h} g_i Y_1$, where Y_1 is a Gaussian random variable with zero mean and unity variance, while the remaining terms are given in Eq. (6) of Ref. 39.

According to Stratonovich [40–42], the integral in the sense of Stratonovich can be converted into an Ito integral by adding one correction term. That is, if the SDE is modified and is given in the sense of Stratonovich as

$$dX_i = \left[f_i(\mathbf{X}) - \frac{1}{2}\sum_J \frac{\partial g_i(\mathbf{X})}{\partial \mathbf{X}_J} g_J(\mathbf{X}) \right] dt + g_i(\mathbf{X})\,dw(t) \tag{13}$$

the system evolution of (13) by using the above numerical algorithm is statistically equivalent to the evolution of (11) in the sense of Ito [41,42], which is desired here because the stochastic term $n(t)$ is modeled as a true white noise. We use this algorithm to resimulate the robust self-synchronization ability for white noise, and the simulation results are shown Fig. 7. Now, simulation results are consistent using different integration step sizes, which is necessary for a valid system modeling. [20] also provided the error probabilities of CSM and DFM communication systems using the appropriate stochastic integration algorithm as described above (see Fig. 8).

2.5. CSK, DCSK, and FM-DCSK

As seen in Fig. 8, the extremely poor performances of antipodal DFM and CSM systems motivate a fundamental consideration of the use of chaotic waveforms for digital communications. Since the chaotic waveform used in a coherent receiver, obtained by self-synchronization, is known to be extremely sensitive to channel noise, one approach to improve the system performance is to consider the use of different versions of CSK modulations. For an antipodal CSK transmitter, let $x(t)$ be a chaotic waveform and the modulated bits 0 and 1 be given by $s_0(t) = x(t),\ 0 \le t \le T$, and $s_1(t) = -x(t),\ 0 \le t \le T$, respectively. Of course, at the receiver, in order to perform coherent correlation operation, the chaotic waveform $x(t)$ is not available if the self-synchronization property of the chaotic waveform is

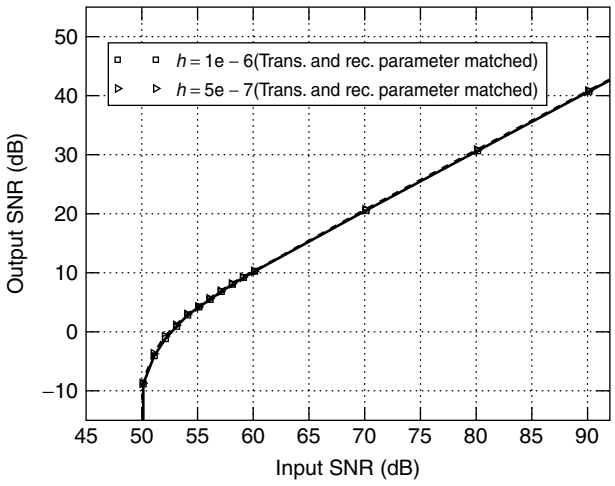


Figure 7. Robust ability for Mannella algorithm simulation with $K = 2505$.

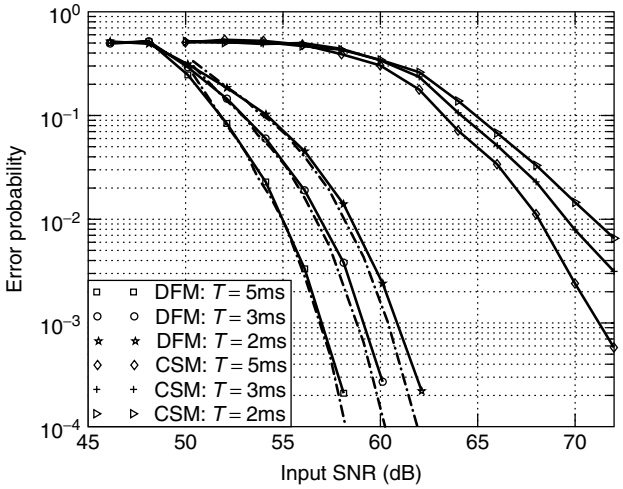


Figure 8. Comparison of BER versus input SNR for DFM and CSM systems.

not used. One way to circumvent this problem is to use the differential CSK (DCSK) modulation scheme as advocated by Kolumban [22]. For an antipodal DCSK transmitter, the two waveforms are given by

$$s_0(t) = \begin{cases} x(t), 0 \le t \le \dfrac{T}{2}, \\ x\left(t - \dfrac{T}{2}\right), \dfrac{T}{2} \le t \le T \end{cases}$$

$$s_1(t) = \begin{cases} x(t), 0 \le t \le \dfrac{T}{2} \\ -x\left(t - \dfrac{T}{2}\right), \dfrac{T}{2} \le t \le T \end{cases} \tag{14}$$

Then the received waveform is given by

$$r(t) = s_i(t) + n(t) = \begin{cases} x(t) + n(t), 0 \le t \le \dfrac{T}{2}; x\left(t - \dfrac{T}{2}\right) \\ \qquad +n(t), \dfrac{T}{2} \le t \le T, H_0 \\ x(t) + n(t), 0 \le t \le \dfrac{T}{2}; -x\left(t - \dfrac{T}{2}\right) \\ \qquad +n(t), \dfrac{T}{2} \le t \le T, H_1 \end{cases} \tag{15}$$

where $n(t)$ is the AWGN channel noise. For a differential coherent detection scheme, the first portion of the received chaotic waveform is used as a reference to perform coherent integration of the second portion of the received waveform. Then the output of the integrator is given by

$$\eta = \int_0^{T/2} r(t) r\left(t + \frac{T}{2}\right) dt = \pm \int_0^{T/2} x^2(t)\, dt \pm \int_0^{T/2} x\left(t + \frac{T}{2}\right) n(t)\, dt + \int_0^{T/2} x(t) n\left(t + \frac{T}{2}\right) dt + \int_0^{T/2} x\left(t + \frac{T}{2}\right) n(t)\, dt \tag{16}$$

In a conventional differential coherent BPSK system, the output of the integrator has the same form as that given in (16) except for the first term on the right-hand side (RHS). In a conventional system, $x(t)$ is a known deterministic waveform, and thus the first term is a known constant. However, for a chaotic $x(t)$ waveform that first term is not a constant, but only the expectation of that term (i.e., $E\{\int_0^{T/2} x^2(t)\, dt\}$) is a constant [22]. The nonconstancy of this term causes an additional system performance degradation of the DCSK system relative to the differential coherent BPSK system. We also note that, in the DCSK scheme, since the first portion of the waveform over a duration of $T/2$ seconds carries no information, the transmission rate of the system is one-half that of a nondifferential CSK system.

One possible remedy to eliminate the nonconstancy problem is to frequency-modulate the chaotic $x(t)$ waveform, which results in still another chaotic waveform but has constant energy. This leads to the FM-DCSK modulation scheme [24,25]. The demodulation of the FM-DCSK waveforms is performed in the same manner as that of the DCSK system, except now FM chaotic waveforms are used. Performances of non-band-limited BPSK, FSK, noncoherent FSK, and FM-DCSK are compared in Fig. 9. Some advantages of the FM-DCSK system [25] include the following:

1. There is no need to use the self-synchronization property of the chaotic waveform, which yields low performance in the presence of noise; it is relatively insensitive to channel distortions, and thus nonlinear amplifier can be used.
2. Since the FM-DCSK waveforms are wideband, they are relatively immune to frequency-selective fading in multipath scenarios.
3. Unlike the case in conventional DS-CDMA systems, since chaotic waveforms are used here, no distinct spectral lines are present; multiuser capability also exists.

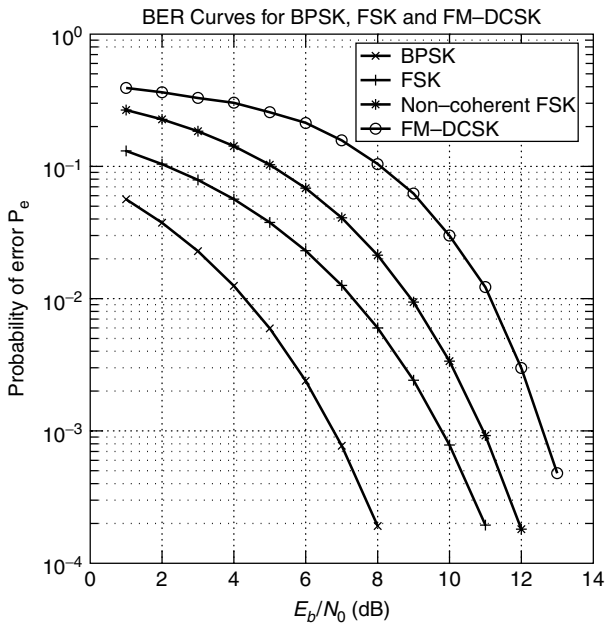


Figure 9. Comparison of BER versus E_b/N_0 of BPSK, FSK, noncoherent FSK, and FM-DCSK systems.

FM-DCSK is probably one of the more practical communication systems using chaos-based carrier modulations.

Besides frequency modulation, the basic DCSK scheme can be extended in various directions. Just as QPSK and QAM are extensions of BPSK in order to transmit more information data per baud, multilevel quadrature CSK scheme has also been proposed by Galias and Maggio [37].

3. SYMBOLIC DYNAMICS

A quite different chaotic communication technique using symbolic dynamics was originally proposed by Hayes et al. [27]. This approach attempts to provide a bridge between the theory of the chaotic system and information theory to design a communication system using chaotic dynamics. From the formalism of *symbolic dynamics* [63], Hayes et al. considered a chaotic system to be a natural digital information source with some constraint, denoted as the *grammar*. They showed that the symbolic dynamics of a chaotic oscillator can be made to follow a desired symbol sequence by using small perturbations. The mechanism behind the chaos control can be explained in an abstract but simple way [28]. Consider the Bernoulli map given by $x_{n+1} = 2x_n \bmod 1$. If x is represented by a binary fraction with finite precision, say, $x_n = .10101010$, then either $x_{n+1} = .01010101$ or $x_{n+1} = .01010100$ can be obtained by changing the eighth significant bit, representing a change of about 0.004 in base 10. By repeatedly changing the eighth bit, the change will show up in the most significant bit, which determines whether $x \geq \frac{1}{2}$ or $x < \frac{1}{2}$, a large-scale and easily observable signal attribute. Therefore, any message that can be encoded in a sequence of bits can be transmitted by controlling the symbolic dynamics of the chaotic system. The receiver is a simple level threshold detector.

For the continuous-time and continuous-state system, which is described by a set of ordinary differential equations, the symbolic dynamics of the system can be constructed using the Poincaré section concept by running the system without control [27]. Then we associate each binary sequence generated when the corresponding trajectory, starting from the initial condition, crosses the Poincaré surface with a real number, denoted by the *coding function*. The information message can be embedded into the chaotic dynamics by applying control pulses according to the coding function. The receiver observes the transmitted chaotic waveform and make a decision by observing the points on the Poincaré surface to see which side of the surface the crossing point lies in.

As stated earlier, in a DCSK system, the first portion of the chaotic waveform is used for reference purpose, carries no information, and yields an inefficient use of the channel bandwidth. Maggio and Galias [43] encoded some information onto the first portion of the chaotic waveform through the use of symbolic dynamics and increased the effective throughput of the system. Ciftci and Williams [44] proposed the use of optimum Viterbi estimation and channel equalization algorithms to estimate sequences encoded using symbolic dynamics over a channel with distortion. Other advanced symbolic dynamics and related analytic tools relevant to chaotic communications include those described in Refs. 45–47.

4. ANALOG CHANNEL ENCODING AND ESTIMATION

The use of chaotic dynamics as a channel encoder was proposed by Chen and Wornell [30]. A novel analog code based on the tent map dynamics and having a fast decoding algorithm was proposed for use on unknown, multiple, and time-varying channels with different SNRs. These practical chaotic codes having recursive receiver structures and important performance advantages over conventional codes were demonstrated. A convolutional encoder and multiresolution codes using chaotic systems are also developed in this article.

The basic idea of this chaotic encoder is to encode the analog message into the initial condition of the chaotic system, and an estimation technique was used to retrieve the message. Optimal state estimation for chaotic sequences is in general a difficult problem. Papadopoulos and Wornell [29] derived the maximum-likelihood (ML) estimator for the tent map sequences in stationary AWGN and showed that it can be implemented by a forward recursive filter followed by a backward recursive smoother. The forward recursive filter developed by Papadopoulos and Wornell [29] is identical to the Kalman filter.

5. CHAOTIC PULSE POSITION MODULATION

Instead of transmitting the chaos waveform, a chaotic pulse position modulation (CPPM) scheme was proposed Sushchik et al. [31]. This proposed system is similar to an ultra-wide-bandwidth impulse radio system of Win [32] that offers a promising communication method, especially in a severe multipath environment. A pulse position method is used to modulate binary information onto the carrier. The separation between the adjacent pulses is chaotic, arising from a dynamical system with irregular behavior.

The communication scheme is built around a chaotic pulse regenerator (CPRG), as shown in Fig. 10. With a pulse train of interpulse intervals T_i as its input to the CPRG, the nth incoming pulse is produced after a delay time given by $\Delta T_n = F(T_{n-1}, \ldots, T_{n-k})$ where $F(.)$ is a chaotic map. Thus, the system is expected to generate a pulsetrain with chaotic interpulse intervals. The binary information message is modulated at the output of the CPRG by delaying the pulse of a fixed time if binary 1

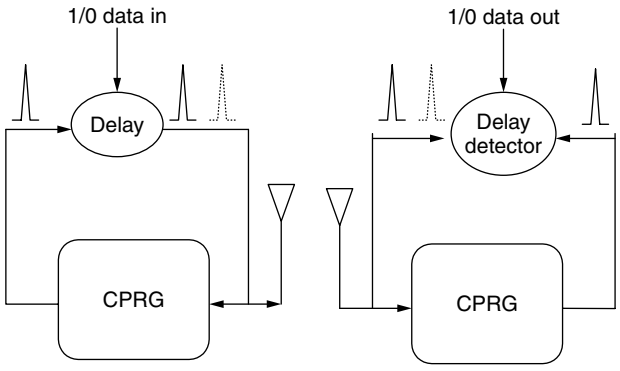


Figure 10. Chaotic pulse position modulation.

is being transmitted, or leaving unchanged if binary 0 is being transmitted. The received signal is fed into the identical CPRG at the receiver. Since the inputs to both CPRGs at the transmitter and receiver are identical, the outputs of CPRGs are expected to be identical. By computing the timing difference between the received signal and the output signal of the CPRG at the receiver, the embedded message can be retrieved.

This communication system may have a lower probability of intercept due to the aperiodicity of the chaos signal. This system performs well compared to other chaos-based convert communication schemes. Rulkov et al. have further analyzed the CPPM system with application to multiuser communication [64].

6. CHAOTIC SPREAD-SPECTRUM SEQUENCES

There has been an increasing interest in spread-spectrum communications, particularly in code-division multiple-access (CDMA) format. Some of the operations, concepts, and advantages of direct-sequence CDMA (DSCDMA) have been described [48]. System performance of DSCDMA critically depends on the auto- and cross-correlation properties of the *spreading sequences*. The use of chaotic sequence/waveform as spread sequences for DSCDMA communication systems has been proposed [49–51,53–58]. The chaotic signals have low nonzero shift autocorrelation and all cross-correlation properties due to the intrinsic broad nature of the spectrum and sensitivity to initial conditions. The use of the correlation function characteristics of some specific chaotic sequences and its comparison to *m* sequences/Gold sequences has been studied.

The chaotic sequence, generated by a logistic map, was first proposed for DSCDMA in 1992 [49]. The correlation properties of these logistic sequences are similar to random white noise. By simulations, the performance of chaotic sequences in the DSCDMA system is shown to be similar to that of PN sequences. Furthermore, due to their real values instead of binary values, chaotic sequences outperform PN sequences in low probability of intercept (LPI). Umeno et al. [51] used Chebyshev sequences for a synchronous CDMA system and analyzed the system performance using ergodic theory [7,52].

The statistical properties of binary sequences generated by a class of ergodic maps with some symmetric properties have been discussed [53]. Simple methods were used to generate a sequence of i.i.d. binary random sequence. The correlation functions of various types of chaotic binary sequences have been evaluated exactly by ensemble-averaging technique based on the Perron–Frobenius operator theory [7]. They also obtained a sufficient condition for a binary function to produce a sequence of i.i.d. binary random variables.

Mazzini et al. [54,55] proposed chaotic complex spreading sequences for asynchronous DSCDMA. They provided rigorous analysis of DSCDMA system performance bounds using chaotic complex spreading sequences. The simulation results in these papers show that systems based on chaotic spreading sequences perform generally better than the *Gold sequences*. Moreover, by treating the spreading sequences as random processes, Mazzini et al. [57] have shown the optimal *ensemble-averaged* autocorrelation of spreading sequences with minimum achievable interference variance decays nearly exponentially. They also proposed a chaotic map to generate the sequences with nearly exponential autocorrelation function. Without the assumption of independence and stationarity of the spreading sequences required by Mazzini et al. [57], Chen and Yao [60] provided a methodology to derive the general results on the partial autocorrelation function of the spreading sequences to minimize interference variance as well as a real-valued spreading sequence implementation that is at the same time optimal and practical by using the ergodic theory. For an asynchronous DSCDMA system, results in Refs. 57 and 60 show that these sequences generated based on chaos and ergodic theory concepts, can support approximately 15% more users for a fixed amount of interferences compared to Gold codes. Equivalently, for a fixed number of users, these sequences produce ~15% fewer interferences. In Fig. 11, there are nine BER curves for sequences of length 64 in AWGN. But there are essentially only three sets of curves, with each set having about the same performance. The lowest set of curves include the two optimal sequences for asynchronous CDMA operation (induced from optimal filtering of either a second- or third-order Chebyshev sequence); the middle set of curves include the Gold code and the original unfiltered second- or third-order Chebyshev sequence in asynchronous CDMA operation; and the upper set of curves represent operations under the chip-synchronized CDMA operations. There is an approximate 15% improvement in the lower set of curves compared to the middle set of curves. Various other related issues on sequences generated based on chaos theory and their implementations have also appeared in the literature [59,61,62].

7. CHAOS IN LASER COMMUNICATION AND MODELING OF RADAR AND RADIO PROPAGATION EFFECTS

There are many applications of chaotic nonlinear dynamics to various system problems. We consider only two such

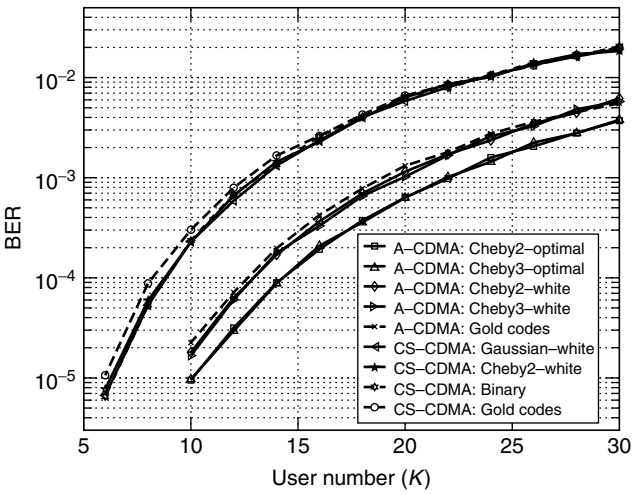


Figure 11. BER versus user number for nine chaotic spreading sequence asynchronous DSCDMA scenarios.

applications in communications. Semiconductor lasers are the most important light sources for optical communications. Unlike the previously considered electronic chaotic communication systems, where the nonlinearities are often introduced intentionally to create chaos, single-mode semiconductor lasers operate with various complex intrinsic nonlinearities due to the physical behaviors of the devices. Thus, chaos may or may not be avoidable in such devices. Many papers have dealt with the chaotic behaviors of lasers. The paper by Liu et al. [65] exploits the chaotic behaviors of the lasers and deals with the dynamics, synchronization, and message encoding and decoding of two optical laser communication systems. One system uses optical injection and the other uses delayed optoelectronic feedback. From these numerical simulation and experimental measurement works, the basic concept of using chaotic optical communication to transmit and receive hundreds of gigabit rate data has been demonstrated as feasible and practical.

Radar and radiofrequency propagation effects due to scattering, reflection, and shadowing are known in various scenarios to severely limit the performance of these systems. The modeling and understanding of these propagation effects are of great theoretical and practical interest. In this section we consider two physical problems. The first problem considers sea clutter as the backscattered returns from a patch of the sea surface illuminated by a transmitted radar pulse. Sea clutter waveforms are quite complicated. Traditionally, they have been modeled by statistical means such as through the marginal densities of the amplitude of the waveform. Often there is little insight or justification for these statistical characterizations such as lognormal, or k distribution. Since the sea clutter waveforms are functions of sometimes turbulent wave motions on the sea surface, it is not unreasonable to conjecture that perhaps nonlinear dynamics may be in operation. Perhaps the apparent seemingly randomness of the sea clutter may be modeled by deterministic chaos. Haykin and colleagues [66,67] have collected considerable amount of sea clutter data and used the method of Grassberger and Procaccia [68] and found nonintegral fractal dimensions and positive Lyapunov exponents from these data. Thus, they conjectured that under certain conditions, sea clutter may be modeled as deterministic chaos. But it is also known that a nonintegral fractal dimension together with a positive largest Lyapunov exponent obtained by computational means is not a sufficient condition. Specifically, the $1/f$ fractal random process (which is not a deterministic chaos) may have a nonintegral fractional dimension and also a positive Lyapunov exponent. Gao and Zheng [69,70] provided a more stringent test for chaos by showing that for a chaotic sequence, a plot of the time-dependent exponent $\Lambda(k)$–time index k forms a common envelope over different shells. The slope of these envelope is the largest Lyapunov exponent. This common envelope property is shown in Fig. 12 for the well-known chaotic Lorenz sequence. On the other hand, for a nonchaotic sequence, such as that for a white sequence (Fig. 13), the common envelope property does not hold. For the 130 s of sea clutter, Haykin et al. [71] obtained the results shown in Fig. 14, which does not manifest the common

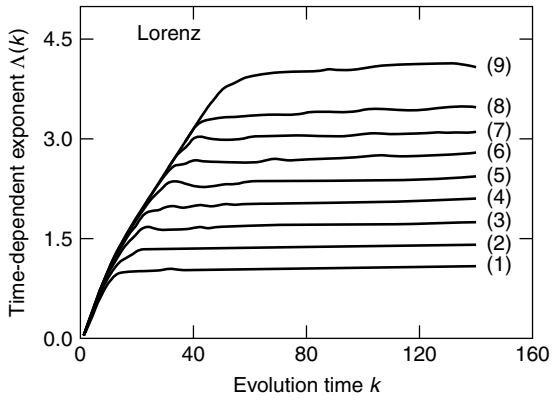


Figure 12. Exponent versus evolution time for different shells of a chaotic Lorenz sequence.

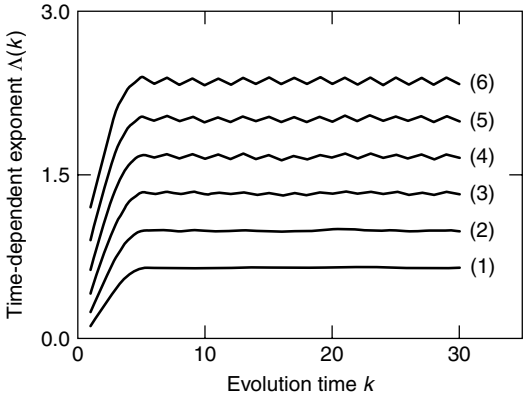


Figure 13. Exponent versus evolution time for different shells of a white-noise sequence.

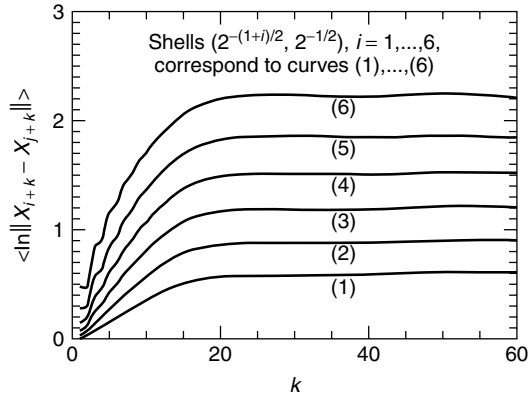


Figure 14. Exponent versus evolution time for different shells of a sea clutter sequence.

envelope property. This would seem to indicate that sea clutter data are not deterministic chaos. To obtain deeper insight from these sea clutter data, a multifractal analysis was performed [71]. Without going into detail, random multiplicative multifractal theory [76,77] shows that if a sequence of data has infinitely many power-law scaling relationships, then plots of $\log_2 M_q(\varepsilon)$ versus $-\log_2 \varepsilon$ for different values of q should form a straight line and the amplitudes must be lognormally distributed. Here the

moment $M_q(\varepsilon) = \sum_i w_i^q$, $\varepsilon = 2^{-N}$ at stage N, where q is a real number, and the weights $w_i, i = 1, \ldots$ are obtained directly from the measured data. For this set of sea clutter data, Fig. 15 shows the amplitude and particularly the envelopes form essentially sets of straight lines. Furthermore, Fig. 16 shows the amplitude and particularly the envelope indeed satisfies the lognormal distribution. Indeed, lognormal distribution for radar clutter amplitude has been known for many years, although no known justification has been given until now. It is also interesting that some work by Cowper and Mulgrew [72] and even Haykin et al. [73] shows reservations about the proper modeling of sea clutter as deterministic chaos. Clearly, more data should be collected and analysis performed in order to construct a valid propagation model for radar clutter returns.

The second problem considered in this section deals with the modeling of the fading radio propagation phenomena in wireless communication. There have been many propagation measurements in different frequency bands, and various statistical models have been proposed. Tannous et al. [74] applied the method of Grassberger and Procaccia [68] to some indoor 0.915-GHz radio propagation data, and found nonintegral fractal dimensions together with positive largest Lyapunov exponents and concluded that the propagation channel may be modeled by chaos. However, due to the limited data length and highly nonstationary character of their measured data, it is not obvious that a deterministic chaos conclusion can be made from these data. Gao et al. [75] reported on short timescaled analysis of some indoor 1-GHz propagation data using the method due to Gao and Zheng [69,70]. Figure 17 shows a

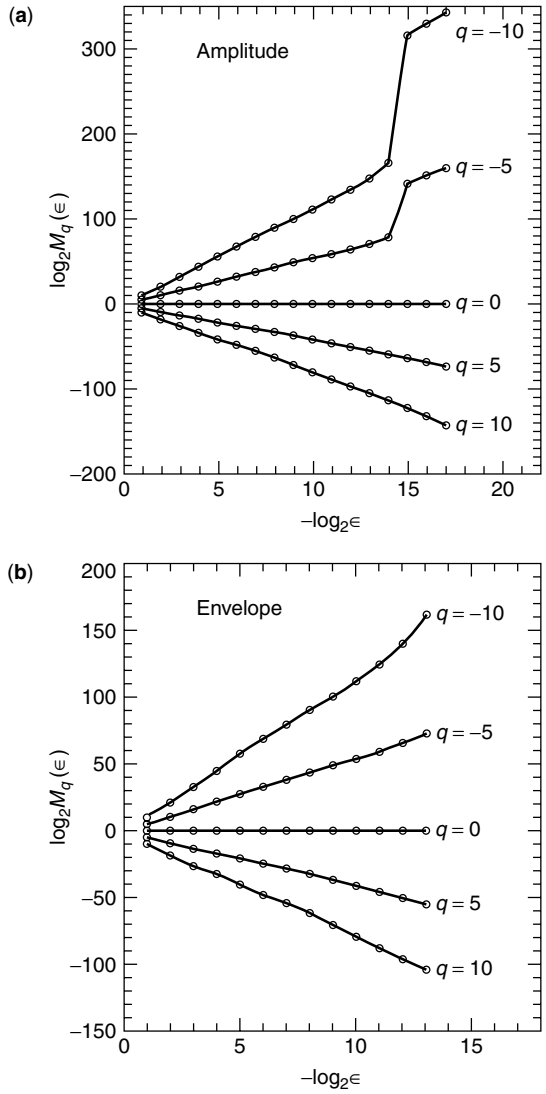


Figure 15. Multifractal scaling law for amplitude and envelope of a sea clutter sequence.

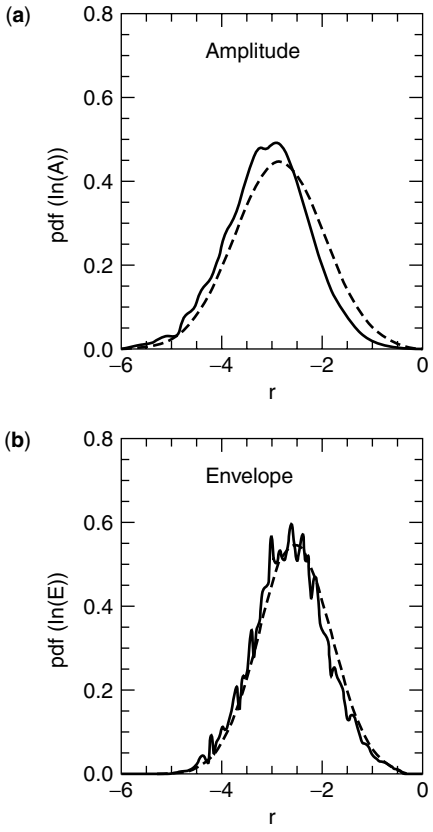


Figure 16. Probability distribution functions of log of amplitude and envelope versus values.

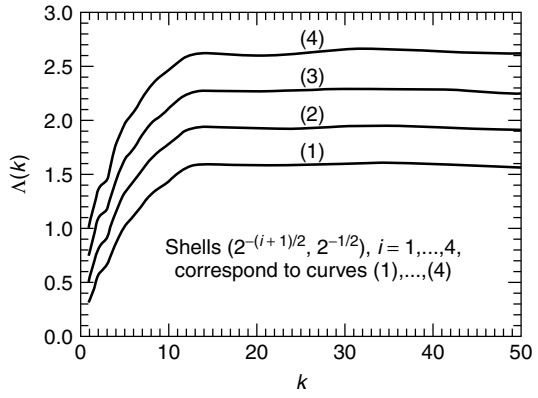


Figure 17. Exponent versus evolution time for different shells of a measured radio propagation data.

typical plot of the time-dependent exponent $\Lambda(k)$ versus time index k and reveals no common envelope property. From the earlier comments made with regard to Figs. 12 and 13, we may conclude that at least for these measurements, deterministic chaos were not present. Further analysis based on fractional Brownian motion process (FBMP) [78] were reported [75]. A FBMP is a Gaussian process with zero-mean and stationary increments. Its variance has the form of t^{2H}, and its power spectral density has the form of $f^{-(2H+1)}$, where H is the Hurst parameter. For $0.5 < H < 1$, the process has long-range dependence. For $H = 0.5$, the process becomes the standard BMP (whose formal derivative is the standard white Gaussian noise used in communication system analysis). For $0 < H < 0.5$, the adjacent values of the process have highly negative correlations and the process fluctuates widely. Details are omitted here, but if the data are from an FBMP, then the set of variance–time $\log_2 F_q(m)/\log_2 m$ curves in Fig. 18 should form straight lines. For two sets of measured data, the estimated Hurst parameters have almost constant values of about 0.4 and 0.45. For these limited number of measurements, the data may suggest a fractal FBM-like process modeling. However, since the RF band around 1 GHz is crowded with various radio transmitters and microwave ovens, drawing definitive conclusion about the true nature of the observed data is delicate. It is interesting that researchers in electromagnetics [79] have also proposed the modeling of RF scattering based on fractal theory. Clearly, more careful data collection and advanced tools for analyzing these data must be performed before valid models can be established and these results can be applied to practical wireless communication applications.

8. CONCLUSIONS

In this article, we have summarized various aspects of deterministic chaos to the analysis, design, and modeling of communication systems and channels. We first provided some background on the history of chaos. Then early chaotic modulation techniques utilizing self-synchronization of the chaotic waveform were discussed. Evaluation of the error probability of these systems cannot be performed analytically or even by conventional numerical simulation. A numerical simulation method for a nonlinear stochastic system is discussed, and an accurate stochastic integration algorithm is provided. Because of the sensitivity of the self-synchronization process to channel distortion and noise, the performances of these early chaotic communication systems were poor. The FM-DCSK modulation scheme was proposed and shown to be practical for implementation. It is competitive to conventional communication system performance and may possess certain additional desirable properties. Chaotic pulse position modulation particularly with application to UWB systems may also be practical and useful. More recently, various chaotic nonlinear dynamics and ergodic theory concepts have been proposed to create CDMA sequences that can perform better than known sequences in the asynchronous CDMA mode. Preliminary investigations show these results are practical for implementation.

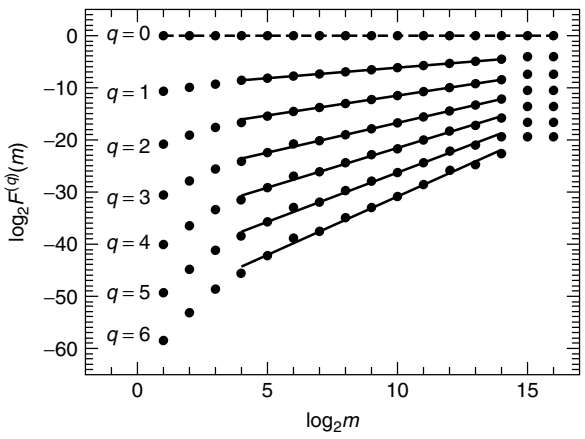


Figure 18. Power versus time for different scalings for a measured radio propagation data.

It is quite clear that many of the researchers in "chaotic communications" are more interested in exploiting various complex and often interesting properties of chaotic nonlinear dynamics rather than using them for explicit communication purpose. Ultimately, "chaotic communication" schemes must be compared to comparable conventional communication schemes with respect to bandwidth, data rate, energy per bit, error probability, and complexity of the transmitter and receiver. It remains a challenge to exploit the complexity and richness of chaos and related analytic tools to understand, analyze, design, and model communication systems and channels.

Acknowledgment

This work is partially supported by an ARO-MURI grant on "chaotic communication," a UC-CoRe grant sponsored by ST Microelectronics, and NASA/Dryden grant NCC4-153.

BIOGRAPHIES

Kung Yao received the B.S.E. (Highest Honors), M.A., and Ph.D. degrees in electrical engineering, all from Princeton University, Princeton, New Jersey. He was a NAS-NRC Post-Doctoral Research Fellow at the University of California, Berkeley. Presently, he is a Professor in the Electrical Engineering Department at UCLA. In 1969, he was a Visiting Assistant Professor at the Massachusetts Institute of Technology. In 1985–1988, he served as an Assistant Dean of the School of Engineering and Applied Science at UCLA. His research and professional interests include sensor array systems, digital communication theory and systems, smart antenna and wireless radio systems, chaos communications and system theory, digital and array signal and array processing, systolic and VLSI algorithms, architectures and systems, radar systems, and simulation. He has published over 250 journal and conference papers. Dr. Yao received the IEEE Signal Processing Society's 1993 Senior Award in VLSI Signal Processing. He was the co-editor of a two-volume series of an IEEE reprint book entitled *High Performance VLSI Signal Processing*, IEEE Press, 1997. He was a Guest Associate Editor of a Special Issue on "Applications of Chaos in Modern Communication Systems" of the *IEEE Transactions on Circuits and Systems* — Part I, December 2001. He is a Fellow of IEEE.

Chi-Chung Chen was born in Taiwan in 1970. He received the B.S. and M.S. degrees in control engineering from National Chiao Tung University, Taiwan, in 1993 and 1995, respectively, and the Ph.D. degree in electrical engineering from the University of California at Los Angeles in 2000. He was employed as an Associate Researcher on the design of controller in MIRL, ITR1, Taiwan from 1995 through 1996. From 1997 to 2000, he was a Research Assistant at UCLA. His current research interests include chaotic communications, spread-spectrum COMA systems, pseudorandom sequences, adaptive systems, and wireless communication systems. Dr. Chen is a member of the Phi Tau Phi Scholastic Honor Society. Since 2001, he has been with Accton Technology Corporation in Tainan, Taiwan.

BIBLIOGRAPHY

1. E. Lorenz, Deterministic nonperiodic flow, *J. Atm. Sci.* **20**: 130–141 (1963).
2. B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, 1982.
3. *IEEE Trans. Circuits Syst.* (Special Issue on Chaos in Nonlinear Electronic Circuits) **40**(10): (1993).
4. J. Gleick, *Chaos: The Amazing Science of the Unpredictable*, Random House, 1988.
5. G. P. Williams, *Chaos Theory Tamed*, Joseph Henry Press, 1997.
6. R. C. Hilborn, *Chaos and Nonlinear Dynamics*, Oxford Univ. Press, 1994.
7. A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, 2nd ed., Springer-Verlag, 1994.
8. L. M. Pecora, Overview of chaos and communications research, *SPIE* **2038**: (1993).
9. M. Hasler, Synchronization of chaotic systems and transmission of information, *Int. J. Bifurc. Chaos* **8**(4): 647–659 (1998).
10. A. V. Oppenheim, G. W. Wornell, S. H. Isabelle, and K. M. Cuomo, Signal processing in the context of chaotic signals, *Proc. IEEE ICCASP* **4**: 117–120 (March 1992).
11. M. P. Kennedy, R. Rovatti, and G. Setti, *Chaotic Electronics in Telecommunications*, CRC Press, 2000.
12. *IEEE Trans. Circuits Syst. — I* (Special Issue in Noncoherent Chaotic Communications) **47**: (Dec. 2000).
13. *IEEE Trans. Circuits Syst. — I* (Special Issue on Applications of Chaos in Modern Communication Systems **48**: (Dec. 2001).
14. *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: (May 2002).
15. L. M. Pecora and T. L. Carroll, Synchronization in chaotic systems, *Phys. Rev. Lett.* **64**: 821–824 (Feb. 1990).
16. L. M. Pecora, T. L Carroll, G. A. Johnson, and D. J. Mar, Fundamentals of synchronization in chaotic systems, concepts, and applications, *Int. J. Bifurc. Chaos* **7**: 520–543 (1997).
17. K. M. Cuomo, A. V. Oppenheim, and S. H. Strogatz, Synchronization of Lorenz-based chaotic circuits with applications to communications, *IEEE Trans. Circuits Syst. — II* **40**: 626–633 (1993).
18. V. Milanovic and M. E. Zaghloul, Improved masking algorithm for chaotic communications systems, *IEE Electron. Lett.* **32**: 11–12 (1996).
19. L. Kocarev, K. Halle, K. Eckert, and L. Chua, Experimental demonstration of secure communication via chaotic synchronization, *Int. J. Bifurc. Chaos* **2**: 709–713 (Sept. 1992).
20. C. C. Chen and K. Yao, Stochastic-calculus-based numerical evaluation and performance analysis of chaotic communication systems, *IEEE Trans. Circuits Syst. — I* **47**: 1663–1672 (Dec. 2000).
21. H. Dedieu, M. Kennedy, and M. Hasler, Chaos shift keying: modulation and demodulation of a chaotic carrier using self-synchronizing Chua's circuits, *IEEE Trans. Circuits Syst. — II* **40**: 634–642 (Oct. 1993).
22. G. Kolumban, B. Vizvari, W. Schwarz, and A. Abel, Differential Chaos shift keying: A robust coding for chaotic communication, *Proc. NDES*, 1996, pp. 87–92.
23. G. Kolumban, M. P. Kennedy, and L. O. Chua, The role of synchronization in digital communications using chaos — Part II: Chaotic modulation and chaotic synchronization, *IEEE Trans. Circuits Syst. — I: Fund. Theory Appl.* **45**: 1129–1140 (1998).
24. M. P. Kennedy and G. Kolumban, Digital communication using chaos, *Signal Process.* **80**: 1307–1320 (2000).
25. G. Kolumban, G. Kis, Z. Jako, and M. P. Kennedy, FM-DCSK: A robust modulation scheme for chaotic communication, *IEICE Trans. Fund. Electron. Commun. Comput. Sci.* **E81-A**: 1798–1802 (Oct. 1998).
26. G. Kolumban, M. P. Kennedy, Z. Jako, and G. Kis, Chaotic communications with correlator receiver: Theory and performance limits, *Proc. IEEE* **90**: 711–732 (May 2002).
27. S. Hayes, C. Grebogi, and E. Ott, Communicating with chaos, *Phys. Rev. Lett.* **70**: 3031–3034 (May 1993).
28. S. Hayes, *Communicating with Chaos: A Physical Theory for Communication via Controlled Symbolic Dynamics*, Ph.D. thesis, Univ. Maryland, 1994.
29. H. C. Papadopoulos and G. W. Wornell, Maximum likelihood estimation of a class of chaotic signals, *IEEE Trans. Inform. Theory* **41**: 312–317 (1995).
30. B. Chen and G. W. Wornell, Analog error-correcting codes based on chaotic dynamical systems, *IEEE Trans. Commun.* **46**: 881–890 (1998).
31. M. Sushchik et al., Chaotic pulse position modulation: A robust method of communicating with chaos, *IEEE Commun. Lett.* **4**: 128–130 (April 2000).
32. M. Z. Win and R. A. Scholtz, Impulse radio: How it works, *IEEE Commun. Lett.* **2**: 360–363 (1998).
33. *IEEE Trans. Commun.* (Special Issue on Spread Spectrum Communications) **25**(8): (1977).
34. *IEEE Trans. Commun.* (Special Issue on Spread Spectrum Communications) **30**(5): (1982).
35. K. S. Halle, C. W. Wu, M. Itoh, and L. O. Chua, Spread spectrum communication through modulation of chaos, *Int. J. Bifurc. Chaos* **3**: 409–477 (1993).
36. G. Kolumban, M. P. Kennedy, and L. O. Chua, The role of synchronization in digital communications using chaos — Part I: Fundamentals of digital communications, *IEEE Trans. Circuits Syst. — I: Fund. Theory Appl.* **44**(10): 927–936 (1997).
37. Z. Galias and G. M. Maggio, Quadrature chaos-shift keying: theory and performance analysis, *IEEE Trans. Circuits*

Syst. — I (Special Issue on Applications of Chaos in Modern Communication Systems) **48**: 1510–1518 (Dec. 2001).

38. N. J. Kasdin, Runge-Kutta algorithm for the numerical integration of stochastic differential equations, *J. Guidance, Control, Dynamics* **18**: 114–120 (1995).
39. R. Mannella and V. Palleschi, Fast and precise algorithm for computer simulation of stochastic differential equations, *Phys. Rev. A* **40**: 3381–3386 (Sept. 1989).
40. R. L. Stratonovich, A new representation for stochastic integrals and equations, *SIAM J. Control* **4**: 362–371 (1966).
41. R. E. Mortensen, Mathematical problems of modeling stochastic nonlinear dynamic systems, *J. Stat. Phys.* **1**: 271–296 (1969).
42. L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Wiley, 1973.
43. G. M. Maggio and Z. Galias, Enhanced differential shift keying using symbolic dynamics, *Proc. IEEE GLOBECOM*, Nov. 2001, Vol. 2, pp. 1157–1161.
44. M. Ciftci and D. B. Williams, Optimal estimation and sequential channel equalization algorithms for chaotic communication systems, *EURASIP J. Appl. Signal Process.* **2001**: 249–256 (Dec. 2001).
45. D. Lind and B. Marcus, *Introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press, 1995.
46. T. Kohda, Information sources using chaotic dynamics, *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: 641–661 (May 2002).
47. G. Setti, G. Mazzini, R. Rovatti, and S. Callegari, Statistical modeling of discrete-time chaotic processes-basic finite-dimensional tools and applications, *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: 662–690 (May 2002).
48. A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, 1992.
49. G. Heidari-Bateni and C. D. McGillem, Chaotic sequences for spread spectrum: An alternative to PN-sequences, *Proc. IEEE Int. Conf. Selected Topics in Wireless Communications*, 1992, pp. 437–440.
50. G. Heidari-Bateni and C. D. McGillem, A chaotic direct-sequence spread-spectrum communication system, *IEEE Trans. Commun.* **422**(2–4): 1524–1527 (1994).
51. K. Umeno and K. I. Kitayama, Improvement of SNR with chaotic spreading sequences for CDMA, *Proc. IEEE Information Theory Workshop*, South Africa, June 1999, p. 106.
52. R. L. Adler and T. J. Rivlin, Ergodic and mixing properties of Chebyshev polynomials, *Proc. Am. Math. Soc.* **15**: 794–796 (1964).
53. T. Kohda and A. Tsuneda, Statistics of chaotic binary sequences, *IEEE Trans. Inform. Theory* **43**: 104–112 (1997).
54. G. Mazzini, G. Setti, and R. Rovatti, Chaotic complex spreading sequences for asynchronous DS-CDMA — Part I: System modeling and results, *IEEE Trans. Circuits Syst. — I* **44**: 937–947 (Oct. 1997).
55. R. Rovatti, G. Setti, and G. Mazzini, Chaotic complex spreading Sequences for asynchronous DS-CDMA — Part II: Some theoretical performance bounds, *IEEE Trans. Circuits Syst. — I* **44**: 937–947 (Oct. 1997).
56. R. Rovatti and G. Mazzini, Interference in DS-CDMA systems with exponentially vanishing autocorrelations: Chaos-based spreading is optimal, *IEE Electron. Lett.* **34**: 1911–1913 (October 1998).
57. G. Mazzini, R. Rovatti, and G. Setti, Interference minimization by auto-correlation shaping in Asynchronous DS-CDMA systems: Chaos-based spreading is nearly optimal, *IEE Electron. Lett.* **35**: pp. 1054–1055 (June 1999).
58. T. Yang and L. O. Chua Chaotic digital code-division multiple access (CDMA) communication systems, *Int. J. Bifurc. Chaos* **7**: 2789–2805 (1997).
59. L. Cong and L. Shaoqian, Chaotic spreading sequences with multiple access performance better than random sequences, *IEEE Trans. Circuits Syst. — I* **47**: 394–397 (March 2000).
60. C. C. Chen, E. Biglieri, and K. Yao, Design of spread spectrum sequences using chaotic dynamical systems and ergodic theory, *IEEE Trans. Circuits Syst. — I* **48**: 1110–1113 (Sept. 2001).
61. G. Mazzini, R. Rovatti, and G. Setti, Chaos-based asynchronous DS-CDMA systems and enhanced Rake receivers: Measuring and improvements, *IEEE Trans. Circuits Syst. — I* **48**: 1445–1454 (Dec. 2001).
62. C. C. Chen, K. Yao, and E. Biglieri, Optimal spread spectrum sequences — constructed from Gold codes, *Proc. IEEE GLOBECOM*, Nov. 2000, pp. 867–871.
63. B.-L. Hao, *Elementary Symbolic Dynamics and Chaos in Dissipative Systems*, World Scientific, Singapore, 1989.
64. N. Rulkov, M. Sushchik, L. Tsimring, and A. Volkvskii, Digital communication using chaotic pulse position modulation, *IEEE Trans. Circuits Syst. — I* (Special Issue on Applications of Chaos in Modern Communication Systems) **48**: 1436–1444 (Dec. 2001).
65. J. M. Liu, H. F. Chen, and S. Tang, Optical communication systems based on chaos in semiconductor lasers, *IEEE Trans. Circuits Syst. — I* (Special Issue on Applications of Chaos in Modern Communication Systems) **48**: 1475–1483 (Dec. 2001).
66. S. Haykin and S. Puthusserypady, Chaotic dynamics of sea clutter, *Int. J. Bifurc. Chaos* **7**: 777–802 (1997).
67. S. Haykin, *Chaotic Dynamics of Sea Clutter*, Wiley, New York, 1999.
68. P. Grassberger and I. Procaccia, Characterization of strange attractors, *Phy. Rev. Lett.* **50**: 346 (1983).
69. J. B. Gao and Z. M. Zheng, Local exponent divergence plot and optimal embedding of a chaotic time series, *Phys. Lett. A* **181**: 153–158 (1993).
70. J. B. Gao and Z. M. Zheng, Direct dynamical test for deterministic chaos and optimal embedding of a chaotic time series, *Phys. Rev. E* **49**: 3807–3814 (1994).
71. J. B. Gao and K. Yao, Multifractal features of sea clutter, *Proc. 2002 IEEE Radar Conf.*, April 2002, pp. 500–505.
72. M. R. Cowper and B. Mulgrew, Nonlinear processing of high resolution radar sea clutter, *Proc. IJCNN*, July 1999, Vol. 4, pp. 2633–2638.
73. S. Haykin, R. Bakker, and B. W. Currie, Uncovering nonlinear dynamics — the case study of sea clutter, *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: 860–881 (May 2002).
74. C. Tannous, R. Davies, and A. Angus, Strange attractors in multipath propagation, *IEEE Trans. Commun.* **38**: 629–631 (May 1991).

75. J. B. Gao et al., Can sea clutter and indoor radio propagation be modeled as strange attractors? *Proc. 7th Experimental Chaos Conf.*, Aug. 2002.

76. J. F. Gouyet, *Physics and Fractal Structure*, Springer-Verlag, 1995.

77. B. B. Mandelbrot, *Fractals and Scaling in Finance*, Springer-Verlag, 1997.

78. B. B. Mandelbrot and J. W. Van Ness, Fractional Brownian motions, fractional noises and applications, *SIAM Rev.* **10**: 422–437 (Oct. 1968).

79. D. L. Jaggard, On fractal electrodynamics, in H. N. Kritikos and D. L. Jaggard, eds., *Recent Advances in Electromagnetic Theory*, Springer-Verlag, 2001, pp. 183–224.

CHARACTERIZATION OF OPTICAL FIBERS

GILBERTO M. CAMILO
OmniGuide Communications
Cambridge, Massachusetts

1. INTRODUCTION

At present there are many practical applications of optical fibers of different technologies. Optical fibers are used primarily (90% of total applications) in telecommunications because optical signals can be transmitted over long distances at high speed in optical fibers. Other applications are also important, such as military, medical, and industrial. The economic advantage of the optical fiber technology and its exploding implementation has pushed the manufacturing process from 60 m/min in 1978 to 1800 m/min in 2002.

Different glasses and coating materials have been developed since the early 1980s, but most of the optical fibers installed in the field are silica glass optical fibers protected by a polymeric coating. The coating is applied during the manufacturing process and is used to protect the surface of the glass from defects caused by abrasion, thus avoiding premature breaks. The high purity of the silica glass being used can guarantee very low attenuation of the light radiation being transmitted and also good mechanical qualities.

Other optical fiber technologies not based in silica glass are under study or already in use as all-plastic optical fibers [1], used in illumination systems and short-distance data communication; and infrared optical fibers [2,3], manufactured using chalcogenide glasses, or crystal glasses, which are used in sensor systems and telecommunication systems.

The optical and mechanical characterization methods used in optical fiber qualification are in general the same for all optical fiber types. Depending on which optical fibers are being tested, the appropriate light wavelength is used during the measurement process. Optical fibers can operate from the visible range (600 nm) to the infrared range (10,000 nm).

Measurement methods for silica glass optical fibers, coated by a primary and a secondary polymeric material, with operating range 1300–1550 nm, are considered here. Multimode silica optical fibers will be discussed briefly.

A silica glass optical fiber has an overall polymeric coating diameter of 250 μm and a centered glass with diameter around 125 μm. The glass part is composed of a core and a cladding, one glass cylinder inside another with a slight difference in refraction index. In the lightpulse that travels through the single-mode optical fiber, the core guides only the fundamental mode. The diameter of the core in these optical fibers is around 9 μm. In multimode optical fibers, hundreds of modes propagate at the same time in the lightpulse, and the optical fiber core is around 50 μm in diameter.

Presently the most widely used coating polymers are ultraviolet-cured acrylates. For mechanical and optical reasons, two polymer layers with distinct mechanical properties are applied to the glass surface. The soft inner polymer is in direct contact with the glass and absorbs small mechanical stresses. The harder external polymer has much higher elastic modulus compared to the internal polymer and is intended to resist the environment and abrasion.

2. CHARACTERIZATION METHODS

The National Standard Committees formed by industries, universities, and government standard institutions propose and discuss various characterization methods. These methods are used to specify and qualify optical fiber systems, including optical fibers as a telecommunication product. Fiberoptic test procedures (FOTPs) are developed to provide a uniform plan of action during the tests of optical fiber system components. The Telecommunications Industry Association and the Electronic Industries Alliance (TIA/EIA) specify the optical fiber and optical fiber cable standards in the United States. Many other countries use these standards as well. In Europe, Japan, and other foreign markets the International Electrotechnical Commission (IEC) determine alternative telecommunication standards studies.

Among these organizations, in the United States and Europe, a homogenization process of procedures and specifications for optical fibers and optical fiber cables is currently under way.

The characterization methods and techniques discussed here are in accordance with national and international standard procedures. The optical fiber characterization can be divided in three different aspects:

- Geometric characterization
- Transmission characterization
- Mechanical characterization

2.1. Geometric Characterization

An optical fiber is an electromagnetic waveguide operating at a very short wavelength, around 1300–1550 nm. The micrometer optical fiber geometric characteristics are essential to maintain the integrity of the optical signal, which carries the information. According to the standard specifications, many parameters need to be checked and must be maintained inside narrow ranges; these include:

- Diameters of the core and of the cladding
- Core–cladding concentricity

- Ellipticity of the core and of the cladding
- Length of the optical fiber
- Numerical aperture
- Primary and secondary coating diameters

Numerical aperture is the cone angle light acceptance in front of the optical fiber. If the incident light radiation is inside this cone angle, and this light consists of a wavelength compatible with the geometric guidance characteristics of the optical fiber, this radiation will be coupled in the core and transmitted through the length of the fiber. If it is outside the cone angle, the radiation will be reflected by the glass in the extremity of the optical fiber or refracted and spread in the cladding, and from the cladding to the outside of the optical fiber.

Different techniques can be used to measure these optical fiber geometric parameters. The most common are

- Refracted near-field method
- Transmitted near-field method
- Transmitted far-field method
- Microscopy method

These techniques use the radiation pattern that is refracted or transmitted at a very short distance in the extremity of the optical fiber in order to measure the geometric parameters. In the refracted near-field method [4], a laser coupled to a very precise stepper motor is used to focus a lightbeam at different angles in the core of the optical fiber. By analyzing the cladding refracted light, it is possible to measure precisely the dimensions of the core–cladding and cladding–coating interfaces.

In the *transmitted field method*, after the light is transmitted by a small piece of optical fiber, very sensitive optical detectors, or charge-coupled device (CCD) cameras, measure its dimensions.

In the microscopic method, a microscope with a photographic machine, or a videocamera, inspects directly the transmitted radiation in the extremity of a small piece of optical fiber and performs the geometric measurement.

To measure the length of the optical fiber, it is most common to use optical time-domain reflectometry (OTDR) [5]. This technique was developed as a modification of a similar technique that was used during decades in metallic telecommunication cables. The great advantage of this technique is that it is necessary to have access to only one extremity of the optical fiber. In this method, modulated laser light is injected in the core of the optical fiber and during its propagation through the optical fiber length part of the radiation is reflected and spread by defects in the glass structure. A small backscattered portion of the reflected radiation returns to the extremity and brings back the information necessary to localize precisely (to within a few centimeters), those defects. By capturing the light that reflects in the other extremity it is possible to determine the length of the optical fiber.

In an OTDR, which is the equipment that uses such a technique in optical fibers, a small percentage of the incident radiation returns to the detector and defines the maximum length range to be analyzed. This technique can still be used to measure the attenuation of the optical signal after it has been transmitted through the optical fiber length. Actually, a commercial OTDR can measure the optical attenuation, the length, and localize defects to within a range of a few meters over 200 km of an optical fiber length.

It must be pointed out that some of these methods are still being refined, and the commercial apparatus that use them have significantly improved since the early 1980s.

2.2. Transmission Characterization

The transmission characteristics are related to capacity of the optical fiber and its ability to maintain the optical signal integrity after being conducted through the length of the fiber. Different characteristics of the optical fiber can affect its capacity to perform well at long distances. The most important factors that cause power depletion can be subsumed in what is called the attenuation of the optical fiber.

The optical fiber materials absorb part of the light radiation when the lightpulse travels through the glass and a part is spread in glass imperfections, causing attenuation. Most of the spread light goes from the core into the cladding, and from there it is lost. Absorption and spreading are caused by several different phenomena, including nonuniformity of the glass composition, particles and gas bubbles trapped in the glass structure, cracks and other mechanical defects in the glass, nonuniformity of the glass geometry, and defective coatings. The majority of these imperfections originate during the optical fiber manufacturing process and others result from installation and use of the optical fiber cables.

The methods most commonly used to measure the attenuation of the optical fiber consist of measuring the power loss, using power detectors. Another important technique is based on optical time-domain reflectometry (OTDR). OTDR measuring apparatus are compact and versatile devices, easily transported to the field, and have the capability to measure very long lengths of optical fiber. Attenuation can be measured for a specific operating wavelength of the optical fiber, as in OTDRs, or in a broader wavelength range, called *spectral attenuation*.

Figure 1 shows the schematics of an apparatus to measure the optical fiber spectral attenuation by power

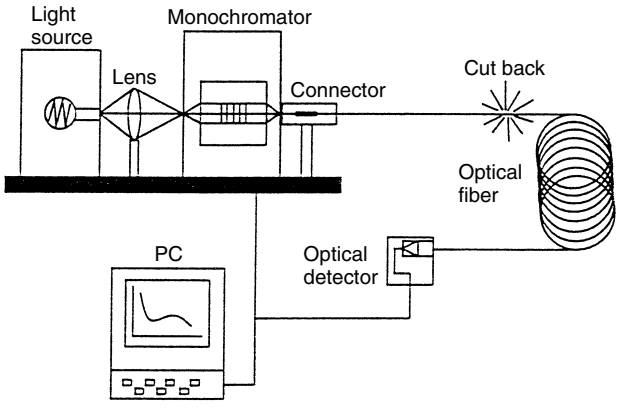


Figure 1. Spectral attenuation apparatus schematics.

measurement. In this test the wavelength is selected using a computer-controlled monochromator, and the light source consists of a white-colored, powerful device. After the monochromator the monochromatic selected light is injected into the optical fiber extremity. This extremity of the optical fiber must be cleaned and cut carefully in order to guarantee that the maximum light is injected in the optical fiber core. The other extremity of the optical fiber, clean and with a good quality cut, is in contact with the power detector. For silica glass optical fibers, the first power measurement series is done with the monochromator scanning at each few nanometers in the range 600–1600 nm. After the first power scanning, the optical fiber is cut at approximately 2 m from the launching extremity and inserted again in the power detector. Another power measurement is performed as described above, and the attenuation at each wavelength is the quotient between the second power measurement and the first power measurement divided by the length of the optical fiber. It is convenient to use the common logarithm of the power quotient to express the attenuation in decibels per kilometer (dB/km).

Because the cut is made close to the power source, this technique is also known as the *cutback method*. This method is precise and reliable and is used as a reference for other attenuation measurements methods.

In single-mode optical fibers the loss of light radiation that affects the intensity and spreading of the lightpulse can be caused by other phenomena. Light sources, such as lasers and light-emitting diodes (LEDs), are designed to emit a specific wavelength, but they really emit light in a wavelength range. A lightpulse generated by those sources has many components with different wavelengths. In the core of the optical fiber different wavelengths travel at different velocities causing a time spreading of the pulse. The phenomenon in which the velocity of propagation of an electromagnetic wave is wavelength-dependent is called *dispersion*, in which different wavelengths are connected with different colors. This phenomenon is thus called *chromatic dispersion*.

At least two methods are used to measure chromatic dispersion: the spectral group delay method and the phase-shift method [6,7]. These methods measure the time dispersion of the lightpulse that occurs after the pulse has traveled the length of the optical fiber.

The optical fiber core does not have a perfectly symmetric cylindrical shape; the purpose of this design is to avoid *polarization mode dispersion* (PMD). In practice, the core diameter and shape vary slightly in a random fashion during the optical fiber manufacturing process. In PMD, internal stresses induced by thermal expansion and external forces induced by the environment through handling and cabling adds more stress fields inside the optical fiber core. Those perturbations induce the two orthogonally polarized modes that travel at different group velocities in a single-mode optical fiber, and the lightpulse is broadened and distorted. At frequency transmissions above 10 gigabits per second (Gb/ps), PMD is a limiting factor for lightwave transmission in optical fiber systems.

Nowadays systems operating at very high transmission rates are commonly installed in practice. The PMD limitation has motivated many efforts to understand and quantify the phenomenon. Since the early 1990s, at least six methods have been proposed to measure PMD; they are divided in two groups: (1) methods that involve performing measurements in the time domain and (2) methods involving measurements in the frequency domain. In the time domain, three methods are used: the modulation phase-shift method, the pulse delay method, and the interferometric method. In the frequency domain the methods are the fixed analyzer method; the Poincare arc method, which is also called the *Muller matrix* method, and the Jones matrix method.

The fixed analyzer method is the most commonly used method [8] (Fig. 2). In this method polarized light is injected into an optical fiber and then the optical power transmitted through the optical fiber and then through a polarizer, as a function of the wavelength, is measured. As the wavelength is changed, the power transmitted through the polarizer goes up and down. By counting the number of maximums and minimums or counting the number of zero crossings it is possible to determine the average PMD. When these zero crossings are counted, the test actually consists in measuring the rate at which the output state of polarization changes with the wavelength.

It is possible to use an OTDR to measure the PMD in optical fibers by interferometric methods [9].

In multimode optical fibers a wide range of wavelengths constitutes the lightpulse and hundreds of modes travel together. The electromagnetic interference of those mode components reduces the power intensity being transmitted and, because the modes all travel at different velocities, the pulse spreads in time. If the transmission rate of pulses being transmitted increases, the pulse spreading causes adjacent pulses to overlap in time, resulting in interpulse interference and errors in signal detection. The maximum frequency at which it is still possible to recover the information is called the *bandwidth* of the multimode optical fiber. Measurement of this parameter is done in two domains; (1) the time domain, where the pulse spreading time is measured; and (2) the frequency domain, where the maximum frequency is measured according to the power loss in system detection [10].

2.3. Mechanical Characterization

Ensuring adequate integrity of optical fibers is complex because of the wide range conditions of temperature, humidity, and mechanical stresses that are present in

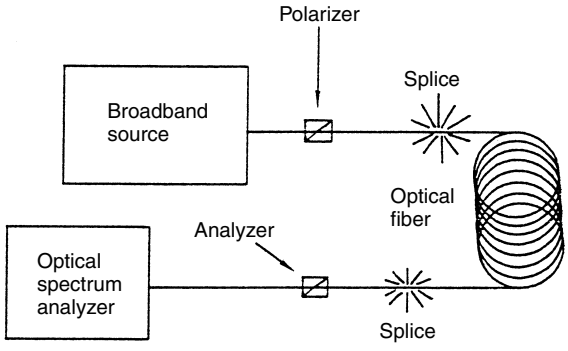


Figure 2. PMD measurement using the fixed analyzer method.

the manufacturing and in the optical fiber field. It is safe to state that the only way to determine the effectiveness of the optical fiber mechanical quality and how long it survives in the presence of moisture in the field is to wait until it breaks. But the lifetime can be very long, and predictability is needed. A short-term mechanical characterization can be useful to qualify the optical fiber, and a strength degradation model is necessary in order to predict failures.

After the manufacturing process it is necessary to embed the optical fibers in stronger structures in order to be installed and reliably operated during a minimum calculated period of time. These structures are called *optical fiber cables*.

Currently hundreds of optical fiber cable structures have been proposed for a large variety of applications, including submarine transoceanic cables, aerial cables, and underground cables. A common optical fiber cable design consists of a few basic elements. A strong tensile member of stranded steel wires, which are used to pull the cable during the installation process; a multilayer polymeric and metallic cover used to protect the optical fiber from contamination due to chemicals from the surrounding environment; and polymeric tubes with optical fibers inside it.

The silica glass chemically reacts with water, causing mechanical degradation of the glass optical fiber in a well-known phenomenon studied since the 1950s [11]. In the presence of stress and humidity a glass surface flaw can accelerate its growth and subsequently cause a fracture. Mechanical laboratory tests have shown that in a free water environment, as in tests performed with the sample immersed in liquid nitrogen, the strength of the optical fiber is at least 3 times higher compared with tests in a humid environment. Some authors found that this water glass corrosion is still possible in a stress-free situation for optical fibers coated with different polymeric materials [12,13].

Flaws in the surface of the glass are classified in two groups. The first group, *extrinsic defects*, represents a serious danger to the optical fiber. These flaws are many micrometers in length and in general originated during the manufacturing process, or were introduced in the glass surface by mechanical abrasion and handling. These defects can usually be identified after the optical fiber breaks, using microscopic techniques to analyze the broken surface [14]. The second group of flaws, called *intrinsic defects*, whose lengths are of the order of the silica glass structure (a few nanometers), cannot be observed using conventional microscopes, but studies using atomic force microscopy have revealed how they can affect the strength degradation process [15].

To mechanically qualify the optical fiber immediately after the manufacturing process, it is possible to do an optical fiber length tensile test in a proof testing machine [16]. The objective of this test is to eliminate large flaws that cause breaks during the optical cable manufacturing process, or in the cable lifetime. The test consists in applying a controlled tensile stress to the entire length of the optical fiber, which is done by using a system with pulleys and belts driven by electric motors (Fig. 3). In this test, when a large flaw is present in the length of the optical fiber that passes through the tensile proof test region, the applied stress activates the crack growth, causing a break in the fiber. The minimum stress level guaranteed for the survival of the optical fiber pieces after this destructive test is a function of the constant stress used in the tensile region, of the crack growth during the unloading time, and of the functional characteristics of the machine, such as the velocity of the fiber passing through the system.

The complete analysis of the proof testing must take into account the additional crack growth during the applied proof stress. Some of these flaws can be taken to their critical fracture size during the test causing breaks, and others can be very close to the critical size after the test. The proof test must be performed at the highest velocity of the fiber passing through the machine, in order to minimize the mechanical strength degradation.

After the proof stress area, the optical fiber is unloaded in the last pulley and an additional crack growth occurs, and this additional flaw growth during the unloading time

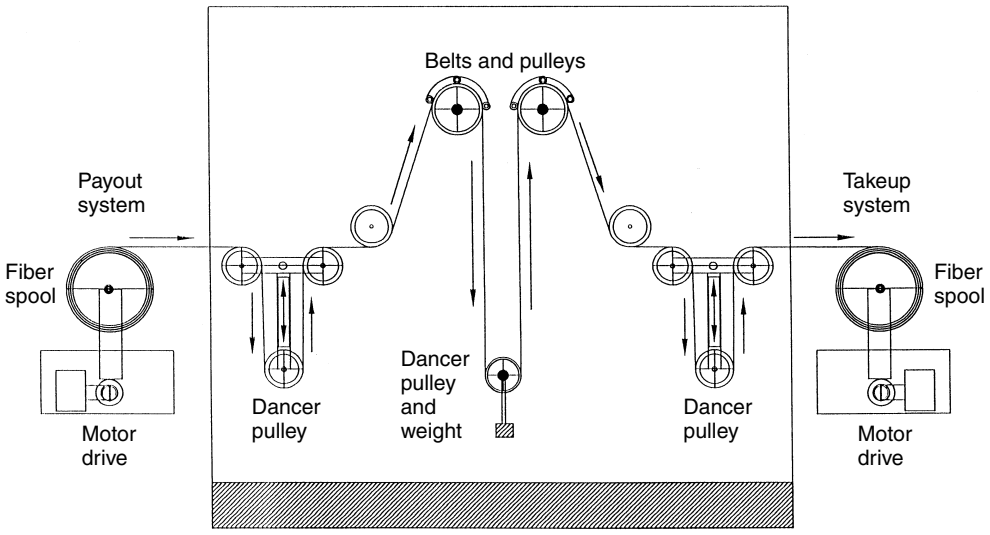


Figure 3. Optical fiber proof test machine.

reduces the guaranteed minimum strength to a level that is under the stress level applied to the tensile region. It is possible to calculate the minimum guaranteed strength by using the crack growth theory, the proof test specifications, and the velocity of the optical fiber passing through the system. The minimum strength after the proof test is a very useful parameter in order to guarantee a reliable cable manufacturing process, the optical fiber integrity during the installation of the cable, and the use of the cable.

After the proof test, additional mechanical characterization consists of measuring the strength of the optical fibers under tensile or bending stresses and their susceptibility to environmental changes. These tests, applied to a few samples removed from one extremity of the optical fiber, consist of applying a crescent force until the break. By assuming that the maximum force is connected with the geometric and physical parameters of the optical fiber, such as its diameter and length, and the elastic modulus of the silica glass, it is possible to calculate the maximum stress in the moment of the fracture. This is called the *strength* of the optical fiber. The results for these few samples removed from one extremity of the fiber are extrapolated to the entire length, assuming that the optical fiber presents the same mechanical characteristics in its entire length. The flaws distributed in the glass surface length can be compared to weak links in a chain, and a specific fracture probability distribution can be developed to describe the fracture event. This distribution is called the *Weibull fracture probability distribution* [17]. The Weibull model is applied to the strength of the tested samples from which the mean strength and the variability of the strength, referred as the m Weibull parameter, are calculated. These parameters are very useful for comparison of distinct optical fiber mechanical qualities.

In the past, mechanical apparatus used to test metal wires and pieces of plastics were adapted to test the tensile strength of pieces of optical fibers. Currently, mechanical apparatus have been developed specifically to test multiple and longer samples simultaneously. Laboratory tensile apparatus normally use a few meters of optical fibers, around 24 samples of 0.5 m each, tested one by one, to characterize many kilometers of an optical fiber (Fig. 4). Once the extremities of the optical fiber sample are securely held, a crescent force is applied until it breaks; this is known as the *dynamic fatigue test*. It is necessary to apply a high intensity force to hold the extremities of an optical fiber sample until it breaks. To avoid fractures to the extremities held by grip devices, it is common to wrap two or three turns around cylindrical pieces of metal, called *mandrels*. The force is applied by pulling the mandrels mechanically attached to the tensile machine. To guarantee no slippage of the optical fiber sample in the mandrel surface, it is necessary to use a double-face tape in the surface of the mandrel. Today this is the most commonly used method.

The problem with this approach is that it is not possible to know exactly what length of fiber is being tested, and part of the force is absorbed by the piece of optical fiber glued to the mandrel surface. In order to do fracture probability extrapolation to the untested piece of optical fiber using the Weibull model, it is essential to know the tested

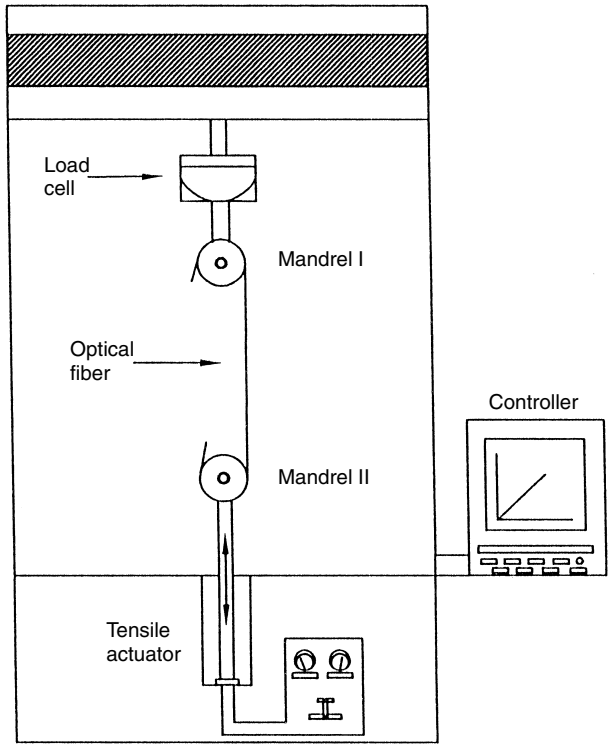


Figure 4. Optical fiber tensile tester.

length. To minimize this problem, one possibility is to test long-length samples of ($\geq$10 m). Another advantage with long length tests is that it gives more knowledge about the extrinsic defects present in the optical fiber [18].

Bending tests are performed in very small pieces of optical fiber samples, around 5 cm in length. In this test an optical fiber sample is introduced between two grooved steel plates. When one plate is pushed against the other, the curvature radius decreases until the optical fiber breaks. This is called the *bending dynamic fatigue test* (Fig. 5). By measuring the optical fiber curvature radius in the moment of the fracture, it is possible to know the maximum bending stress at the fracture (bending strength). This test can be performed on one by one sample or in multigrooved plates, up to 24 samples. Acoustic detectors can be used to precisely measure the distance

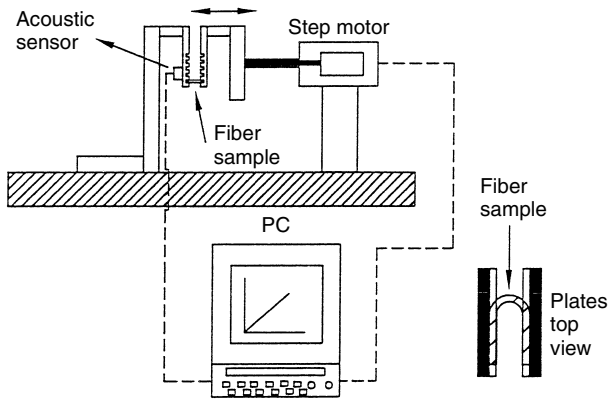


Figure 5. Two-point bending tester with acoustic detector.

of the plates that is connected with the curvature radius during the breaks [19].

The maximum bending stress is applied to a few square micrometers of the glass surface. The probability of finding an extrinsic defect in such areas is small, assuming that the optical fiber is of reasonable quality. This test is intended to measure the intrinsic strength of the glass surface. The test can be used to check how much the coating covers the glass, or the glass corrosion when the optical fiber is under chemical attack in harsh environments.

Another test in bending consists of holding the steel plates at a calculated distance, maintaining the bending stress constant, until the samples break. This is called a *static fatigue in bending*. In such a case, the survival time of the optical fiber under bending stress is measured. This bending test can be easily performed in harsh environments but cannot be done so easily under tension. In order to perform the static fatigue test for tensile, strength it is necessary to have a lot of space in order to accommodate the samples and the expensive equipment that is required to maintain constant environmental conditions in the area of the test.

The dimensions of the optical fiber core and its cladding refractive-index contrast confine the light being transmitted, but slight perturbations in geometry can cause a loss of power. Small mechanical forces acting on the surface of the optical fiber coating can be mechanically transmitted to the core, causing small perturbations in diameter. These perturbations are of the order of the fiber core diameter or less. Pressing a piece of fiber in a rough surface can be sufficient to change the optical fiber attenuation. This phenomenon is called *microbending sensitivity*. In cable design, this parameter must be minimized, taking in to account the maximum and minimum temperature fluctuations that can affect the format of the fiber inside the cable, the number of optical fibers in contact with each other, and the compatibility of the optical fiber with other cable structures. In order to measure the microbending sensitivity, attenuation measurements are used during susceptibility tests, such as temperature fluctuations and mechanical tests, with optical fiber cables.

It is still possible to measure the microbending sensitivity of a piece of optical fiber by monitoring the attenuation in a test where the optical fiber is compressed at known loads between two rough sandpaper sheets. This it can be done for a specific light wavelength using a power detector or for a broader wavelength range, using a spectral attenuation apparatus. Actually there does not exists standard procedure for this test.

When a piece of optical fiber is bent in a radius larger than that of the fiber, a small part of the light radiation can propagate through the core and leak to the cladding, causing the light to be lost. This phenomenon is called *macrobending attenuation*. Using metallic mandrels at different diameters and wrapping the same length of fiber around them, it is possible to measure relatively accurately the relationship between the fiber bent diameter and the optical attenuation. This measurement can be done for a specific wavelength or for a wide wavelength range, using a spectral attenuation apparatus. This is useful information in cable design and installation of optical fiber systems.

Silica glass is susceptible to environmental humidity, and polymer-coated optical fibers reach equilibrium with the environment in less than an hour. The plastic polymer coating works as a net on the surface of the glass, as it is permeable to the water molecules. Environmental humidity fluctuations, around 10% relative humidity (RH), can affect the optical fiber mechanical tests. Other liquids and gases can affect the strength performance of the fiber. Basic substances are more supportive of glass corrosion. All mechanical tests must be performed in a humidity-controlled environment, after the optical fiber is in equilibrium with the environment. Environmental recommendations are present in all standardized mechanical tests.

The glass susceptibility to the environment can be measured by performing strength tests at different stress rates. Stress rates are related to the force variation in time applied to a sample. Using different stress rates, it is possible to isolate the strength degradation caused by the presence of humidity from the strength degradation caused by the stress alone. The parameter that describes how rapidly the strength degradation occurs in the presence of humidity is called *optical fiber fatigue n* [20]. This parameter is obtained using at least four groups of samples tested at four different stress rates. In terms of fiber reliability, this is the most important parameter. The n number is the power that will be used to calculate the optical fiber lifetime.

An important aspect of the optical fiber mechanical degradation is related to the survival time of the optical fiber in the field. This is called the *optical fiber reliability* [21]. For this calculation it is necessary to know well the *fiber stress history* (FSH), which *FSH* is the accumulation of all the stresses applied to the optical fiber during the different transmission system construction and use: optical fiber manufacturing, proof testing, cabling, installation, and application. The FSH parameters using time fracture probability to estimate the fiber lifetime after installation are the minimum guaranteed strength after the proof test, as described above; the optical fiber fatigue parameter n; and the low−stress break flaw distribution, measured using the tensile test For long lengths of the proof tested optical fibers. To complete the necessary calculations, it is necessary to use a crack growth model when the optical fiber is under low stress in the presence of humidity. Normally, a *power-law crack growth model* is used, which assumes a power relation between the crack growth velocity, the parameters of the material, and the applied stress. The fatigue parameter n is the power variable in this model. Other crack growth models were proposed in the last decade but the power law is the most reliable and treatable model. The power-law model is used in national and international standards.

To complete the mechanical characterization, it is necessary to measure the qualities of the optical fiber coatings. The most important characteristics of the fiber coatings are (1) the primary coating must have a good chemical reactivity with the glass surface, to protect the glass from moisture; (2) the primary polymer coating must have lower elastic modulus, to absorb external small stress that causes attenuation by microbending sensitivity; and

(3) the secondary polymer coating must have a higher elasticity modulus to improve the resistance to the abrasion.

It is possible to join the extremities of optical fibers by keeping the attenuation and the mechanical qualities under control. This operation is called *optical fiber splice*. Splices are common in optical fiber cable installation and in optical fiber cable repairs. An important step in the splice procedure consists in removing the coating of the fiber extremities. The glass must be completely clean and well cleaved to perform the thermal fusion of the extremities. If the primary coating is overconnected with the glass, the operation will be difficult and can affect the splice quality. If the primary polymer coating is loose on the surface of the fiber, or not completely cured, it does not promote the necessary protection [22].

Standard procedures to measure coating quality and how it affects splice performance are under study by standards committees and include methods to measure the simplicity of removing the coating of the optical fiber, strip-force and pullout force techniques, and methods to measure the elasticity modulus of the primary coating.

BIOGRAPHY

Gilberto Camilo received his D.Sc. degree from Campinas University, Sao Paulo, Brazil, in 1991. He was employed by PIRELLI Optical Fibers, Sao Paulo, Brazil, between 1987 and 1991. He taught physics and mechanical engineering at Goias and CEFET-Parana Federal University, Brazil, during 1991–1998. He was a Post-Doctoral Fellow at Rutgers University, and worked on the Ceramic and Materials Engineering — Optical Fibers Project during 1994–1995. He was with Furukawa Optical Fiber Cables, Parana State, Brazil, during 1997–1998 and with Alcatel, in charge of Optical Fibers Mechanical Reliability at Claremont, North Carolina, during 1998–2001. Since 2001 he has been with OmniGuide Communications, at Cambridge, Massachusetts, as Optical Fiber Reliability Specialist. He has been active in the TIA/EIA Optical Fibers Standard Committees since 1998. He is a member of the OSA, IEEE, and SPIE. Dr. Camilo has published over 50 papers in the area of optical fiber mechanical Characterization. His main interest areas are optical and mechanical characterization and reliability of optical fibers and optical cables.

BIBLIOGRAPHY

1. T. Kaino, Plastic optical fibers, *Proc. SPIE* **CR63**: 164–187 (1996).
2. J. A. Harrington, Infrared optical fibers, in *Handbook of Optics*, Optical Society of America, McGraw-Hill, 2001.
3. S. G. Johnson et al., Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers, *Optics Express* **9**: 748–779 (2001).
4. W. J. Stewart, A new technique for measuring the refractive index profiles of graded optical fibers, *Proc. IOOC Tech. Digest* 395–398 (1977).
5. R. Girbig and M. Hoffart, Highly accurate backscatter measurement in the quality control of the cabling of single-mode fibers, *Proc. IWCS* **38**: 480–485 (1989).
6. A. J. Barlow, R. S. Jones, and K. W. Forsyth, Technique for direct measurement of single-mode fiber chromatic dispersion, *J. Lightwave Technol.* **LT-5**: 1207–1213 (1987).
7. B. Costa, M. Puleo, and E. Vezzoni, Phase-shift technique for the measurement of chromatic dispersion single-mode fibers using LED's, *Electron. Lett.* **19**: 1074–1076 (1983).
8. C. D. Poole and D. L. Favin, Polarization-mode dispersion measurements based on transmission spectra through a polarizer, *J. Lightwave Technol.* **LT-12**: 917–922 1994.
9. A. J. Rogers, Polarization-optical time domain reflectometry: A technique for the measurement of field distributions, *Appl. Opt.* **20**: 1060–1074 (1981).
10. A. H. Hartog et al., Comparison of measured and predicted bandwidth of graded-index multimode fibers, *J. Lightwave Technol.* **QE-18**: 825–838 (1982).
11. R. J. Charles, Static fatigue of glass I and II, *J. App. Phys.* **29**: 1549–1560 (1958).
12. N. Evanno, M. Poulain, and A. Gouronnec, Optical fiber lifetime in harsh conditions, *Proc. SPIE* **3848**: 70–76 (1999).
13. P. Regio, P. Motta, and S. Apone, Influence of the coating in mechanical behavior of aged optical fiber, *Proc. Eurocable 97*, 1997.
14. L. K. Baker and G. S. Glaesemann, Break source analysis: alternate mirror measurement method, *Proc. IWCS* **47**: 933–937 (1998).
15. G. Camilo, C. Turnbill, and B. Overton, Glass corrosion in commercial optical fibers with defective coatings, *Proc. IWCS* (in press).
16. T. A. Hanson, Analysis of the proof test with power law assumptions, *Proc. SPIE* **2074**: 108–119 (1994).
17. W. Weibull, A statistical theory of the strength of materials, *Proc. Royal Swed. Inst. Eng. Res.* **151**: 1–45 (1939).
18. W. Griffioen, Mechanical lifetime of optical fibers, *Proc. European Fibre Optic Communications and Networks*, 1994, pp. 164–168.
19. M. J. Matthewson, C. R. Kurkjian, and S. T. Gulati, Strength measurement of optical fiber by bending, *J. Am. Cer. Soc.* **69**: 815–821 (1986).
20. V. V. Rondinella and M. J. Matthewson, Ionic effects on silica optical fiber strength and models for fatigue, *Proc. SPIE* **1366**: 1–8 (1990).
21. W. Griffioen, *Optical Fiber Mechanical Reliability*, Eindhoven Univ. Technology, 1994.
22. G. Camilo and B. Overton, Evolution of fiber strength after draw, *Proc. NFOEC* **17**: 143–153 (2001).

CHIRP MODULATION

DIRK DAHLHAUS
Communication Technology Laboratory
Zurich, Switzerland

1. INTRODUCTION

Chirp modulation (CM) represents a special type of spread-spectrum signaling where a carrier signal is modulated in two ways. The primary modulation is carried

out in the complex baseband and constitutes the usual formats such as phase shift keying (PSK), pulse position modulation (PPM), or binary orthogonal keying (BOK). The primary modulation is combined with a secondary modulation for spectrum spreading. For a data rate T^{-1} with T denoting the symbol duration, the occupied Fourier bandwidth B exceeds T^{-1} considerably; that is, in general, the time-bandwidth product $TB \gg 1$. The spreading is advantageous in frequency-selective fading channels often encountered in wireless or mobile radio systems. If the occupied spectrum is larger than the coherence bandwidth of the channel, the transmission is more robust against the fading because of the resulting frequency diversity. In addition, as shown by Berni and Gregg [1], CM is resistant to the Doppler effect arising in time-variant scenarios typically encountered in mobile radio applications. CM signals have been first proposed by Winkler [2] for their high robustness against distortions and different types of interference.

In most cases, "chirp modulation" refers to a sinusoidal signal of duration T whose instantaneous frequency changes linearly in time t between the lower frequency $f_1 = f_0 - B/2$ and the upper frequency $f_2 = f_0 + B/2$, where f_0 denotes the carrier frequency of the signal. In its simplest form, CM is used in combination with binary signaling as primary modulation. To transmit a logical 0 using a binary CM signal, an "upchirp" is used, which corresponds to a linear frequency sweep from f_1 to f_2. A logical 1 is transmitted correspondingly as a "downchirp," a linear frequency sweep from f_2 to f_1. For sufficiently large values of the time-bandwidth product, the upchirp and downchirp signals transmitted in a common band constitute the aforementioned BOK. The name *chirp* has been given to such signals by Bell Telephone Laboratories because of the resemblance to a sound heard in nature [2]. CM signals have their roots in radar applications where one of the most important observations states that range resolution and accuracy are functions of the signal bandwidth, and not of the transmitted pulsewidth [3].

In Section 2, important properties of CM signals are discussed including the form of linear frequency-modulated (FM) signals, the signal spectrum, the matched-filter (MF) characteristics, measures for sidelobe reduction, and modulation schemes. In Section 3, the performance of the different schemes is analyzed. Section 4 describes different ways to implement CM systems including surface acoustic wave (SAW) devices as well as digital baseband techniques. Other aspects related to CM in communication systems are discussed in Section 5.

2. PROPERTIES OF CM SIGNALS

2.1. Time and Frequency Representation

The general form of FM bandpass signals to be considered is given by

$$s(t) = a(t)\cos(\omega_0 t + \theta(t)), \qquad -\frac{T}{2} < t < \frac{T}{2} \tag{1}$$

where $f_0 = \omega_0/2\pi$ and $\theta(t)$ denote the carrier frequency and the signal phase, respectively. The envelope $a(t)$ can be used as a weighting function to improve the autocorrelation properties of $s(t)$ as discussed in Section 2.3. Here, it is first assumed a rectangular function over the interval $[-T/2, T/2]$. The instantaneous frequency is defined by

$$f(t) = \frac{1}{2\pi}\left(\frac{\omega_0 + d\theta}{dt}\right) \tag{2}$$

For linear FM signals, we obtain

$$f(t) = f_0 + \mu t, \qquad -\frac{T}{2} < t < \frac{T}{2} \tag{3}$$

with $\mu \in \mathcal{R}$ denoting the dispersive slope or rate of the chirp signal. From (3), $f(t)$ varies between the lower (resp. upper) frequencies

$$f_1 = f_0 - |\mu|\frac{T}{2}$$

$$f_2 = f_0 + |\mu|\frac{T}{2}$$

over a range $B = |\mu|T$[1], where $\mu > 0$ and $\mu < 0$ denote an upchirp and (resp. downchirp) signal, as shown in Fig. 1. The signal phase results in

$$\theta(t) = \pi\mu t^2 + \theta_0, \qquad -\frac{T}{2} < t < \frac{T}{2} \tag{4}$$

with a suitable initial phase value θ_0. For the usually valid narrowband assumption $f_0 \gg |\mu|T$ [3] and $\theta_0 = 0$, the spectrum of an upchirp $S(\omega) = \int_{\mathcal{R}} s(t)\exp[-\jmath\omega t]\,d\omega$ of $s(t)$ is given by [3]

$$S(\omega) = \frac{1}{2\sqrt{2\mu}}\exp\left[-\jmath\frac{(\omega-\omega_0)^2}{4\pi\mu}\right] \times \left[\mathcal{C}(X_+) + \jmath\mathcal{S}(X_+) + \mathcal{C}(X_-) + \jmath\mathcal{S}(X_-)\right]$$

where

$$\mathcal{C}(X) = \int_0^X \cos\left(\frac{\pi y^2}{2}\right)dy, \qquad \mathcal{S}(X) = \int_0^X \sin\left(\frac{\pi y^2}{2}\right)dy$$

are Fresnel integrals and the integral limits are given by

$$X_\pm = \frac{\pi\mu T \pm (\omega - \omega_0)}{\pi\sqrt{2\mu}}$$

On substituting

$$\mu = \frac{B}{T}, \qquad \omega - \omega_0 = n\pi B$$

with the normalized frequency n, we obtain

$$X_\pm = \frac{1 \pm n}{\sqrt{2}}\sqrt{TB}$$

[1] The bandwidth occupied by $s(t)$ is larger than B, but approaches B for $TB \to \infty$.

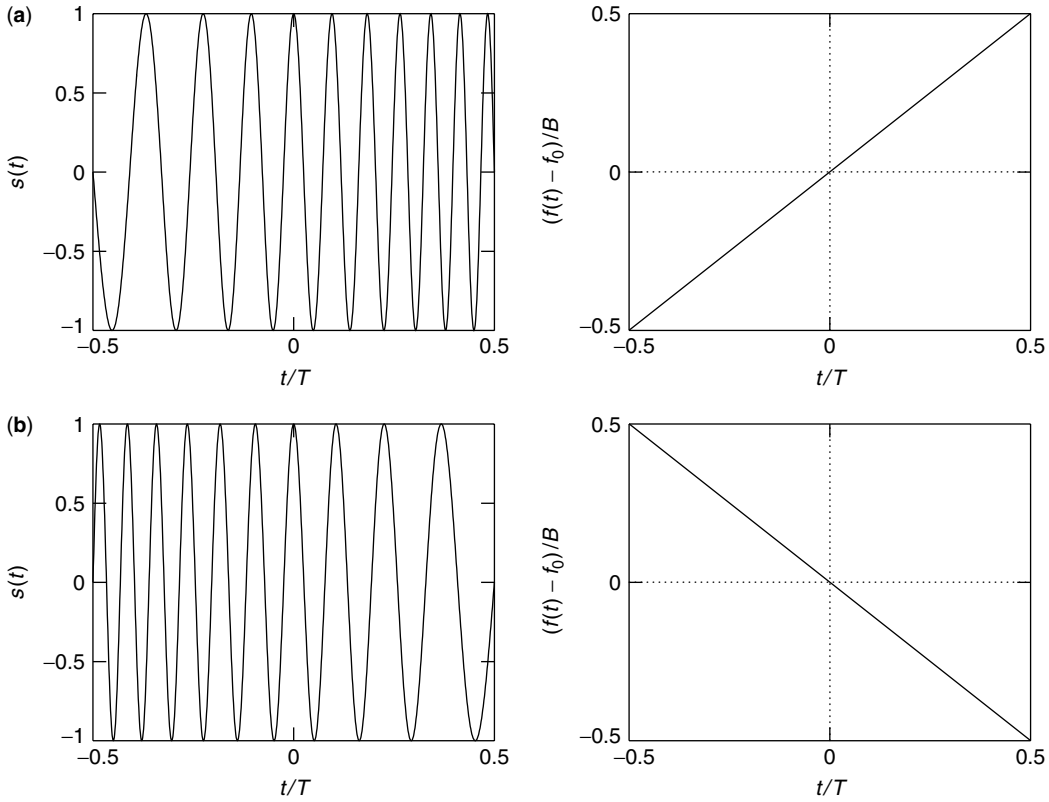


Figure 1. Chirp signals and corresponding instantaneous frequencies: (**a**) upchirp; (**b**) downchirp.

specifically, the amplitude spectrum of $s(t)$

$$|S(\omega)| = \frac{1}{2\sqrt{2\mu}}\sqrt{\left[\mathcal{C}(X_+) + \mathcal{C}(X_-)\right]^2 + \left[\mathcal{S}(X_+) + \mathcal{S}(X_-)\right]^2} \tag{5}$$

in Fig. 2 depends solely on the time–bandwidth product TB. The spectrum characteristics in the band center are mainly determined by the so-called Fresnel ripples arising from the Fresnel integrals. The height of the ripples increase for decreasing TB while the spectrum is asymptotically rectangular for $TB \to \infty$. It can be shown that the amplitude spectrum in (5) is valid also for a downchirp if μ is replaced by $|\mu|$ in all terms containing the dispersive slope variable. More information on the phase spectrum for upchirp and downchirp signals can be found in the treatise by Cook and Bernfeld [3]. For $TB \gg 1$, the described binary modulation is sometimes termed BOK since the normalized cross-correlation of upchirp and downchirp signals is almost zero. The correlation properties of the chirp signals are considered in the next section.

2.2. Matched Filtering

The MF providing a sufficient statistic for symbol detection in additive white Gaussian noise (AWGN) and maximizing the signal-to-noise ratio (SNR) for a perfectly synchronized receiver is termed a *compression filter* in linear FM systems. For an upchirp signal in (1), the impulse response of the MF is a downchirp given by

$$h(t) = ks(-t) = k\cos\left(\omega_0 t - \pi\mu t^2\right), \qquad -\frac{T}{2} < t < \frac{T}{2}$$

where $k = 2\sqrt{\mu}$ is chosen for a unity gain of the MF at $f = f_0$. Here, the MF output is considered for a channel with a Doppler shift f_D. This situation studied extensively in radar applications [3] arises in mobile communications, such as in time-varying line-of-sight channels. The MF output signal in the absence of thermal receiver noise is given by

$$\begin{aligned} g(t, f_D) &= \int_{\mathcal{R}} s(\tau)\, h(t-\tau)\, d\tau \\ &= 2\sqrt{\mu}\int_a^b \cos\left[(\omega_0 + 2\pi f_D)\,\tau + \pi\mu\tau^2\right] \\ &\quad \times \cos\left[\omega_0(t-\tau) - \pi\mu(t-\tau)^2\right] d\tau \\ &= \begin{cases} \dfrac{\sqrt{\mu}}{\pi}\,\dfrac{\sin\left(\pi\left(f_D + \mu t\right)\left(T - |t|\right)\right)}{f_D + \mu t} \\ \quad \times \cos\left(2\pi\left(\dfrac{f_0 + f_D}{2}\right)t\right), & -T < t < T \\ 0 & |t| \geq T \end{cases} \end{aligned} \tag{6}$$

with

$$a = -\frac{T}{2} + t, \quad b = \frac{T}{2} \qquad \text{for } t \geq 0$$

$$a = -\frac{T}{2}, \qquad b = \frac{T}{2} + t \quad \text{for } t < 0$$

The frequency shift of $f_D/2$ in (6) can be easily understood from considering the spectra of the input and MF signals [3]. Figure 3 shows the envelope of $g(t) =$

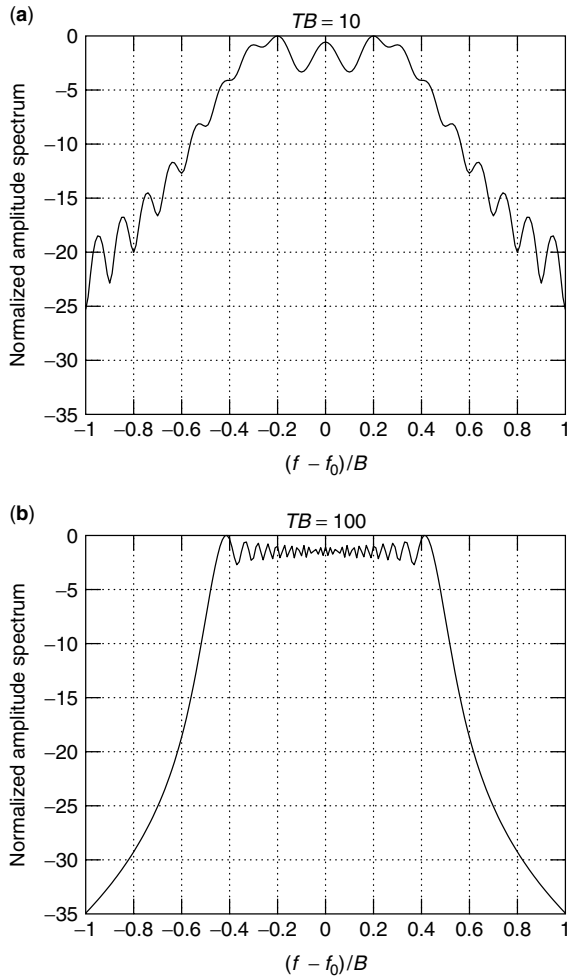


Figure 2. Amplitude spectra for different time–bandwidth product values: (**a**) $TB = 10$; (**b**) $TB = 100$.

$g(t, f_D = 0)$ for different values of TB. As can be concluded from (6), the height of the correlation peak is $T\sqrt{\mu} = \sqrt{TB}$. Thus, the *compression gain*, defined as the ratio of the peak value to the chirp amplitude, is $G = 10\log(TB)$ dB. For $TB \gg 1$, it can be shown that the width of the mainlobe is $2/B$ while the minimum sidelobe suppression, expressed by the ratio G_{SL} of the mainlobe and the first sidelobe values, is approximately 13.3 dB. This ratio essentially determines the system robustness against intersymbol interference (ISI) in frequency-selective fading channels arising from multipath propagation of the transmitted signal. Measures for reducing the sidelobes are discussed in the next section. Figure 4 shows the shape of $g(t, f_D)$ for different values f_D/B. Obviously, the correlation peak is shifted, attenuated, and spread as compared to Fig. 3. Proceeding as in (6), it is readily shown that the cross-correlation for $f_D = 0$ and $t = 0$ between upchirp and downchirp equals $(C\sqrt{TB})$, which approaches $\frac{1}{2}$ for large TB. In this case, the ratio of the cross-correlation and autocorrelation equals $1/(2\sqrt{TB})$ which justifies the assumption of BOK for large TB.

2.3. Sidelobe Reduction

One way to increase the dynamic range of the pulse compression is to modify the signal spectrum of $g(t)$

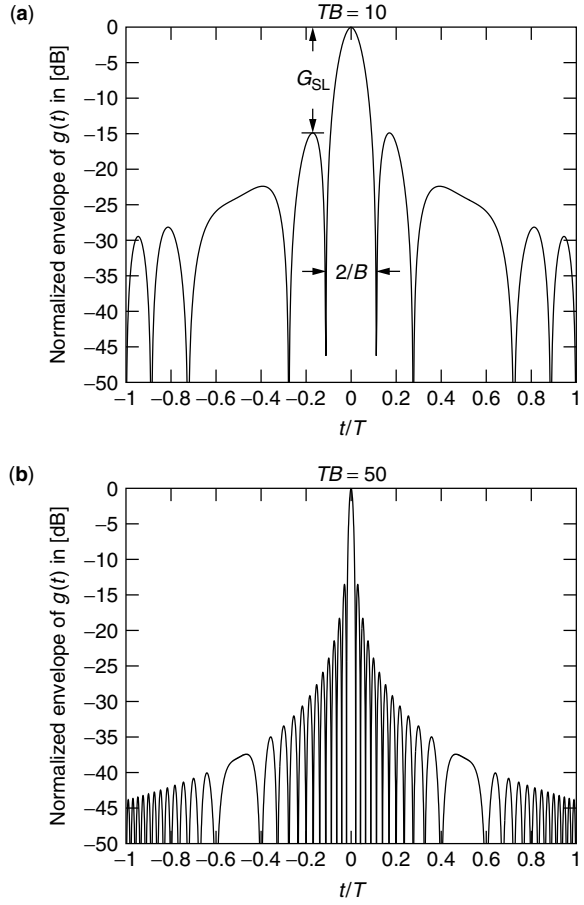


Figure 3. Normalized envelope of the MF output signal $g(t)$ for different time–bandwidth product values: (**a**) $TB = 10$ (**b**) $TB = 50$.

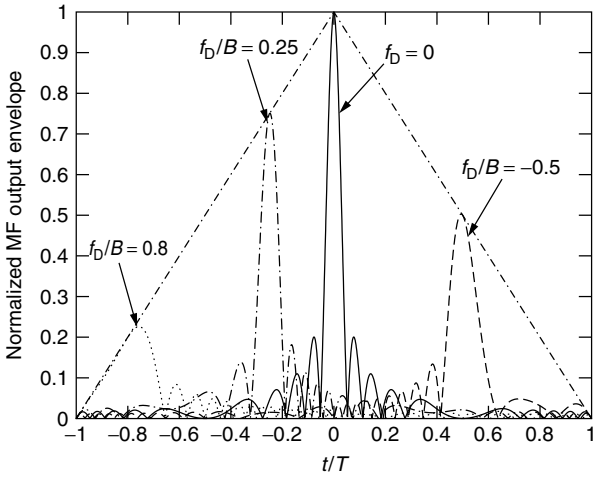


Figure 4. MF output signal $g(t, f_D)$ for different values f_D/B and $TB = 20$.

for decreasing the sidelobe levels. The corresponding weighting can be implemented in the frequency or time domain. The optimum distribution function can be derived from an analogous problem in antenna theory treated by Dolph [4] and Van der Mass [5] that targets the narrowest beamwidth of a broadside antenna array for a desired

sidelobe level. However, the resulting Dolph–Chebyshev weighting function has infinite power and is thus physically not realizable. Approximations to the optimum solution are provided by the Taylor functions and modified Taylor functions [3]. The latter contain the so-called Hamming function

$$W_H(f) = 0.08 + 0.92\cos^2\left(\pi\frac{f-f_0}{B}\right), \qquad |f-f_0| \le \frac{B}{2}$$

as a special case. It can be shown that this weighting results in a sidelobe suppression of −42.8 dB, a spreading of the mainlobe by a factor of 1.47 and a loss in compression gain of 1.34 dB. The influence of the Fresnel ripples on Hamming weighted chirp compression is treated by Kowatsch and others [6,7] who considered both time and frequency weighting as well as Doppler shifts. A considerable reduction of the Fresnel ripples can be achieved by using a Tukey window, where the undesirable mainlobe width increase is only moderate [6].

2.4. Modulation Schemes

Although the spectrum spreading in CM can in principle be combined with any baseband signaling, the choice of the primary modulation in a practical system is restricted by, For example, performance requirements, an efficient implementation, and component imperfections. Here, some modulation schemes are described that can be used in CM and are discussed in Section 3 in terms of the achievable bit error rate (BER). As in Proakis' text [8], linear modulation is distinguished from nonlinear modulation with memory.

2.4.1. Linear Modulation. The chirp pulse in (1) can be implemented efficiently by a SAW device (see Section 4). The properties of this technology rule out certain modulation schemes, such as amplitude shift keying (ASK). The problem with the latter scheme is the high dependence of the output power on the rising time of the broadband pulse exciting the filter. This problem is usually circumvented if PSK, BOK, or PPM is employed [9,10]. El-Khamy and Shaaban [11], match μ to the dispersion parameters of the communication channel. They show that for a channel with a second-order polynomial phase spectrum and a partially coherent detection, μ should be chosen according to $TB = 2.65$, which minimizes the BER. The aforementioned BOK requires upchirp and downchirp filters at both the transmitter and the receiver. Clearly, the bit error performance depends critically on the sidelobe level and the cross-correlation properties of the employed chirp signals as well as the delay dispersion of the channel. Some of the problems can be solved if PPM signals are employed using only one chirp signal type (e.g., an upchirp). In PPM [9], the binary signals are orthogonal, the cross-correlation problem does not arise, and only one chirp filter has to be implemented in the transmitter and the receiver. In case of a logical 1, the chirp is sent Δt before the reference clock, while for a logical 0, the chirp is delayed by Δt. The system performance is determined by the value Δt and the channel delay dispersion. Another standard modulation method is differentially encoded quaternary PSK (DQPSK), which allows for a differential demodulation without carrier phase estimation. Usually, $\pi/4$ DQPSK offering reduced envelope fluctuations as compared to ordinary DQPSK is employed. Again, the performance limiting factors are the sidelobe levels and the channel delay dispersion. In general, the data rate can be increased by applying overlapping signal pulses that can be resolved at the receiver for sufficient sidelobe reduction capabilities of the employed MF output signals.

2.4.2. Nonlinear Modulation with Memory. CM for binary signaling Hirt and Pasupathy combined with full-response (FR) continuous phase modulation and termed FR *continuous-phase chirp modulation* (FR-CPCM). The idea of FR-CPCM is the improvement of the independent bit-by-bit detection in conventional CM systems by observing the phase-constrained received signal over two or more bit intervals prior to bit detection. In CPCM, instead of (4), the transmitted signal phase for $t \ge 0$ is given by

$$\theta_k(t) = a_k\psi(t-kT) + \pi q\sum_{r=0}^{k-1} a_r + \theta_{-1}, \quad kT \le t \le (k+1)T \tag{7}$$

where $a_i = \pm 1$ denotes the binary data, $k = 0, 1, \ldots$ and the phase function is defined as

$$\psi(t) = \begin{cases} 0, & t \le 0, t > T \\ \pi\left(h\dfrac{t}{T} - w\left(\dfrac{t}{T}\right)^2\right), & 0 \le t \le T \\ \pi q = \pi(h-w), & t = T \end{cases}$$

where h and w represent the normalized initial peak-to-peak frequency deviation and the frequency sweep width, respectively. For the case of coherent detection, it has been shown [12] that a receiver with an observation of 2 bits provides a good compromise between signal-to-noise ratio (SNR) gain and system complexity.

As conjectured by Hirt and Pasupathy [12], the system performance can be further improved by considering multimode continuous phase systems. Raveendra [13] investigated the approach of varying the modulation of a continuous-phase FSK. Here, the phase in (7) is replaced by

$$\theta(t) = \sum_{i=1}^{n} a_i\psi_i(t-(i-1)T), \qquad 0 \le t \le nT \tag{8}$$

where the phase functions in (8) depend now on the symbol interval i according to

$$\psi_i(t) = \begin{cases} 0, & t \le 0 \\ \pi\left(h_i\dfrac{t}{T} - w_i\left(\dfrac{t}{T}\right)^2\right), & 0 \le t \le T \\ \pi q_i = \pi(h_i - w_i), & t \ge T \end{cases}$$

While in conventional monomode continuous-phase chirp transmission $q_i = q$ and $w_i = w$ for $i = 1, 2, \ldots$, the (q_i, w_i) now form a sequence of sets with period K, specifically, $(q_i, w_i) = (q_{i+K}, w_{i+K})$.

Fonseka [14] employed partial-response CPCM (PR-CPCM) signals in an attempt to increase the minimum distance $d_{\min}$ of the signals that determines the system performance. At the same time, the spectrum is to be kept flat so that the system is robust against jammers in the transmission band. The increase in the number of states in PR-CPCM as compared to FR-CPCM is an important issue. If in the latter $(h - w)$ is expressed as the ratio of two relatively prime integers as $h - w = \ell/m$, the m possible states during any interval can be represented by m evenly spaced phase states. In PR-CPCM with LT denoting the support of the baseband frequency pulse, the number of phase states depends on the individual values of h and w. When h and w are expressed as ratios of integers $h/L = \ell_1/m$ and $w/L^2 = \ell_2/m$ with the smallest common denominator m, the number of phase states during any interval is m. In view of the symbol states arising from the $(L - 1)$ previous symbols, the total number of states in PR-CPCM is $m2^{(L-1)}$.

2.4.3. Multiple Access. The aforementioned modulation schemes are designed for the case where receiver thermal noise (and possibly some narrowband interfering signals) represent the only disturbances in the bandwidth occupied by the signal. If the spectrum is shared among M simultaneously transmitting users, the resulting multiple-access interference (MAI) among the users increases the BER as compared to the single-user system. To avoid complex signal processing schemes for interference mitigation at the receiver, the signal formats have to be chosen carefully in order to limit the MAI. Different approaches based on CM have been proposed and are discussed below.

Takeuchi and Yamanouchi [15] assigned consecutive bits transmitted by CM with DPSK to different users; thus, time-division multiple access (TDMA) is applied here. The sidelobe level suppression and processing gain are 30 and 19 dB, respectively, and the system is implemented using SAW devices. Nonlinear CM is applied in order to obtain a flat amplitude spectrum within B.

A more sophisticated approach [16] for multiple access assigns $2M$ different instantaneous frequency functions $f(t)$ [cf. Eq. (2)] to M users employing binary signaling. The chirp duration T is split into two intervals of length $T/2$, and each of the proposed multiuser chirp signals of duration T is characterized by two different slopes in the two intervals. This approach is a straightforward extension of the chirp signals used in BOK to a M-user system where all signals occupy a common bandwidth B. Since the detection of the M symbols requires $2M$ different chirp matched filters, however, the complexity of the receiver is relatively high as compared to that reported by Takeuchi and Yamanochi [15].

Frequency-hopped code-division multiple access (FH-CDMA) is employed in another study [17]. Using basically the multiuser chirp signals of Ref. 16 in each time–frequency (TF) hop, the collision of different FH-CDMA user signals containing binary frequency-shift keying (FSK) symbols in the same TF hop can be resolved. Still, however, $2M$ different chirp matched filters are required and all chirp filters have to be changed after a new user has entered the system.

Improved flexibility and a simple receiver design are the main objectives of the multiple-access scheme in another paper [18], where transmission from a base station to multiple users over a multipath channel is considered. Here, the advantages of synchronous binary direct-sequence (DS) CDMA are combined with chirp signaling. The positive and negative chip pulses $\pm c(t)$ normally used in DS-CDMA for spreading the signal spectrum are replaced by the upchirp and downchirp signals in Section 2.1. It is shown that the quasiorthogonality of these signals allows for a noncoherent detection followed by a postdetection integrator (PDI) whose output is sampled to provide estimates of the superimposed code sequences of the different users. Finally, the information bit of a certain user is estimated from the correlation of the sample sequence with the user code. Only two different chirp filters are required, and the multiple access can be fully controlled by assigning codes in the digital domain.

3. PERFORMANCE ANALYSIS

For the case of linear modulation, the BER of a BOK system has been evaluated [10] for nonoverlapping chirp signals with $T = 2\,\mu\text{s}$, $B = 80\,\text{MHz}$, and a sidelobe suppression of $G_{SL} = 42\,\text{dB}$, resulting in a bit rate of 500 kbps (kilobits per second). To increase the data rate to 2 Mbps, the chirp pulses are allowed to overlap. Figure 5 shows the resulting BER as a function of the SNR $\gamma = E_b/N_0$ in an AWGN channel where E_b and N_0 denote the bit energy and the noise power spectral density, respectively. It has been pointed out [10] that multipath propagation limits the BER performance where sequences of consecutive upchirp or downchirp signals are to be avoided. BER results corresponding to Fig. 5 for $\pi/4$-DQPSK can be found in Gugler's thesis [9].

Hirt and Pasupathy [12] investigated the BER performance of FR-CPCM for coherent detection. It has been shown that a receiver with an observation of 2 bits provides a good compromise between SNR gain and system complexity. In this case, the optimum choice $(q, w) = (0.28, 1.85)$ gives an SNR gain of 1.75 dB over the optimum coherent 1-bit chirp receiver with $(q, w) = (0.35, 1.55)$.

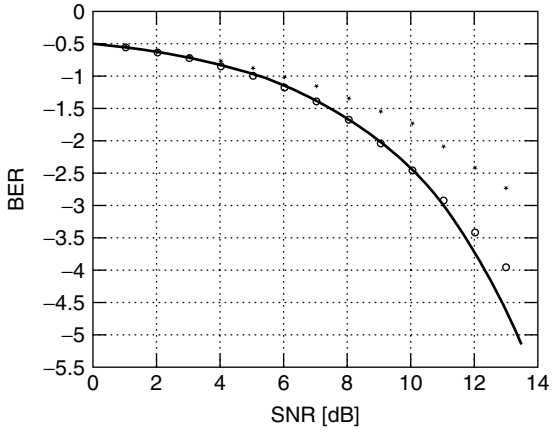


Figure 5. BER for BOK and different data rates: lower bound for orthogonal signals [8] (−), nonoverlapping 2-μs-long chirp signals (○), and overlapping 2-μs-long chirp signals with a data rate of 2 Mbps (∗) (figure taken from Ref. 10).

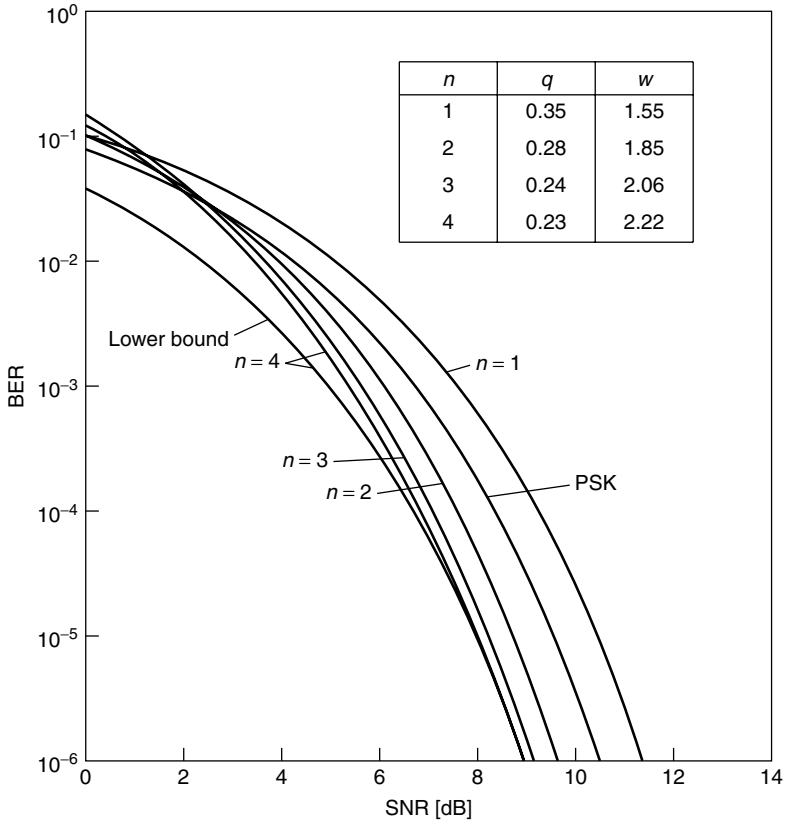


Figure 6. BER bounds for a coherent FR-CPCM receiver (figure taken from Ref. 12).

Upper BER bounds for different values of the observation length n together with the optimal values for q and w are shown in Fig. 6. The lower bound for $n = 4$ indicates the tightness of the bounds for increasing SNR values. In the context of a possible implementation, a simple suboptimum average matched filter (AMF) is shown to provide binary PSK performance for an optimum 2-bit observation. In another study, Hirt and Pasupathy [19] investigated the noncoherent detection case for FR-CPCM and showed the 3-bit noncoherent AMF receiver is to yield a 3-dB SNR gain over a wide range of signal parameters.

An investigation of multimode transmission performance [13] reveals that for an observation interval with $n = 5$, the optimum dual-mode chirp system, namely, $K = 2$, with $(q_1, w_1) = (0.3, 1.68)$ and $(q_2, w_2) = (0.5, 1.68)$ outperforms the optimum coherent 1-bit chirp receiver with $(q, w) = (0.35, 1.55)$ by 3.4 dB.

For Fonseka's [14] PR-CPCM with $L = 4$ and 24 states, the value of $d^2_{\min}$ can be increased by a factor of 2.15 as compared to the case $L = 1$ [14]. Furthermore, the PR-CPCM signals, are shown to have better spectral variations than conventional CM signals which indeed leads to the required robustness against jamming.

A CM TDMA scheme with DPSK [15] shows a BER that increases is only marginally for an increasing number of simultaneous users M. It is shown in simulations that the SNR loss for $M = 9$ as compared to $M = 1$ is only about 1 dB, where the latter case is about 2 dB worse than the lower BER bound for DPSK transmission.

The CM multiple-access scheme with the $2M$ different instantaneous frequency functions has been analyzed [16] for an AWGN channel using upper BER bounds for different values of M. As in the case of the CM TDMA scheme with DPSK, the bounds are relatively robust against different values of M. For $TB = 500$, the SNR loss of $M = 16$ as compared to $M = 1$ is only about 1 dB, and the BER decreases for increasing values of TB.

In another study [17], FH-CDMA with multirate chirp rate (MRC) signals is compared with a FH-CDMA scheme with FSK for an AWGN channel. It is observed that MRC-FH-CDMA is at least 2 dB better than the FSK-FH-CDMA scheme. For a BER of 10^{-1}, an increase of M to $M + 5$ results in a loss of 0.6 dB for MRC-FH-CDMA and 1.5 dB for FSK-FH-CDMA, respectively.

Kocian and Dahlhaus [18] observed bounds on the BER performance of the CDMA scheme described at the end of Section 2.4.3 with PDI and derived a square-law envelope detector in a non-frequency-selective channel (NFSC). The BER in a frequency-selective channel (FSC) with Rayleigh fading and an exponential power delay profile is depicted in Fig. 7 for $M = 1$ and $M = 16$ users, respectively, for a CDMA codelength of $N_c = 16$. For comparison, the BER of the optimum noncoherent detector (ND) has been included in Fig. 7. If the data rate is increased, the decision variable is corrupted by ISI and MAI, and the BER starts to saturate for increasing γ.

The preceding analysis has assumed a perfect channel state information at the receiver. Partially coherent detection of CPCM signals is considered in another study [20], while parameter estimation of chirp signals has been treated by Kay and others [21,22].

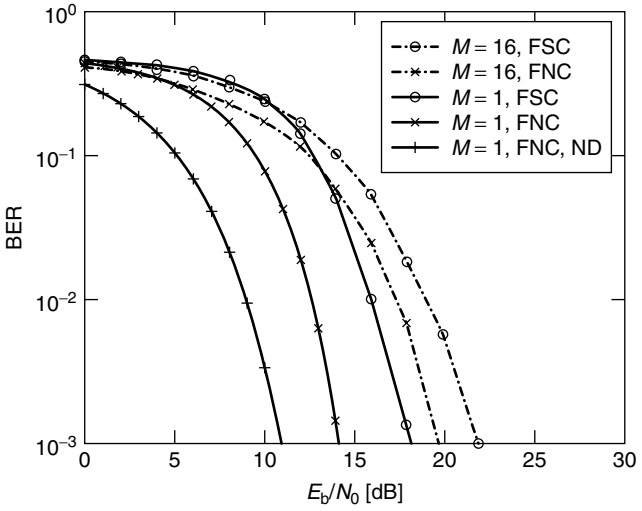


Figure 7. Mean BER for $N_c = 16$ in a multipath fading channels for different numbers M of users (figure taken from Ref. 18).

4. IMPLEMENTATION ISSUES

There are different ways to implement communication systems using CM. The two most prominent ones, namely SAW filters and direct digital frequency synthesizers (DDFS), are outlined below. Approaches based on voltage-controlled oscillators with appropriate function generators as well as excitation of a conjugate MF network with an impulse are described in another treatise [3].

In SAW filters there is no need for complex digital baseband signal processing. SAW filters are well suited for today's wireless communications because of their high performance, small size, and low cost [23]. On the other hand, since the CM parameters specify the form of the filter, SAW devices are not flexible and cannot be applied in systems where the parameters are subject to changes. SAW filters operate at an intermediate frequency (IF), and a mixer is used to upconvert the signal to the radiofrequency (RF). For the system reported by Koller et al. [23], IF = 348.8 MHz, RF = 2.45 GHz, $B = 80$ MHz, and $T = 0.5\,\mu$s. For the $\pi/4$-DQPSK modulation described by Gugler [9], a suitable pulse for exciting the filter is located at the IF center frequency and has a rectangular shape, the length of which equals four periods of the IF. Unlike in conventional systems where the IF is modulated by a $\pi/4$-DQPSK signal, the IF pulse exciting the SAW filter is modulated. Because of the sensitivity of the output power on the rising time of the broadband pulse exciting the filter, ASK is not suited for CM with SAW filters. Other examples of CM systems based on SAW filters can be found in the literature [15,24,25].

Unlike SAW filters, DDFS are highly flexible since the parameters of the CM signal can be set by a digital controller. Salous et al. [26] have presented a digital chirp sounder for mobile radio applications. In particular, the time and frequency resolutions are fully programmable. This is important for the different multiple-access schemes in Section 2.4.3, where the CM signal format depends on the number of simultaneous users. Clearly, the computational effort is large, but it is expected for decreasing costs of digital components such as digital–analog converters and dedicated signal processors that the DDFS will be preferred to the SAW approach in the future. A commercially available digital channel sounder with CM is described in Ref. 27 and a digital local oscillator for CM generation, in Ref. 28. Allen et al. [29] have direct digital synthesis of CM signals for a light detection and ranging application.

5. OTHER ASPECTS RELATED TO CM IN COMMUNICATION SYSTEMS

CM has been described in Section 2.4 as a means of transmitting information in form of a spread-spectrum signal over a linear channel. In semiconductor lasers (SCL), frequency chirping arises as an undesired effect in fiberoptic transmission using current modulation. As pointed out in another study [30], when the device current is modulated at frequencies approaching a few gigahertz, the dynamic response of SCL leads to an increased linewidth of an individual longitudinal mode where the line broadening is proportional to the linewidth enhancement factor (also termed the *antiguiding parameter*) β_c. The resulting chirp has its origin in the carrier-induced refractive-index change that accompanies any gain change in SCL. Agrawal and Dutta [30] described several measures that lead to light emission of SCL predominantly in a single longitudinal mode even under high-speed modulation.

Concerning the use of CM in communication systems, there are presently only very few applications. One example is the Consumer Electronic Bus (CEBus) EIA/IS60 powerline communications standard where CM can be used with a data rate of 10 kbps in an unlicensed frequency band, 100–450 kHz, for home networking. Although proposals have been made to use CM in wireless communication systems, especially in wireless local-area networks, standardization bodies have preferred other types of modulation to CM. With the advent of more advanced CM systems based on DDFS in combination with software defined radio concepts, however, CM might be an interesting alternative modulation format for future communication systems operating in frequency-selective channel environments.

BIOGRAPHY

Dirk Dahlhaus received the Dipl.-Ing. degree in electrical engineering from Ruhr-Universität Bochum, Germany, in 1992, and the Ph.D. degree from Swiss Federal Institute of Technology (ETH) Zurich, Switzerland, in 1998. Since April 1999, he has been assistant professor for mobile radio systems at the Communication Technology Laboratory of ETH Zurich. He was president of the 2002 International Zurich Seminar on Broadband Communications. His main research interests include radio channel modelling, digital signal processing and link adaptation in multiuser wireless and mobile radio communication systems.

BIBLIOGRAPHY

1. A. J. Berni and W. D. Gregg, On the utility of chirp modulation for digital signaling, *IEEE Trans. Commun.* **COM-21**: 748–751 (1973).

2. M. R. Winkler, Chirp signals for communications, *Proc. Western Electronic Show and Convention* (*WESCON*), Los Angeles, Aug. 21–24, 1962, Vol. 14.2.

3. C. E. Cook and M. Bernfeld, *Radar Signals*, Artech House, Norwood, MA, 1993.

4. C. L. Dolph, A current distribution for broadside arrays which optimizes the relationship between beamwidth and sidelobe level, *Proc. IRE* **34**: 335–348 (1946).

5. G. J. Van der Maas, A simplified calculation for Dolph-Tchebycheff arrays, *J. Appl. Phys.* **25**: 121–124 (1954).

6. M. Kowatsch, *Codierte Nachrichtenübertragung mit Chirp-Modulation*, Ph.D. thesis (in German), Technical Univ. Vienna, Vienna, Austria, 1981.

7. M. Kowatsch, H. R. Stocker, F. J. Seifert, and J. Lafferl, Time sidelobe performance of low time-bandwidth product linear FM pulse compression systems, *IEEE Trans. Sonics Ultrasonics* **28**(4): 285–288 (July 1981).

8. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

9. W. Gugler, *Untersuchung von hochratigen OFW-basierten Chirp-Übertragungssystemen*, Ph.D. thesis (in German), J. Kepler Univ. Linz, Linz, Austria, 2000.

10. A. Springer et al., A robust ultra-broad-band wireless communication system using SAW chirped delay lines, *IEEE Trans. Microwave Theory Tech.* **46**(12): 2213–2219 (Dec. 1998).

11. S. E. El-Khamy and S. E. Shaaban, Matched chirp modulation: Detection and performance in dispersive communication channels, *IEEE Trans. Commun.* **36**(4): 335–348 (April 1988).

12. W. Hirt and S. Pasupathy, Continuous phase chirp (CPC) signals for binary data communication — Part I: Coherent detection, *IEEE Trans. Commun.* **COM-29**(6): 836–847 (June 1981).

13. K. V. Raveendra, Digital transmission using multimode phase-continuous chirp signals, *IEE Proc. Commun.* **143**(2) (April 1996).

14. J. P. Fonseka, Partial response continuous phase chirp modulation, *IEE Electron. Lett.* **35**(6): 448–449 (March 1999).

15. Y. Takeuchi and K. Yamanouchi, A chirp spread spectrum DPSK modulator and demodulator for a time shift multiple access communication system by using SAW devices, *Microwave Symp. Digest*, 1998 IEEE MTT-S International, 1998, Vol. 2, pp. 507–510.

16. S. E. El-Khamy, S. E. Shaaban, and E. A. Thabet, Efficient multiple access communications using multi-user chirp modulation signals, *Proc. IEEE 4th Int. Symp. Spread Spectrum Techniques and Applications* (*ISSSTA'96*), Mainz, Germany, 1996, Vol. 3, pp. 1209–1213.

17. C. Gupta and A. Papandreou-Suppappola, Wireless CDMA communications using time-varying signals, *Proc. 6th Int. Symp. Signal Processing and Its Applications*, 2001, Vol. 1, pp. 242–245.

18. A. Kocian and D. Dahlhaus, Downlink performance analysis of a CDMA mobile radio system with chirp modulation, *Proc. 49th IEEE Vehicular Technology Conf.* (*VTC'99*) *Spring*, Houston, TX, 1999, Vol. 1, pp. 238–242.

19. W. Hirt and S. Pasupathy, Continuous phase chirp (CPC) signals for binary data communication — Part II: Noncoherent detection, *IEEE Trans. Commun.* **COM-29**(6): 847–858 (June 1981).

20. S. E. El-Khamy, S. E. Shaaban, and E. A. Thabet, Partially coherent detection of continuous phase signals, *Proc. 13th Nat. Radio Science Conf.*, Cairo, Egypt, 1996, pp. 1–11.

21. P. M. Djuric and S. M. Kay, Parameter estimation of chirp signals, *IEEE Trans. Acoustics Speech Signal Process.* **38**(12): 2118–2126 (Dec. 1990).

22. S. Saha and S. M. Kay, Maximum likelihood parameter estimation of superimposed chirps using Monte Carlo importance sampling, *IEEE Trans. Signal Process.* **50**(2): 2118–2126 (Feb. 2002).

23. R. Koller et al., A SAW based high-speed spread-spectrum WLAN using chirp $\pi/4$-DQPSK modulation, *Proc. IEEE 2000 Ultrasonics Symp.*, 2000, pp. 367–370.

24. J. Q. Pinkney, A. B. Sesay, S. Nichols, and R. Behin, A robust high speed indoor wireless communications system using chirp spread spectrum, *Proc. 1999 IEEE Canadian Conf. Electrical and Computer Engineering*, Edmonton, Alberta, Canada, May 1999, Vol. 1, pp. 84–89.

25. Y. R. Tsai and J. F. Chang, The feasibility of combating multipath interference by chirp spread spectrum techniques over Rayleigh and Rician fading channels, *Proc. IEEE 3rd Int. Symp. Spread Spectrum Techniques and Applications* (*ISSSTA'94*), 1994, Vol. 1, pp. 282–286.

26. S. Salous, N. Nikandrou, and N. F. Bajj, Digital techniques for mobile radio chirp sounders, *IEE Proc. Commun.* **145**(3): 191–196 (June 1998).

27. *http://www.gage applied.com/resource/newslett/07.3/Real-World.htm* (2002).

28. *http://www.spectrumsignal.com/support/_and_training/3_manuals/tim-ddc.pdf* (2002).

29. C. Allen, Y. Cobanoglu, S. K. Chong, and S. Gogineni, Performance of a 1319nm laser radar using RF pulse compression, *Proc. 2001 Int. Geoscience and Remote Sensing Symp.* (*IGARSS '01*), July 2001.

30. G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*, Kluwer, Boston, 1993.

COCHANNEL INTERFERENCE IN DIGITAL CELLULAR TDMA NETWORKS

Savo G. Glisic
Pekka Pirinen
University of Oulu
Oulu, Finland

1. INTRODUCTION

In cellular TDMA networks cochannel interference is generated in surrounding cells using the same carrier frequency. For this reason a careful planning of sectors and surrounding layers allowed to reuse the same frequency is required. In addition to sectorization (three sectors per cell), narrower antenna lobes can be used to further reduce the angular sectors of the receiving antennas so that the interference can be spatially filtered.

Usually none of these measures are efficient enough to warrant additional action to deal with the interference by using different cancellation techniques in either time,

frequency, or spatial domain. Having this in mind, we can represent the residual interference signal power as

$$I_r(r,\theta,f,t) = (1-C_f)(1-C_p)(1-C_\theta)(1-C_t)I(r,\theta,f,t)$$
$$= (1-C_r)(1-C_\theta)(1-C_t)I(r,\theta,f,t) \quad (1)$$

where C_f, C_p, C_θ, and C_t are frequency, propagation (distance + shadowing + fading), angle (space), and time isolation coefficients, respectively. $I(r,\theta,f,t)$ is the interference signal power without any suppression techniques. For perfect isolation, at least one of these coefficients is equal to one and the interference has no influence on the received signal. In practice, it is rather difficult and economically impractical to reach the point where $C_i = 1$. Instead, the product $(1-C_r)(1-C_\theta)(1-C_t)$ depending on these coefficients should be kept as low as possible with an affordable effort measured by cost, power consumption, and physical size of the hardware required for the solution.

Coefficient C_f is related to frequency assignment in the cellular network, while coefficient C_p is related to the propagation conditions. $C_f = 1$ if the interfering signal frequency is different from the frequency of the useful signal. $C_p = 1$ if, as a result of propagation losses, the interfering signal cannot reach the site of the useful reference signal. In general, the same frequency can be used in two cells only if the propagation losses between the two cells are high enough that the interfering signals are attenuated to the acceptable level. This will be characterized by the frequency reuse coefficient C_r defined as $(1-C_r) = (1-C_f)(1-C_p)$ and will be discussed in Section 2. Coefficient C_θ is related to antenna beamforming, and possibilities of reducing the interference level by spatial filtering are discussed in Section 3. Finally, interference cancellation and equalization in time domain, which is included in coefficient C_t, will be discussed in Section 4.

2. NETWORK PLANNING AND FREQUENCY REUSE

Depending on the cell size, three different categories of cellular networks can be distinguished. *Macrocells* are the largest, with a cell radius of 1 km up to 35 km or more. Base station antennas are located well above the rooftop level. The commonly used macrocellular modeling structure assumes a uniform grid of hexagonal cells [1]. Part of the hexagonal cellular layout is illustrated in Fig. 1. In Fig. 1a frequencies are reused in each cluster of seven cells and in Fig. 1b the cluster size is 3.

The hexagonal grid is optimal in the sense that there is no overlap between cells. In addition, hexagons closely approximate circles. This kind of modeling is highly theoretical since effective cell coverage areas vary considerably depending on factors such as terrain, buildings, weather, and time. Cellular models can be further classified according to base station antenna directivity. In the case of omnidirectional antennas, base stations can be located in the cell centers as illustrated in Fig. 2a. When directional antennas are used, cells can be divided into widebeam sectors as shown in Fig. 2b. Directional antenna patterns can also be modeled by corner-illuminated base stations with three narrow antenna lobes per base station. One advantage of this approach is lower cost. Fewer base stations are required over a certain geographical area than with direct sectorization. Corner-illuminated cells or the so-called "three leaf clover" structure is illustrated in Fig. 2c.

Microcells are smaller than macrocells with a typical cell radius of 20–300 m. In this scenario base station antennas are usually below the mean rooftop level. In urban areas microcells are often characterized as having a Manhattan type of grid, where the base stations are in the crossings of linear streets as shown in Fig. 1c. *Picocells* or indoor cells usually cover indoor areas (rooms, halls) with typical cell radius of 5–30 m. These scenarios are not covered in this article.

Frequency reuse is an essential element in cellular networks. It means that the same frequencies are reused in the system within a certain distance depending on the reuse factor. This reuse factor can be represented as a cluster size, which includes the group of cells where all different available channels are used. Regular cluster sizes K [1] can be calculated according to

$$K = i^2 + ij + j^2 \quad (2)$$

where i and j are nonnegative integers. Equation (2) leads to balanced cluster sizes $K = 1, 3, 4, 7, 9, 12, \ldots$. If D is defined as the distance between the closest cochannel centers and R is the cell radius (see Fig. 1a), the frequency reuse factor D/R and the cluster size K are related as [1]

$$\frac{D}{R} = (3K)^{1/2} \quad (3)$$

In order to increase network capacity, cluster size must be reduced. The more aggressive reuse (smaller the cluster size), the higher the level of cochannel interference that will be generated and vice versa. For these reasons frequency reuse has been studied extensively in the literature [1–4]. The problem becomes even more challenging if macrocells and microcells are overlaid [5].

2.1. Cochannel Interference Distributions

From the previous discussion one can see that no matter what cluster size is chosen, a certain level of cochannel interference (CCI) can not be avoided. For the analysis of cochannel interference statistics, the CCI distribution function is required. Cochannel interference can be seen as a superposition of distance-dependent attenuation (path loss), short-term fluctuations, and long-term variations. The long-term or large-scale signal variation (shadowing, slow fading) can be characterized by the lognormal distribution. The short-term signal variation (fast fading), on the other hand, may fit to some other distributions such as Rayleigh, Rice, or Nakagami. An overview of fading distributions related to CCI can be found in the paper by Yacoub [6]. In the sequel, the lognormal shadowing is assumed.

The total interference power is often accumulated from several cochannel signals. Unfortunately, there is no known closed-form expression for the distribution of

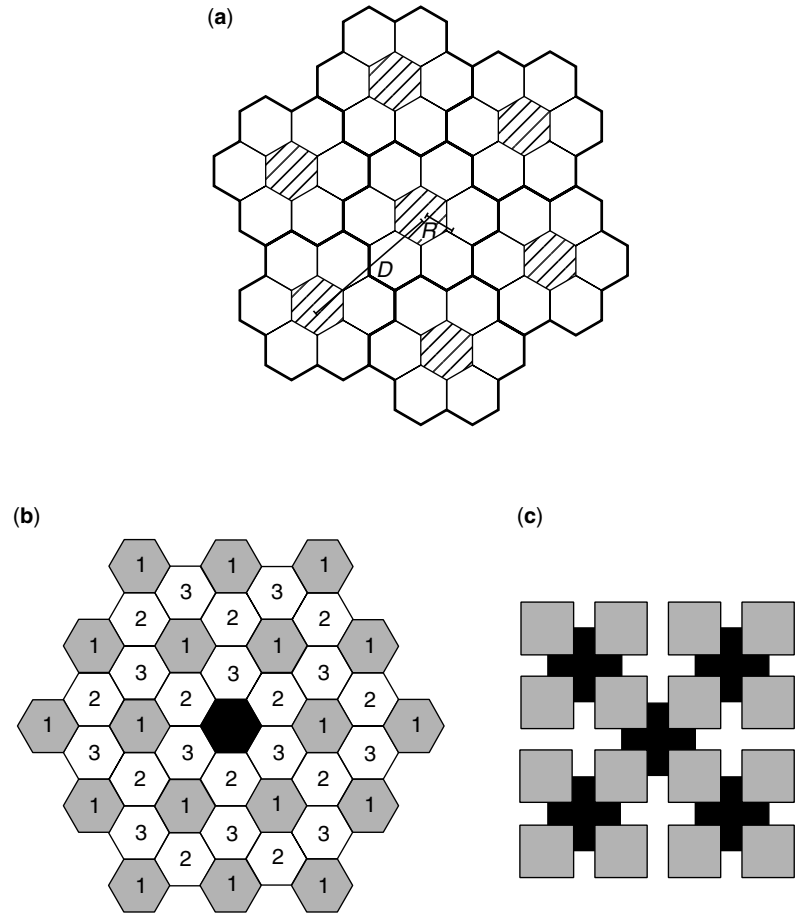


Figure 1. Cellular layouts: (**a**) uniform hexagonal cellular layout with reuse 7; (**b**) macrocell layout with reuse 3; (**c**) street microcell layout with reuse 2.

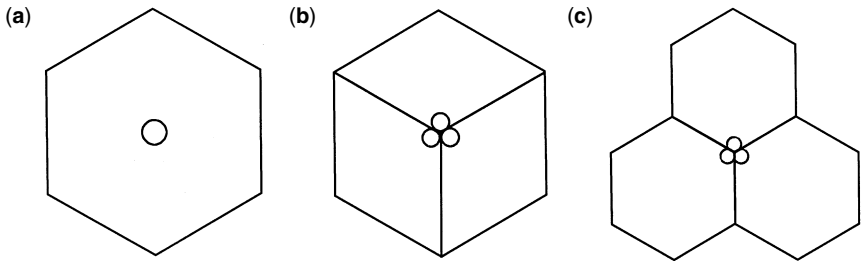


Figure 2. Cell types considered: (**a**) omnicell; (**b**) sectored cell; (**c**) corner-illuminated cells.

the sum of lognormally distributed random variables. However, several approximations have been derived. A common feature for all these approximations is that they estimate the sum of lognormal random variables by another lognormally distributed random variable [7]. This can be represented as

$$L = \sum_{i=1}^{n} e^{y_i} = \sum L_i \cong e^z \tag{4}$$

where y_i represents Gaussian random variables. In the Fenton–Wilkinson (FW) approximation [7–10], the mean m_z and the standard deviation σ_z of z are derived by matching the first two moments of the both sides of Eq. (4). If the first moment of $(L_1 + L_2 + \cdots + L_n)$ is denoted by u_1 and the second by u_2, the following expression is obtained after moment matching [9]

$$m_z = 2 \ln u_1 - \tfrac{1}{2} \ln u_2 \tag{5}$$

$$\sigma_z^2 = \ln u_2 - 2 \ln u_1. \tag{6}$$

The Fenton–Wilkinson approach is applicable when the standard deviations of the lognormal components are lower than 4 dB for uncorrelated signal components [11]. For higher deviation values, this approximation tends to underestimate the mean and overestimate the variance of the sum distribution. When there is correlation between the components, the FW approximation is quite accurate for higher deviation values ($\leq$12 dB), too [9].

The Schwartz–Yeh (SY) method [7,9–11] is also based on the assumption that the power sum is lognormally

distributed. The SY approximation is different in the use of the exact expressions for the first two moments of the sum of two lognormal random variables. Nesting and recursion techniques are then used to extend the approach to a larger number of cumulative random variables. Originally, the SY method was developed for the sum of independent lognormal random variables. However, it has been extended to the case of correlated lognormal random variables with some modifications [9].

The Schwartz–Yeh approximation can be best applied when the range of the standard deviation is $4 \leq \sigma \leq 12$ dB. If all components in the summation are identically distributed, this approximation tends to underestimate the variance in the resulting signal distribution. The error increases as a function of the number of added components.

In addition to the Fenton–Wilkinson and Schwartz–Yeh approaches, there are some other approximations for the sum of lognormal components. For example, Farley's approximation is a strict lower bound for the cumulative distribution function (CDF) of a sum of independent lognormal random variables [7]. For further studies on lognormal sum approximations the reader is referred to the additional reading listed at the end of this article.

2.2. Cochannel Interference and Outage Probabilities

Following Refs. 10 and 12, cochannel interference probability is defined as

$$P(I_c) = \sum_n P(I_c|n)P_n(n) \tag{7}$$

where $P_n(n)$ is the probability of n cochannel interferers being active and $P(I_c|n)$ is the corresponding conditional CCI probability.

The conditional CCI probability can be defined as

$$P(I_c|n) = P\left(\frac{C}{I} < \alpha\right) \tag{8}$$

where C is the instantaneous power of the desired signal (carrier), I is the joint interference power from n active cochannel users, and α is the specified cochannel interference protection ratio.

$P_n(n)$ can be represented by the binomial distribution in terms of carried traffic per channel

$$P_n(n) = \binom{N}{n} a_c^n (1 - a_c)^{N-n} \tag{9}$$

where N is the number of effective cochannel interferers ($N = 6$ if only the closest ring cochannel interferers are taken into account) and $a_c = m_1/m_t$ is carried traffic per channel (erlangs per channel). Parameters m_1 and m_t are discussed in more detail in the next Section 2.3. It is assumed that the number of traffic channels is equal for all cells.

The outage probability P_{out} for the desired user can be defined as the probability of failing to achieve a bit error probability P_e lower than a fixed threshold P_{e0}, namely

$$P_{\text{out}} = P(P_e > P_{e0}) \tag{10}$$

If only the effects of cochannel interference are taken into account, the received carrier-to-interference ratio C/I is the key parameter. If the minimum required carrier-to-interference ratio is α and it corresponds to the bit error probability $P_e = P_{e0}$, the outage probability is the same as the conditional CCI probability defined by (8).

Following the procedure in [9], the outage probability of the lognormally distributed signals can be represented in the form

$$P_{\text{out}} = P(I_c|n) = 1 - Q\left(\frac{\ln\alpha - \ln\xi_d + m_{z_n}}{(\sigma_d^2 + \sigma_{z_n}^2 - 2r_{yz}\sigma_d\sigma_{z_n})^{1/2}}\right) \tag{11}$$

where ξ_d is the area mean desired signal power, m_{z_n} is the area mean joint interference power of n interferers, σ_d is the standard deviation of the desired signal, σ_{z_n} is the standard deviation of the joint interference from n interferers, and r_{yz} is the correlation coefficient of the desired signal and joint interference. The initial mean single interferer power in the worst geometric case can be approximated by

$$m_{z_{1w}} = \ln[((3K)^{1/2} - 1)^{-\beta}] \tag{12}$$

In the average geometric case, the exact interferer power is of the form

$$m_{z_{1a}} = \ln[(3K)^{-\beta/2}] \tag{13}$$

The desired signal area mean power ξ_d can be represented as

$$\xi_d = \left(\frac{r}{R}\right)^{-\beta} \tag{14}$$

In Eqs. (12)–(14), β denotes the path loss exponent, K is the cluster size, and $r/R \in (0, 1]$ is the normalized distance between the desired mobile station and the base station.

The combined effect of frequency allocation and propagation conditions, characterized implicitly by the parameter $(1 - C_r)$, is illustrated in Fig. 3. The figure shows the outage probability defined by (11) with the maximum number of first-tier interferers as a function of cluster size (worst and average case geometries) with variable path loss exponents β. The standard deviation of each lognormal component is 6 dB. All signals are uncorrelated. The Fenton–Wilkinson method has been used for the mean and variance approximations.

It can be noted that in free-space propagation conditions ($\beta = 2$), cochannel interference can be very severe even for large cluster sizes. On the other hand, in dense urban areas, where the path loss attenuation slope is steep, small cluster sizes can be supported. That allows larger system capacity for highly populated cities where it is the most desirable. The outage probability is very sensitive to the changes in the propagation environment.

Figure 4 shows the strong impact of normalized mobile distance (14) on conditioned full load CCI probability (outage probability) in the presence of lognormal shadowing ($\sigma = 6$ dB, FW approximation). It can be seen that without power control, outage events are more likely near the cell edges. Larger cluster sizes guarantee lower outage probabilities. The gap between worst and average

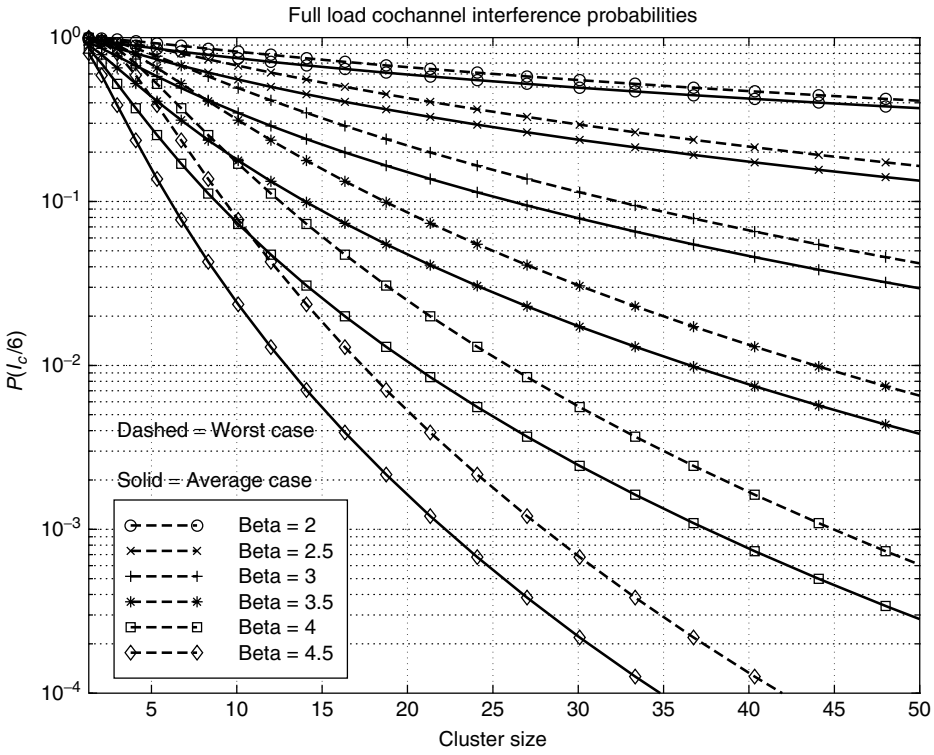


Figure 3. Effect of path loss exponent variation to the outage probability.

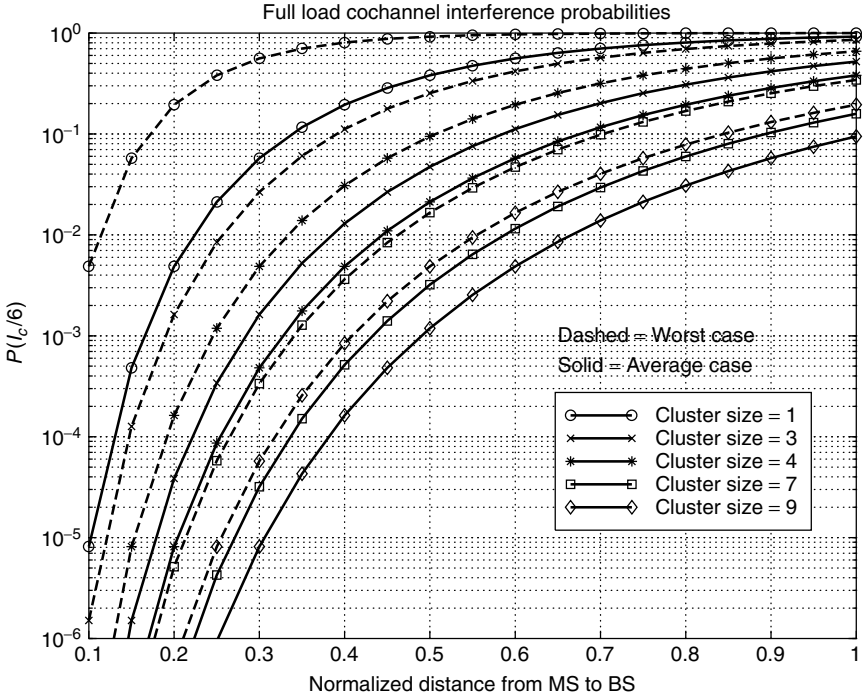


Figure 4. Full-load outage probabilities at variable cluster sizes.

case interference geometries diminishes as the cluster size increases.

2.3. Spectrum Efficiency

Spectrum efficiency describes how effectively a system can utilize limited frequency resources. In general, spectrum efficiency can be seen as a ratio between benefit (number of traffic channels, data rate) and cost (bandwidth) [13]. Spectrum efficiency is usually measured in units related to system capacity. Problems may arise if system capacities of different systems have been calculated with different assumptions or if they are represented in different units. A fair comparison of different systems is often cumbersome. Falciasecca et al. [13] discuss the effect of some working assumptions on spectrum efficiency, and ways to allow a fair comparison.

The system capacity of a voice-oriented network is related to the grade of service by the Erlang-B formula

$$P_B = \frac{{m_1}^{m_t}/m_t!}{\sum_{n=0}^{m_t} {m_1}^{m_t}/n!} = \Im(m_t, m_1) \tag{15}$$

where P_B is blocking probability, m_1 is the offered traffic (capacity) (in erlangs), and m_t is the total number of traffic channels. The blocking probability P_B refers to the probability that a new call attempt will not find an available channel in a trunk of channels and is dropped. Thus, there is no queueing in the system. The Erlang formula was originally developed for wired telephone traffic. It is not strictly applicable to cellular systems, because it does not take into account the handover traffic. An additional assumption of this model is that the total offered traffic is constant, which is not valid in the cell where the traffic is time-varying as a result of moving subscribers. Despite the limitations of the Erlang formula, it can be used for relative comparison purposes; however, one must be careful in interpreting the absolute values.

The spectrum efficiency and capacity evaluation are based on the radio capacity m_t introduced by Lee [14]. The radio capacity of the omnicellular TDMA system is defined as

$$m_t = \frac{N_s B_t}{B_c \left(\frac{2}{3}\left(\frac{C}{I}\right)_s\right)^{1/2}} = \frac{M_t}{\left(\frac{2}{3}\left(\frac{C}{I}\right)_s\right)^{1/2}} = \frac{M_t}{K} \quad \text{(frequency channels/cell)} \tag{16}$$

where B_t is the total allocated spectrum for the system, B_c is the channel bandwidth, $(C/I)_s$ is the minimum required carrier-to-interference ratio, M_t is the total number of frequency channels multiplied by the number of TDMA slots N_s, and K is the cluster size.

Radio capacity can be represented in different units as presented in Lee's paper [14]. These new measures can be derived from the general radio capacity definition and depend on issues such as service area, call duration, number of calls, and total bandwidth. Other commonly used units for spectrum efficiency are erlangs per MHz/km^2 and erlangs per cell/MHz.

For the system with parameters given in Table 1, maximum capacity obtained from Eq. (15) is presented in Table 2. By using (15) and (16), Fig. 5 illustrates real Erlang capacities for TDMA omnicellular uplink with

Table 1. Essential Parameters for the Capacity Evaluation of TDMA System

B_t (MHz)	R_C (kHz)	B_c (kHz)	α	$M = B_t/B_c$	M_t
10	270.8	200	9 dB	50	400

Table 2. Maximum Radio Capacities of Compared Cluster Sizes ($P_B = 0.02$)

TDMA (K)	m_t (Traffic Channels)	m_1 (erlangs/cell)	$a_c = m_1/m_t$
TDMA(7)	57	46.8	0.821
TDMA(9)	44	34.7	0.789
TDMA(12)	33	24.6	0.745
TDMA(21)	19	12.3	0.679

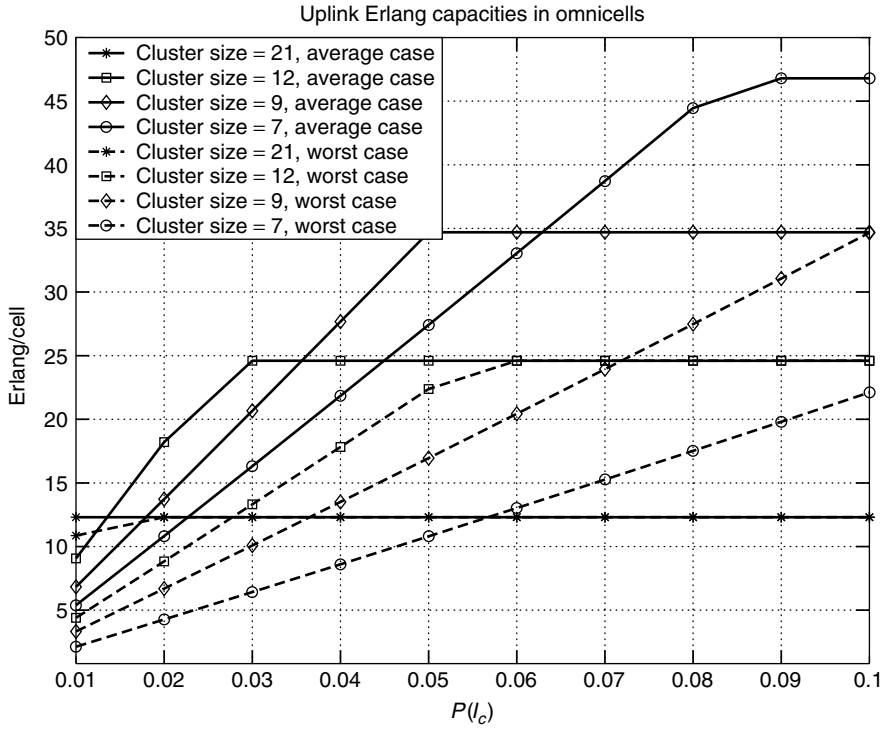


Figure 5. Uplink Erlang capacities in omnicells (Rayleigh fading only).

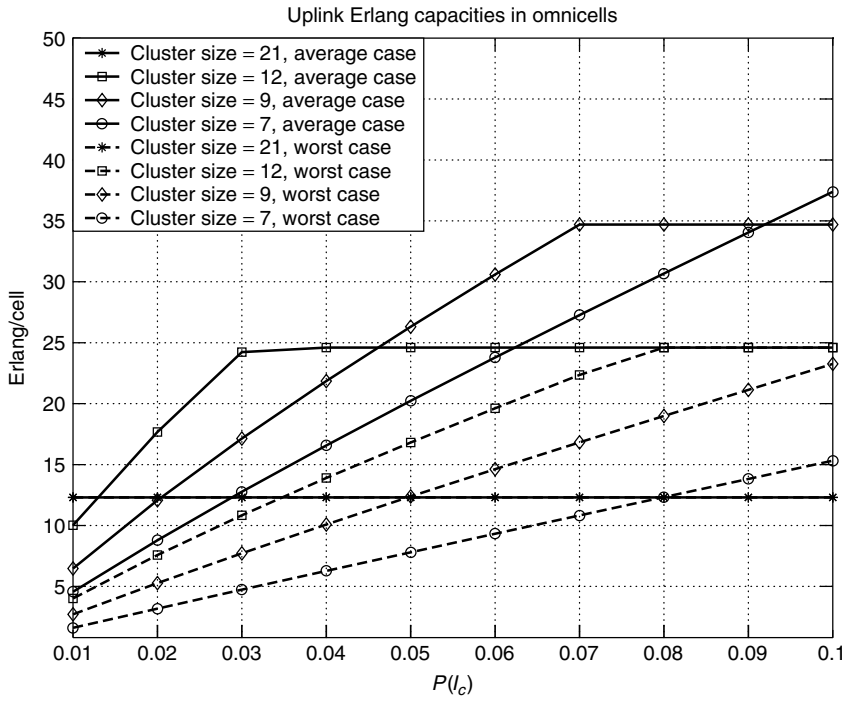


Figure 6. Uplink Erlang capacities in omnicells (lognormal shadowing only, $\sigma = 6$ dB).

several cluster sizes in a Rayleigh fading environment. Figure 6 presents uplink Erlang capacities when both the desired signal and cochannel interferers are lognormally shadowed with standard deviation $\sigma = 6$ dB.

Curves in Fig. 6 show that the performance is very close to the purely Rayleigh case. For larger cluster sizes, lognormal shadowing only ($\sigma = 6$ dB) will give more optimistic results than the purely Rayleigh case. Horizontal parts of the curves indicate that the maximum capacity limit m_1, for the particular cluster size, has been reached (hard capacity limit). Elsewhere, the Erlang capacity is softly limited by the highest tolerated CCI probability.

Alouini and Goldsmith consider a slightly different definition of spectrum efficiency, the area spectral efficiency (ASE) [15]. It is better suited for variable rate data transmissions, where the total throughput is of interest. The measure for average area spectral efficiency is the sum of the maximum average data rates divided by the bandwidth and the unit area for the system: bandwidth per second per Hz/m^2. The analytical framework provides theoretical limits for area spectral efficiency versus reuse distance for adaptive data rate cellular systems, where the rate adaptation depends on fading and interference conditions. Users' random locations, impact of propagation parameters, cell size, carrier frequency, and sectorization in both macrocells and microcells are taken into account. Furthermore, the loading in the cells is varied. Best and worst case analytical results are verified via average case Monte Carlo simulations.

Results based on the worst-case interference geometry show that the optimal reuse distance is close to 4. The area efficiency decreases as an exponential of a fourth-order polynomial relative to the cell size. Shadowing and fading reduce the absolute ASE, but do not affect the relative behavior as the function of reuse distance. Moreover, it is noted that the fading parameters of the desired user have stronger contribution on the ASE than do the fading parameters of the interferers [15].

3. SPATIAL FILTERING

Spatial domain processing included in the term $(1 - C_\theta)$ can be used to combat cochannel interference. At least at the base station there is a possibility to steer radiation/reception to the desired directions: (1) spatially directed transmission can enhance signal coverage and quality and (2) interference coming outside from the antenna main lobe is suppressed significantly in the reception.

3.1. Sectorization

One conventional way to improve cellular system capacity is cell splitting, that is, subdividing the coverage area of one base station to be covered by several base stations (smaller cells) [1]. Another simple and widely applied technique to reduce interference spatially is to divide cells into sectors, for example, three 120° sectors. These sectors are covered by one or several directional antenna elements. Effects of sectorization to spectrum efficiency have been studied [16]. Chan [16] concluded that sectorization reduces cochannel interference and improves signal-to-noise ratio of the desired link at the given cluster size. However, at the same time the trunking efficiency is decreased. Because of the improved link quality, a tighter frequency reuse satisfies the performance criterion in comparison to the omnicellular case. Therefore, the net effect of sectorization is positive at least for large cells and high traffic densities.

3.2. Adaptive Antennas

By using M-element antenna arrays at the base station the spatial filtering effect can be further improved.

The multiple beam adaptive array would not reduce the network trunking efficiency unlike sectorization and cell splitting [17]. These adaptive or "smart" antenna techniques can be divided into switched-beam, phased-array, and purely adaptive antenna systems. Advanced adaptive systems are also called *spatial division multiple access* (SDMA) systems. Advanced SDMA systems maximize the gain toward the desired mobile user and minimize the gain toward interfering signals in real time.

According to Winters [18], by applying a four-element adaptive array at the TDMA uplink, frequencies can be reused in every cell (three-sector system) and sevenfold capacity increase is achieved. Correspondingly, a four-beam antenna leads to reuse of 3 or 4 and doubled capacity at small angular spread.

Some practical examples of the impact of the use of advanced antenna techniques on the existing cellular standards have been described [19,20]. Petrus et al. [19] use the AMPS reference system and Mogensen et al. [20] use GSM. The Petrus et al. analysis [19] uses ideal and flat-top beamformers. The mainlobe of the ideal beamformer is flat and there are no sidelobes whereas the flat-top beamformer has a fixed sidelobe level. The ideal beamformer can be seen as a realization of the underloaded system; that is, there are less interferers than there are elements in the array. The overloaded case is better modeled by the flat-top beamformer because all interferers cannot be nulled and sidelobe level is increased. Performance results show that a reuse factor of 1 is not feasible in AMPS, but reuse factors of 4 and 3 can be achieved with uniform linear arrays (ULA) with five and eight elements, respectively.

Mogensen et al. [20] concentrate on the design and performance of the frequency-hopping GSM network using conventional beamforming. Most of the results are based on simulated and measured data of eight-element ULA. The simulated C/I improvement follows closely the theoretical gain at low azimuth spreads. In urban macrocells the C/I gain is reduced from the theoretical value 9 dB down to approximately 5.5–7.5 dB. The designed direction of arrival (DoA) algorithm is shown to be very robust to cochannel interference. The potential capacity enhancement is reported to be threefold in a $\frac{1}{3}$ reuse FHGSM network for an array size of $M = 4$–6.

4. INTERFERENCE CANCELLATION IN TIME DOMAIN

For the purpose of this section, the overall received signal, which is a superposition of N cochannel components received through M antennas, can be represented in the simplified case, when all signals are received bit synchronously as

$$\begin{aligned} \boldsymbol{r} &= (r^{(1)}, r^{(2)}, r^{(3)}, \ldots, r^{(M)})^{-1} \\ r^{(m)} &= \sum_{n=1}^{N} a^{(n)} h^{(m,n)} + n^{(m)} \\ h^{(n,m)} &= c^{(m,n)} * g \end{aligned} \tag{17}$$

where $a^{(n)}$ is data of user n, g is the pulseshape, and $c^{(m,n)}$ is the channel transfer function between the cochannel signal source n and receiving antenna m. The corresponding vector representation is

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{a} + \boldsymbol{n} \tag{18}$$

where $\boldsymbol{a}$ and $\boldsymbol{n}$ are vectors with components $a^{(n)}$ and $n^{(m)}$, respectively, and $\boldsymbol{H}$ is the matrix with elements $h^{(n,m)}$. In a real situation the received cochannel signals are not bit synchronous, and (18) should be modified to include two additional components representing the impact of the previous and subsequent bits accordingly, similar to asynchronous CDMA detectors [21]. Details may be found in the paper by Valenti and Woerner [22]. In accordance with Eq. (14), the system capacity can be increased by using more advanced demodulation techniques that provide the same quality of service (QoS) with lower $(C/I)_s$. A number of algorithms have been presented in the literature.

When used for DSCDMA systems, the objective of multiuser detection (MUD) is primarily to jointly detect signals that originate from the same cell (intracell interference) because the most critical interference comes from there. This is quite the opposite of the situation in a TDMA network where the interest is focused on the interference coming from the adjacent cells.

An optimum detector has been introduced [23] as a joint demodulator of cochannel signals. This detector is based on already known solutions for optimum single-user detection with intersymbol interference and Gaussian noise for *M-input/M-output* (MIMO) systems [24,25] and joint maximum-likelihood sequence estimation (MLSE) [26]. A similar approach is used for CDMA systems [27]. Practical results for the Japanese PDC system [28] and for GSM [29] have been shown. The joint MLSE for a hybrid CDMA/TDMA based on the GSM system has also been presented [30].

MLSE-type detectors are so complex and impractical that a number of blind detector algorithms must be considered. These algorithms do not require knowledge of the other signal parameters. See the "Further Reading" section for further information about blind CCI cancellation.

The latest results include cochannel interference suppression with successive cancellation [31], which is a technique already well established in CDMA systems. Performance results show that the cancellation succeeds poorly when the signal levels are comparable. Timing differences can be used for initial signal separation in order to improve performance. Soft subtraction provides further improvement.

When M signals $r^{(m)}$ from Eq. (17) are combined by using maximal ratio combining, then a reliable estimate of the channel coefficients is required. One of the latest references dealing with this problem is that by Grant and Cavers [32].

The most recent development in Turbo decoding has inspired research in the field of iterative multiuser detection, macrodiversity combining, and decoding for the TDMA cellular uplink [22]. In this approach, as the first step, each base station (BS) in a cluster of cochannel cells performs soft-output multiuser detection of the desired signal (originating from its cell) and the interfering

signals (originating from other cells in the cluster). So the multiuser detector will produce a loglikelihood ratio (LLR) for each mobile in the cluster. These LLRs for each user are then summed up across the BS cluster, which in effect produces a diversity combining signal. After that the signal may be deinterleaved and decoded. If the decoder also produces soft outputs, this may be reinterleaved and fed back to the multiuser detector to be used as a priori information in the next iteration. Once again one should be aware that the system places an additional burden on the backhaul links. Since soft information is now shared among BSs more capacity is needed on the links between BSs and the base station controller (BSC).

BIOGRAPHIES

Savo Glisic is Professor of Electrical Engineering at the University of Oulu, and Director of Globalcomm Institute for Telecommunications. He was a Visiting Scientist at Cranfield Institute of Technology, Cranfield, England (1976/77) and the University of California at San Diego (1986/87). He has been active in the field of spread-spectrum and wireless communications for 20 years and has published a number of papers and five books. He is doing consulting in this field for industry and government. He has served as Technical Program Chairman of The Third IEEE ISSSTA'94, The 8 IEEE PIMRC'97, and IEEE ICC'01. Dr Glisic was Director of IEEE ComSoc MD programs.

Pekka Pirinen received the M.S. and Lic.Tech. degrees in electrical engineering from the University of Oulu, Oulu, Finland, in 1995 and 1998, respectively. He started his career as Research Assistant in the Telecommunication Laboratory, University of Oulu, in 1994. Since 1995, he has been with the Telecommunication Laboratory and Centre for Wireless Communications, University of Oulu, as a Research Scientist in various wireless communication research projects, including the European ACTS project FRAMES. His research interests are focused on multiple access techniques, radio network planning, modeling, capacity, and performance evaluation issues. Mr. Pirinen is currently pursuing a doctoral degree in electrical engineering at the University of Oulu.

BIBLIOGRAPHY

1. V. H. MacDonald, The cellular concept, *Bell Syst. Tech. J.* **58**: 15–41 (1979).
2. I. Katzela and M. Nagshineh, Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey, *IEEE Pers. Commun.* **3**: 10–31 (1996).
3. S. W. Halpern, Reuse partitioning in cellular systems, *IEEE Trans. Vehic. Technol.* **32**: 322–327 (1983).
4. P. M. Blair, G. C. Polyzos, and M. Zorzi, Plane cover multiple access: A new approach to maximizing cellular system capacity, *IEEE J. Select. Areas Commun.* **19**: 2131–2141 (2001).
5. R. Coombs and R. Steele, Introducing microcells into macrocellular networks: A case study, *IEEE Trans. Commun.* **47**: 568–576 (1999).
6. M. D. Yacoub, Fading distributions and co-channel interference in wireless systems, *IEEE Antennas Propag. Mag.* **42**: 150–159 (2000).
7. G. L. Stüber, *Principles of Mobile Communication*, Kluwer, Norwell, MA, 1996.
8. A. A. Abu-Dayya and N. C. Beaulieu, Outage probabilities in the presence of correlated lognormal interferers, *IEEE Trans. Vehic. Technol.* **43**: 164–173 (1994).
9. L. F. Fenton, The sum of log-normal probability distributions in scatter transmission systems, *IRE Trans. Commun.* **CS-8**: 57–67 (1960).
10. R. Prasad and A. Kegel, Improved assessment of interference limits in cellular radio performance, *IEEE Trans. Vehic. Technol.* **40**: 412–419 (1991).
11. S. Schwartz and Y. S. Yeh, On the distribution function and moments of power sums with log-normal components, *Bell Syst. Tech. J.* **61**: 1441–1462 (1982).
12. R. Muammar and S. C. Gupta, Cochannel interference in high-capacity mobile radio systems, *IEEE Trans. Commun.* **30**: 1973–1978 (1982).
13. G. Falciasecca, C. Caini, G. Riva, and M. Frullone, General approach for the comparison of spectrum efficiency of digital mobile radio systems, *Eur. Trans. Telecommun.* **5**: 77–83 (1994).
14. W. C. Y. Lee, Spectrum efficiency in cellular, *IEEE Trans. Vehic. Technol.* **38**: 69–75 (1989).
15. M.-S. Alouini and A. J. Goldsmith, Area spectral efficiency of cellular mobile radio systems, *IEEE Trans. Vehic. Technol.* **48**: 1047–1066 (1999).
16. G. K. Chan, Effects of sectorization on the spectrum efficiency of cellular radio systems, *IEEE Trans. Vehic. Technol.* **41**: 217–225 (1992).
17. S. C. Swales, M. A. Beach, D. J. Edwards, and J. P. McGeehan, The performance enhancement of multibeam adaptive base station antennas for cellular land mobile radio systems, *IEEE Trans. Vehic. Technol.* **39**: 56–67 (1990).
18. J. H. Winters, Smart antennas for wireless systems, *IEEE Pers. Commun.* **5**: 23–27 (1998).
19. P. Petrus, R. B. Ertel, and J. H. Reed, Capacity enhancement using adaptive arrays in an AMPS system, *IEEE Trans. Vehic. Technol.* **47**: 717–727 (1998).
20. P. E. Mogensen et al., Performance of adaptive antennas in FH-GSM using conventional beamforming, *Wireless Pers. Commun.* **14**: 255–274 (2000).
21. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press (1998).
22. M. C. Valenti and B. D. Woerner, Iterative multiuser detection, macrodiversity combining, and decoding for the TDMA cellular uplink, *IEEE J. Select. Areas Commun.* **19**: 1570–1583 (2001).
23. W. van Etten, Maximum likelihood receiver for multiple channel transmission systems, *IEEE Trans. Commun.* **COM-24**: 276–283 (1976).
24. D. Forney, Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363–378 (1972).
25. G. Ungerboeck, Adaptive maximum likelihood receiver for carrier modulated data transmission systems, *IEEE Trans. Commun.* **COM-22**: 624–636 (1974).

26. K. Giridhar et al., Nonlinear techniques for the joint estimation of cochannel signals, *IEEE Trans. Commun.* **COM-45**: 473–484 (1997).

27. S. Verdú, Minimum probability of error for asynchronous Gaussian multiple access channels, *IEEE Trans. Inform. Theory* **IT-32**: 85–96 (1986).

28. H. Yoshino, K. Fukawa, and H. Suzuki, Interference canceling equalizer (ICE) for mobile radio communication, *IEEE Trans. Vehic. Technol.* **46**: 849–861 (1997).

29. S. W. Wales, Technique for cochannel interference suppression in TDMA mobile radio systems, *IEE Proc. Commun.* **142**: 106–114 (1995).

30. J. Blanz, A. Klein, M. Nasshan, and A. Steil, Performance of a cellular hybrid C/TDMA mobile radio system applying joint detection and coherent receiver antenna diversity, *IEEE J. Select. Areas Commun.* **12**: 568–579 (1994).

31. H. Arslan and K. Molnar, Cochannel interference suppression with successive cancellation in narrow-band systems, *IEEE Commun. Lett.* **5**: 37–39 (2001).

32. S. J. Grant and J. K Cavers, Multiuser channel estimation for detection of cochannel signals, *IEEE Trans. Commun.* **49**: 1845–1855 (2001).

FURTHER READING

Cochannel Interference Distributions

Abu-Dayya A. A. and N. C. Beaulieu, Outage probabilities of cellular mobile radio systems with multiple Nakagami interferers, *IEEE Trans. Vehic. Technol.* **40**: 757–768 (1991).

Cardieri P. and T. S. Rappaport, Statistical analysis of co-channel interference in wireless communications systems, *Wireless Commun. Mobile Comput.* **1**: 111–121 (2001).

French R. C., The effect of fading and shadowing on channel reuse in mobile radio, *IEEE Trans. Vehic. Technol.* **VT-28**: 171–181 (1979).

Ho C.-L., Calculating the mean and variance of power sums with two log-normal components, *IEEE Trans. Vehic. Technol.* **44**: 756–762 (1995).

Lee C.-C. and R. Steele, Signal-to-interference calculations for modern TDMA cellular communication systems, *IEE Proc. Commun.* **142**: 21–30 (1995).

Prasad R. and A. Kegel, Effects of Rician faded and log-normal shadowed signals on spectrum efficiency in microcellular radio, *IEEE Trans. Vehic. Technol.* **42**: 274–281 (1993).

Punt J. B. and D. Sparreboom, Summing received signal powers with arbitrary probability density functions on a logarithmic scale, *Wireless Pers. Commun.* **3**: 215–224 (1996).

Safak A., Statistical analysis of the power sum of multiple correlated log-normal components, *IEEE Trans. Vehic. Technol.* **42**: 58–61 (1993).

Schleher D., Generalized Gram-Charlier series with application to the sum of lognormal variates, *IEEE Trans. Inform. Theory* **23**: 275–280 (1977).

Outage Probability

Caini C., G. Immovilli, and M. L. Merani, Outage probability for cellular mobile radio systems: Simplified analytical evaluation and simulation results, *Electron. Lett.* **28**: 669–671 (1992).

Caini C., G. Immovilli, and M. L. Merani, Outage probability in FDMA/TDMA mobile communication networks, *Eur. Trans. Telecommun.* **5**: 59–68 (1994).

Immovilli G. and M. L. Merani, Simplified evaluation of outage probability for cellular mobile radio systems, *Electron. Lett.* **27**: 1365–1367 (1991).

Linnartz J.-P. M. G., Exact analysis of the outage probability in multiple-user mobile radio, *IEEE Trans. Commun.* **40**: 20–23 (1992).

Sowerby K. W. and A. G. Williamson, Outage probability calculations for multiple cochannel interferers in cellular mobile radio systems, *Proc. IEE Commun.* **135**: 208–215 (1988).

Sowerby K. W. and A. G. Williamson, Outage probability calculations for mobile radio systems with multiple interferers, *Electron. Lett.* **24**: 1073–1075 (1988).

Sowerby K. W. and A. G. Williamson, Outage probabilities in mobile radio systems suffering cochannel interference, *IEEE J. Select. Areas Commun.* **10**: 516–522 (1992).

Yeh Y.-S. and S. C. Schwartz, Outage probability in mobile telephone due to multiple log-normal interferers, *IEEE Trans. Commun.* **32**: 380–388 (1984).

Spectrum Efficiency

Clark M. V., V. Erceg, and L. J. Greenstein, Reuse efficiency in urban microcellular networks, *IEEE Trans. Vehic. Technol.* **46**: 279–288 (1997).

Nagata Y. and Y. Akaiwa, Analysis for spectrum efficiency in single cell trunked and cellular mobile radio, *IEEE Trans. Vehic. Technol.* **35**: 100–113 (1987).

Prasad R. and J. C. Arnbak, Comments on "Analysis for spectrum efficiency in single cell trunked and cellular mobile radio," *IEEE Trans. Vehic. Technol.* **37**: 220–222 (1988).

Spatial Filtering

Au W. S., R. D. Murch, and C. T. Lea, Comparison between the spectral efficiency of SDMA systems and sectorized systems, *Wireless Pers. Commun.* **16**: 15–67 (2001).

Godara L. C., Applications of antenna arrays to mobile communications, Part I: Performance improvement, feasibility, and system considerations, *Proc. IEEE* **85**: 1031–1060 (1997).

Howitt I. and Y. M. Hawwar, Evaluation of outage probability due to cochannel interference in fading for a TDMA system with a beamformer, *Proc. IEEE Vehicular Technology Conf.*, 1998, 520–524.

Litva J. and T. K.-Y. Lo, *Digital Beamforming in Wireless Communications*, Artech House, Boston, (1996).

Paulraj A. J. and C. B. Papadias, Space-time processing for wireless communications, *IEEE Signal Process. Mag.* **14**: 49–83 (1997).

Wang L.-C., K. Chawla, and L. J. Greenstein, Performance studies of narrow-beam trisector cellular systems, *Int. J. Wireless Inform. Networks* **5**: 89–102 (1998).

Winters J. H., Optimum combining in digital mobile radio with cochannel interference, *IEEE Trans. Vehic. Technol.* **VT-33**: 144–155 (1984).

Zetterberg P. and B. Ottersten, The spectrum efficiency of a base station antenna array system for spatially selective transmission, *IEEE Trans. Vehic. Technol.* **44**: 651–660 (1995).

Zetterberg P., A comparison of two systems for downlink communication with base station antenna arrays, *IEEE Trans. Vehic. Technol.* **48**: 1356–1370 (1999).

Interference Cancellation in Time Domain

Batra A. and J. R. Barry, Blind cancellation of co-channel interference, *Proc. IEEE Global Telecommunications Conf.*, 1995, pp. 157–162.

Berangi R. and P. Leung, Indirect cochannel interference cancelling, *Wireless Pers. Commun.* **19**: 37–55 (2001).

Fukawa K. and H. Suzuki, Blind interference canceling equalizer for mobile radio communications, *IEICE Trans. Commun.* **E77-B**: 580–588 (1994).

Grant S. J. and J. K. Cavers, Performance enhancement through joint detection of cochannel signals using diversity arrays, *IEEE Trans. Commun.* **46**: 1038–1049 (1998).

Lo B. C. W. and K. B. Letaief, Adaptive equalization and interference cancellation for wireless communication systems, *IEEE Trans. Commun.* **47**: 538–545 (1999).

Ranta P. A., A. Hottinen, and Z.-C. Honkasalo, Co-channel interference cancelling receiver for TDMA mobile systems, *Proc. IEEE Int. Conf. Communications*, 1995, pp. 17–21.

Ranta P. A., Z.-C. Honkasalo, and J. Tapaninen, TDMA cellular network application of an interference cancellation technique, *Proc. IEEE Vehicular Technology Conf.*, 1995, pp. 296–300.

CODE-DIVISION MULTIPLE ACCESS

Branimir R. Vojčić
Raymond L. Pickholtz
George Washington University
Washington, District of Columbia

1. INTRODUCTION

Multiple-access communications is a means by which many individual, geographically dispersed users access a shared medium and/or resources in order to transmit/receive information. Multiple access is used for local-area networks (LAN), satellite and cellular terrestrial radio networks, and other applications. Conventional multiple access schemes include random access such as ALOHA and its successor CSMA/CD, which is used in LANs, and structured orthogonal signals multiple access such as *frequency-* and *time-division multiple access* (FDMA, TDMA), which are used in satellite systems and terrestrial cellular radio. *Code-division multiple access* (CDMA) is a multiuser communications method, which uses spread-spectrum signals with uniquely addressable signature waveforms that permit the separation of each signal at the receiver. The main paradigm change in spread spectrum is that all users use the entire available spectrum simultaneously. It is possible, in some instances, to arrange for the signature waveforms to be orthogonal at the receiver, so that this separation can be affected by means of a linear correlator, wherein the desired users' signal is extracted while the other users' signals are completely suppressed. However, even when orthogonality is not perfect, for design or practical reasons, the spread-spectrum processing gain causes undesired multiuser signals to be significantly suppressed. In conventional CDMA receivers, such linear correlators are used and, to the extent that the *multiple-access interference* (MAI) is suppressed by the processing gain, it is tolerated. Naturally, this effect results in the characteristic "MAI-limited" channel whose capacity in terms of number of users is thereby limited according to what performance objective is specified. The two principal CDMA approaches are based on *direct-sequence* (DS) spreading and *frequency hopping* (FH). The FH transmission is conceptually very similar to conventional narrowband schemes, except that the carrier frequency is changed pseudorandomly over the spread spectrum bandwidth. In addition, it appears that DSCDMA is a less costly approach for commercial applications. Consequently, in our discussion we will mainly address DSCDMA and provide only a brief discussion of FHCDMA.

A major shortcoming of conventional detection CDMA systems is that, since each user contributes interference in proportion to their received power level, users that either generate excessive power, or whose power is received as larger than the desired signal (e.g., by virtue of being close to the receiver), degrade performance. This effect, sometimes called the "near–far problem," is a major impediment to practical CDMA using conventional detectors. The near–far problem is usually mitigated by exercising tight, closed-loop power control on all users so that all the received signals are of equal power at the receiver.

In Section 2 we first address the fundamental principles of CDMA using the notion that spread spectrum may be viewed as a way of embedding a signal into a high-dimensionality signal space and show how conventional CDMA receivers deal with MAI and its effects on both performance and user capacity. Next, in Section 3, we introduce some well-established performance measures by which, in addition to *signal-to-noise-and-interference ratio* (SNIR) and *bit-error probability* (BER), we can assess the behavior of multiuser systems. In Section 4, we examine the effect of the near–far problem and demonstrate, by an example, the consequences of imperfect power control. The obvious question is that since all the "signature" waveforms are presumable known at the receiver, why must we tolerate the MAI as if they were not known? Indeed, the optimum, maximum likelihood multiuser detector attempts to demodulate all received signals *jointly*. In Section 5, we will examine both the performance and complexity of this optimum approach and subsequently examine several suboptimal schemes, which exhibit significantly reduced complexity, and their performance. It appears that some of the multiuser detection schemes are practically feasible and, as such, can be exploited to improve the CDMA capacity, especially in situations in which tight power control is not achievable.

2. DIMENSIONALITY, PROCESSING GAIN, AND MULTIPLE ACCESS

A fundamental issue in CDMA is how this technique affords multiple simultaneous transmissions using a common bandwidth. The underlying principle is that of distributing relatively low-dimensional data signals in a high-dimensional environment. This is accomplished by means of spreading codes (signature waveforms), unique for each user so that all multiple-access signals are mutually nearly orthogonal. The idea of using quasiorthogonal signature waveforms (noiselike waveforms) for multiple access is originally due to Claude Shannon [2,3]. He perceived it as a democratic way of sharing the frequency spectrum: "If more people (signals) were there (in the

crowded radio spectrum), gradually the noise level would increase on each channel. But everyone could still talk, even though it might be a pretty noisy 'cocktail party' by that time" (this is now called a graceful degradation in CDMA).

In the "standard" problem of digital transmission, the set of M signaling waveforms $\{s_i(t), 0 \le t \le T, 1 \le i \le M\}$, known to both transmitter and receiver, is used to transmit $(\log_2 M)/T$ bps (bits per second). If, for example, $s_i(t)$ is sent, the received signal is $r(t) = s_i(t) + n_w(t)$, $0 \le t \le T$, where T denotes the symbol duration and $n_w(t)$ is additive, white Gaussian noise (AWGN) with two-sided power spectral density $N_0/2$ W/Hz. It is well known that the signal set can be completely specified by a linear combination $D \le M$ orthogonal basis functions. The dimensionality D, of the signal waveforms, is approximately equal to $2B_DT$, where B_D is the total (approximate) spectral occupancy of the employed signal set [4]. If the total *available* bandwidth is B_N, corresponding to an N-dimensional signal space, the maximum number of simultaneously active users, each one using D dimensionality, with orthogonal multiplexing is $K = N/D$. With quasiorthogonal multiplexing, however, it is possible to accommodate more than K users in the same bandwidth, but with some mutual multiple-access interference. In addition to sharing the bandwidth, the quasiorthogonal users share interference as well. The quasiorthogonal multiplexing is usually accomplished by means of *pseudonoise* (PN) spreading (signature) sequences, which have desired cross-correlation properties. At least one sequence is available to the cooperating transmitter and receiver, which may or may not know the PN sequences employed by other transmitter/receiver pairs.

A general model, which conveys the idea of CDMA, is as follows. Consider K simultaneous binary antipodal ($D = 1$) transmissions embedded in an N-dimensional signal space. Assuming jointly synchronous transmission, the aggregate of all transmitted signals can be represented by

$$x(t) = \sum_{i=1}^{K} x_i(t), \quad 0 \le t \le T \tag{1}$$

where the transmitted signal of the ith user is

$$x_i(t) = \sqrt{W_i} b_i s_i(t), \quad 0 \le t \le T \tag{2}$$

where W_i represents the signal energy per bit, b_i is the bit value (± 1) and $s_i(t)$ is the unit energy signature waveform of the ith user, defined as

$$s_i(t) = \sum_{k=1}^{N} s_{ik} \varphi_k(t), \quad 0 \le t \le T \tag{3}$$

where

$$s_{ik} = \int_0^T s_i(t) \phi_k(t)\, dt \tag{4}$$

and where $\{\phi_k(t), 1 \le k \le N\}$ is an orthonormal basis spanning the space:

$$\int_0^T \phi_l(t) \phi_m(t)\, dt = \delta_{lm} = \left\{ \begin{array}{ll} 1, & l = m \\ 0, & l \ne m \end{array} \right\} \tag{5}$$

Note that each binary antipodal signal requires one dimension ($D = 1$) for transmission, while N dimensions are employed to generate K distinct signaling waveforms. The factor N is usually called *spreading factor* or *processing gain*. The term spreading factor reflects the fact that the actual transmission bandwidth B_N (Fourier bandwidth), is N times larger than the Shannon bandwidth[1] of the modulated signal. The latter term, processing gain, stems from the capability of the spread signal to suppress interference by exploiting spectral redundancy and will be discussed subsequently.

In general, the PN sequence of the ith user, $\{s_{il}, \ldots, s_{iN}\}$, is chosen so as to have minimal possible cross-correlation with PN sequences of other users $\{s_{jl}, \ldots, s_{jN}\}, j = 1, \ldots, K$ and $j \ne i$. Here we assume, for the time being, that the sequences are random such that

$$E[s_{ij}] = 0, \quad \forall i, j, \quad i = 1, \ldots, K, \quad j = 1, \ldots, N \tag{6}$$

$$E[s_{il} s_{im}] = \frac{1}{N} \delta_{im}, \quad \forall i \tag{7}$$

$$E[s_{il} s_{jm}] = 0, \quad i \ne j, \quad \forall l, m \tag{8}$$

Although the spreading sequences are generated randomly, they are known to the communicators (at least to communicating pairs).

Consider next the conventional DSCDMA receiver, where the received signal is given by $r(t) = x(t) + n_w(t)$. The output of the ith correlation receiver is given by

$$\begin{aligned} U_i &= \int_0^T r(t) s_i(t)\, dt \\ &= \sum_{k=1}^{n} \left(\sqrt{W_i} b_i + \sum_{\substack{i=1 \\ l \ne i}}^{K} \sqrt{W_l} b_l s_{lk} s_{ik} + s_{ik} n_k \right), \end{aligned} \tag{9}$$

where

$$n_k = s_{ik} \int_0^T n_w(t) \phi_k(t)\, dt. \tag{10}$$

In conventional CDMA receivers the multiple access interference (MAI) at the output of correlation receiver is tolerated and a decision is made according to $b_i = \text{sgn}(U_i)$. Under these circumstances, a measure of performance which is monotonically related to the bit error rate (BER) is the SNIR, which can be expressed as

$$\begin{aligned} \text{SNIR}_i &= \frac{E^2[U_i \mid b_i]}{Var[U_i \mid b_i]} \\ &= \left[\frac{N_0}{2W_i} + \frac{1}{NW_i} \sum_{\substack{i=1 \\ l \ne i}}^{K} W_l \right]^{-1} \end{aligned} \tag{11}$$

The first term is due to thermal noise, the second term is due to MAI. It can be seen from the second term that

[1] By Shannon bandwidth we mean one-half the minimum number of orthogonal functions per T seconds that are required in a basis for a signal space in which signal can be represented [12].

MAI is suppressed by the factor N (spreading factor), which explains why the spreading factor is often called the processing gain. If we let $W_i/N_0 \to \infty$ (MAI dominates) and assume $W_i = W_l, \forall i, l$, then

$$\lim_{W_i/N_0 \to \infty} \text{SNIR}_i = \frac{N}{K-1} \tag{12}$$

This last result reveals the fundamental difference between orthogonal and quasiorthogonal multiplexing when conventional, single-user detection is employed; even for vanishingly small noise, the SNIR is finite. Hence, in a "cocktail party" of this example, to use Shannon's words, the number of simultaneous transmissions is $K \leq N/\text{SNIR}_{\text{REQ}} (K \gg 1)$, where SNIR_{REQ} corresponds to the desired transmission quality. Depending upon whether $\text{SNIR}_{\text{REQ}} < 1$ or $\text{SNIR}_{\text{REQ}} > 1$, more than N or less than N users, respectively, can transmit simultaneously.

Consider now the probability of error when the sign decision on U_i is employed. For large N we can invoke the central-limit theorem and assume that the MAI is Gaussian. Then it is easy to see that the conditional probability of error for the ith user is given by

$$P(i)\Big|_{\rho_{il}} = \frac{1}{2^{K-1}} \sum_{\text{all } b_l} Q\left[\sqrt{\frac{2W_i}{N_0}}\left(1 - \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} b_l \rho_{il}\right)\right] \tag{13}$$

where the first summation is over all possible combinations of data bits of interfering users and

$$\begin{aligned} \rho_{il} &= \int_0^T s_i(t) s_l(t)\, dt \\ &= \sum_{k=1}^{n} s_{ik} s_{lk} \end{aligned} \tag{14}$$

is the (random) cross-correlation between signature waveforms of the ith and lth user signature waveforms. For vanishingly small N_0, the probability of error is dominated by the term corresponding to the worst combination of interfering bits, so that

$$P(i)_{\text{wc}}\Big|_{\rho_{il}} = \frac{1}{2^{K-1}} Q\left[1 - \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} |\rho_{il}|\right]. \tag{15}$$

Since random spreading sequences were assumed, we obtain the following, after averaging over all signature waveforms, according to Jensen's inequality:

$$P(i)_{\text{wc}} \geq \frac{1}{2^{K-1}} Q\left[\sqrt{\frac{2W_i}{N_0}}\left(1 - \frac{1}{N} \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}}\right)\right]. \tag{16}$$

Hence, we can see that if

$$\left(1 - \frac{1}{N} \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}}\right) \leq 0$$

$P(i)_{\text{wc}}$ does not vanish as $W_i/N_0 \to \infty$. This phenomenon is not inherently a CDMA characteristic, but is rather a consequence of suboptimum detection. We will return to this in a subsequent section.

In a synchronous CDMA system, it is possible to choose spreading sequences, which are mutually orthogonal, as long as the number of signals does not exceed the dimensionality of the signal space. In that case we have orthogonal multiplexing (MAI free) and the probability of error is given by

$$P(i)_{\text{ort}} = Q\left(\sqrt{\frac{2W_i}{N_0}}\right) \tag{17}$$

Again with the Gaussian assumption[2] on MAI which we will relate below. Moreover, this is possible to achieve as long as the spreading sequences of K users are linearly independent [11]. At this point we abandon the analysis tool of random signature sequences and merely assume that they have *known* cross-correlations.

We now show that the signals can be processed at both transmitter and receivers so that the $\text{sgn}(U_i)$ is optimal and individual probabilities of error are given by Eq. (17).

Consider the vector of sufficient statistics (vector of correlation receivers' outputs in (9) for demodulation of all K signals jointly), given by

$$\mathbf{U} = \mathbf{RWb} + \mathbf{N}_w \tag{18}$$

where $\mathbf{R} = \{\rho_{ij}\}_{K \times K}$ is the cross-correlation matrix of signature waveforms $\mathbf{W} = \text{diag}(\sqrt{W_i})$ is a diagonal $K \times K$ matrix of signal amplitudes, $\mathbf{b}$ is the $K \times 1$ data vector, and $\mathbf{N}_w$ is a $K \times 1$ zero-mean Gaussian noise vector with covariance matrix $\mathbf{R}_N = \mathbf{R}N_0/2$. Hence, each component of the vector $\mathbf{U}$ contains the desired signal component, multiuser interference and a Gaussian noise component. When the matrix $\mathbf{R}$ is positive-definite, which is equivalent to the linear independence of K signature waveforms, there exist a unique Cholesky decomposition [18] of the matrix $\mathbf{R}$, such that $\mathbf{R} = \mathbf{G}^T\mathbf{G}$, where $\mathbf{G}$ is an upper triangular matrix. Consider a linear transformation $\mathbf{G}^{-1}$ at the transmitter such that the transmitted signal $x(t)$ is given by

$$x(t) = \mathbf{s}(t) T \mathbf{G}^{-1} \mathbf{W} \mathbf{b} \tag{19}$$

where $\mathbf{s}(t)$ is $K \times 1$ vector of signature waveforms. Then the vector of correlation receiver outputs is given by

$$\mathbf{U} = \mathbf{G}^T \mathbf{W} \mathbf{b} + \mathbf{N}_w \tag{20}$$

and by applying the linear transformation $(\mathbf{G}^T)^{-1}$ on the vector $\mathbf{U}$ in the receiver, we obtain

$$\mathbf{U}_0 = \mathbf{W}\mathbf{b} + \mathbf{Z} \tag{21}$$

where $\mathbf{Z}$ is a $K \times 1$ zero-mean Gaussian noise vector with covariance matrix $\mathbf{R}_Z = \mathbf{I}N_0/2$ and where $\mathbf{I}$ is the identity

[2] The additive noise is always assumed to be Gaussian in this regard.

matrix. Hence, by means of a pair of linear transformations $\mathbf{G}^{-1}$ and $(\mathbf{G}^T)^{-1}$, in the transmitter and the receiver, respectively, K users are decoupled and the resultant noise vector $\mathbf{Z}$ is white. Thus, individual sign decisions on the components of $\mathbf{U}_0$ will be optimal. Moreover, it is easy to see that the total transmit energy per bit interval is independent of the vector $\mathbf{b}$ and is equal to $W_{\text{tot}} = \sum_{i=1}^{K} W_i$, as it would be if orthogonal multiplexing were used in the first place. Indeed, the linear transformation $\mathbf{G}^{-1}$ in the transmitter gives rise to K orthogonal signature waveforms $\mathbf{p}(t)^T = \mathbf{s}(t)^T\mathbf{G}^{-1}$ and, similarly, by employing a bank of corresponding matched filters in the receiver, an orthogonal *multiuser communication system* results. A block diagram of the resulting end-to-end system is shown in Fig. 2. Actually, one could use the signature waveforms $\mathbf{p}(t)$ in the first place, or another set of orthogonal waveforms, in many multiple access scenarios, but there may be advantages to forcing this condition. Other decompositions of the correlation matrix, such as $\mathbf{R} = \mathbf{R}^{1/2}\mathbf{R}^{1/2}$ and $\mathbf{R} = \mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix, yield the same result as Cholesky decomposition, but the Cholesky decomposition is preferred due to its numerical stability. An adaptive scheme for joint transmitter–receiver optimization suitable for asynchronous channels is described elsewhere [38].

The described method of coordinated linear transformations in the transmitter and receiver in synchronous multiuser channels orthogonalizes the multiple access users, with no transmit power or noise enhancement penalties whatsoever, when users are linearly independent. It will be always possible to choose $\mathbf{s}(t)$ to have linearly independent users when the number of users is such that $K \leq N$. For $K > N$, this will not be possible for then the matrix $\mathbf{R}$ is singular so that the described coordinated transmitter/receiver linear transformations are not feasible. However, the use of joint transmitter/receiver processing with other criteria is not precluded.

At this point we would like to generalize the assertion presented above to the case where modulated signals may have different Shannon bandwidths but occupy the same Fourier bandwidth, such that the spreading factor of the ith user is $\gamma_i, i = 1, \ldots, K$. This generalization is succinctly summarized by the following proposition [12].

Proposition 1. Suppose that K users send their modulated signals to a single receiver, using a common Fourier bandwidth. Then zero interuser interference (IUI) is possible at the receiver only if the users transmit spread-spectrum signals whose spreading factors satisfy

$$\sum_{i=1}^{K} \frac{1}{\gamma_i} \leq 1 \tag{22}$$

It should be noted that in many practical multiple access channels, characterized by multipath propagation and loss of synchronization, for example, the single-user performance (no IUI) may not be achievable even when the conditions stated above are satisfied.

3. PERFORMANCE MEASURES

The SNIR and BER, introduced in the previous section, are the most common performance measures in digital communications. In multiuser communications, these performance measures very often do not admit analytic evaluation, whereas some asymptotic performance measures (corresponding to vanishingly small noise) may be readily found. To facilitate the comparison of different schemes in subsequent sections, we introduce the *asymptotic multiuser efficiency* (AME) and *near–far resistance*, originally proposed by Verdú [17].

Definition 1. The asymptotic multiuser efficiency of a multiuser detector, characterized by the probability of error for the ith user equal to $P(i)$, is given by

$$\eta_i = \sup\left\{ 0 \leq r \leq 1 : \lim_{N_0 \to 0} \frac{P(i)}{Q\left(\sqrt{r\frac{2W_i}{N_0}}\right)} < +\infty \right\} \tag{23}$$

or equivalently

$$\eta_i = \lim_{N_0 \to 0} \frac{W_{eff}(i)}{W_i} \tag{24}$$

where $W_{eff}(i)$ represents the effective signal energy of the ith user, reduced by the presence of MAI, such that the corresponding probability of error can be expressed as $P(i) = Q(\sqrt{2W_{eff}(i)N_0})$.

Definition 2. The near–far resistance of a multiuser detector for the ith user represents the minimum asymptotic efficiency over the relative energies of all other users:

$$\overline{\eta}_i = \inf_{\substack{W_j > 0 \\ j \neq i}} \eta_i \tag{25}$$

The AME measures the slope at which $P(i)$ goes to 0 in the high signal-to-noise ratio region for a given set of signal amplitudes of the desired and interfering users. On the other hand, the near–far resistance represents the AME for the worst-case combination of interfering signal amplitudes relative to the desired signal amplitude.

In the next sections, we use the AME and the near–far resistance to demonstrate the relative performance of conventional CDMA compared to optimal, or near-optimal, CDMA detectors.

4. CDMA, NEAR–FAR EFFECT, AND POWER CONTROL

Consider again the synchronous CDMA system of Section 2, whereby users employ *deterministic* spreading sequences with the cross-correlation between the ith and the jth spreading waveform given by ρ_{ij}. The probability of error for the ith user, when the conventional correlation receiver is employed, similarly as in (13), is given by

$$P(i) = \frac{1}{2^{K-1}} \sum_{\text{all } b_l} Q\left[\sqrt{\frac{2W_i}{N_0}}\left(1 - \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} b_l \rho_{il}\right)\right] \tag{26}$$

The corresponding AME is found as

$$\eta_i = \left[\max \left(1 - \sum_{\substack{i=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} |\rho_{il}| \right) \right]^2 \tag{27}$$

It can be easily seen from (27) that the AME of the conventional detector can take the value 0 for a user with relatively small W_i. Indeed, the near–far resistance of the; conventional receiver is 0 if at least one $\rho_{il} \neq 0, i \neq l, il = 1, \ldots, K$. The situation in which a strong interferer overwhelms the desired signal is usually referred to as the near–far effect. To provide the same performance to all receivers, power control is used so as to have all received signal energies the same ($W_i = W_l, \forall i, l$). The effect of imperfect power control on conventional CDMA, when the aggregate MAI was approximated by equivalent Gaussian noise,[3] was thoroughly analyzed in an earlier study [19]. To illustrate the near–far effect and the effect of power control error, consider a two-user example.

Let the cross-correlation between signature waveforms of users 1 and 2 be equal to ρ. Then, the probability of error for user 1 is given by

$$P(1) = \frac{1}{2} Q \left[\sqrt{\frac{2W_1}{N_0}} \left(1 - \sqrt{\frac{W_2}{W_1}} \rho \right) \right] + \frac{1}{2} Q \left[\sqrt{\frac{2W_1}{N_0}} \left(1 + \sqrt{\frac{W_2}{W_1}} \rho \right) \right]$$

and the corresponding AME is given by

$$\eta_1 = \left[\max \left(1 - \sqrt{\frac{W_2}{W_1}} \rho \right) \right]^2$$

Hence, we can see that $\eta_1 = 0$ for $W_2/W_1 \geq 1/\rho^2$. To provide the same performance to both users, we need $W_1 = W_2$, in which case $\eta_i = (1 - \rho)^2, i = 1, 2$.

To illustrate the effect of power control error, without examining specific power control mechanisms, we consider a simplified model of power control error. Let us assume that the received energy per bit for the ith user is $\alpha_i W_i, i = 1, 2$, where α_1 and α_2 are independent and identically distributed with the probability mass function given by

$$f_\alpha(\alpha_l) = \left\{ \begin{array}{ll} \Lambda, & \text{with probability } \dfrac{1}{4} \\ 1, & \text{with probability } \dfrac{1}{2} \\ \dfrac{1}{\Lambda}, & \text{with probability } \dfrac{1}{4} \end{array} \right\}, \quad i = 1, 2$$

[3] It should be noted that the Gaussian approximation for the MAI yields satisfactory performance estimation accuracy for the conventional correlation receiver, for low to moderate values of W_i/N_0 and when the number of interfering users and/or the processing gain are relatively large. This should not be mistaken for near optimality of the correlation receiver in the presence of MAI, which is in fact non-Gaussian.

Then, it is easy to calculate the AME in the presence of the power control error, by comparing the two-user performance with the single-user performance with the same distribution of the power control error; it is given by

$$\eta_1 = \left[\max \left(0, 1 - \sqrt{\frac{W_2}{W_1}} \rho \Lambda \right) \right]^2$$

The detrimental effect of power control error is apparent. Essentially, as can be seen from the formula for the AME, the power control error has an equivalent effect on the performance as an increase in the correlation coefficient between signature waveforms or/and an increase in the number of users; thus, it eats up the available capacity. In the next section we will see that multiuser detectors do not exhibit such a sensitivity to the imbalance in received signal amplitudes, and that the performance/capacity can be significantly improved, even compared to the conventional detector with perfect power control. Moreover, some multiuser detectors achieve optimal performance when the received energies are quite dissimilar.

5. MULTIUSER DETECTION AND INTERFERENCE CANCELLATION

As indicated in the previous section, the near–far effect is detrimental to conventional CDMA. This is a consequence of suboptimum detection, which ignores the interference from other users. When interfering signals are accounted for in the detection process, the adverse near–far problem can be eliminated. Moreover, these schemes, which we refer to as *multiuser detection* or *interference cancellation*, provide better performance than does a conventional correlation receiver, even when all signals arrive at the same power level at the common receiver (no near–far effect). The correlation receiver (optimum for AWGN channels) was often considered near optimum, based on the conjecture that the aggregate multiple access interference is approximately Gaussian, which may not be a good approximation for a finite-user population with dissimilar power levels. The non-Gaussian nature of the MAI is what enables optimum or near-optimum multiuser detectors to outperform, in some cases significantly, the conventional single-user correlation receiver. To obtain some insight into possible performance improvements and the incurred complexity, we will discuss some characteristic multiuser detection schemes. More detailed surveys of multiuser detection and interference cancellation schemes can be found in the literature [13,29,42].

Without loss of generality, we continue to use the synchronous model introduced in Section 2. The likelihood function (conditional probability density function of the channel output given the binary data vector $\mathbf{b}$ from all the K users) is given by

$$f_{r(t)}[r(t), t \in [0, T] | \mathbf{b}] = C \exp \left\{ -\frac{1}{N_0} \int_0^T \left[r(t) - \sum_{i=1}^{K} \sqrt{W_i} b_i s_i(t) \right]^2 dt \right\}. \tag{28}$$

The maximum-likelihood multiuser detector selects the most likely hypothesis given observation, choosing $\hat{b}$, which maximizes the likelihood function:

$$\hat{b} = \arg \min_{b \in \{-1,1\}^K} \int_0^T \left[r(t) - \sum_{i=1}^{K} \sqrt{W_i} b_i s_i(t) \right]^2 dt$$
$$= \arg \min_{b \in \{-1,1\}^K} (2\mathbf{U}^T - \mathbf{b}^T \mathbf{W}\mathbf{R})\mathbf{W}\mathbf{b} \tag{29}$$

To make a decision, 2^K hypotheses must be examined. Hence, the optimum multiuser detector has exponential complexity in the number of users, which may be restrictive for moderate and large values of K. In the asynchronous case, the optimum detector is the maximum-likelihood sequence detector operating on the sequences of MF outputs, again characterized with exponential complexity with respect to the number of users. It is the complexity of the optimum detector that has motivated research for suboptimum multiuser detectors with polynomial complexity.

One of the best-known multiuser detectors with linear complexity in the number of users is the *decorrelator* detector, first proposed in 1979 [16], but later thoroughly analyzed and correctly characterized [17]. The decorrelator detector follows immediately from the vector of sufficient statistics in (20). By applying the linear transformation $\mathbf{T} = \mathbf{R}^{-1}$ on the vector of sufficient statistics, we obtain

$$\mathbf{U}_0 = \mathbf{R}^{-1}\mathbf{U} = \mathbf{W}\mathbf{b} + \mathbf{Z} \tag{30}$$

where the transformed noise vector has the covariance matrix $\mathbf{R}_Z = \mathbf{R}^{-1}N_0/2$ and the MAI is decoupled. The receiver is completed by applying the sign rule, $\hat{\mathbf{b}} = \text{sgn}(\mathbf{U}_0)$, which is not optimal because the noise vector is not white. The MAI is completely eliminated at the expense of noise enhancement. It should be noted that this detector does not need the knowledge of received amplitudes, unlike the optimum detector, and achieves optimum near–far resistance. Moreover, the decorrelator is the maximum likelihood solution when signal amplitudes are not known. The probability of error for the decorrelator detector has the simplest form of all multiuser detectors and is given by

$$P_d(i) = Q\left(\sqrt{\frac{2W_i}{N_0} \frac{1}{R_{ii}^1}} \right) \tag{31}$$

where R_{ii}^1 represents the ith diagonal element of $\mathbf{R}^{-1}$. The corresponding AME is

$$\eta_d(i) = \frac{1}{R_{ii}^1} \tag{32}$$

It can be shown that the AME of the decorrelator detector is bounded, when the signature waveforms are linearly independent, according to [15]

$$\frac{4\lambda_{\max}\lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2} \le \frac{1}{R_{ii}^1} \le 1 \tag{33}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ represent the minimum and maximum eigenvalues of $\mathbf{R}$, respectively. Hence, as the system becomes heavily loaded and/or eigenvalue spread increases, the lower bound on the asymptotic efficiency decreases. When the matrix R becomes singular (loss of linear independence of signature waveforms), the near–far resistance of the decorrelator detector becomes equal to 0.[4] A generalization of the decorrelator detector to the asynchronous case can be found elsewhere [20].

For the two-user case of Example 1, the AME of the decorrelator detector is $\eta_d = 1 - \rho^2$ while the AME of the optimum detector was obtained [17] as $\eta_{ml} = \min[1, 1 + W_2/W_1 - 2\rho\sqrt{W_2/W_1}]$. In Fig. 1 we compare the AMEs of these two multiuser detectors with that of the conventional receiver; in all cases, perfect power control was assumed and $\rho = 0.7$. The advantage of multiuser detectors over the conventional receiver is apparent, even in the absence of the near–far effect and power control error ($W_1 = W_2$). It is a simple exercise to see that the AME of the decorrelator detector remains unchanged in the presence of power control error, and hence, the near–far resistance of the optimum detector does not degrade, either.

Another linear multiuser detector that is closely related to decorrelator detector is the *minimum mean-square-error* (MMSE) detector, originally proposed in the context of multiuser detection for asynchronous channels [21]. For the synchronous case, it is defined by Proposition 2:

Proposition 2. The MMSE detector for the synchronous multiuser system corresponding to the vector of sufficient statistics in (18) is defined by

$$\mathbf{T}^* = \left(\mathbf{R} + \frac{N_0}{2}\mathbf{W}^{-2} \right)^{-1} \tag{34}$$

$$\hat{\mathbf{b}} = \mathbf{T}^*\mathbf{U} \tag{35}$$

Proof: The MMSE criterion leading to the desired estimator is given by

$$\mathbf{T}^* = \arg \min_{\mathbf{T} \in R^{K*K}} E_{\mathbf{b},\mathbf{N}} \|\mathbf{T}\mathbf{U} - \mathbf{b}\|^2 \tag{36}$$

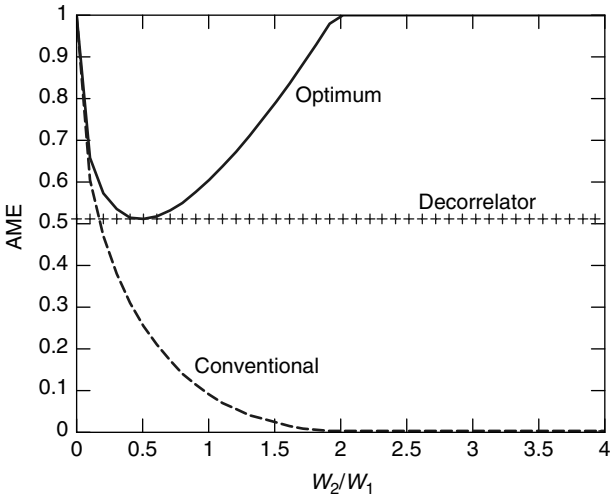


Figure 1. Comparison of AME for decorrelator, conventional, and optimum detectors.

[4] When signature waveforms are not linearly independent, no multiuser detector is near–far-resistant, in the synchronous case [17].

By applying the orthogonality principle, the optimum detector is obtained from

$$E[(\mathbf{TU} - \mathbf{b})\mathbf{U}^T] = 0 \tag{37}$$

from which we obtain

$$\mathbf{T}^* = \mathbf{W}^{-1}\left(\mathbf{R} + \frac{N_0}{2}\mathbf{W}^{-2}\right)^{-1} \tag{38}$$

and since the sign decision on $\mathbf{T}^*\mathbf{U}$ suffices, the factor $\mathbf{W}^{-1}$ is irrelevant and (34) follows. It should be noted that for vanishingly small noise the MMSE detector tends to the decorrelator detector, that is $\lim_{N_0 \to 0} \mathbf{T}^* = \mathbf{R}^{-1}$.

The most important feature of the MMSE detector is its suitability for adaptive implementation, whereby no information about interfering signals is required. Only the timing of the desired signal and a training sequence is required for the adaptive receiver to converge to its optimum setting. This adaptive MMSE detector was analyzed, in various forms, in Refs. 22–24 and references cited therein. Since the adaptive MMSE receiver does not need the knowledge of the signal attributes of interfering users, it represents a natural choice for mobile receivers in cellular or packet radiocommunications. Its performance in the steady state is superior to that of a conventional CDMA receiver, especially in near–far scenarios. Even in the presence of perfect power control, this receiver achieves roughly 2 times larger communication capacity than the conventional receiver [23].

A blind adaptive multiuser detector, closely related to the adaptive MMSE receiver was proposed [25]. The main advantage of this blind receiver is in that the training mode is not required and only an approximate knowledge of the signature waveform of the desired user is needed. The latter characteristic is important in mobile channels in which transmitted waveforms usually suffer from distortion. For a survey of adaptive multiuser detection schemes, the reader is referred to Verdu [26].

Finally, we would like to discuss an important class of multiuser detectors that is based on the cancellation of the estimated MAI in a feedback fashion [27–29]. Although nonlinear in structure, these detectors are characterized by linear complexity in the number of users. A simplified block diagram of one such detector, proposed by Varanasi and Aazhang, is shown in Fig. 2; for simplicity the first two detection stages are shown completely only for user 1.

This multistage detector employs the decorrelator receiver in the first stage to get an initial estimate of interfering bits. These bit estimates are multiplied by the corresponding correlation coefficients and amplitudes of interfering signals to reconstruct the MAI. The estimated MAI is subtracted from the matched filter output of the corresponding user, user 1, and a new decision is made on the thus obtained decision statistics. Since the matched filter output is used to obtain the decision statistics for the second stage, the noise enhancement effect of the decorrelator detector is not explicitly present in later stages. However, the effect of noise enhancement in the first stage propagates into subsequent stages through the tentative first stage decisions. The original version of the multistage detector employed conventional receivers in the first stage [27]. The version with the decorrelator

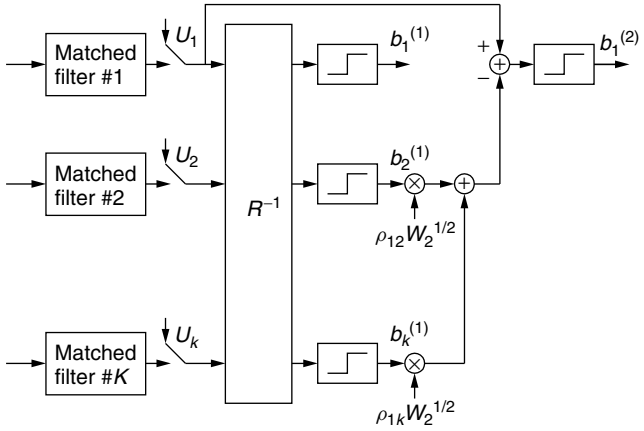


Figure 2. Block diagram of a two-stage detector with the decorrelator receiver in the first stage.

detector in the first stage achieves better performance and admits simpler analysis. The tentative decisions in the first stage are obtained by applying the sign decision on the vector of decision statistics defined in (30), that is, $\mathbf{b}_{(l)} = \text{sgn}(\mathbf{R}^{-1}\mathbf{U}_0)$. The vector of decision statistics for the second stage is formed as

$$\mathbf{Y} = \mathbf{Wb} + (\mathbf{R} - \mathbf{I})\mathbf{W}(\mathbf{b} - \hat{\mathbf{b}}) \tag{39}$$

The second-stage decisions are made according to $\mathbf{b}^{(2)} = \text{sgn}(\mathbf{Y})$. It is apparent that this detector results in the isolated transmission performance in the second stage when the first-stage tentative decisions are perfect. However, when the decision errors are made in the first stage, the corresponding interference terms double, thus adversely affecting the decisions in the second stage.

Returning to our two-user example, it is easy to show, by exploiting the results in Ref. 28, that the probability of error for user 1 in the second stage is given by

$$\begin{aligned} P(1)^{(2)} = {} & Q\left(\sqrt{\frac{2W_1}{N_0}}\right) \times \left[1 - Q\left(\sqrt{\frac{2W_2(1-\rho^2)}{N_0}}\right)\right] \\ & + \frac{1}{2}Q\left(\sqrt{\frac{2W_2(1-\rho^2)}{N_0}}\right) \\ & \times \left\{Q\left[\sqrt{\frac{2W_1}{N_0}}\left(1 - 2\rho\sqrt{\frac{W_2}{W_1}}\right)\right]\right. \\ & \left. + Q\left[\sqrt{\frac{2W_1}{N_0}}\left(1 + 2\rho\sqrt{\frac{W_2}{W_1}}\right)\right]\right\} \end{aligned}$$

It can be shown that the AME in this case is

$$\eta_{\text{ms}} = \min(\alpha, 1)$$

where

$$\alpha = (1-\rho^2)\frac{W_2}{W_1} + \left[\max\left(0, 1 - 2\rho\sqrt{\frac{W_2}{W_1}}\right)\right]^2$$

To minimize the effect of doubling the interference when wrong decisions are made, soft interference cancellation

can be employed. The idea here is to make soft tentative decisions at the decorrelator output in such a way that decisions are somehow weighted according to their reliability. It was shown [30] that two-stage detection with soft interference cancellation can significantly outperform its hard-interference cancellation counterpart, and in the two-user case achieves optimum AME. Specifically, for $K = 2$, it was shown that a linear clipper as a soft weighting nonlinearity achieves the optimum AME when the threshold of the clipper is chosen according to

$$\delta\left(\rho, \sqrt{\frac{W_2}{W_1}}\right) = \sqrt{W_1}\max\left[0, \frac{|\rho|\left(\sqrt{\frac{W_1}{W_2}} - |\rho|\right)}{1-\rho^2}\right] \quad (40)$$

Hence, when the decorrelator output is larger than δ, a hard decision is made, otherwise the decorrelator output is scaled proportionally to the linear part of the clipper, before feedback cancellation. In the K user case the soft limiting nonlinearity is optimized on a pairwise basis according to (40). For numerical results on possible improvements with soft interference cancellation, the reader is referred to Ref. 30. Similar performance improvements for the asynchronous case have been demonstrated [32].

BIOGRAPHIES

Branimir R. Vojčić is a professor in, and chairman of, the Department of Electrical and Computer Engineering at the George Washington University, Washington, D.C. He has received his D.Sc. degree from the University of Belgrade, Yugoslavia. His current research interests are in the areas of communication theory, performance evaluation and modeling mobile and wireless networks, code division multiple access, multiuser detection, adaptive antenna arrays and space-time coding and ad-hoc networks. He has also been an industry consultant in these areas and has published and lectured extensively in these areas. Dr. Vojcic is a senior member of IEEE, was an associate editor for IEEE Communications Letters and a receipient of 1995 National Science Foundation CAREER Award.

Raymond L. Pickholtz is a professor in, and former chairman of, the Department of Electrical and Computer Engineering at The George Washington University, Washington, D.C., and received his Ph.D. in electrical engineering from the Polytechnic Institute of Brooklyn, New York. He was an editor of the *IEEE Transactions on Communications*, and guest editor for special issues on computer communications, military communications spread spectrum systems and social impacts of technology. He is currently the coeditor of chief of the *Journal of Communications and Networks*. He has published scores of papers (several award winning), acts as a consultant to industry, and holds six U.S. patents.

Dr. Pickholtz is a fellow of the IEEE, AAAS, and the Washington Academy of Sciences. He was elected president of the IEEE Communications Society in 1991. He received the Donald W. McLellan Award in 1994. He was a visiting Erskine fellow at the University of Canterbury, Christchurch, New Zealand, 1997. He was awarded the IEEE Third Millenium Medal in 2000.

BIBLIOGRAPHY

1. A. Viterbi, *CDMA Principles of Spread Spectrum Communications*, Addison-Wesley, Reading, MA, 1995.
2. J. R. Pierce, A conversation with Claude Shannon, *IEEE Commun. Mag.* **22**: 123–126 (May 1984).
3. J. Costas, Poisson, Shannon and the radio amateur, *IEEE Proc.* **47**: 2058–2068 (Dec. 1959).
4. J. M. Wozencraft and I. M. Jacobs, *Principles of Communications Engineering*, Waveland Press, Prospect Heights, IL, 1990.
5. R. Dixon, *Spread Spectrum Systems*, Willey-Interscience, New York, 1984.
6. M. Simon, J. Omura, R. Scholtz, and B. Levitt, *Spread Spectrum Communications Handbook*, McGraw-Hill, New York, 1995.
7. R. Peterson, R. Ziemer, and D. Borth, *Introduction to Spread Spectrum Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
8. J. Proakis, *Digital Communications*, McGraw-Hill, New York, 1995.
9. R. Pickholtz, D. Schilling, and L. Milstein, Theory of spread spectrum communications — a tutorial, *IEEE Trans. Commun.* **COM-30**(5): (May 1982).
10. R. Pickholtz, D. Schilling, and L. Milstein, Theory of spread spectrum communications — a tutorial (revisions), *IEEE Trans. Commun.* **COM-32**(2): (Feb. 1984).
11. B. R. Vojcic and R. L. Pickholtz, Joint transmitter receiver optimization in synchronous multiuser communications, Information Theory Workshop, Rydzyna, Poland, 1995.
12. J. L. Massey, Spectrum spreading and Multiple accessing, Information Theory Workshop, Rydzyna, Poland, 1995.
13. S. Verdu, Recent progress in multiuser detection, in N. Abramson, ed., *Multiple Access Communications*, IEEE Press, New York, 1993.
14. A. Duel-Hallen, J. Holtzman, and Z. Zvonar, Multiuser detection for CDMA systems, *IEEE Pers. Commun.* **2**: 46–58 (April 1995).
15. R. Lupas, *Near-Far Resistant Linear Multiuser Detection*, Ph.D. thesis, Princeton Univ., Princeton, NJ, 1989.
16. K. S. Schneider, Optimum detection of code division multiplexed signals, *IEEE Trans. Aerospace Electric Syst.* **AES-15**: 181–185 (Jan. 1979).
17. R. Lupas and S. Verdu, Linear multiuser detectors for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **IT-34**: (1988).
18. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1983.
19. B. R. Vojcic, R. L. Pickholtz, and L. B. Milstein, Performance of DS-CDMA with imperfect power control operating over a low earth orbiting satellite link, *IEEE J. Select. Areas Commun.* **12**: (May 1994).
20. R. Lupas and S. Verdu, Near–far resistance of multiuser detectors in asynchronous channels, *IEEE Trans. Commun.* **COM-38**: 496–508 (April 1990).

21. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multiuser communications, *IEEE J. Select. Areas Commun.* 683–690 (May 1990).

22. P. B. Rapajic and B. S. Vucetic, Adaptive receiver structures for asynchronous CDMA systems, *IEEE J. Select. Areas Commun.* 685–697 (May 1994).

23. S. L. Miller, An adaptive direct-sequence code-division multiple-access receiver for multiuser interference rejection, *IEEE Trans. Commun.* 1746–1755 (Feb.–April 1995).

24. U. Madhow and M. L. Honig, MMSE interference suppression for direct-sequence spread-spectrum CDMA, *IEEE Trans. Commun.* 3178–3188 (Dec. 1994).

25. M. Honig, U. Madhow, and S. Verdu, Blind adaptive multiuser detector, *IEEE Trans. Inform. Theory* 944–960 (July 1995).

26. S. Verdu, Adaptive multiuser detection, *Proc. IEEE Int. Symp. Spread Spectrum Theory and Applications*, Oulu, Finland, July 1994.

27. M. Varanasi and B. Aazhang, Multistage detection in asynchronous code-division multiple-access communications, *IEEE Trans. Commun.* **COM-38**(4): (April 1990).

28. M. Varanasi and B. Aazhang, Near-optimum detection in Synchronous CDMA systems, *IEEE Trans. Commun.* **COM-39**: (May 1991).

29. A. Duel-Hallen, Decorrelating decision-feedback multiuser detector for asynchronous code-division multiple-access channel, *IEEE Trans. Commun.* **COM-41**: 285–290 (1993).

30. V. Vanghi and B. Vojcic, Soft interference cancellation in multiuser communications, *Int. J. Wireless Pers. Commun.* (Special Issue on Signal Separation and Interference Cancellation for Personal, Indoor and Mobile Radio Communications), **3**: 111–128 (1996).

31. P. Patel and J. Holtzman, Performance comparison of a DS/CDMA system using a successive interference cancellation (IC) scheme and a parallel IC scheme under fading, *Int. Conf. Communication*, New Orleans, 1994, pp. 510–514.

32. X. Zhang and D. Brady, Soft-decision multistage detection for asynchronous AWGN channels, *Proc. 31st Annual Allerton Conf. Communication, Control and Computing*, Allerton House, Urbana-Champaign, IL, Oct. 1993.

33. B. R. Vojcic, Transmitter precoding for synchronous multiuser communications, Workshop on Mobility Management, George Mason University, Fairfax, VA, Oct. 1994.

34. B. R. Vojcic, Transmitter precoding in multiuser communications, *Proc. 1995 IEEE IT Workshop on Information Theory, Multiple Access and Queueing*, St. Louis, April 1995.

35. Z. Tang and S. Cheng, Interference cancellation for DS-CDMA systems over flat fading channels through predecorrelating, *Proc. PIMRC'94*, Hague, The Netherlands, 1994.

36. Y. Yasuda, K. Kashiki, and Y. Hirata, High-rate punctured convolutional codes for soft decision Viterbi decoding, *IEEE Trans. Commun.* **COM-32**: 315–319 (March 1984).

37. B. Vojcic and R. Pickholtz, Spectral shaping in DS CDMA on a satellite link, *Proc. AIAA Conf.*, Washington, DC, Feb. 1996.

38. P. Rapajic and B. Vucetic, Linear adaptive transmitter-receiver structures for asynchronous CDMA systems, *Eur. Trans. Telecommun.* **6**: 21–27 (Jan.–Feb. 1995).

39. W. M. Jang and B. Vojcic, Transmitter precoding in synchronous multiuser communications over multipath channels, *Proc. Symp. Interference Rejection and Signal Separation in Wireless Communications*, New Jersey Institute of Technology, Newark, NJ, March 1996.

40. J. Hui, Throughput analysis for code division multiple accessing of the spread spectrum channel, *IEEE J. Select. Areas Commun.* **SAC-2**: 482–486 (July 1984).

41. M. Pursley, Performance evaluation of phase-coded spread spectrum multiple-access communication — system analysis, *IEEE Trans. Commun.* **COM-25**: 795–799 (Aug. 1977).

42. S. Verdu, *Multiuser Detection*, Cambridge Univ. Press, Cambridge, UK, 1998.

CODING FOR MAGNETIC RECORDING CHANNELS

LIH-JYH WENG
Maxtor Corporation
Shrewsbury, Massachusetts

1. INTRODUCTION

Since 1990, the areal magnetic recording density for rigid disks has been growing at a rate of 60% annually and the trend is accelerating [1]. This is one of the most important factors for the magnetic recording capacity of more recent rigid disks to double every nine months. To maintain this pace of density increase and to meet the stringent requirement of today's digital storage systems in both the data integrity and recording density increase, error-correcting codes (ECCs) become both indispensable in achieving the low postdecoding bit error rate and effective in overall recording density optimization. An ECC encoder first encodes the user data into ECC codewords; these code words are then mapped, using a modulation code, into a form suitable for the write circuit to record the encoded data to the disk surfaces. During a read, the process is reversed; the readback signal is first processed to recover the codewords for the modulation code; the role of the ECC is to correct all errors that may have occurred during the entire read/write process. Figures 1 and 2 give the logical flow of the entire process. The main purpose of the ECC is to ensure that the user data in Fig. 1 are identical to the user data of Fig. 2. In more conventional communication systems, the main emphasis is to communicate reliably

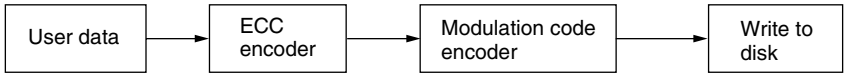


Figure 1. Logic flow of writing data to disks.

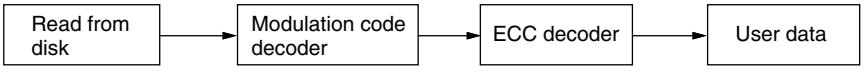


Figure 2. Logic flow of retrieving data from disks.

from one point to another with an acceptable amount of delay. On the other hand, disk and tape drives serve the purpose of storing data at one time instant and reading them back reliably at another instant.

Some of the modern modulation codes have the capability of performing a limited amount of error correction by themselves [2]. However, the error rate at the output of the modulation code decoder, in most cases, does not achieve the stringent low error rates required of present storage systems. An ECC is a simple and effective way to bring the error rate at the output of the modulation code decoder to the level specified by the user. Furthermore, with the separation of the modulation code and ECC, each component can be more effectively designed to achieve the overall optimality of the entire recording system.

2. REED–SOLOMON CODES: ORIGIN AND BASIC CHARACTERISTICS

In the early days, the recording density was relatively low, and, hence, the main concern was a single burst in a sector. The single-burst correcting codes and codes that correct a small number of errors were the dominant ECCs at the time. Reference 2 is a good source for early ECC implementations. When the recording density increases, the most likely errors are not confined to single bursts; therefore, codes that can deal with both random and burst errors are necessary to adequately protect the written data. Reed–Solomon codes have such characteristics and also possess efficient encoding and decoding algorithms, which can be readily implemented. Consequently, they are presently the most widely used codes for magnetic disk drives. All the techniques discussed here are applicable to rigid disk drives and tape drives as well as optical drives. When complexity is a main concern, tape drives also use other coding techniques. On tape drives, the error bursts tend to be much longer than those observed in disk drives. Some structure is usually introduced among code blocks to deal with long bursts. Specifically, tape drives often employ so-called two-dimensional codes. This can be pictured with data arranged in a two-dimensional array; both the rows and the columns of the array are protected by ECCs, or some sort of redundancy is introduced both vertically and horizontally. Furthermore, redundancies can be introduced diagonally or along any line of a given slope in the array. The main reason for using codes along more than two directions is to use very simple codes, such as single-parity-check binary codes along each direction to provide sufficient protection. Array codes [4] provide such advantages.

The Reed–Solomon codes employed in recording systems need to satisfy some special constraints, which may be different from those in other applications. In the following discussion, several important considerations for applying Reed–Solomon codes in magnetic recording systems are addressed. First, a simple example is presented to introduce some terminology commonly used in Reed–Solomon codes as well as more generally in ECCs. Other related issues concerning the code applications to magnetic recording such as implementation, block synchronization, interleaving, performance, and error detection are discussed. Finally, some remarks concerning tape drives, RAID (reductant array of independent disks) systems, soft decoding and increasing the sector size are also addressed.

2.1. An Example of a Reed–Solomon Code

A Reed–Solomon (10,3,8) code over Galois Field code GF(2^4) is selected for illustration purposes. The meaning of the parameters 10, 3, and 8 will be given shortly. First, the field is a Galois field, commonly denoted as GF(2^m), where m is 4 in this example. This field is used to perform all arithmetic operations needed for the Reed–Solomon code. Therefore, the definition and essential properties for the GF(2^m) should be given first. GF(2^m) is often referred to as an *extension field* of GF(2) = {0, 1}. A more concrete way of viewing a GF(2^m) field is to list the field elements as all possible m-bit binary representation of the integer $0, 1, 2, \ldots, 2^m - 1$. For this example, $m = 4$, the field elements of GF(2^4) are the set {0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111}. A field needs two basic operations: addition (+) and multiplication (∗) (symbol ∗ used to indicate Galois field multiplication). The addition of two elements (a, b, c, d) and (e, f, g, h) are defined to be $(a + e, b + f, c + g, d + h)$, where a, b, c, d, e, f, g, h are either one or zero, the elements of GF(2) = {0, 1}. The same symbol + is used for addition over GF(2) and addition over GF(2^m). The addition over GF(2) is modulo-2 addition, which follows the rule that $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, and $1 + 1 = 0$. For example, $(1010) + (1100) = (0110)$. The multiplication rule over GF(2) is defined by $0 * 0 = 0 * 1 = 1 * 0 = 0$, and $1 * 1 = 1$. In more familiar engineering terms, the addition is equivalent to an EXOR (exclusive OR) operation and the multiplication is the same as the AND operation commonly seen in logic. To define the multiplication rules for GF(2^m), an irreducible binary polynomial of degree m is needed. A degree m binary polynomial, where the coefficients are either 0 or 1, is said to be irreducible if it is not divisible by any binary polynomial of degree lower than m except the trivial degree 0 polynomial, specifically, the constant 1. In the case of $m = 4$, $p(x) = x^4 + x + 1$ is an irreducible polynomial. Each element of GF(2^m) is associated with a unique polynomial of degree $m - 1$ or lower. The elements of GF(2^4) can be expressed in two convenient ways: GF(2^4) = {0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, …} = $\{0, x^0, x, x + 1, x^2, x^2 + 1, x^2 + x, x^2 + x + 1, x^3, x^3 + 1, \ldots\}$. Let $a(x)$ and $b(x)$ be two elements of GF(2^m); then the multiplication rule is given by $c(x) = a(x) * b(x) \bmod p(x)$. Since the degree of $p(x)$ is m, the degree of $c(x)$, which is the remainder of $a(x) * b(x)$ divided by $p(x)$, is less than m. Next an element of GF(2^m) is selected whose powers generate every possible nonzero element of the field by multiplying itself many times. In this selected field, any of the following nonzero elements can be used as such a generating element: $x, x^2, x + 1, x^2 + 1, x^3 + 1, x^3 + x + 1, x^3 + x^2 + 1$, and $x^3 + x^2 + x$. Let the generating element be x, then $x^0 = 1, x^1 = x, x^2 = x^2, x^3 = x^3, x^4 = x + 1, x^5 = x * x^4 = x * (x + 1) = x^2 + x, x^6 = x * x^5 = x * (x^2 + x) = x^3 +$

$x^2, x^7 = x * x^6 = x * (x^3 + x^2) = x^4 + x^3 = x + 1 + x^3 = x^3 + x + 1, x^8 = x * x^7 = x * (x^3 + x + 1) = x^4 + x^2 + x = x + 1 + x^2 + x = x^2 + 1, x^9 = x * x^8 = x * (x^2 + 1) = x^3 + x,$ .... Expressed as powers of the generating element, the field elements are written as $\mathrm{GF}(2^4) = \{0000, 0001, 0010, 0100, 1000, 0011, 0110, 1100, 1011, 0101, 1010, \ldots\} = \{0, x^0, x^1, x^2, x^3, x^4, x^5, x^6, x^7, x^8, x^9, \ldots\}$. In this case, all the nonzero elements of the field can be expressed as distinct powers of x. There are exactly $2^m - 1$ distinct powers $x^j \bmod p(x)$ for $j = 0, 1, 2, \ldots, 2^m - 2$. The class of polynomials, powers of whose roots generate all nonzero field elements, is collectively called *primitive polynomials*. A nonprimitive polynomial can also be used to form the field. For example, the polynomial $p'(x) = x^4 + x^3 + x^2 + x + 1$ is irreducible but not primitive. This can be easily checked by the fact that $x^4 = x^3 + x^2 + x + 1 \bmod p'(x)$ and $x^5 = x * x^4 = x * (x^3 + x^2 + x + 1) = x^4 + x^3 + x^2 + x = (x^3 + x^2 + x + 1) + x^3 + x^2 + x = 1 \bmod p'(x)$. Since x is not a primitive element for the field defined by $p'(x)$, a different element must be used. Let $\alpha = x + 1$, then the successive powers of $\alpha^j \bmod p'(x)$ for $j = 0, 1, 2, \ldots, 2^m - 2$, are all distinct. Therefore, another $\mathrm{GF}(2^4)$ can be generated by $p'(x)$ using α^j as a nonzero element. It is convenient to represent $\mathrm{GF}(2^m)$ as $\{0, \alpha^0, \alpha^1, \alpha^2, \alpha^3, \ldots\}$, where α is a primitive element. For the primitive irreducible $p(x)$ shown above, $\alpha = x$. For all primitive polynomials, the generating element α can be selected to be x. For the remaining discussion, the binary primitive polynomial $p(x)$ is used. The generating element α is sometimes referred to as the *primitive element*.

With α^j defined for $j = 0, 1, 2, \ldots, 2^m - 2$, the product of two elements α^i and α^j can be obtained in at least two ways: (1) express α^i and α^j as polynomials, find the product as the polynomial multiplication of $\alpha^j * \alpha^j \bmod p(x)$; (2) using the simple exponent addition rule, namely $\alpha^i * \alpha^j = \alpha^k$ with $k = i + j \bmod 2^m - 1$. There are $2^m - 1$ nonzero distinct elements, each corresponding to a distinct power $0, 1, 2, \ldots, 2^m - 2$. It should be noted that $\alpha^s = \alpha^0$ if $s = 2^m - 1$. The identity element for $\mathrm{GF}(2^m)$ multiplication is α^0. The identity element for $\mathrm{GF}(2^m)$ addition is the zero element. In $\mathrm{GF}(2^m)$, the subtraction operation is the same as the addition operation and the "division" a/b can be considered as $a * (b^{-1})$ for any nonzero element b. The element b^{-1} can be obtained from b by examining the exponent of b; namely, if $b = \alpha^j$, then $b^{-1} = \alpha^k$ such that $j + k = 0 \bmod 2^m - 1$. For example, in the example of $\mathrm{GF}(2^4)$, to find the inverse of $x^3 + x$, which is α^9, the exponent is 9; it is a simple matter to determine that $9 + 6 = 15 = 0 \bmod 2^4 - 1$. Therefore, $(\alpha^9)^{-1} = \alpha^6$. References 5 and 6 provide more detailed properties of Galois fields.

In software or firmware implementations, the most common method of computing the inverse of a field element is by making use of a logarithm table and an antilogarithm table. The logarithm table associates every nonzero element with an exponent; for example, in the $\mathrm{GF}(2^4)$ above, the logarithm table gives the elements 0100 and 1001 their respective exponents 3 and 14. On the other hand, the antilogarithm table takes the exponents 3 and 14 as inputs and produces the elements 0100 and 1001, respectively. The approach of finding the inverse is very similar to the real-number computation using a logarithm table and an antilogarithm table [6]. In a hardware implementation, the logarithm table and the antilogarithm table are seldom provided. To find the inverses, a special algorithm is employed; the algorithm is often dependent on the symbol size m selected and sometimes depends on the irreducible polynomial generating the field [7,8].

2.2. Reed–Solomon Encoder

A Reed–Solomon code can now be defined over the field $\mathrm{GF}(2^m)$. One way to specify a Reed–Solomon code is to define the roots of the generator polynomial of the code. The generator polynomial can be written as

$$g(x) = (x + \alpha^L) * (x + \alpha^{L+1}) * (x + \alpha^{L+2}) * \cdots * (x + \alpha^{L+R-1}) \tag{1}$$

This is the generator polynomial of a Reed–Solomon (n, k, d) code, where n is the code length, k is the number of information symbols, and d is the minimum distance (or Hamming distance) among all possible codewords, where the distance between two codewords is equal to the number of symbols at which these two codewords differ. The code rate of a code is defined as the ratio k/n, which is a number between 0 and 1. In magnetic recording, high-rate codes are frequently employed. The value L can be selected arbitrarily; n is at most $2^m - 1$ for easy hardware implementation; and $d = n - k + 1 = R + 1$, where R is equal to the number of redundant symbols of the code or the degree of the generator polynomial. In other words, there are n symbols in a codeword, among which k symbols can be assigned arbitrarily as information symbols or data symbols. The degree of the generator polynomial is equal to $R = n - k$. The one requirement for the roots is that they must be consecutive roots $\alpha^L, \alpha^{L+1}, \alpha^{L+2}, \ldots, \alpha^{L+R-1}$. The choice of the value L does not change the code minimum distance. Let $n = 10, k = 3, R = 7$ or $d = 8$, and set $L = 11$. Then the generator polynomial is given by

$$\begin{aligned} g(x) &= (x + \alpha^{11}) * (x + \alpha^{12}) * (x + \alpha^{13}) * (x + \alpha^{14}) \\ &\quad * (x + \alpha^0) * (x + \alpha^1) * (x + \alpha^2) \\ &= \alpha^0 * x^7 + \alpha^1 * x^6 + \alpha^3 * x^5 + \alpha^{12} * x^4 + \alpha^{11} * x^3 \\ &\quad + \alpha^0 * x^2 + \alpha^{11} * x + \alpha^8 \end{aligned} \tag{2}$$

All the codewords of this Reed–Solomon (10,3,8) code can be expressed in polynomial form such as

$$c(x) = c_9 * x^9 + c_8 * x^8 + c_7 * x^7 + c_6 * x^6 + c_5 * x^5 + c_4 * x^4 + c_3 * x^3 + c_2 * x^2 + c_1 * x + c_0 \tag{3}$$

The requirements for $c(x)$ to be a codeword are (1) its degree is no higher than 9 and (2) it must be a multiple of the generator polynomial [i.e., $c(x) = m(x) * g(x)$]. To simplify the notation, the polynomial is often written as a vector:

$$c(x) = (c_9, c_8, c_7, c_6, c_5, c_4, c_3, c_2, c_1, c_0) \tag{4}$$

For example, $g(x) = (0, 0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8)$ and $x * g(x) = (0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0)$. It is clear that both $g(x)$ and $x * g(x)$ are codewords because they are both multiples of the generator polynomial $g(x)$. Another way of specifying a code is by a generator matrix. The generator matrix of this (10,3,8) code is given as

$$G = \begin{bmatrix} 0, 0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8 \\ 0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0 \\ \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0, 0 \end{bmatrix} \quad (5)$$

With the help of the generator matrix, a codeword can be expressed as

$$(c_9, c_8, c_7, c_6, c_5, c_4, c_3, c_2, c_1, c_0) = (m_2, m_1, m_0) * G \quad (6)$$

where m_2, m_1, m_0 are arbitrary elements of GF(2^m). It should be noted that not all codes possess generator polynomials or generator matrices. A class of codes, called *cyclic codes*, can be specified by generator polynomials. Another class of codes, called *linear codes*, have generator matrices. The Reed–Solomon codes are both cyclic and linear; therefore, they have rich algebraic properties and structure, which provide the foundation of being well studied and understood. As a result, Reed–Solomon codes are widely employed. As mentioned previously, the distance between two codewords is the number of symbols at which these two codewords differ. For example, the distance between $g(x)$ and $x * g(x)$ is 9 (symbols.) If all possible codewords of this Reed–Solomon (10,3,8) code are listed, every pair of distinct codewords differs in at least eight (8) symbols. Therefore, any combination of three or fewer symbol errors will not change a codeword closer to another codeword. Consequently, this code can correct all possible combinations of three or fewer errors. In addition, any error pattern of four symbol errors cannot change a codeword closer to another codeword; in the worst case, the codeword corrupted by a four-error pattern can be at the same distance from many codewords. As a result, the decoder cannot decode the erroneous code word to a unique codeword. In this case, the errors are detected but not corrected. In general, a distance d Reed–Solomon code is capable of correcting any combination of t symbol errors per code block if $t < or = \lfloor d/2 \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x ($\lfloor x \rfloor$ is referred to as the "floor function" of x). A more interesting way of employing a distance d Reed–Solomon code is to use a distance d code to correct t or fewer errors and detect with certainty if there are $t+1, t+2, \ldots, e$ symbol errors per code block provided $t + e < d$ and $t < e$ [6,9–11]. In this application, it is very important that the value t be used to determine the postdecoding symbol or block error rate and the value e be designed to ensure that the miscorrection probability meets the stringent requirement for data storage applications. It should also be noted that for highly redundant Reed–Solomon codes, the value of e can be set to be very close to or the same as t. In other words, very little or no additional symbol detection is needed to achieve the level of miscorrection probability demanded by the data storage systems.

In the last paragraph, the distance is measured by symbols. Because every symbol contains m bits, it is also possible to measure the distance in bits. For example, the Reed–Solomon (10,3,8) code over GF(2^4) is also a binary (40,12,10) code. The minimum distance of the binary code is now measured in bits. This binary code can correct four (4) random bit errors and detect five (5) random bit errors with certainty. The minimum symbol distance of a Reed–Solomon code can be easily determined as $d = n - k + 1$. However, the minimum distance of its binary expansion version cannot be determined easily except that the lower bound of the minimum distance is d, the same as the symbol minimum distance. This lower bound may not be very tight, as can be seen from the example presented above.

The encoding process of a Reed–Solomon (n, k, d) code is often achieved through a division process, which ensures that the codeword $c(x)$ is a multiple of the generator polynomial $g(x)$. It is possible to achieve this with the additional condition that the data symbols of $c(x)$ are unaltered. This class of codes is called *systematic codes*.

2.3. Reed–Solomon Decoder

In decoding, the first step is also by division. For error-detection purposes, a corrupted codeword $c'(x)$ is divided by $g(x)$. If the remainder of the division is zero, the decoder assumes that the codeword is error-free; otherwise, the decoder assumes the codeword is corrupted. Instead of dividing by the generator polynomial $g(x)$, an equivalent method is to divide the corrupted codeword by each factor $x + \alpha^{L+j}$ of $g(x)$. The results of the divisions are called syndromes S_{L+j}. Therefore, for an uncorrupted codeword, all the syndromes S_{L+j} must be zero. The syndrome polynomial $S(x)$ is a degree $R-1$ polynomial defined as

$$S(x) = S_L + S_{L+1} * x + S_{L+2} * x^2 + S_{L+3} * x^3 + \cdots + S_{L+R-1} * x^{R-1} \quad (7)$$

In most algebraic coding textbooks [6,9–11], the starting point of decoding Reed–Solomon codes is the syndrome polynomial, which contains all the necessary information to find error locations and error values. In Reed–Solomon codes over GF(2^m), the symbol positions within a code block are denoted by $\alpha^0, \alpha^1, \alpha^2, \ldots, \alpha^{n-2}, \alpha^{n-1}$, with α^p corresponding to symbol position P. The first data symbol occupies position $n-1$, and the last redundant symbol occupies position 0. The common decoding algorithms for high-rate (n, k, d) Reed–Solomon codes are divided into the following four major steps:

1. Compute the syndrome polynomial $S(x)$ from a corrupted codeword $c'(x)$.
2. Solve the key equations $\sigma(x) * S(x) = \omega(x) \bmod x^R$, where $R = n - k$ and $\sigma(x)$ is the error locator polynomial containing all the error location information and $\omega(x)$ is the error evaluator polynomial, which can be used in conjunction with the error locator polynomial to find the error values.
3. Find the roots of $\sigma(x) = (\alpha^{a1} * x + 1) * (\alpha^{a2} * x + 1) * \cdots * (\alpha^{at} * x + 1)$. The following conditions can be used to abort the decoding process: (a) if the number of roots found is less than the degree of $\sigma(x)$, (b) any

roots found does not correspond to a location between 0 and $n-1$, and (c) if repeated roots are found.

4. For each root α^{ai} of $\sigma(x)$ find the error values with the help of $\sigma(x)$, $\omega(x)$ using the Forney formula. The error value at location α^{ai} is given by $\omega(x) * x^{(L-1)} * /\sigma'(x)$ evaluated at $x = \alpha^{ai}$, where $\sigma'(x)$ is the formal derivative of $\sigma(x)$. Over GF(2^m), if $\sigma(x). = \sigma_0 + \sigma_1 * x + \sigma_2 * x^2 + \sigma_3 * x^3 + \sigma_4 * x^4 + \sigma_5 * x^5 + \sigma_6 * x^5 + \sigma_7 * x^7 + \cdots$, then $\sigma'(x). = \sigma_1 + \sigma_3 * x^2 + \sigma_5 * x^4 + \sigma_7 * x^6 + \cdots$. In other words, $\sigma'(x)$ contains only even-powered terms of x.

To complete the decoding process, the errors values computed in step 4 should be added to the corrupted code symbol indicated by the respective decoded error locations.

The most time-consuming steps are steps 1 and 3. The most difficult step to understand is step 2. Again, most coding textbooks provide detailed information about solving the key equations [6,9–11]. Two competing algorithms for solving the key equations are the Berlekamp–Massey algorithm and the Euclidean algorithm. The former may use fewer gates to implement, but the latter can be understood more easily for first-time readers. In addition, there are other decoding algorithms, which can be found in Refs. 12 and 13.

The most frequently used technique for finding the roots of the error locator polynomial is the well-known Chien search [6,9–11], which is a systematic and efficient trial-and-error root testing technique. Every possible root, one root for each possible error location, is tested as a possible solution to the error locator polynomial $\sigma(x)$. Therefore, this is usually a very time-consuming process.

It may be helpful to continue the example with errors introduced. Let the codeword be $c(x) = x * g(x) = (0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0)$. Let the two introduced errors be one at position 0 and one at position 8. The error value at position 0 is α^3 and the error value at position 8 is α. Then the corrupted codeword $c'(x)$ becomes $c'(x) = x * g(x) = (0, \underline{\alpha}^4, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, \underline{\alpha}^3)$. The underlined symbols are erroneous. The first step of the decoding process is to compute the syndrome polynomial, which is $S(x) = \alpha^0 + \alpha^4 * x + \alpha^{14} * x^2 + \alpha^{13} * x^3 + \alpha^9 * x^4 + \alpha * x^5 + \alpha^6 * x^6$. Either the Berlekamp–Massey or the Euclidian algorithm will produce the error locator polynomial $\sigma(x) = 1 + \alpha^2 * x + \alpha^8 * x^2$ and the error evaluator polynomial $\omega(x) = 1 + \alpha^{10} * x$. It can be checked that $\sigma(x) * S(x) = \omega(x) \bmod x^R$. The next step is to find the roots of $\sigma(x)$, which is equivalent to the complete factorization of $\sigma(x)$. The result of this process is $\sigma(x) = (x+1) * (\alpha^8 * x + 1)$. The two roots α^0 and α^{-8} correspond to errors at symbol locations 0 and 8, respectively. To find the error values, the Forney formula is used, as follows. At symbol location i, the error value is given by $Y_i = \omega(x)x^{L-1}/\sigma'(x)$ evaluated at $x =$ root i of $\sigma(x)$. In this case, the results are $Y_0 = \alpha^3$ and $Y_1 = \alpha^1$.

2.4. Comparison of Decoding Algorithms

The Reed–Solomon codes employed in most magnetic storage devices are high-rate codes. As mentioned previously, the major steps involved in the decoding algorithms are (1) syndrome computation (2) error locator polynomial and error evaluator polynomial determination, and (3) error value computation. Step 2 has the most variations. When the maximum number of errors to be corrected is small — typically, four or fewer errors — the most efficient way of obtaining the error locator polynomial is to compute all the coefficients of the error locator polynomial directly from the syndromes using a special and simple technique [14]. For higher numbers of errors, usually the well-known Berlekamp–Massey algorithm or the Euclidean algorithm is used. Most of the algebraic coding books give detailed descriptions of these two algorithms [6,9–11]. These two algorithms are both iterative and equivalent but they have two interesting opposite characteristics, which may be the determining consideration in algorithm selection. In the Berlekamp–Massey algorithm, the degrees of the polynomials used in the iteration increases as the procedure continues and the total number of iterations is a fixed number. On the other hand, in the Euclidean algorithm, the maximum polynomial degree decreases from iteration to iteration and the number of iterations to complete the procedure is variable. Another notable thing is that the Berlekamp–Massey algorithm can be used to find either the error locator polynomial alone or both the error locator polynomial and the error evaluator polynomial. On the other hand, in the Euclidean algorithm, both of these polynomials are computed at the same time. It is not very efficient to use the Euclidean algorithm to compute the error locator polynomial alone.

3. OTHER ASPECTS OF REED–SOLOMON CODING

3.1. Hardware Versus Firmware Implementations

The entire encoding and/or decoding process can be implemented either by microprocessor firmware or by hardware. Typically, the encoding is done in hardware to improve the speed of the process of writing to disk. Also, the complexity of an all-hardware encoder is much lower than that of the hardware decoder. Disk drives have the luxury of rereads, which are equivalent to requests for retransmission in communication systems. When hardware gates were expensive, the ECCs employed in magnetic recording systems made use of rereads to reduce the cost of hardware implementation. The hardware complexity is a function of the number of symbol errors corrected. Therefore, one way to save gates or hardware complexity is to correct only a smaller number of symbol errors than the designed error-correcting capability in hardware; when the hardware decoder fails, the firmware decoder is then used to correct all the errors. For example, when the ECC is designed to correct six errors per interleave, the hardware decoder can be designed to correct two or fewer errors and the firmware decoder to correct three, four, five, and six errors per interleave. Data throughput is affected every time the firmware algorithm is employed. The hardware encoder can also be designed to perform the syndrome computation with simple modification [15]. This design further reduces the total complexity of the hardware decoding.

There is a mode of error correction that may be different from other communication systems. Occasionally, a small particle may get in contact with the read head resulting in an increase in temperature. The heat produced by the contact may cause a long burst of errors to occur. This phenomenon is called a thermal asperity. Fortunately, the location of a thermal asperity can be detected by the signal processing system prior to ECC decoding. As a result, a more powerful error–erasure decoding algorithm can be used to deal with the long burst in the event that the normal error-only decoding algorithm fails to correct the errors. This decoding algorithm is capable of correcting e erasures and t symbol errors when $2 * t + e < d$, where an erasure is an error with known location. This algorithm is also given in most coding textbooks [6,9–11].

3.2. Block Missynchronization Detection

As can be seen from the example, a Reed–Solomon code word is likely decoded to a different codeword if a shift in the symbol position occurs. As mentioned previously, the Reed–Solomon code is a cyclic code, which means that a cyclic rotation of a nonshortened codeword is also a codeword. If $(c_{n-1}, c_{n-2}, c_{n-2}, \ldots, c_2, c_1, c_0)$ is a codeword of a nonshortened Reed–Solomon code over GF(2^m), that is, $n = 2^m - 1$, then both cyclic shifted versions $(c_{n-2}, c_{n-2}, \ldots, c_2, c_1, c_0, c_{n-1})$ and $(c_0, c_{n-1}, c_{n-2}, c_{n-2}, \ldots, c_2, c_1)$ are also codewords of the same code. As a result, Reed–Solomon codes are susceptible to miscorrection when symbol synchronization errors occur. To minimize this possibility, block synchronization must be assured for each sector prior to ECC decoding. This is accomplished with a specially designed sequence called an *address mark*. In other words, the data of a sector are preceded by an address mark. The address mark is designed to ensure correct block synchronization before the ECC attempts error correction. There is a preamble preceding the address mark. The main purpose of the preamble is for training the phase-locked loop (PLL) to acquire the bit synchronization. The secondary purpose of the preamble is to assist the address mark to establish a distinct position in the bitstream, which, in turn, enables the system to find the beginning of the first symbol of the ECC block. To facilitate the PLL training, a repeated pattern is used as the preamble. This pattern is often a long sequence of identical bits, such as a sequence of ones. In the following discussion, let a long sequence of ones (1s) serve as the preamble, which precedes the address mark. Let the address mark be a 12-bit sequence 000010100110 with the bit on the left preceding the bit on the right in time. The pattern, including the preamble and address mark, recorded on the disk is $\cdots$ 111111111111000010100110, where the underlined bits are the address mark. As soon as the bit synchronization is established by the phase-locked loop, a circuit compares every 12 consecutive bits with the address mark. The Hamming distance between the address mark and the 12 ones is 8 (bits), because there are 8 zeros in the address mark. Let the distance from the address mark to an out-of-phase sequence be $d_a(s)$. An out-of-phase sequence consists of s ones of the preamble followed by the first 12 s bits of the address mark. It can be checked that $d_a(s)$ is 7,8,9,8,8,7,6,7,7,7,7,0 for $s = 11,10,9,8,7,6,5,4,3,2,1,0$, respectively. Other than the case of $s = 0$, the minimum distance for this address mark from the out-of-phase sequences is 7. Therefore, this address mark can tolerate three errors. In other words, with three or fewer errors in any span of 12 consecutive bits prior to the last bit of the address mark, the position of the address mark can be correctly established. This address mark is called a "3-error-tolerance address mark." Let t be any threshold with t less than or equal to 3. Then the correct address mark is assumed found if any span of 12 consecutive bits differs from the correct address mark by t or fewer bits. By varying the value t, the performance of the address mark can be controlled to a certain degree. There are two important probabilities associated with an address mark for a specified threshold t. The probability of failure to synchronize is the probability that the system fails to find the correct synchronization position; this occurs when there are more than t errors on the address mark retrieved from the disk. This probability for the case of independent errors is given by the dominant term of the probability

$$P_{\text{failure\_to\_synchronize}} \approx C(L, t+1)p^{(t+1)}(1-p)^{(L-t-1)} \tag{8}$$

where L is the length of the address mark in bits, p is the raw bit error rate, t is the threshold, and $C(L, t+1)$ is the binomial coefficient of $x^{(t+1)}$ in the expansion of $(x+1)^L$. For most cases of interest, p is a small number. Therefore, the $1-p$ term approaches 1. Consequently, the probability of failure to synchronize is decreased as the threshold is increased. The other important probability is the probability of false synchronization. This is the probability that a span of L bits prior to the correct address location is mistakenly identified as the address mark. This probability is lower bounded by

$$P_{\text{false\_synchronization}} \approx C(d_a, d_a - t)p^{d'}(1-p)^t \tag{9}$$

where d_a is the minimum distance between the address mark and any span of L bits prior to the correct address mark position and $d' = d_a - t$. As can be seen, the probability is proportional to $p^{d'}$, and this probability increases as the threshold decreases.

Therefore, these two probabilities put conflicting requirement on the threshold t. One obvious solution is to increase the length of the address mark L, which leads to higher values of error tolerance; as a result, there are more values of threshold to select to meet the demand for both probabilities. Unfortunately, a longer address mark means larger overhead and more complicated circuits for detecting the address mark. The other choice is not to increase the length of the address mark and rely on the ECC to help in detecting the false synchronization. Using the 12-bit address mark with a raw bit error rate of $p = 1.0 \times 10^{-4}$ and $t = 3$ as an example, the two probabilities are $P_{\text{failure\_to\_synchronize}} \approx 9.92 \times 10^{-21}$ and $P_{\text{false\_symchronization}} \approx 7.00 \times 10^{-15}$. The probability of failure to synch is acceptable but the probability of false synchronization is

too high for the data storage application. Reed–Solomon codes over GF(2^m) are susceptible to synchronization errors when the bit slippage is a multiple of m, the symbol size. Therefore, the miscorrection probability is of the order of $P_{\text{false\_symchronization}}/m$. With a false synchronization probability of 7.00×10^{-15}, the miscorrection probability is of the order of 1.00×10^{-18}, which is too high. A solution to this situation is to use a coset instead of the code itself for recording purposes.

A coset consists of the set $\text{cl}(x) + c(x)$, where $c(x)$ is any codeword of the Reed–Solomon code and $\text{cl}(x)$ is a fixed sequence or vector of length n, called the coset leader. During recording, a member of the coset is recorded in the disk. When a sector is retrieved, the coset leader is added to the sequence resulting in $c(x)$, the codeword, assuming no errors and perfect synchronization. When out of synchronization, the retrieved sector contains $\text{cl}'(x) + \text{c}'(x)$, where $\text{cl}'(x)$ and $\text{c}'(x)$ are respectively shifted versions of $\text{cl}(x)$ and $\text{c}(x)$ with proper truncation and padding under error-free conditions. With $\text{cl}(x)$ added before decoding, the decoder sees $c'(x)$, a codeword plus possible errors due to truncation and padding, with $\text{cl}(x) + \text{cl}'(x)$. Therefore, with proper selection of $\text{cl}(x)$, $\text{cl}(x) + \text{cl}'(x)$ contains many nonzero terms. Each of these nonzero terms appears as an error to the ECC decoder. A well-designed $\text{cl}(x)$ results in more errors than the ECC correcting capability most of the time. Therefore, the decoder cannot correct these "errors." Consequently, the miscorrection probability is substantially lower than the value $P_{\text{false\_symchronization}}/m$. In fact, it is given by $\{P_{\text{false\_symcronization}}/m\} * \{\text{probability of decoding } \text{cl}(x) + \text{cl}'(x) + c'(x) \text{ as a codeword}\}$. With proper design of $\text{cl}(x)$ and sufficiently long redundant symbols, the probability of decoding $\text{cl}(x) + \text{cl}'(x) + \text{c}'(x)$ as a codeword can be made very small. Therefore, the miscorrection probability is significantly reduced. Reference 16 provides a way to find the coset leaders. In the event that the bit slippage is not a multiple of the symbol size m, the miscorrection probability caused by the false synchronization is given by $\{(1-m) * P_{\text{false\_synchronization}}/m\} * \{\text{probability of decoding } \text{cl}(x) + \text{cl}'(x) + \text{c}'(x) \text{ as acodeword }\}$; however, in this case, $c'(x)$ looks like a random sequence to the decoder, and so does the sequence $\text{cl}(x) + \text{cl}'(x) + \text{c}'(x)$. Therefore, the probability that a randomlike sequence is decoded as a codeword can be easily computed or approximated [17] to be a small number. Consequently, with a properly designed coset leader, the threshold t for the address mark can be simply set according to the requirement for $P_{\text{failure\_to\_synchronize}}$ alone. From a different point of view, a shorter length for address mark can be used with the help of the coset leader. Another similar approach for avoiding false synchronization is to initialize the encoder to a nonzero state prior to the encoding. This approach is more appropriate for cases where that the decoder needs to compute the remainder of the corrupted codeword as the first decoding step. More detailed information can be found in Ref. 18. Either approach will enhance the performance of the address mark, because the address mark deals mainly with one probability, the probability of failure to synchronize.

3.3. Interleaving Versus Noninterleaving

Let the symbols in a sector be orderly numbered as 0,1,2,3,4,5,6,7,8,9,.... Assuming that there are three codewords in a sector, there are many ways to associate the symbols with code symbols: $a[0], a[1], a[2], \ldots, b[0], b[1], b[2], \ldots, c[0], c[1], c[2], \ldots$, where $a[k], b[k]$, and $c[k]$ are kth symbols of code a, code b, and code c, respectively. Let the code length of each code be L. Then the first way of interleaving is $a[0], a[1], a[2], \ldots, a[L-1], b[0], b[1], b[2], \ldots, b[L-1], c[0], c[1], c[2], \ldots, c[L-1]$. Namely, $a[0], a[1], \ldots, a[L-1]$ are the symbols $0, 1, \ldots, L-1$ of the sector, respectively; $b[0], b[1], \ldots, b[L-1]$ are the symbols $L, L+1, \ldots, 2L-1$, of the sector, respectively; and $c[0], c[1], \ldots, c[L-1]$ are the symbols $2L, 2L+1, \ldots, 3L-1$ of the sector, respectively. This arrangement of three codewords has the main advantage that the single decoder can be used sequentially. However, the error-correction power of Reed–Solomon coding is not "enhanced." The more commonly used arrangement is $a[0], b[0], c[0], a[1], b[1], c[1], a[2], b[2], c[2], \ldots$. This arrangement is called *interleaving*. The number of code words involved is denoted by I and called the *depth of interleaving*. The main advantage of this arrangement is that a burst of I symbols can affect at most one symbol for each codeword. Therefore, from the entire sector point of view, the I code words in a sector form a t-burst correction code if each codeword can correct t-symbol errors.

A primary consideration for code selection involves the decision for determination of the degree of interleaving. The length of a Reed–Solomon code over GF(2^m) is limited to $2^m - 1$ symbols or $(2^m - 1) * m$ bits. Let the degree of interleaving be I. Then the minimum depth of interleaving is given by the equation $I * (2^m - 1) * m >$ or $= 4096 +$ number of redundant bits per sector. For $m = 8$, the minimum value of I is 3. The smallest value of m for I as 1 (the case of noninterleaving) is 9. In addition to code length, the ease of decoding sometimes plays an important role in determination of interleaving depth. For example, for the case of $m = 9$ and $I = 2$, each interleave needs to correct t errors when there are t bursts, where each burst corrupts no more than two 9-bit symbols. This design is much simpler than the design of using a single code to correct $2t$-symbol errors, especially when t is a small value such as 1, 2, 3, or 4 [14]. It should be noted that the ECC can correct any combination of $2t$ symbol errors for the noninterleaved case and that the interleaved case can correct only the "well behaved" $2t$ or fewer symbol errors. By a "well-behaved" pattern, the $2t$ or fewer errors are distributed in such a way that each interleave sees t or fewer symbol errors. As a result, the noninterleaved case is more powerful in terms of ECC correction power. However, the interleaved implementation may be the right design if the errors are frequently in bursts and complexity is the main concern.

3.4. Performance

The main purpose of using an ECC is to bring the raw bit error rate at the output of the modulation code decoder to an acceptable level. The raw bit error rate at the output of the modulation code decoder is often targeted at 1.0×10^{-6}

or higher. The required bit error rate at the output of the ECC is usually below 1.0×10^{-15}. Traditionally, the error rate specification is given in terms of bit error rate. However, for Reed–Solomon codes, it is easier to perform the error rate computation on the basis of the symbol error rate and the code block error rate. Therefore, the definitions and relationships among various error rates should first be clarified.

3.5. Error Rate Definitions

The commonly used definition of bit error rate is as follows:

$$\text{Bit error rate} = \frac{\text{total number of bits in error}}{\text{total number of bits observed}} \tag{10}$$

Similarly, symbol error rate is defined as

$$\text{Symbol error rate} = \frac{\text{total number of symbols in error}}{\text{total number of symbol observed}} \tag{11}$$

$$\text{Block error rate} = \frac{\text{total number of blocks in error}}{\text{total number of blocks observed}} \tag{12}$$

The denominators of these three definitions have very simple relations:

$$\text{Total number of bits observed} = m * (\text{total number of symbols observed}) \tag{13}$$

$$\text{Total number of symbols observed} = n * (\text{total number of blocks observed}) \tag{14}$$

$$\text{Total number of bits observed} = n * m * (\text{total number of blocks observed}) \tag{15}$$

where m is the number of bits per symbol and n is the number of symbols per code block. If the relationship among the numerators can be made as simple, then the conversion from one error rate to another becomes straightforward. With different modulation codes and signal processing techniques, the relationships among the numerators may not be simple. However, bounding techniques and approximations can be used to obtain estimated results, which often provide sufficient information from a designer's point of view. The relationships among the various error rates at the output of the modulation code decoder can be obtained more readily. The error rates at the output of the modulation code decoder are often measurable quantities as the total number of bits to be observed or counted is no more than a few million if the bit error rate is 1.0×10^{-6} or worse. The relationships among various error rates at the output of an ECC system are often difficult to quantify. Therefore, the following discussion deals with the decoded error rates.

When a t-symbol error-correcting code is employed, the most likely errors for which the ECC fails to correct are error patterns containing $(t+1)$-symbol errors. Therefore, the following relationship can be established:

$$\text{The total number of symbols in error} \approx (t+1)(\text{the total number of blocks in error})$$

The number of bits in error in an erroneous symbol can be anywhere from 1 to m, where m is the total number of bits per symbol. In fact, this number depends on the modulation code used. If this number can be obtained from the modulation code, it should be a number between 1 and m. For a completely random system the number is approximately $m/2$. A good modulation code tends to make this number less than $m/2$. Therefore, the following relationship can be established:

$$\begin{aligned}\text{Total number of bits in error} &= am(\text{total number of symbols in error}) \\ &\approx am(t+1)(\text{the total number of blocks in error})\end{aligned}$$

(with $0 < a \leq 1$, for random data $a = 0.5$). With these approximations, it is straightforward to show that

$$\begin{aligned}\text{Bit error rate} &\leq a * (\text{symbol error rate}) \quad (\text{with } 0 < a \leq 1) \\ \text{Symbol error rate} &\approx \frac{t+1}{n}(\text{block error rate}) \\ \text{Bit error rate} &\leq a\frac{t+1}{n}(\text{block error rate}) \quad (\text{with } 0 < a \leq 1).\end{aligned}$$

For the commonly used symbol sizes of $m = 8$ and 10, whether the true value of a or its upper bound is used in the preceding relationship, the results differ less than one order of magnitude. Therefore, in the following discussion, the symbol error rate and the block error rate are used; the results can be readily converted to the desired bit error rate with either the accurate estimate of a or its upper bound. The starting point is the symbol error rate.

Let the raw symbol error rate be P_s at the input of the t-symbol-correcting ECC decoder. For a system in which symbol errors are statistically independent, the decoded block error rate P_b is given by

$$P_b = \sum C(n,j)P_s^j(1-P_s)^{(n-j)} \tag{16}$$

where the summation limits are from $j = t+1$ to $j = n$ and $C(n,j)$ is the binomial coefficient for expanding the polynomial $(1+x)^n = \sum C(n,j)x^j$. Knowing the decoded block error rate, we can obtain the decoded symbol error rate. In fact, the decoded symbol error rate P_d can be computed directly from the formula

$$P_d \cong \frac{1}{n}\sum C(n,j)jP_s^j(1-P_s)^{(n-j)} \tag{17}$$

where the summation limits are from $j = t + 1$ to $j = n$. The results in Eqs. (16) and (17) can be obtained using combinatorial arguments. The reason Eq. (17) is an approximation is that when there are more than t-symbol errors, the decoder may introduce additional errors. Reference 17 provides the formula for the exact expression for Reed–Solomon codes under the condition that the symbol errors are statistically independent. Equations (16) and (17) both indicate that the decoded error rate is proportional to $P_s^{(t+1)}$. This can be seen from the observation that every term in the summation contains $P_s^{(t+1)}$ as a factor. Therefore, as t increases, the decoded error rate decreases. However, using more powerful ECC results in a lower code rate, which has the effect of making P_s worse. As a result, an ECC system cannot arbitrarily increase the correcting power to gain better overall decoded error rates. For the cases where the symbol errors are not statistically independent, the preceding arguments also hold. For more detailed information, see Ref. 19, where a soft bit error rate (SBER) is defined as (total number of error events)/(total bits read that are protected by the ECC system); therefore, it is equal to 1/b times the bit error rate defined by Eq. (10), where b is the average number of bits in error per error event.

3.6. Separate EDC or Embedded EDC

When a disk drive or a tape drive is used to store data, the miscorrection probability must be many orders of magnitude better than the probability that the ECC cannot decode the code block. There are two common ways of achieving this goal. The first approach is to use a separate error-detecting code (EDC) and let the error-correcting code decode to the limit of its error correcting capability. In this case, an odd-distance error-correcting code is often employed and the code corrects t errors, where t is equal to $(d-1)/2$, where d is the minimum distance of the code. The second approach is to correct t errors only with $t < (d-1)/2$. Let the number of bits in the EDC for the first approach be r_1 and the relationship for the second approach be $r_2 = (d - 1 - 2t) * m$, where m is the symbol size. Then r_2 can be considered as the effective number of bits in the second approach reserved for error detection. When $r_1 = r_2$, and t is the same for both approaches, the miscorrection probabilities for both approaches are about the same. The main consideration as to which approach to adopt is listed below:

1. Approach 1 is more flexible. The value of r_1 can be more or less arbitrarily selected to achieve any desirable miscorrection probability. For approach two, the value for r_2 needs to be a multiple of m, the symbol size.
2. Approach 1 needs the additional step for checking the correctness of EDC after the error-correction procedure. This verification process is often time-consuming and complex. For approach 2, the EDC is automatically satisfied at the end of the decoding algorithm if the algorithm is properly designed and employed.
3. When it is required to provide the data block to the host computer, which may require that the data block have its own EDC, it may be mandatory to use approach 1.

In practice, both approaches are sometimes employed simultaneously. Both embedded EDC and separated EDC are used. The embedded EDC assures the low probability of miscorrection, while the separate EDC is used to communicate with the host computer. Often, the separate EDC itself is covered by the ECC with embedded EDC. As a result, the separate EDC is not checked as the embedded EDC provides the necessary detection power.

The separate EDC has another advantage in that it may not be affected by cyclic shifts of symbols. Therefore it can reduce the miscorrection probability due to symbol shift. As a result, the selection of the coset leader for missynchronization detection may be easier if the EDC is of sufficient length.

3.7. Tape Drive ECC

All the coding techniques employed in disk drives can be readily applied to tape drives. Tape drives support multiple tracks; in other words, a read head reads several tracks simultaneously. This provides the opportunity to apply a code across the tracks. Usually, the ECC along a single track corrects a limited number of symbol errors or none at all, but with sufficient redundancy to make sure that the miscorrection or mis-detection probability is extremely small. Whenever a track fails to correct, the errors are detected. The ECC across the track then corrects the errors indicated by the tracks whose ECC fail to correct. In other words, the symbols on the tracks, which are known to contain erroneous symbols as indicated by the ECC, are treated as erasures for the ECC across the tracks. An erasure is an error with known location. The more powerful error–erasure decoding algorithm is then used to correct both errors and erasures across the tracks. An example for this approach can be found in Ref. 20.

Another way of protecting data on tape drives is to use array codes [4]. This can be understood again by visualizing the two-dimensional structure of tape recording. Simple codes, preferably the single-parity-check codes, are introduced on both the longitudinal direction (the direction following the head movement) and the direction across the tracks as well as the directions that are at constant angle with respect to the longitudinal direction. Therefore, a symbol is protected by several simple parity-check equations, one for each direction. When one or more of these directions contains one or no errors, the symbol can be correctly recovered [4].

3.8. Codes for RAID

In redundant array of independent disks (RAID), inter-relations are introduced to several disk drives to form redundancy among sectors, one sector from each drive. The codes are in fact so-called concatenated codes. These are also two-dimensional codes. The code in one dimension

is the Reed–Solomon code presently used in each sector. The code used across the sectors is usually a very simple single-parity-check code. Namely, if the information symbols are $c[n-1], c[n-2], \ldots, c[2], c[1]$, then the redundant symbols are given by $c[0] = c[1] + c[2] + \cdots + c[n-1]$. The reason this is used is because the recovery of a failed drive can be achieved simply. The only limitation for this simple scheme is that it can protect against the case that a single drive fails. More sophisticated distance d Reed–Solomon codes can also be used to protect against the case that $d-1$ or fewer drives fail simultaneously. However, the reconstruction of the failed sectors often involves many reads and writes of the related sectors from different drives; also, the data recovery from the failed sectors needs more complicated operations than the simple exoring. Therefore, there is no reason not to employ more powerful Reed–Solomon codes across the drives. The time delays and extensive read and write operations may be the deciding factor for future adoption of more powerful codes.

3.9. Decoding Beyond $(d-1)/2$ for Reed–Solomon Codes

The decoding algorithms for Reed–Solomon codes presently employed in magnetic recording are the so-called hard-decoding algorithm and the error–erasure algorithm. In a hard-decoding algorithm, the channel outputs a bit-stream to the ECC decoder, where each bit is either 1 or 0. Each bit can be either right or wrong. The symbols at the input to the ECC decoder are symbols with predetermined "hard" values. In the error–erasure decoding algorithm, the channel also provides the locations of symbols, which are erasures whose values may not be correct. Therefore, in the case of error–erasure decoding, each symbol is associated with a reliability information, that is, with reliability information of 1 bit. For example, a one indicates that the symbol is reliable and a zero indicates that the symbol is not reliable; therefore, it is an erasure.

The present "hard" error-only decoding algorithm for a minimum distance d Reed–Solomon codes attempts to correct all possible error patterns, which are at a distance no more than $(d-1)/2$ symbols away from a codeword. New algorithms have been designed to provide a list of codewords, which are no more than $n*(1-r^{1/2})$ symbols away from the n symbol sequence to be decoded [21]. (Here $r = k/n$ is the code rate.) Conventional hard decoding can correct $(d-1)/2 = (n-k)/2 = n*(1-r)/2$ symbols, which is smaller than $n*(1-r^{1/2})$ symbols. The advantage of this new decoding algorithm is more significant when the code rate r is low. With a properly selected algorithm and code rate, this approach has the potential to enhance the performance for Reed–Solomon codes. More information can be found in Refs. 21 and 22.

Another direction is soft-decision decoding. Instead of providing a symbol with one bit of reliability information like the erasure information, many bits of reliability information are associated with a symbol. The decoding algorithm makes use of this reliability information to further enhance code performance. There are many encouraging results for soft-decision decoding algorithms employed with binary codes. References 23 and 24 are examples for soft-decision decoding for nonbinary codes. The performance gain for soft-decision decoding in Reed–Solomon codes over hard-decision decoding is an interesting and important area of research in magnetic recording. If it proves to provide a significant improvement, the algorithm no doubt will soon be adopted in magnetic recording applications.

3.10. Larger Sector Size

Strictly from a coding point of view, the sector size for hard-disk drives should be larger than the present 512 bytes. For the same code rate, code performance improves as block size increases. For example, the Reed–Solomon (440,410,31) code over GF(2^{10}) can correct 15 symbol errors among 440 symbols, while the (880,820,61) code can correct 30 symbols among 880. By Eq. (16), the shorter code has the decoded block error rate proportional to P_s^{16}, while the longer code has the decoded block error rate proportional to P_s^{31}. It can be easily seen that for a very large range of P_s, the longer code gives lower decoded block error rate for the same raw symbol error rate P_s. In addition, there are other overhead savings such as preambles and address marks. Unfortunately, many operating systems assume a 512-byte sector size. A change in disk drive sector size would require corresponding modification to these operating systems. At present, the 512-bytes data constitute the one sector size that the disk industry must provide.

BIBLIOGRAPHY

1. H. J. Richter, Longitudinal recording at 10 to 20 gbit/inch^2 and beyond, *IEEE Trans. Magn.* **35**(5): 2790–2795 (1999).
2. J. Lee and V. K. Madisetti, Error correcting run-length limited codes for magnetic recording, *IEEE Trans. Magn.* **31**(6): 3084–3086 (1995).
3. N. Glover, *Practical Error Correction Design for Engineers*, Data Systems Technology Corp., Broomfield, CO, 1982.
4. M. Blaum and R. M. Roth, New array codes for multiple phased burst correction, *IEEE Trans. Inform. Theory* **39**(1): 66–77 (1993).
5. R. J. McEliece, *Finite Fields for Computer Scientists and Engineers*, Kluwer, Norwell, MA, 1987.
6. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
7. U.S. Patent 4,975,867 (1990), L. J. Weng, Apparatus for dividing elements of a Galois field of GF(2$\hat{Q}$M).
8. U.S. Patent 6,044,389 (2000), L. J. Weng and B. A. Shen, System for computing the multiplicative inverse of a field element for Galois field without using tables.
9. R. E. Blahut, *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, MA, 1983.

10. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2nd ed., MIT Press, Cambridge, MA, 1972.

11. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

12. U.S. Patent 4,633,470 (1986), L. R. Welch and E. R. Berlekamp, Error correction for algebraic block codes.

13. U. Sorger, A new Reed–Solomon decoding algorithm based on Newton's interpolation, *IEEE Trans. Inform. Theory* **39**: 358–365 (1993).

14. U.S. Patent 5,710,782 (1998), L. J. Weng, System for correction of three and four errors.

15. G. Fettweis and M. Hassner, A combined Reed–Solomon encoder and syndrome generator with small hardware complexity, *1992 IEEE Int. Symp. Circuit Syst.* **4**: 1871–1874 (1992).

16. U.S. Patent 5,528,607 (1996), L. J. Weng, B. Leshay, and D. Langer, Method and apparatus for protection of data from mis-synchronization.

17. Z. McC. Huntoon and A. M. Michelson, On the computation of the probability of post-decoding error events for block codes, *IEEE Trans. Inform. Theory* **23**: 399–403 (1977).

18. U.S. Patent 4,989,211 (1991), L. J. Weng, Sector mis-synchronization detection method.

19. C. M. Riggle and S. G. McCarthy, Design of error correction systems for disk drives, *IEEE Trans. Magn.* **34**(4): 2062–2371 (1998).

20. U.S. Patent 5,136,592 (1992), L. J. Weng, Error detection and correction system for long burst errors.

21. V. Guruswami and M. Sudan, Improved decoding of Reed–Solomon and algebraic-geometric codes, *IEEE Trans. Inform. Theory* **45**: 1755–1764 (1999).

22. M. Sudan, Decoding of Reed–Solomon codes beyond the error correction bound, *J. Complexity* **12**: 180–193 (1997).

23. E. R. Berlekamp, Bounded distance+1 soft-decision Reed–Solomon decoding, *IEEE Trans. Inform. Theory* **42**: 704–721 (1996).

24. G. D. Forney, Jr., Generalized minimum distance decoding, *IEEE Trans. Inform. Theory* **12**: 125–131 (1966).

COMMUNICATION SATELLITE ONBOARD PROCESSING

STEVEN BERNSTEIN
MIT Lincoln Laboratory*
Lexington, Massachusetts

1. INTRODUCTION TO ONBOARD PROCESSING

The primary purpose of communication satellites is to receive signals from one or more sources and relay them to intended recipients. With the exception of early passive on-orbit reflectors, the relay function is accomplished by some form of active processing carried out on board the satellite. The most common form of active processing is the "translating repeater" or "transponder," which simply amplifies whatever is received in a given uplink frequency band and retransmits it in a different downlink frequency band. While this could be called "onboard processing," we will reserve this term for techniques that manipulate the signals passing through a satellite in more complex ways.

Digital circuit and signal processing technology has progressed to the point that it is feasible to consider very ambitious onboard processing functions. In this article we will concentrate on concepts, confident that technology will soon enable virtually all the ambitious techniques to be described.

1.1. Conventional (Nonprocessing) Communication Satellites

Figure 1 is a block diagram of a translating repeater, the workhorse of conventional (nonprocessing) communication satellites. This is also sometimes called a "bent pipe" because it simply takes in a band of frequency spectrum on its uplink and bends it back to earth. An onboard oscillator and mixer is used to translate the uplink band

* This work was sponsored by the Air Force under Air Force Contract FI9628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Air Force.

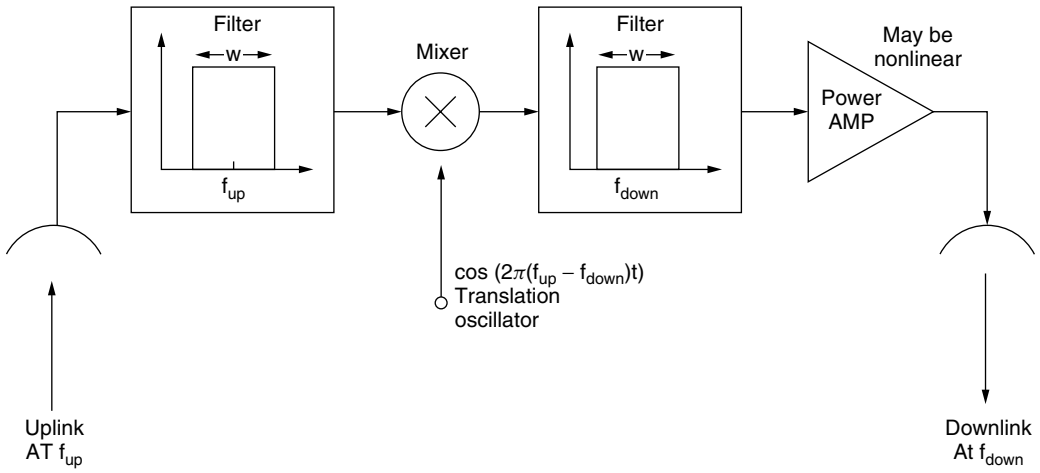


Figure 1. Translating repeater.

to a different downlink band in order to isolate the input and output communication signals. For example, many C-band commercial satellites operate with uplink signals around 6 GHz and downlink signals around 4 GHz. Most such satellites are built with a number of translating repeaters, each operating at a slightly different frequency.

The amplifier shown may be linear or nonlinear. Linear amplifiers are often used in order to minimize crosstalk due to intermodulation products when there are multiple signals within the operating band. Automatic gain control (AGC) is sometimes used to keep the amplifier in a linear region. Nonlinear amplifiers, often operating in a saturated mode, are used when it is desired to maximize prime-to-RF power efficiency. In the nonlinear case, using signals that rely on amplitude modulation may not be feasible; constant envelope phase modulated signals would be more appropriate.

Also shown is a representation of a single antenna beam on the uplink and a single one on the downlink. The antenna beam patterns may be quite complex (e.g., shaped to cover a single country), but all signals in each direction share a common beam. (The uplink and downlink beam shapes and positions, however, may be different.)

1.2. Limitations of Conventional Satellite Architecture

While appealingly simple (and useful), the translating repeater has a number of limitations. Many of these stem from the fact that the power of the single RF amplifier must be shared.

1.2.1. Multiple Access. If a number of communication signals are being relayed simultaneously as in an FDMA (frequency division multiple access) system, then care must be taken to ensure that each user signal arrives at its downlink destination with sufficient signal-to-noise ratio. In order to achieve this, coordination is required among the uplink transmitting stations, especially with regard to power control. For example, an uplink signal that is 3 dB stronger than necessary may capture twice as much transponder power as it entitled to, thus acting as though it were two users. (In this case, even if the amplifier were operating in a linear mode, the amplifier gain, and consequently the downlink power transmitted, would need to be reduced.)

In a system where the user population is heterogeneous, such as operating with different rates and with different terminal antenna sizes, the power control problem becomes quite complex and real-time system monitoring and control is used. This is summarized in Fig. 2.

An approach that reduces the power control complexity is to use TDMA (time-division multiple access). In this design only one user transmits through the transponder at a time and hence may use full uplink power. However, power control in the form of burst data rate and duty cycle is still needed so that each downlink signal obtains at least the minimum required signal-to-noise ratio at the receiving terminal.

1.2.2. Uplink Interference. In some systems uplink interference (intentional or natural) is a major factor to consider. In order to illustrate this, consider the simple scenario consisting of a single user signal and uplink interference that is equivalent to flat Gaussian noise over the system bandwidth, W Hz. Assume that the transponder signal power is shared proportionately between the communication signal and the interference.

Define other needed parameters as

Total satellite power received at the downlink terminal = Pr

Background noise density at the downlink terminal = N_0

Uplink signal power received at the satellite = S

Uplink interference power received at the satellite (which would include the satellite's own receiver noise) = I

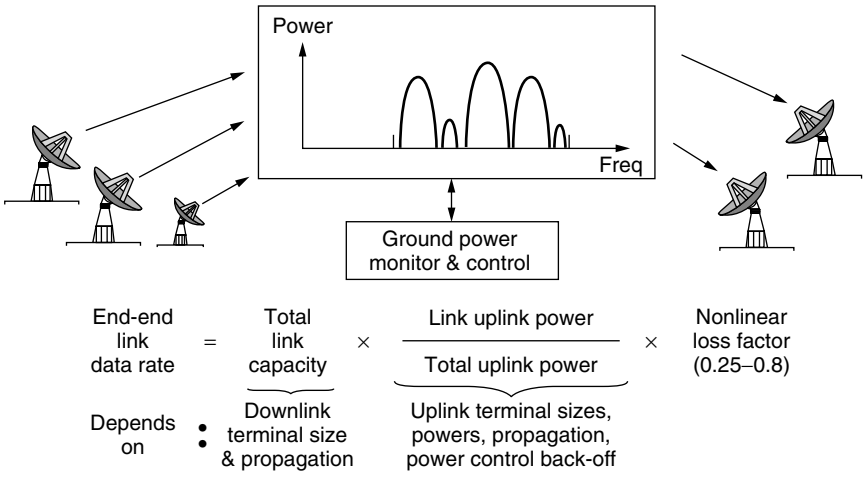


Figure 2. Transponder, multiple access.

Required signal-energy-to-noise-ratio per bit determined by the modulation/coding scheme used $= (E_b/N_0)_{\text{req}}$

The useful signal power received at the downlink terminal is $(S/(S+I))\text{Pr}$, and the interference power retransmitted by the satellite will be $(I/(S+I))\text{Pr}$. The total noise density at the downlink terminal will be the sum of the background and retransmitted amounts $N_0 + (I/(S+I))\text{Pr}/W$. Hence the effective carrier power : noise density ratio will be

$$\left(\frac{\text{Pr}}{N_0}\right)_{\text{eff}} = \frac{\dfrac{S}{S+I}\dfrac{\text{Pr}}{N_0}}{N_0 + \dfrac{I}{S+I}\dfrac{\text{Pr}}{W}}$$

From this we can find the data rate that can be supported as

$$R = \frac{\left(\dfrac{\text{Pr}}{N_0}\right)_{\text{eff}}}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}$$

$$= \frac{W}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}\frac{S}{I}\frac{\dfrac{\text{Pr}}{N_0 W}}{1+\dfrac{\text{Pr}}{N_0 W}} \quad \text{bps}$$

This relationship is shown qualitatively in Fig. 3 as a function of the total downlink signal-to-noise ratio, $\text{Pr}/N_0 W$. We assume that $I \gg S$.

We observe that when $\text{Pr}/N_0 W \gg 1$, the supportable data rate is given by

$$R = \frac{W}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}\frac{S}{I} \quad \text{bps}$$

which is not a function of the downlink signal to noise ratio; however, it is a function of system bandwidth, W. This is called the *bandwidth-limited case* because increasing the bandwidth W would directly improve performance; there would be no improvement if the satellite power were increased. This is a goal sought after in designing systems to counter uplink interference, such as jamming. It will be shown later that onboard processing can approach this goal.

On the other hand, if $\text{Pr}/N_0(W) \ll 1$, then the supportable data rate is

$$R = \frac{\left(\dfrac{\text{Pr}}{N_0}\right)}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}\frac{S}{I} \quad \text{bps}$$

which is not a function of the system bandwidth, W. This is called the *power-limited* case because the only satellite parameter we can change to improve performance is to increase the downlink power. We see that in this case the interference is working by an effect called "power robbing." This is a serious limitation of transponder-based systems. We will return to this example when discussing onboard processing techniques.

It should be noted that in either case, a nonlinear transponder can degrade performance by an additional 1 dB through a small-signal suppression effect. If the interference had a constant envelope, the degradation could reach 6 dB.

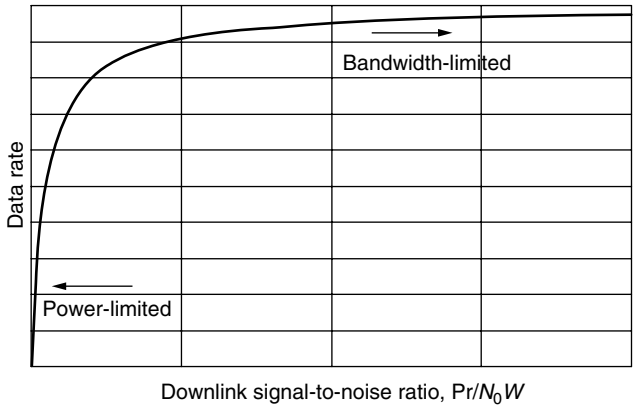


Figure 3. Power- and bandwidth-limited performance of transponder with uplink interference.

1.2.3. User Interconnection. Many satellite systems have multiple antenna beams on both the uplink and downlink. Multiple beams can provide more gain in the direction of ground terminals and permit the reuse of the same frequencies in different parts of the earth. Observe, however, from Fig. 1 that no provision in a simple translating repeater is made for anything more than a simple beam-to-beam connection. Although this is often acceptable, it is a significant limitation in other scenarios.

1.3. What Is Onboard Processing?

Onboard processing includes a wide variety of techniques: analog and digital, RF and baseband, circuit-oriented and packet-oriented. Following the discussion above it is instructive to relate these techniques to the limitations of the model of nonprocessing, the translating repeater.

Table 1 lists some of the main onboard processing techniques that address the limitations, which will be discussed further below. (Note that not all systems with onboard processing necessarily employ all of these techniques.)

2. ONBOARD PROCESSING TECHNIQUES

This section describes some of the principal onboard processing techniques.

2.1. Antenna Beam Switching

Many satellites are deployed with multiple antenna beams. A principal reason for doing so is to increase capacity by taking advantage of the high gain of narrow beams. This is illustrated in Fig. 4, which shows typical link capacities for several diameter geosynchronous satellite

Table 1. Onboard Processing Techniques

Translating Repeater Limitation	Onboard Processing Technique
Limited means of sharing resources without significant user cooperation	Demodulation–remodulation
	Access control
	Circuit multiplexing and switching
	Packet switching
Significant vulnerability to uplink interference	Demodulation–remodulation
	Adaptive antennas
	Despreading of spread-spectrum signals
Fixed interconnectivity	Antenna beam switching and crosslinks
	Adaptive antennas
	Demodulation–remodulation
	Access control
	Circuit multiplexing and switching
	Packet switching

spot beams as a function of ground terminal diameter. (It is assumed that the link is limited by signal-to-noise ratio, not bandwidth.) The link capacity varies inversely with the square of the diameter of the beam on the earth. The motivation for using spot beams should be clear.

A second reason for using spot beams is to reuse frequency assignments in different (usually nonneighboring) beams. The smaller the beam size, the greater number of times the frequency can be reused in a region of coverage.

The presence of multiple beams raises the question of how to interconnect users in different beams to each other. One way to do this is with RF crossbar switches as illustrated in Fig. 5. This is the architecture used in a portion of the NASA ACTS satellite [1]. The crossbar switch can be controlled quasistatically from the ground or, as shown, can respond dynamically to a signal structure such as TDMA. The beam–beam switch can also be combined with frequency filters to switch bands of the frequency spectrum to designated beams.

While uplink and downlink beams are being used for illustration, the same technique can be combined with satellite-to-satellite crosslinks.

2.2. Adaptive Antennas

While many satellites are launched with antenna systems that form carefully designed beam shapes to cover a specific country or region, few can adapt their patterns to meet changing user needs. Pattern adaptation can take several forms, including

- Steerable narrow beams for both uplink and downlink
- Formation of shaped uplink and downlink beams to provide enhanced gain to one or more specific regions
- Formation of spatial uplink "nulls" to reduce the effect of interference sources

The steering of individual narrow beams is usually implemented with mechanically steered dishes. The implementation of more complex shapes can be achieved with

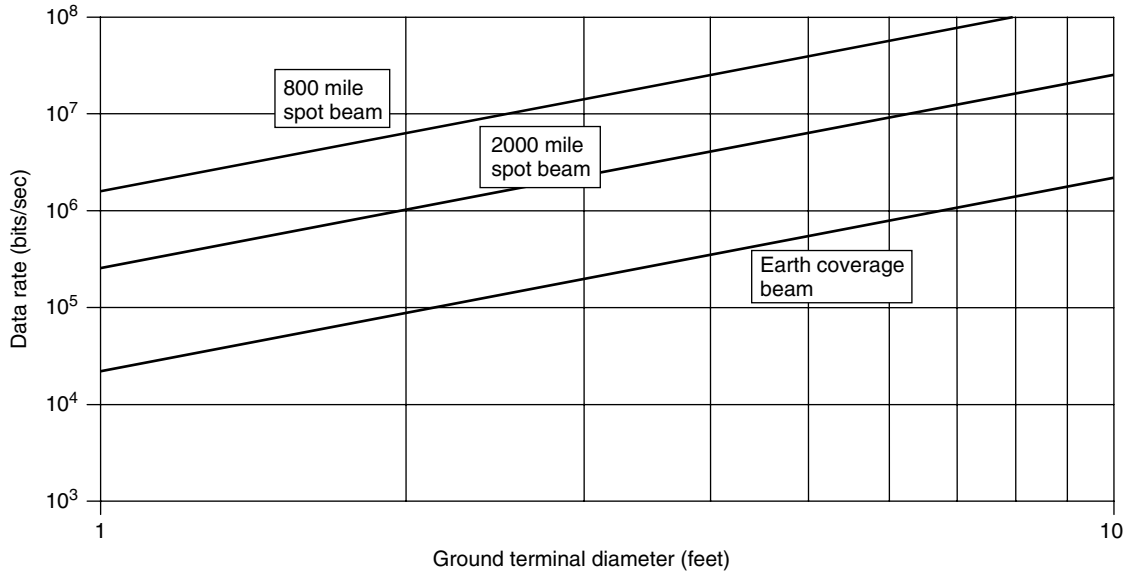


Figure 4. Typical total link capacity for geosynchronous satellites.

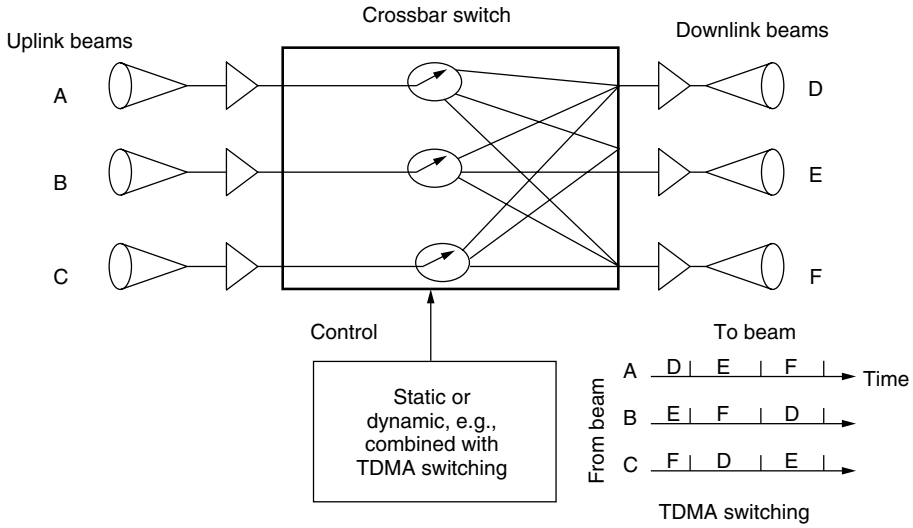


Figure 5. Interconnecting spot beams.

microwave lens antennas or phased arrays [2]. The control of the antenna patterns is usually performed on the ground, but some systems are autonomous in space. More recent approaches to adaptive designs often use *digital beamforming*, whereby the output of each antenna element is downconverted and sampled for subsequent digital signal processing.

It should be noted that adaptive antenna processing could be applied to both analog and digital communication systems.

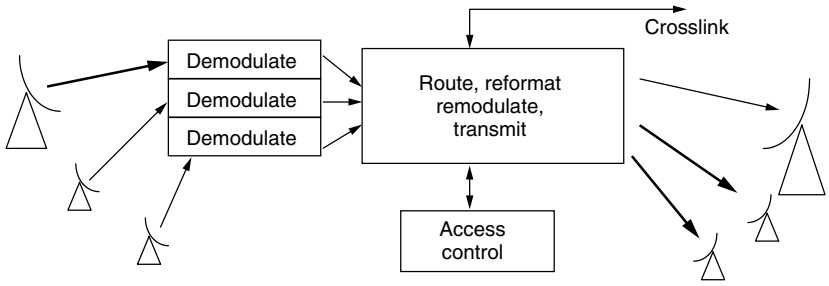
2.3. Demodulation–Remodulation

Demodulation–remodulation is one of the most powerful onboard processing techniques. This is illustrated in Fig. 6, which shows each user uplink being demodulated to a bit stream. The user bit streams are then processed by a digital switching subsystem that can route and reformat the streams and finally remodulate them onto one or more downlinks. (Specific system designs might take advantage of only a subset of these functions. There are also a number of intermediate degrees of processing, such as demodulating user datastreams for error-control coding of parity symbols without decoding to information bits.)

The advantages of this approach are numerous, particularly compared with those of translating repeaters:

1. *Satellite transmitter power is used more efficiently*. Satellite processing permits the renormalization of downlink power sharing. For example, an uplink signal that is received at a power level 3 dB higher than that of all the other signals would capture 3 dB more downlink power than its "fair share" in a translating repeater system. However, with a processor on board, the downlink signal can be renormalized to its fair value. (Of course, some degree of uplink power control must be utilized to guarantee that the uplink signal-to-noise : interference ratio on the uplink permits onboard demodulation.) In addition, an onboard processor can adjust the amount of power devoted to each downlink stream to match the capabilities of each downlink terminal's receiver. Thus a broad system capacity maximization can be approached with minimal complexity imposed on individual terminals.

2. *Downlink power is not wasted on retransmitted noise*. Uplink demodulation, in effect, strips off uplink noise that may come from natural sources, uplink interference, or the satellite's receiver front end. Translating repeaters would



- Reduces need to power balance uplink
- Doesn't waste downlink power retransmitting uplink noise
- Allocates downlink power where needed
- Connects users in different narrow beams
- Optimizes uplink and downlink resources independently

Figure 6. Onboard demodulation–remodulation.

repeat such sources of noise, wasting downlink power. Uplink signals that are below a specified quality measure, such as bit error rate, can be rejected by the satellite processor and not transmitted on the downlink. This property is advantageous for multiple access systems and for systems countering uplink interference such as jamming. The effect is to bring the system into the *bandwidth-limited* regime discussed previously since the power-robbing effect is virtually eliminated. This advantage in power sharing is illustrated in Fig. 7.

3. *Users in different antenna beam ranges can be interconnected*. The routing capability of an onboard processor permits uplink and downlink users in different satellite antenna beams to interconnect. This gives the system designer significant degrees of freedom in the design of antenna patterns. For example, a satellite with numerous high-gain narrow beams can be utilized without limiting connectivity to users to in the same beam. Various interconnectivity structures can also be accommodated, including point-to-point, broadcast, multicast, and many-to-one connections.

4. *Uplink and downlink signal structures can be independently optimized*. Onboard processing effectively permits the uplink and downlink signal structures (and antenna designs) to be optimized independently. This is a very powerful degree of freedom for the system designer.

A generic example that combines several of the concepts discussed above is shown in Fig. 8. In this example a number of simultaneous uplink antenna beams are used to permit small user terminals to transmit at low power into high-gain satellite receive beams. Uplink user signals are prevented from interfering with each other by a combination of FDMA (frequency-division multiple access)

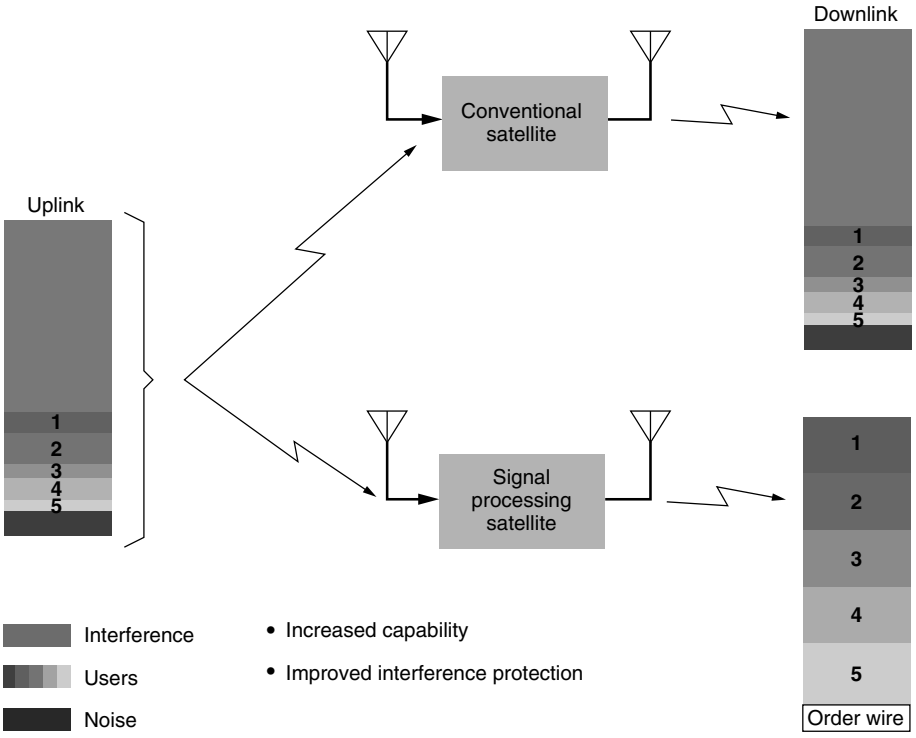


Figure 7. Power sharing advantage of onboard signal processing.

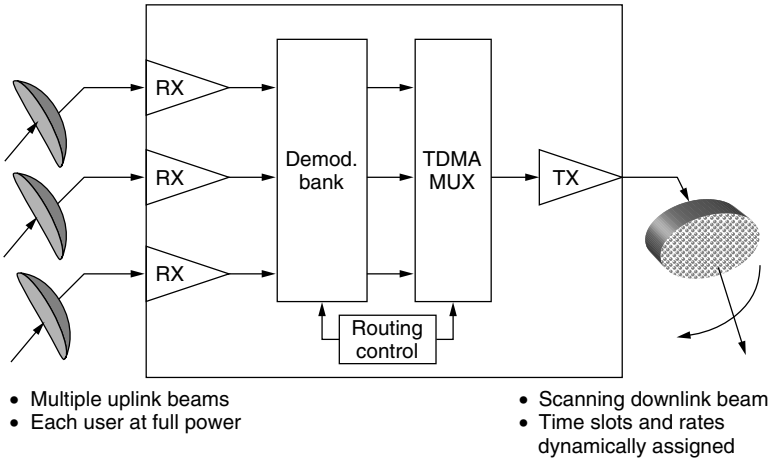


Figure 8. Onboard demodulation–remodulation example with beam switching.

for users in the same beam and low beam-to-beam sidelobe ratios. The satellite processor includes a bank of receivers and demodulators. The downlink consists of a single high-rate TDMA (time-division multiple access) datastream that carries all the downlink data. The burst rate of the data destined for an individual user can be adjusted in accordance with the receive capability of the user terminal. The TDMA downlink beam position is rapidly switched in synchronism with the TDMA stream to place downlink power on the location of the intended user. A downlink satellite transmitter can be used at high efficiency if the TDMA stream uses constant envelope modulation.

2.4. Despreading of Spread-Spectrum Signals

Additional techniques related to demodulation-remodulation can also be used to advantage. One is the onboard despreading of spread-spectrum signals, without complete demodulation.

Consider, for example, a frequency-hopped spread-spectrum system utilizing system bandwidth W Hz. An onboard processor consisting of a synchronized frequency-hopped local oscillator followed by a filter could be used to despread the uplink signal to its baseband bandwidth, approximately equal to the data rate, R bps. The resulting (analog) signal could then be transmitted on the downlink. Note that any broadband uplink noise or interference would be reduced by the ratio R/W, thus reducing the amount of retransmitted noise by the same factor. Although not quite as effective as complete demodulation, this would greatly reduce the "power robbing" caused by uplink interference. In a similar manner, a direct-sequence spread-spectrum system could be designed with an onboard despreader.

2.5. Access Control

Onboard access control places some of or all the functions of user system admission, such as interconnection control or usage monitoring, on the satellite. It provides the designer with a number of degrees of freedom to split these necessary functions between space and ground implementations. Its advantages lie with such factors as system security and survivability, system response times, and efficient use of control links. While some satellite systems require little real-time access control, such as those servicing television broadcasting, others must respond in real-time to user demands (e.g., telephone voice calls). A fully implemented spaceborne access control system would be a "switchboard in the sky."

System security and survivability can be enhanced by exposing fewer links that reach to ground control centers. An autonomous space system reduces the need for such links. In a hostile environment this could be an important factor.

System response times are reduced by not requiring a lot of control data traffic to make two ground-to-space-to-ground round trips between users and ground control sites. Whether this is a important consideration depends on the response time needed by the user community.

Efficient use of control links would reduce the requirements for uplink and downlinks devoted mainly to control. Virtually all satellite systems require control links for satellite on-orbit health maintenance. However, systems that must respond in real time to user demands have greater need for control dataflow. Placing control function on board reduces the data load on the control links, which could be advantageous in some system designs.

2.6. Circuit Multiplexing and Switching

Circuit switching was discussed in the context of onboard demodulation–remodulation of datastreams. Here we want to elaborate on the point that onboard processing can also serve as an *add–drop multiplexer*, a standard subsystem of the ground telecommunications industry. An uplink datastream from an individual user terminal might actually consist of a number of multiplexed substreams. Each of these substreams could have a different ground terminal as its destination. The onboard processors could demultiplex the uplink substreams, switch each substream to its intended downlink, and multiplex all the downlink substreams intended for a given terminal before transmission. By matching each substream to its intended downlink beam and terminal, satellite resources are used most efficiently.

2.7. Packet Switching

In data transmission technology, there is a strong trend toward the use of packet transmission protocols, particularly the IP (Internet Protocol) suite. Onboard processing is directly applicable to this approach. Indeed, aside from specific points made regarding data circuits or datastreams in the preceding sections, the basic advantages of onboard processing apply, and even more strongly to packet-switched systems.

Consider the following advantages to packet switching onboard the satellite:

- Packets can be routed to downlink users and satellite beams as needed, thus making efficient use of downlink resources without requiring dedicated circuits that might be used only in bursts.
- Multicast routing (one-to-many) is easily accomplished.
- With sufficient caching of packets on board, latency can be reduced by halving the user-to-user satellite transmission time.
- Data-link-layer protocols can be utilized to mitigate satellite link impairments, including path blockage due to the motion of ground terminals.

A major decision that needs to be made in the design of an onboard packet-switched system is the choice of protocol layer or layers to be implemented. If only the *physical* and *data-link layers* (layers 1 and 2 of the OSI Reference Model) are implemented, the satellite system is analogous to the connectivity of a LAN (local-area network). The system could directly interconnect satellite users to each other, but would rely on ground-based gateways for connectivity to users that are external to the system and possibly to satellite users that are connected to different satellites of a global system.

Onboard processing systems that include routers that operated at the next higher protocol layer, namely, the *network layer* (OSI layer 3), would be able to interconnect users in more capable ways within and external to the system in accordance with global networking standards. It would also be more effective in routing traffic between users within a global system that are attached to different satellites.

Operating the onboard processor at higher protocol layers could also be considered. For example, operation at the *transport layer* (OSI layer 4) would permit the caching of TCP packets that could be retransmitted with only a single round trip to the satellite to reduce latency. Operation at even higher layers, for example, the *application layer* (OSI layer 7), might also be considered with email and Web servers on board. (Store-and-forward messaging is a simplified version of this.)

2.8. Ground Processing Tradeoffs

The point of view taken in this article has been to illuminate the advantages and capabilities of onboard processing. However, it must be acknowledged that onboard processing comes with a price, namely, the need to place hardware on the satellites with the corresponding burdens of additional complexity, power, and weight. Whether this tradeoff is worthwhile depends on the system application. For example, satellites used for wide-area television broadcasting would gain little from processing; high-power translating repeaters generally suffice.

Another major tradeoff issue is where to place the processing functions that may be desired — in space or on the ground? Figure 9 illustrates some of these tradeoffs. Strong feeder links could, in principle, be used to relay all signals (from all antenna beams) received at the satellite in analog (or sampled) form to the ground. Virtually all processing functions (demodulation–remodulation, antenna beamshaping, packet routing, etc.) could then be done on the ground and relayed on strong links back to the satellite for subsequent retransmission, although at the cost of additional latency.

Designers of different systems may come to different conclusions regarding the use of processing and where to implement it. The next section presents selected system implementations that represent a range of processing applications and the range of design choices made.

3. EXAMPLES OF ONBOARD PROCESSING SYSTEMS

This section provides examples of satellite systems that have included onboard processing to varying degrees. Each will be described briefly to relate it to the types of processing described in the preceding sections.

3.1. Government Systems

A number of pioneering onboard processing systems were developed under U.S. government sponsorship:

- **Lincoln Laboratory Experimental Satellites 8 and 9 (LES-8 and 9)**, shown in Fig. 10, launched in 1976, were among the first to include onboard demodulation–remodulation and spread-spectrum despreading. These feature were included to demonstrate satellite-to-satellite crosslinks, UHF–Ka-band frequency crossbanding, and antijamming protection for a variety of military platforms. A more complete description can be found in Ward's study [3].
- The **Defense Satellite Communication System (DSCS)** [4], operating in the SHF region of the spectrum since the early 1970s, is an example of a system that utilizes flexible antenna patterns both to shape coverage areas and reject interference. Microwave lens antennas and steerable parabolic dishes are used. The signal processing that supports the beamforming is shared between the satellite and the ground.
- The NASA **Advanced Communication Technology Satellite (ACTS)** [1] has several onboard processing features. It includes a number of narrow beams, some of which can be rapidly steered. It also carries a signal processor very much like that shown in Fig. 5.
- The **Fleetsat EHF Packages (FEPs)** built by Lincoln Laboratory were launched in 1986 and 1989. The FEPs took jamming-resistant onboard processing a significant step forward with the

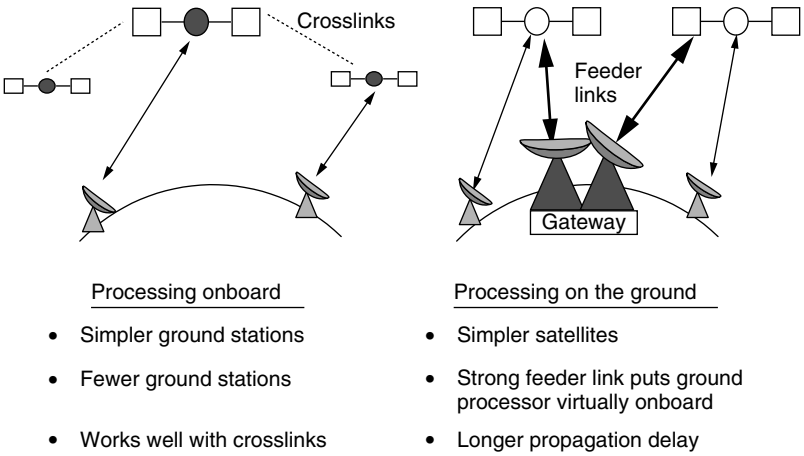


Figure 9. Processing onboard compared to processing on the ground.

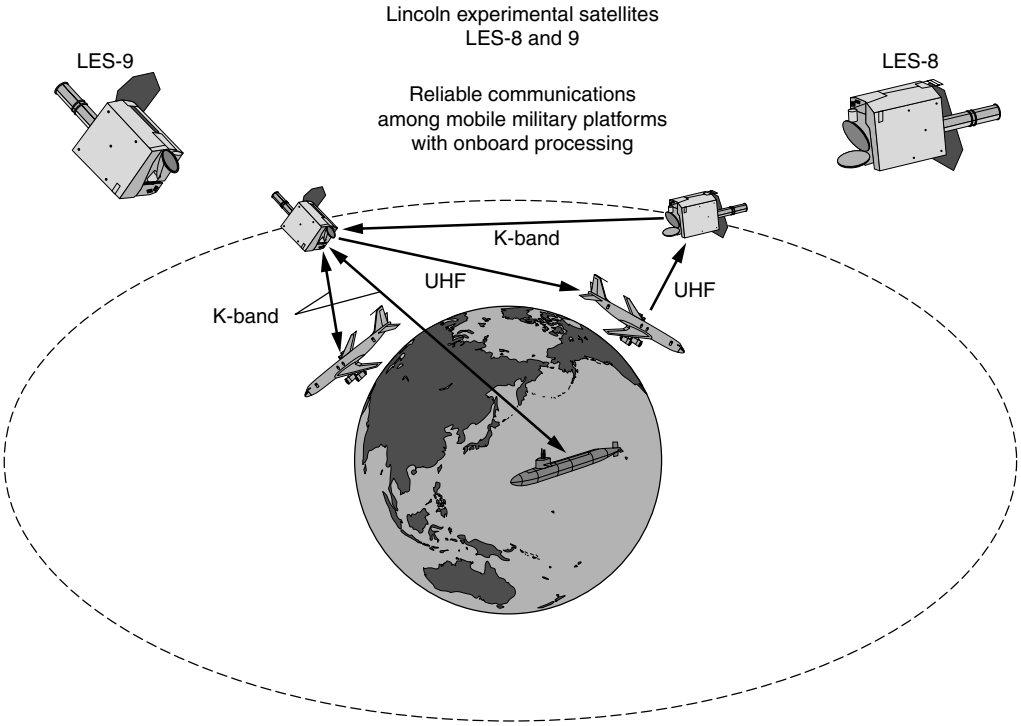


Figure 10. Lincoln laboratory experimental satellites 8 and 9 (LES-8 and 9).

inclusion of extensive demodulation–remodulation of multiple-user onboard access control realizing a true "switchboard in the sky" [5] and a steerable antenna. The FEPs can operate as a self-contained demand-assigned circuit-switched system. The FEPs are the forerunners of the Milstar [6] system, which includes more antennas, more channels, higher data rates, and satellite-to-satellite crosslinks.

3.2. Commercial Systems

A growing number of commercial system are using or are expected to use onboard processing.

Iridium provides worldwide digital voice and paging services to handheld user terminals [7]. The system makes extensive use of onboard processing. The spaceborne part of the system consists of 66 low-orbit (780-km-altitude) satellites in six orbital planes. Each satellite projects 48 narrow beams onto the earth; each beam is about 30 mi in diameter. In order to provide worldwide connectivity, the satellites are crosslinked to each other. A user signal (at L band) emanating from a handheld terminal is demodulated on board and crosslinked (at K band) around the Iridium constellation until it is downlinked to a gateway terminal connected to the terrestrial telecommunication infrastructure.

The **Thuraya** system provides voice and data services to most of Europe and portions of Asia and Africa. Each of its geosynchronous satellites includes a 12.25-m antenna that is used to form 250–300 spot beams. Its onboard processing is described as [8] follows:

- On-board digital signal processing (DSP) to facilitate interconnectivity between the common feeder link coverage and the spot beams to make effective use of the feeder link band and to facilitate mobile to mobile links between any spot beams
- Digital beamforming capability that will allow Thuraya to reconfigure beams in the coverage area, to enlarge beams, and to activate new beams. It also allows the system to maximize coverage of "hot spots," or those areas where excess capacity is required
- The flexibility to allocate 20% of the total power to any spot beam
- The flexibility to reuse the spectrum up to 30 times and therefore use the spectrum efficiently

The **Teledesic** [9] and **Astrolink** [10] systems have been under development since the mid-1990s. They both aim at providing broadband data services to small terminals around the world. As originally conceived, the Teledesic would have included a constellation of numerous low-orbiting satellites with extensive onboard demodulation–remodulation, routing, and crosslinking. Astrolink uses geosynchronous satellites with onboard ATM (asynchronous transfer mode) switches. Both systems' satellites have multibeam antennas.

It is too soon to predict the commercial success or failure of any these ventures. However, they are all pioneers in the technology of onboard processing, which will inevitably become a key part of the global information infrastructure.

BIOGRAPHY

Steven Bernstein received the S.B. and S.M. Degrees in Electrical Engineering from the Massachusetts Institute of Technology in 1964 and the Degree of Electrical Engineer from MIT in 1966. In 1966 he joined the Communication Division in of MIT Lincoln Laboratory, where he has

worked on and led a wide variety of communication and networking projects. His early assignments included work on a pioneering Air Force frequency-hopping UHF satellite communication system and the development of an ELF system for communication to submerged submarines. Following these staff assignments, he managed a succession of satellite communication projects at UHF and EHF for the U.S. Navy, Army, and Air Force. Following a one-year leave of absence to teach graduate level courses in digital communication at ENST (Paris, France), he returned to Lincoln and was given the responsibility to lead the development of a new satellite operations center. He also has managed the Optical Communication Technology Group, which developed novel techniques for optical ground and space communication. He is currently the Associate Leader of the Applied Communications and Information Technology Group. His areas of technical interest are satellite communication systems, channel coding/decoding, optical transmission, and applying these technologies to practical problems.

BIBLIOGRAPHY

1. F. M. Naderi and S. J. Campanella, NASA's Advanced Communications Technology Satellite (ACTS)—an overview of the satellite, the network, and the underlying technologies, *AIAA Int. Communication Satellite Systems Conf.*, 1988, pp. 204–224; AIAA paper 88-0797.
2. L. J. Ricardi, Communication satellite antennas, *Proc. IEEE* 336–369 (March 1977).
3. W. W. Ward, *Developing, testing and operating Lincoln Experimental Satellites 8 and 9* (*LES-8/9*), Lincoln Laboratory Technical Note 1079-3, Jan. 16, 1979.
4. R. Donovan, R. Kelley, and K. Swimm, Evolution of the DSCS Phase III satellite through the 1990's, *IEEE Int. Conf. Communication*, 1983, Vol. 2, pp. 611–619.
5. M. D. Semprucci, The first "Switchboard in the Sky": An autonomous satellite-based access/resource controller, *Lincoln Lab. J.* **1**(1): 5–18 (1988).
6. L. F. Kwiaykowski et al., The Milstar system, *AIAA Int. Communications Satellite Systems Conf.*, 1994, pp. 744–748; AIAA paper 94-1013.
7. Iridium, World Wide Web (WWW) Website *http://www.iridium.com*.
8. Thuraya, Website *http://www.thuraya.com*.
9. Teledesic, Website *http://www.teledesic.com*.
10. Astrolink, Website *http://www.astrolink.com*.

COMMUNICATION SYSTEM TRAFFIC ENGINEERING

APOSTOLOS K. KAKAES
Cosmos Communications
Consulting Corporation
Centreville, Virginia

1. INTRODUCTION AND MOTIVATION

At its core, the traffic engineering problem is very easy. At least stating the problem is! Indeed, traffic engineering falls into the relatively small set of problems that happen to be relatively easy to articulate, but whose solutions provide a window to a world that is new and in many ways surprisingly complex. Even though our current emphasis is on communication systems and networks, the problems and associated solutions have much broader applicability. We will occasionally refer to these other areas of applications, as they can also aid in the understanding of some of the underlying principles.

In a somewhat abstract sense, the problem can be formulated as follows (see Fig. 1). To a given pool of resources, consisting of N "servers," "customers" arrive with the intention (or need) to use one of the servers for a certain amount of time. Indeed, if one of the resources is available, it is "held" or "occupied" by the arriving customer. If no resources are available, the arriving customer is blocked. Thus we observe that the entire "offered load" is "split" into what becomes "served load" and "blocked load." The basic question then appears to be simple and final: "What is the probability that the system will be full, and thus will be unable to serve a potential customer, that is, how much of the offered load is served and how much is blocked?" Once this relationship is fully understood, it can be exploited to design, or size, a given system to meet certain performance objectives, or at least determine what the performance may be for a given set of conditions.

Note that the set of resources can be, for example

- Channels between points A and B (as is often the case in communication networks)
- Tellers at the bank
- Check-in counters at an airport

Also note that the language is not always helpful in uniquely identifying the problem! Arriving "customers" can be telephone calls or passengers who need to board a plane. Similarly, one would not ordinarily state that "a customer at a bank held (or occupied) the bank teller." In an abstract sense, however, these statements all refer to the same situation, where the resource is being used by an arriving entity, thus is unavailable for use by others, until, of course, it is "released." Once again, in this article, as a rule, we will refer to arrivals as *calls* that attempt to use one *channel* for a certain amount of time, called the *holding time*. Similarly, the terms "resource," "channel," and "server" will be used interchangeably, as indeed they are so used in the real world.

As we will see in detail later, one of the most important quantities that arise in the analysis of such a system,

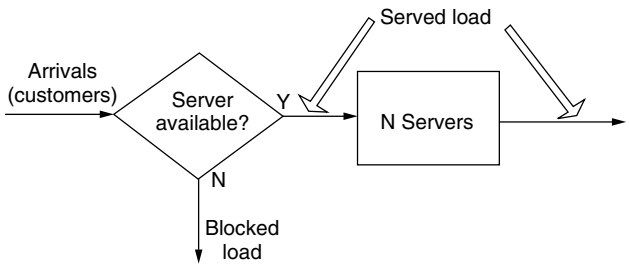


Figure 1. Schematic diagram of a service facility.

is the product of the average call arrival rate, commonly denoted by λ, and the average holding time, also commonly denoted by $1/\mu$. We thus define the *offered load*, A, by

$$A = \lambda \times \frac{1}{\mu} = \frac{\lambda}{\mu}$$

What are the units of the offered load? λ has the units of time^{-1} ("calls" is dimensionless), whereas $1/\mu$ clearly has the units of time. Thus A is dimensionless.[1]

It is in honor of the Danish mathematician A. K. Erlang, the father of traffic engineering, that a load A is referred to as "A erlangs." The basic problem can be stated very concisely as follows. An offered load of A erlangs is offered to a pool of N servers. What is the blocking probability, that is, the probability that an arbitrary arrival finds no server available? As it turns out, the process of answering this seemingly simple question entails more complexity than one might think! It is intuitively clear that, for a given value of N, the blocking probability will increase with increasing offered load, A, thus with increased values of the arrival rate and/or the average service time. What is not necessarily as obvious, is that the *arrival statistics* play a vital role. One can easily appreciate this by considering two particularly simple examples. These examples will also allow us to introduce a number of concepts used later in this article. We will assume that the

- Number of channels, $N = 1$
- Average arrival rate, $\lambda = 1$ arrival/min
- Average holding time, $1/\mu = 1$ min, and it is *constant* for each and every call

Examples 1 and 2 differ only in the way in which calls arrive to our one and only channel, as follows. We should emphasize that for now, we will not worry about how realistic (or unrealistic) these examples are. This point will be amply discussed later when we discuss modeling arrival processes in realistic situations.

Example 1. Calls arrive every minute on the minute. One might think of this as being "organized" by an external entity that coordinates the call arrivals to occur in this fashion: on the minute every minute.

Since each arrival occupies the channel for precisely one minute, it departs just as the next call arrives, and the new call occupies the channel.

We can easily see that the fraction of calls that find the system unavailable is 0, that is, that the blocking probability is 0.

Similarly, it is just as easy to see that the system utilization is 1, that is, the system is fully utilized, or as the Chief Financial Officer would say "we are making money at 100% of the potential to make money." From a practical perspective, the question of utilization is often of importance, as it is the *utilization* of the system that generates revenue!

This is perfection! No user is disappointed (blocked) and we (the network provider) make as much money as possible.

Example 2. All 60 calls arrive within the first minute following the first call's arrival, say, at the beginning of an hour. Once again, one might think of this as being "organized" by an external entity that coordinates the call arrivals to occur in this fashion: a whole bunch of them close together and then nothing for the remaining hour.[2]

It is clearly seen that the first arrival in the hour occupies the channel, and all remaining 59 calls find the channel occupied, thus are blocked. Thus the fraction of calls that find the system unavailable is $\frac{59}{60}$, with a blocking probability of almost 1.

Similarly, it is just as easy to see that the system utilization is $\frac{1}{60}$, that is, we are making money a mere 1.67% of the potential to make money. This is about as bad as it can get! Almost all users are disappointed (blocked), and we (the network provider) make essentially no money — this is one way to get into bankruptcy!

We can summarize Examples 1 and 2 in a manner that will also prove useful later, as shown in Fig. 2. Note that this is a two-dimensional space representing all the possible values of blocking probability, in the range [0,1] and utilization, also in the range [0,1]. Any point (x, y) in this space represents a system for which the blocking probability is x and the utilization is y. As a result, the points (0,1) and (1,0) represent the limiting cases discussed in the two previous examples, respectively. Clearly, in general, we would like to be operating at a point (x, y) with x as small as possible and y as large as possible, specifically, in the upper left quadrant of the unit square, as shown in Fig. 2.

What is the *only* difference in the two examples? The *arrival process*. Thus it becomes important that we undertake a study of the arrival process statistics in order to investigate the traffic engineering problem. We do so

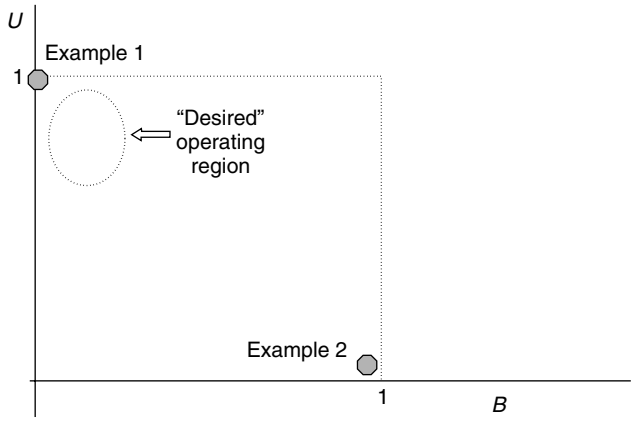


Figure 2. Utilization and blocking for different arrival statistics.

[1] It is worthwhile to note that, as we will see later, μ is called the "service rate" and then the load becomes the ratio of two rates: the arrival rate and the service rate.

[2] Note that the average arrival rate is still 1 call per minute and the average holding time is still 1 minute, thus the offered load is still 1 erlang, exactly as in Example 1.

in Section 2. In Section 3 we analyze the most common models used in traffic engineering: the Erlang B model, followed by the Erlang C and the Poisson models. The fundamental similarity and difference of these models lies in the way in which blocked arrivals are treated: In Erlang B it is assumed that blocked calls depart the system unserved. In Erlang C, it is assumed that arrivals that find the system full, thus are blocked, enter a queue, which is served in a first-in first-out (FIFO) manner. However, it is important to note early on that the Erlang C model assumes that blocked arrivals are willing to wait an unbounded amount of time for service. Thus, one might consider the Erlang B and the Erlang C as diametrically opposite models: One assumes that the arrivals are willing to wait 0 time, while the other assumes that the blocked arrivals are willing to wait as long as it takes. Clearly, there ought to be a model for some sort of in-between case! Indeed, the Poisson model is precisely such a model. We will elaborate on these three models, individually and collectively, in Section 3.

Finally, in Section 4 we discuss the problem of random access, which is of a somewhat different flavor, as articulated in time; while often not associated with conventional traffic engineering problems, the random access problem is indeed a traffic engineering problem and thus needs to be discussed herein.

2. MODELING THE ARRIVAL PROCESS

Clearly, the tools of statistics and random processes are needed in order to characterize the arrival process. This can get very technical and, while these technicalities are important, one can miss the essence of the problem. Our present discussion is limited to an intuitive level and the technicalities can be found in any number of books on probability theory and on random processes, such as the classic books by Feller [1,2]. We discuss the more critical concepts, in particular as they impact traffic engineering issues.

2.1. Stationarity

As noted earlier, load is measured in erlangs, a dimensionless quantity. First and foremost, defining the load has little to do with a period of time. Many practicing engineers, unfortunately, think of one Erlang as being "one call occupying one channel for one full hour." While this statement can be interpreted correctly (we will not attempt to do that here), it is very open to misinterpretation, which unfortunately happens too often. This is one reason why we will not make such usage here and we will use the definition of load in a way that leaves no room for misinterpretation. However, to do that, we need the concept of stationarity. Indeed, implicit in most of traffic engineering is the assumption that the problem is stationary.

Although the mathematical definition of stationarity is not too difficult, we will avoid such technical issues, appealing to one's intuition instead. Simply put, we require that the statistical nature of the underlying problem remain the same over time. As a trivial but helpful example is that of a sequence of flipping the same coin. While each outcome is in general different from the other outcomes, stationarity implies that the probability distributions that govern these outcomes remain the same. Note that we are *not* saying that it is a fair coin — it can be any coin. Stationarity merely says that whatever the statistical behavior is, it remain the same over time.

In a way the question is simple: Is the statistical nature of calls arriving to our resources (say a switch) the same at all times? Clearly, from a purist's perspective the answer must be "no," since things do change, people behave differently over time, and so on. Thus we do not have stationarity. But from a practical perspective, we can ask a slightly different question: Is it realistic to assume that over some nontrivial amount of time, the statistical nature of the problems remains constant? If so, then we need to ensure that such a period of time is long enough so that enough events take place, and yet not too long, so that the underlying statistical dynamics of the problem remain constant. In general, how long is long enough and not too long? While the technical definition of stationarity could be called on, we will appeal more to one's intuition. Clearly, what happens at 3 A.M. is different than what happens at 9 A.M. But how about 8 A.M. and 9 A.M. — is it sufficiently the same, from a statistical nature point of view, or is it changed and thus needs to be considered separately? The traditional answer to the question has been to not pay too much attention to it! In the early days of developing telecommunication networks (when the voice network was the prime example) the computing power to do a more elaborate analysis was simply not available, thus the technical community used a one-hour period as being long enough and not too long to be considered as a period of stationarity. Clearly, then, the hourly division of the 24-h day made some sense. Although some of the original premises are no longer valid, we still tend to use hourly data. To some extent, this has become a chicken-and-egg problem: switch manufacturers tend to default their software programs and so on, into one-hour periods because practicing engineers are used to that. Practicing engineers, in turn, tend to use hourly data because the switch provides these data! Even though in many of today's switches there are options of keeping and reporting other statistics, most engineers do not know that such options even exist, let alone how they might use it. So that brings us to the traditional hourly engineering. We, however, will not make any such assumptions. We will treat the problems from a slightly more abstract point of view, and simply assume that stationarity exists and not be concerned over how long it lasts for.

2.2. Busy-Hour Engineering

The discussion above naturally brings us to the notion of *busy-hour engineering*. As discussed above, there is nothing particular or important about a one-hour period. It was just convenient at one time. Second, the definition of *busy hour* is straight forward, albeit with its own difficulties. If we accumulate load on an hourly basis, then there are 24 measurements for a day, and the highest one is the busy hour. Assuming that the statistical nature of the problem does not change over days (which may or may

not be true), we can aggregate the same-hour data over a number of days and thus generate the concept of "busy hour" over periods of many days, weeks, or even months. Once again, stationarity can play an important part here: In an area serving ski resorts, clearly the statistics change from winter to summer months, so we need to be aware of the stationarity issues both in the small and the larger scale of time constants! As we indicated earlier, we will follow the usual approach and assume that the problem is stationary. A detailed analysis of the complexities arising from considering the *day-to-day variation* can be found in the book by Ash [3].

Two different fundamental questions are

- *Why* do we do busy-hour engineering?
- *Should* we do busy-hour engineering?

The answer as to *why* lies in history! Without getting into a deep historical retrospective, suffice it to say that in the (traditional) regulated monopoly environment, there were both regulatory and business reasons that were pointing toward taking a busy hour engineering approach to the network design problem. It is intuitively obvious that such an approach resulted in a network that was more expensive than it might otherwise be. This observation notwithstanding, the combination of business and regulatory issues in a regulated monopoly provided sufficient justification to indeed adopt such practices quite widely.

The question of continuing to do busy hour engineering in the new days of deregulation of telecommunications on a worldwide basis is a harder one to answer! Indeed, we will not answer it, as it becomes an economic analysis. Any such analysis must be interdisciplinary by nature. In this article we provide the traffic engineering tools needed to carry out such an analysis, but the point should be raised as too often methodologies are adopted only because "that's how it has always been done."

2.3. Definition of Load

As indicated earlier, the proper measure of load is the product of the average arrival rate times the average holding time. This product is referred to as "*A* erlangs." Even though on its way out, an old-fashioned measure still exists, so one should be aware of it: 100 call-seconds per hour (CCS). (The acronym is derived as follows: C, from the Latin initial for hundred; C, call; S, seconds; per hour, by common agreement or convention.) Some, especially older, systems still use CCS, but its usage is strongly discouraged. As will become apparent later, one can easily convert CCS to erlangs: 36 CCS = 1 erlang. Furthermore, the relationship of CCS to erlangs is analogous to that of degrees to radians in measuring angles; the first is arbitrary and based on some agreed-on convention. (Why does a complete rotation correspond to an angle of 360°? Why 360 and not, say 480? The number 360 is just as arbitrary as 480 or any other number! It only "feels" better because we have been used to it!) Much as radians are dimensionless measures of angles and defined as the ratio of two lengths, the erlang is dimensionless and is the ratio of two rates, which is an item we will return to shortly, For now, suffice it to say, that we will use only erlangs from here on, but one must be aware of the possibility of encountering CCS!

2.4. The Poisson Arrival Process

One of the most commonly made assumptions is that the arrivals occur according to a *Poisson process*. What does that mean? What are the practical implications of such an assumption? What are the practical and theoretical implications if this assumption does not hold? Before we attempt to answer some of these questions, we must develop at least some understanding of this concept, which is indeed of paramount importance.

Consider an arrival process of, for example, calls to a switch that satisfies one of the following three properties. We will consider them one at a time, and then use them to define the arguably most fundamental modeling assumption in traffic engineering: The Poisson process.

Property 1: The Infinitesimal Generator. Suppose that the arrival process is such that over an infinitesimal period of time h, the probability of exactly one arrival and exactly zero arrivals are given respectively by[3]

$$P(1 \text{ arrival}) = \lambda h + o(h)$$

$$P(0 \text{ arrivals}) = 1 - \lambda h + o(h)$$

Note that, in a very general sense, for arbitrarily small amount of time h, specifically, in the limit as h goes to 0, the probability of 1 arrival goes to 0 and the probability of no arrivals goes to 1! What we are saying here is much more profound, albeit not obviously so! We are saying that, in the limiting sense as time goes to 0, P (1 arrival) goes to 0, *linearly* with time, with constant of proportionality λ. Similar statement can be made about the probability of zero arrivals.

It is a direct consequence of these assumptions that the probability of two or more arrivals is $o(h)$. In other words, the probability of two or more arrivals over a small interval of time *goes to 0 very rapidly*. As it turns out, from a practical point of view $o(h)$ functions can be disregarded without any harm.

Thus, we are stipulating that over an arbitrarily small interval of time h (and thus the term "infinitesimal generator"), "essentially" (i.e., we ignore $o(h)$ terms) the following happen:

- One arrival occurs with probability λh.
- No arrivals occur with probability $1 - \lambda h$.
- Two or more arrivals occur with probability 0.

[3] The notation $o(h)$ (read as "little o of h") is common and is used to designate any function that satisfies the following limiting property:

$$\lim_{x \to 0} \frac{f(x)}{x} = 0$$

i.e., the function $f(x)$ approaches 0 (as x goes to 0) **faster** than x itself goes to 0.

We further assume that the arrivals in disjoint intervals are independent events, thus the number of such arrivals are independent random variables.

This "infinitesimal" property of arrivals may or may not appeal to one's intuition as being a good model for our problems. We will table further discussion of this model/property until we discuss two other options and then come back and compare all three of them.

Property 2: Poisson Distribution of Number of Arrivals. Here we consider an arbitrary period of time T — no longer an infinitesimal one — and ask the question of how many arrivals occur in it. Let's suppose that the number of arrivals K in the time period T has the Poisson distribution with parameter, λ. Recall then, that the probability distribution of the number of arrivals, K, is given by

$$P(K = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \quad k = 0, 1, 2, \ldots$$

We further assume that the number of arrivals in disjoint intervals of times T_1 and T_2 are independent random variables.

Property 3: Exponentially Distributed Interarrival Times. Here we assume yet another property of some arrival process: The time between two successive arrivals (called the *interarrival time*), X are independent random variables, exponentially distributed with some parameter λ. Recall that X has the exponential distribution if its probability density function is given by

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for} \qquad x \geq 0$$

Among other natural and useful questions to ask, from the engineering perspective, are the followings:

- Which of these three models, if any, is a realistic model for traffic arrivals to a switch or to a set of channels?
- How would one go about confirming or rejecting the modeling assumption that one of these models is indeed a good one?
- If one of these models is not quite good enough, could another one of them be better (or worse for that matter), but harder (or easier) to work with?
- If none of these three models is valid in a given practical situation, and we have established methodology to discover that fact, how do we go about finding a better one?
- Continuing with the previous item, how much error in our performance evaluation and/or design are we making by using the wrong model? It may be worthwhile to use an easier model, even if it is not quite right, if the deviations from the more accurate one are small.
- Assuming that we do find a better model, what will the performance be?

This may come as somewhat of a surprise, and it is intuitively not obvious at all, but the following is a fundamental theorem with many practical implications:

Theorem 1. If an arrival process satisfies any one of properties 1, 2, or 3, then it satisfies *all* of them and the arrival process is called a *Poisson arrival process*, or simply a *Poisson process*. The constant of proportionality λ is the same in all three defining properties and is called the average arrival rate of the arrival process. It is also referred to as the *arrival rate*, or even just the *rate*, or the *parameter* of the Poisson process.

In this article we will forego all proofs, but if this is the first time one sees this, it is highly recommended that one try to prove the equivalence of these three properties, as indeed they form the defining properties of the Poisson process.

While we could spend a large amount of time in investigating the deeper properties of the Poisson process, in the interest of brevity, we will not; however, we will briefly point out some of the more important salient points:

2.4.1. "Random" Arrivals. As it turns out, if a large population generates "arrivals" in a manner that involves *no external* control, the overall resulting arrivals will constitute a Poisson arrival process. This is precisely the reason why the Poisson arrival assumption has been used in telephony so much: when a user A makes a call is not "controlled" by any entity. Thus, precisely when user B makes a call is neither influencing nor is being influenced by user's A actions. This is also why some refer to the Poisson process as being one in which arrivals occur "at random." This, however, can be misleading, since "random" arrivals can obey any number of statistical properties and does not necessarily have to be Poisson. We must also emphasize that in practice the external controls *may* be there implicitly, thus "destroying" what might have otherwise been a Poisson arrival process! The requirement is that there be *no controls*, explicit or implicit ones!

2.4.2. Merging of Poisson Processes

Theorem 2. If n independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_n$ are merged, then the resulting process is Poisson with rate $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. In other words, merging several (independent) Poisson processes preserves the Poisson property.

2.4.3. Splitting a Poisson Process. How about the converse? Does splitting the arrivals of a Poisson process into two (or more) subprocesses preserve the Poisson property? As we are about to discover, the converse problem is a bit more complicated, the results are a bit more surprising, and the practical applications a bit more intriguing.

Suppose that a Poisson process with rate λ is "split", or bifurcated, into two (or more) sub processes. The question then is whether each subprocess is or is not Poisson (see Fig. 3). The mechanism labeled "load regulator" routes each arrival to the set of resources (channels) labeled

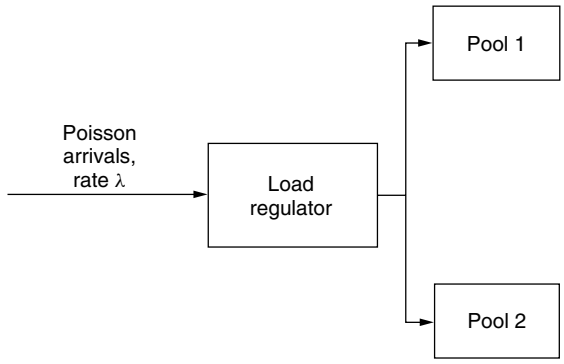


Figure 3. Bifurcation of Poisson arrivals.

"pool 1" or to the one labeled "pool 2." The mechanism by which it decides where to route each call is critical and we will be elaborating on this in the next few paragraphs.

As a first step, assume that the load regulator routes each call to the two pools in an alternating fashion. For convenience, we may assume that the first call is routed to, say, pool 1. Once again, assuming that the original process is Poisson with rate λ, the question is what can be said about each subprocess, i.e., the arrivals as seen by each pool of channels. The answer is provided in the next theorem.

Theorem 3. A stream of Poisson arrivals with rate λ is split into n substreams. Each call is routed to substream i with probability p_i, *independently* of all other calls, or the state of the system (thus of prior arrivals, etc.) Then each sub stream is also Poisson with rate $p_i\lambda$. Conversely, if the routing decision is made in some dependent fashion, including based on the state of the system, then the sub processes are *not* Poisson.

This is a good illustration of the fact that the Poisson process is a "delicate" entity. It is very easy to destroy it! Two practical examples will illustrate the point.

Example 3. Consider a network, a portion of which is as shown in Fig. 4. The direct route between nodes A and B is preferable, so the "load regulator" sends calls to the direct route, unless all channels are occupied in which case it sends the call via the alternative route, namely, via node C. Then, from a practical perspective, it is important to ask if the set of channels on the alternative path, links A–C and C–B are offered a Poisson arrival stream.[4] Since the decision to send the call to the A–B link or the A–C–B one is based on the state of the system, the substream of offered calls to the alternative route is not Poisson, even though the original traffic arrivals is assumed to have satisfied the Poisson assumption.

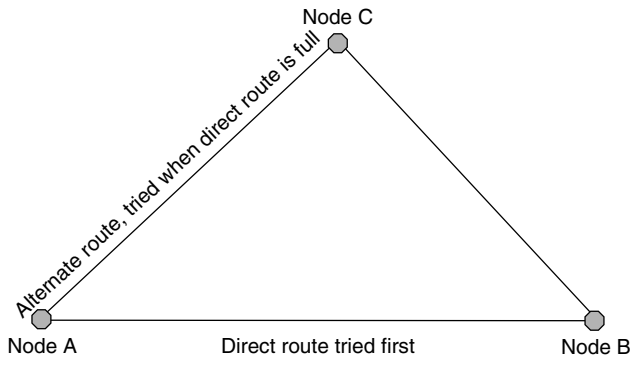


Figure 4. A simple 3-node network.

Example 4. In modern mobile communications networks, it is not uncommon for a given region to be capable of being served by two (or even more) cells. This is an example of the well-established concept of underlay/overlay. We will not get into the pros and cons of such designs or the radio engineering issues surrounding them, but the concept is simple. A given geographic region is divided into small cells as well as large ones (sometimes called "umbrella cells"). A given call is first attempted on the small cell, and if all channels are busy, it is then attempted on the large one. It is obvious that this is a problem equivalent to the one above. The point is the same — the offered load to the alternative set of resources is no longer Poisson.

As one can see, the mechanics and perhaps the reasons by which the load is "split" is rather irrelevant. However, destroying the Poisson property is not all that difficult to do! Both from a theoretical and a practical vantage point, we must be aware of that possibility and take appropriate measures.

2.4.4. Modeling Assumptions. From either a theoretical or a practical perspective, it is often the case that one needs to confirm or refute the assumption that arrivals form a Poisson process. It is often easy to make such an evaluation using one of the three properties of the Poisson process, whereas using the other two properties may be very difficult and/or impractical. We emphasize that these three properties are equivalent and one should avail oneself of whichever one is easier to work with. Furthermore, if one of these properties is shown not to hold, the others do not hold either and the assumption of the arrival process being Poisson is not valid. From the practical perspective this can have very significant implications to which we will return shortly.

2.4.5. If Not Poisson, Then What? If an arrival process is not Poisson, is there anything we can say about it? Yes, indeed there is. The Poisson process is merely one particular point in a continuum of possibilities. What is particular about it can be summarized in one sentence — its peakedness is 1. Peakedness is defined as the ratio of the variance of the number of arrivals to the mean number of arrivals. If the arrivals are Poisson, then we recall that the number of

[4] Clearly in a network environment this is only a portion of the problem. We are specifically and narrowly concerned with this particular portion of the traffic. In network designs, we obviously need to integrate this into a considerably larger set of issues. The reader is referred to Ash's treatise [3] for a comprehensive analysis of the issues in such network designs.

arrivals over any time of duration T satisfies the following conditions:

- Mean number of arrivals = λT.
- Variance of the number of arrivals = λT.
- Thus peakedness = variance/mean = $\lambda T/\lambda T = 1$.

Arrivals for which the peakedness $Z > 1$ is called "peaked" traffic, whereas if the peakedness $Z < 1$ we say that the traffic is "smooth."

We observe that the arrival processes discussed at the outset of this article satisfied:

- When arrivals arrive on the minute every minute, then variance = 0; thus peakedness $= 0 \ll 1$, thus we say that the traffic is very smooth.
- When arrivals came all "bunched up," the variance was large (with the same mean), thus peakedness $Z \gg 1$; thus, we had very peaked traffic.

Recalling our earlier discussion of Poisson traffic being often referred to as "random" traffic, we can make the additional observations that

- If no controlling mechanism exists, then arrivals will be in accordance with the Poisson assumption: $Z = 1$.
- If a controlling mechanism (implicit or explicit) "coordinates" arrivals to occur on a "regular" basis (thus lowering the variance), then the traffic exhibits some smoothness, where $Z = 0$ is the limiting case where there is no variability in the arrival stream.
- If, on the other hand, the controlling mechanism tends to "bunch up" traffic, then it is peaked traffic and the peakedness Z is arbitrarily large, according to the degree to which the "bunching up" takes place.

2.4.6. Time versus Call Congestion. Finally, Poisson versus smooth versus peaked traffic can be seen to have one more impact that needs to be discussed. Two related concepts are those of

- *Call congestion* (CC)—the fraction of call arrivals that find the system full. This is indeed what we have been referring to as probability of blocking.
- *Time congestion* (TC)—The fraction of time that the system is fully occupied and thus can not admit a new arrival.

Note that there is no a priori reason that these two measures of congestion should be equal to each other. Indeed, if the arrivals do form a Poisson process, they are. Otherwise they are not. Figure 5 summarizes the relationship where the time congestion and the call congestion occupy the two sides of a balance beam. In Fig. 5a arrivals are assumed to be Poisson ($Z = 1$); thus CC = TC. Figure 5b illustrates smooth traffic ($Z < 1$). In that case TC > CC. Note that this, in principle, is desirable, as it implies that utilization is higher while blocked arrivals are fewer (than would be the case with $Z = 1$). Of course, the converse is shown if Fig. 5c, where

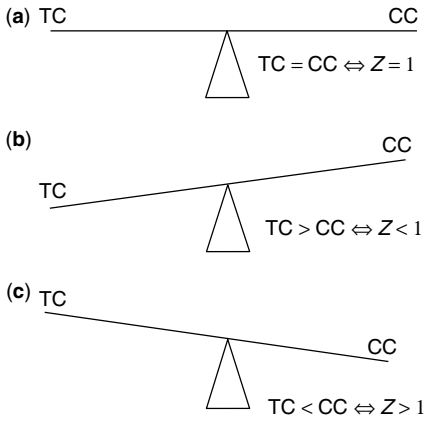


Figure 5. Time congestion versus call congestion.

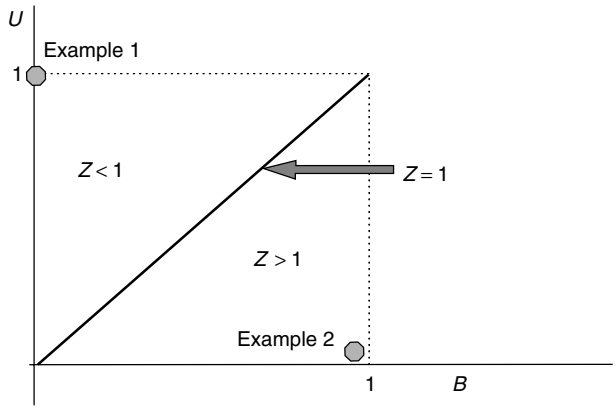


Figure 6. Utilization versus blocking for Poisson arrivals and $N = 1$.

$Z > 1$. Indeed, see Fig. 6, which is a generalization of what we discussed earlier in Fig. 2 for the case that $N = 1$. As the offered load ranges from 0 to infinity, one can see that a path is traced in the unit square of (B, U). The precise path depends on the peakedness. If $Z = 1$, the path is on the "45°" line, as indeed $U = B$ (for the case $N = 1$; we will shortly return to the more interesting case where $N > 1$). This line separates the unit square into two regions: One for smooth traffic ($Z < 1$) and one where traffic is peaked ($Z > 1$) as illustrated in the figure. While the straight line is a direct result of the system having just one channel ($N = 1$), the general concept is valid and will be generalized shortly, after we obtain the so-called result of the Erlang B model.

3. THE ERLANG *B* MODEL

As mentioned earlier, the most common model used for determining the quantitative relationship between the offered load A, the number of channels N, and the blocking probability $B(A, N)$ is the Erlang B model.

The Danish mathematician A. K. Erlang developed what is now known as the Erlang B model early this century, just as telephony was being born.

3.1. Modeling Assumptions and Associated Notation

The Erlang B model is used so widely, often without adequate understanding of the underlying assumptions.

Although a variety of modifications to the basic model are possible, and we will point to some of them, the basic model makes the following fundamental assumptions:

- The "service facility" or the "system" consists of a fixed number of "servers" or channels, N.
- Calls arrive to the system according to a Poisson process with an average arrival rate of λ calls per second.
- If a server, or channel, is available on a call's arrival, the call seizes a server. If more than one server is available, any one of them is seized — it makes no difference which one, as all servers are equivalent in all senses.
- A call occupies the server for an amount of time that is exponentially distributed with parameter, or rate μ, and independent of all other calls, arrivals, system state, or anything else. Thus the average service time, or average holding time is $1/\mu$. It is precisely the desire to express the service statistics in terms of the service rate (rather than the average service time itself) that early on we had the somewhat surprising notation of $1/\mu$ for the average holding time.
- If all servers are occupied at a call's arrival instant, the call departs the system unserved. In particular, it does *not* attempt to be served "a little bit later," try again, or any variant thereof.

Before we proceed, a brief notational comment. The system that we just described is often referred to as an $M/M/N/N$ queueing system. The notation is straight forward, and we present it in a somewhat more general context in the next paragraph.

In general, the notation $F_1/F_2/F_3/F_4/F_5$ implies the following service facility:

- In field F_1, a letter designates the interarrival times, namely, the arrival process. So, for example, M implies that the arrival process is Poisson. Maybe it should be P, you say! It is M, because a Poisson arrival process contributes to the system being a *Markov* chain, thus the letter M. In general, one can have other letters, For example, E_r, implying that the interarrival times have the r-stage Erlangian distribution. G stands for "general" distribution, and so on.
- In field F_2, once again a letter designates the service time distribution. So, for example, M implies that the service time is exponentially distributed. Once again, one might think that it should be E! It is M, because exponential service time is the other component that contributes to the Markovian behavior of the system. In general, one can have other letters; For example, D stands for deterministic (i.e., all service times are the same, etc.).
- Field F_3 is a numeric field, indicating the number of servers available at the service facility.
- Field F_4 is also a numeric field, indicating the "capacity" of the service facility. The term "capacity" is often used in many different ways, so we need to be careful! Field F_4 indicates the maximum number of customers that can be in the service facility, not necessarily being served. So, for example, $M/M/10/15$ represents a system to which arrivals (calls) occur according to a Poisson process, service times are exponentially distributed, there are 10 servers (channels), but there is room for 5 $(15 - 10 = 5)$ arrivals (calls) to be "waiting" for service. In other words, the maximum number of arrivals or users or calls in the system is indeed 15, but only 10 can be in service at any one time.
- Finally, field F_5 is also numeric and it represents the overall population from which arrivals can occur. In our work, and whenever it is much greater than the value in F_4, we just assume it is infinite.
- Also, if any of the numeric fields is missing, it is assumed to be infinite, so in our example above $M/M/10/15$, indeed we assumed that the population base is infinite, indeed a good approximation as long as the population base is much larger than 15.

Note that the $M/M/N/N$ system implies that there is no "waiting room," which is consistent with our assumptions. Indeed, this is an assumption that is often in conflict with human nature. For instance, if I try to make a call and do not get through, I am more likely than not to try again, and that very action violates the assumption at hand. We will return to this point, but for now we'll just say that this assumption, which is somewhat in conflict with human nature, is a weakness of the model and we have to take the model at its face value, else we can not proceed!

3.2. State-Space and State-Transition Diagrams

The mathematical definition of *state space* and the resulting *state-transition diagram* are nontrivial and involve a fair amount of mathematical formalism. The reader is referred to the excellent presentations in Refs. 4 and 5. Here we present a summary of how these tools become useful in analyzing a variety of traffic engineering problems. We accomplish this through the analysis of the $M/M/N/N$ system, which in turn leads us to the development of the so-called Erlang B formula.

As it turns out, for a Markovian system [1,2,4], the number of customers in the system is a state.[5] For non-Markovian systems, the interested reader is referred to the treatise by Akimaru and Kawashinia [6]. The set of states constitutes the state space, as illustrated in Fig. 7. Once all states have been identified, they can be represented in any convenient manner. Traditionally, we simply draw a circle, with the state label inside it, as in Fig. 7a. However, in general, one can think of software tools and therefore other representations of the state space.

The next concept is critical: The *state-transition diagram*. Transitions from one state to another are clearly possible, as indicated on a state-transition diagram (Fig. 7b). The transitions are indicated by an arrow going from or to the respective states. The arrows are labeled

[5] Why this is not necessarily so in general is a difficult and deep question beyond our scope. While not obvious at all, it is the fact that all underlying times (interarrival times and service times) are independent and exponentially distributed that makes this statement possible. For a deeper understanding, the interested reader is strongly encouraged to pursue this in the references.

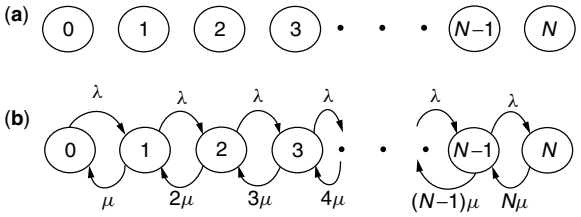


Figure 7. (**a**) State space; (**b**) state-transition diagram for the M/M/N/N system.

with the rate at which a "departure" from a given state to another state occurs. Once this set of states and associated transitions are identified, the entire "thing" is called the state-transition diagram. The state-transition diagram becomes useful in finding the steady-state probability distribution of being in each state. Note that although the state space consists of $N + 1$ states in this example, in general the state space can be finite or infinite. Similarly, drawing the state-transition diagram may be easy or not viable at all, depending on the complexities of the state space and associated transitions.

3.3. Flow Conservation Principle and the Erlang *B* Formula

The fundamental property that makes the state space a useful definition is the notion of flow conservation. As do other conservation laws in physics (conservation of energy, conservation of momentum, etc.), they tend to be rather technical in their proof, but once we accept them, their utility becomes very clear. How many readers have read the proof of conservation of energy? We all learned about it though! Similarly, here we will take the law of conservation of flow as given and work with it. Once again, the interested reader is referred to the references for further reading, specifically the volumes by Kleinrock [4,5].

What is the law of conservation of flow? First, let's define "flow." Referring to Fig. 8, consider a closed surface encompassing state i. Let P_i be the steady-state probability that the system is in state i. Consider a transition from state i to some other state with some rate λ as shown by the arrow (without necessarily indicating which state the transition is into). As it turns out, the natural definition of flow is very easy and quite intuitive: Flow out of state i is the product λP_i.

The flow conservation law, as you might guess, simply states that the flow out of a state is equal to the flow into the state. In fact, it is more general! It states that across any closed surface, containing any set of states, the flow into the closed surface equals the flow out of it. The choice of which states to include in the closed surface is entirely

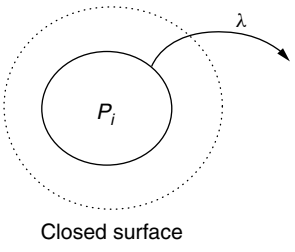


Figure 8. Definition of flow.

up to us! Judicious choices will result in a set of equations that can be solved for the probabilities P_i, i.e., the steady state probability distribution of the various states can be obtained and then applied in a variety of ways.

Once again, the state-transition diagram for the $M/M/N/N$ system is shown in Fig. 7b. Note that transitions from any state $i < N$ to state $(i + 1)$ occur with rate λ, as that is the arrival rate of new calls. Similarly, if the system is in state 1, then the one existing call departs with rate μ, so the transition to state 0 occurs with rate μ. On the other hand, if the system is in state 2, then there are two ongoing calls in the system. Departure of either one of the two, moves the system to state 1. But each call occupies its channel for an independent, exponentially distributed time, with parameter μ. Thus, the departure from state 2 to state 1 occurs with exponential time, with rate $\mu + \mu = 2\mu$. Similarly, if the system is in state $i \leq N$ it departs for state $(i - 1)$ with rate $i\mu$.

The Erlang B formula is essentially one step away! Let us consider the sequence of closed surfaces that enclose

1. State 0
2. States 0 and 1
3. States 0,1, and 2
4. ....
5. States 0,1,2,... $(N - 1)$

Applying the flow conservation principle to these surfaces, one gets the following system of N equations, with the probabilities $P_i (i = 0, 1, 2, \ldots, N)$ as the $(N + 1)$ unknowns:

$$\lambda P_0 = \mu P_1 \tag{1}$$

$$\lambda P_1 = 2\mu P_2 \tag{2}$$

$$\cdots \tag{3}$$

$$\lambda P_{N-1} = N\mu P_N \tag{4}$$

As the reader can see, an easy inductive argument will show that

$$P_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0$$

Finally, the last equation comes from the normalization step, that is, from the fact that all probabilities must sum to 1 which allows us to compute P_0:

$$P_0 = \frac{1}{\sum_{n=0}^{N} \frac{A^n}{n!}}$$

Note that this last normalization step is often one that is quite difficult to perform in more complex systems. Nevertheless, for our problem, we finally get that the probability of being in the blocking state (state N), which is indeed the blocking probability is simply

$$B(A, N) = \frac{\frac{A^N}{N!}}{\sum_{n=0}^{N} \frac{A^n}{n!}}$$

where we have obviously given the term λ/μ the name it deserves. Load A, which as discussed earlier is dimensionless and indeed the ratio of two rates: the arrival rate and the service rate. This is the famous Erlang B formula, a cornerstone of traffic engineering.

3.4. Important Points and Issues

Several points need to be made about the Erlang B formula/result:

1. In Erlang's days, computing power was simply not available. Thus, in the interest of making his result easily applicable and accessible to the engineering community, he developed the *Erlang B tables*, which is a tabulation of the Erlang B formula relating the offered load, the number of available channels, and the resulting blocking probability.
2. For a given number of channels (N) and a desired maximum blocking probability, the table provides the maximum offered load that would result in the said probability of blocking. An illustration of the Erlang B table is presented in Table 1. It should be noted that the granularity of neither N nor $B(A, N)$ is fixed. Some tables provide finer granularity than others, but they are all referred to as "the Erlang B table."
3. While until a few years ago, the Erlang B table was indispensable to a traffic engineer, these days, there are much more efficient numerical techniques that can be implemented in a few lines of code that work much more accurately, for instance, than a table lookup approach. One such approach is the iterative formula is

$$\frac{1}{B(A,N)} = 1 + \frac{N}{A}\frac{1}{B(A,N-1)}$$

It should be noted that there are other approaches, but they are beyond our scope. Notably, the Erlang B formula can be generalized to noninteger values of N, in particular when dimensioning networks. The interested reader is referred to Girard's treatise [7], where, in addition, the dimensioning problem is treated comprehensively.

4. It is clear that the relationship between N, A, and $B(A, N)$, has certain intuitively obvious monotonic behavior e.g., for a fixed number of channels, as the load goes up, the blocking probability goes up. However, the relationship is anything but linear! Interpolations and extrapolations are very dangerous, as the nonlinear behavior of the blocking probability can result in significant errors.
5. What is meant by *capacity*? As is often in engineering problems, this term must be defined in relationship to some performance objective. This is where the commonly made definition of N as capacity is rather inaccurate and possibly misleading. In traffic engineering terms, *capacity* is defined as the maximum offered load that a system can sustain at a given level of performance, that is, at a given blocking probability. For, example, referring to Table 1, we see that a system consisting of five channels and a desired blocking probability of no more than 0.02 (2%) has capacity of 1.66 erlangs. If the number of available channels were to be increased to 10 (i.e., doubled), the capacity would become 5.08 erlangs, about 3.1 times more than with five channels. If one were thinking that the capacity is doubled, the error made would be significant. So one needs to be careful when referring to capacity one must distinguish between the amount of resources (N) and the maximum offered load at a given performance level.[6] Indeed, this nonlinearity is often referred to as *economies of scale*, or *trunking efficiency*, a point to which we return in item 7 below.
6. Unfortunately, increasing the number of channels by a certain factor does not uniformly increase the capacity by the same, or even a constant, factor. The reader should consider what happens if the number of channels is increased from, say, 20 to 40, again a factor of 2, as in the previous example. Similarly, these capacity increases are different, depending on what value we have selected for the blocking probability.
7. It is worth our effort to return to a more general version of Fig. 6, shown in Fig. 9. Recall that as the offered load increases from 0 to infinity, the point (x, y) representing the blocking and utilization of the system traces some path in the $B \times U$ unit square from the point (0,0) to the point (1,1). That

Table 1. The Erlang *B* Table

B(A,N) *N*	1%	2%	5%	10%	30%	50%
5	1.36	1.66	2.22	2.88	5.19	8.44
10	4.46	5.08	6.22	7.51	12.0	18.3
15	8.11	9.01	10.6	12.5	18.9	28.2
20	12.0	13.2	15.2	17.6	25.9	38.2
25	16.1	17.5	20.0	22.8	33.0	48.1
30	20.3	21.9	24.8	28.1	40.0	58.1
35	24.6	26.4	29.7	33.4	47.1	68.1
40	29.0	31.0	34.6	38.8	54.2	78.1
45	33.4	35.6	39.6	44.2	61.3	88.1
50	37.9	40.3	44.5	49.6	68.5	98.1
55	42.4	44.9	49.5	55.0	75.6	108.1
60	46.9	49.6	54.6	60.4	82.7	118.1
65	51.5	54.4	59.6	65.8	89.8	128.1
70	56.1	59.1	64.7	71.3	96.9	138.1
75	60.7	63.9	69.7	76.7	104.1	148.0
80	65.4	68.7	74.8	82.2	111.2	158.0
85	70.0	73.5	79.9	87.7	118.3	168.0
90	74.7	78.3	85.0	93.1	125.5	178.0
95	79.4	83.1	90.1	98.6	132.6	188.0
100	84.1	88.0	95.2	104.1	139.7	198.0

[6] This problem has been quite prevalent in the mobile communications community, where, depending on technology used, and other variables, the number of available channels is increased by a certain factor (e.g., by a factor of 3). It is widely reported, then, that the capacity is increased threefold. As we have seen that is not quite accurate, and as traffic engineers we should not make such oversimplifications.

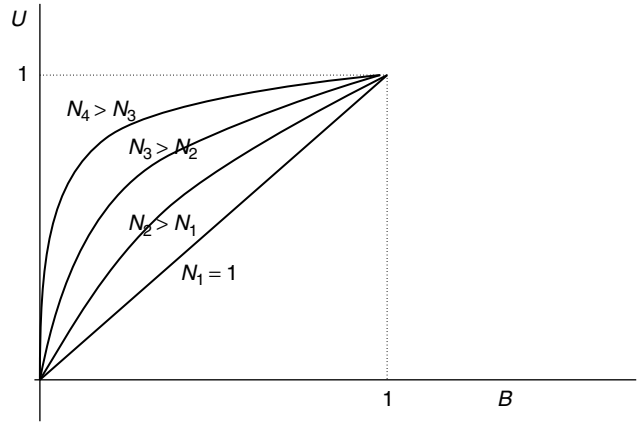


Figure 9. Utilization versus blocking for Poisson arrivals and arbitrary number of channels, N.

"curve" is shown in Fig. 9 for a system in which the number of channels is $N_1 = 1$ (as we did in Fig. 6) to progressively larger values $N_2, N_3, \ldots$. Although not illustrating a precise quantitative relationship, one can see why the larger system is said to be more efficient; for a given level of blocking, the utilization is higher. We ought to note that the curves in Fig. 9, also partition the unit square into two parts: one above and one below the "curve" that joins (0,0) with (1,1) for any given N. Points that lie on that curve represent the situation where arrivals occur according to a Poisson process. As before, the region above that curve represents what would happen if the peakedness is less than 1 and the region below the curve represents the situation where traffic is peaked.

8. Since a fraction of the offered load is blocked, we can now easily introduce the concept of carried load, $C(A, N)$:

$$C(A, N) = A[(1 - B(A, N))]$$

9. Although not necessarily obvious, it is relatively easy to prove that the carrier load is always less than the number of channels N. In fact, it is also easy to show that the ratio $C(A, N)/N$ is a measure of utilization of the system and is indeed a number in the range [0,1). We note that if the offered load A is Poisson, A = 1 erlang and $N = 1$, then $C(1, 1) = \frac{1}{2}$ and the utilization is also $\frac{1}{2}$. Compare this result with what happened when the same 1 erlang of load had peakedness $Z = 0 \ll 1$ and when $Z \gg 1$.
10. Solving the system of equations resulting from the flow conservation equations is not always an easy task. Numeric techniques can be employed, and the interested reader is referred to the treatise by Robertazzi [8], where the problem of a network of such queues is analyzed and numerical techniques are also presented.

3.5. The Erlang *C* Model and Blocked Calls Delayed

While the terminology "blocked calls delayed" and "blocked calls held" is not sufficiently clear and differential, it has been widely adopted by the community at large, so one has to be aware of the differences of the Erlang C model to be discussed here and the Poisson model to be discussed later.

Analysis of the Erlang C model is quite straightforward. Once we realize that the basic assumptions are the same as before (in the Erlang B model), with only one exception, the process is strikingly similar. The exception is that we assume that arrivals that find the system full, simply enter a queue and are served by the service facility on a first-come, first-served basis, else known as *FIFO* (first-in, first-out). We thus have an $M/M/N$ system.[7]

The state-transition diagram is shown in Fig. 10. Obtaining the steady-state probability distribution is quite straightforward as before, but with a couple of important observations:

1. The state space is infinite. This is a direct consequence of assuming an unbounded queueing room capacity.
2. The transition rates are just as before up to state N. Once the system is full, only one of the N customers that are in service can depart, so from that point on, the departure rate (from states k to state $k - 1$, for $k > N$) is always $N\mu$.
3. The fact that the state space is infinite implies that the sum of all probabilities, must converge. As we will see in a moment, the convergence requirement translates into the requirement that the offered load A must satisfy $A < N =$ number of channels in the system.
4. The term "blocking probability" has quite a different meaning in the Erlang B and the Erlang C models. Note that when blocked calls are cleared, they depart the system without being served, so indeed the term "blocking probability" is reflective of what happens. In the Erlang C model, *no* arrival is *blocked* in the sense that *all* arrivals do get served. It is only a question of being served immediately on arrival or having to first enter the queue for some time $T > 0$. Thus when one refers to blocking probability in the Erlang C model, one refers to the probability that an arbitrary arrival has to enter the queue, and thus be subjected to some delay, before being served. By the way, this is also the reason why this model is often referred to as *blocked-calls delayed model*.
5. A natural and important question arises from consideration of paragraph 4. Not only is the "blocking probability" important; probability of the

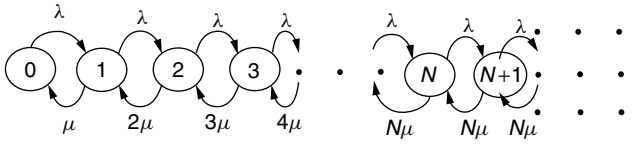


Figure 10. State-transition diagram: Erlang C model.

[7] Recall that the convention is that if a numeric field is absent it is assumed to be infinity. Thus we are making the explicit assumption that the waiting room is infinite.

delay exceeding some threshold may be even more important and critical. Thus we need to quantify the delay distribution, namely, the probability that the waiting time T exceeds some value t, $P(T > t)$. In what follows we show both the probability of finding the system full (blocking probability, or queueing probability) as well as the distribution of the delay. Which of these two expressions is *the* Erlang C formula is somewhat unclear. Some authors refer to one of these expressions — others, to the other one of these expressions — as being the Erlang C formula.

Application of the flow conservation equations reveals that the steady-state probabilities P_k, if they exist, must satisfy

$$P_k = \begin{cases} P_0 \dfrac{A^k}{k!} & \text{if } k \le N \\ P_0 \dfrac{A^k N^{N-k}}{N!} & \text{if } k > N \end{cases}$$

The normalization step requires that the infinite sum of probabilities be summable to 1; in particular, it must converge. As is easy to see, the infinite sum converges if and only if $A < N$. Indeed, from here on we will assume that the load A satisfies $A < N$, that is, that there exists a steady-state solution. While the mathematics forces us directly into making that assumption, it is worthwhile contemplating the practical aspects of this requirement. What would happen if $A > N$? After all, the offered load is whatever it is, and there is no reason why it could not be greater than the number of channels. The answer is strikingly simple — the system never reaches steady state, and thus it becomes unstable. The state-transition diagram of Fig. 10 is helpful in understanding what happens in that case. Recall that we can consider the flows as a "tendency" toward certain states. $A > N$ is equivalent to $\lambda > \mu N$. Note that the arrival rate λ, tends to "push" the system into the higher states. On the other hand, the service rate μ tends to push it into the lower states. In steady state ($\lambda < \mu N$) there is a balance of sorts; the systems drifts up and down and reaches all states, each one with a given probability, precisely the steady state probability of the given state. If, on the other hand, λ is large enough, then μ looses the battle, and the drifting toward the higher states is overwhelming. In such a case, the queue builds up without bound, and every arrival is assured of finding the system full — indeed, the queue is growing without bound as time evolves. The only way to stabilize the system is to reduce the offered load. (or, of course, increase the number of servers N).

Finally, one can show that as long as the system is stable ($A < N$), the complimentary cdf of the waiting time T is given by

$$P(T > t) = \frac{NB(A,N)}{N - A[1 - B(A,N)]} e^{-t\mu(N-A)}$$

where A = offered load
N = number of channels (and recall $N > A$)
μ = service rate (i.e., $1/\mu$ is the average holding time)
$B(A,N)$ = Erlang B blocking for offered load A and system of N channels

It follows trivially that the "Erlang C blocking probability," or as it is sometimes called, the "queueing probability," is given by

$$P_Q(A,N) = P(T > 0) = \frac{NB(A,N)}{N - A[1 - B(A,N)]}$$

Note that it is a simple algebra problem to show that $P_Q(A,N) > B(A,N)$ for all values of A and N. Although this is true, one should be aware of the discussion earlier regarding the direct comparison of these probabilities. Also, depending on the author, either of these two formulas is referred to as the *Erlang C formula*. Since the second one is a special case of the first one, it is the preference of some to refer to the first of these formulas as the Erlang C formula, but the terminology is by no means consistently used.

Note that, technically speaking, if $A < N$ is violated, nothing can be said, since the system is unstable. However, it is often stated that if $A > N$, the blocking probability is 1, thus, in that case, effectively all arrivals enter the queue. In any case, one should also note that as A approaches N, the waiting time grows without bound.

What is the capacity of such a system, operating in accordance with the Erlang C model? Recall that in the Erlang B model, capacity is defined as the maximum offered load that can be sustained at a given blocking probability. Thus, referring to Table 1, a system of 20 channels in which the maximum blocking probability is 2% has capacity of 13.2 erlangs. If the acceptable blocking probability were to be 5%, then the system capacity would be increased to 15.2 erlangs. Capacity is intimately tied to performance. What is the corresponding performance metric in the Erlang C case? Unfortunately, the answer is not as clear-cut. Just as there is some confusion about which formula is *the* Erlang C formula, there is plenty of confusion what performance objectives one should adopt. Some opt to just require that the blocking probability be no more than some value, say, 2%. Others require that some delay threshold be met most of the time; for instance the probability that the delay exceeds $\frac{1}{2}$ normalized units[8] be no more than 2%. As was the case with the Erlang B model, the more relaxed performance requirements we pose, the more capacity the system has. However, if one examines the Erlang C formula a little closer, one will observe that as the delay objective is relaxed from 0 to some positive value, the capacity increases rather sharply. But then, as the delay requirement is relaxed further, the marginal capacity improvement is small. Conversely, one could make the observation that at some point, even small increases of the load, result in large increases in the delay threshold, for a given probability of exceeding that threshold.

[8] Rather than using seconds, minutes, or whatever unit of time one might find convenient in a given situation, the uniformly most convenient unit is one that is normalized to the average service time. Thus a normalized delay of $\frac{1}{2}$ simply refers to the fact that the wait time is $\frac{1}{2}$ of the average service time. Clearly, such a unit is dimensionless.

Table 2. The Erlang *C* Table

Number of Channels	Offered Load	Blocking Probability (%) for Erlang *B*	Blocking Probability (%) for Poisson	Blocking Probability (%) for Erlang *C*
5	2.5	6.97	10.88	13.04
	3.0	11.01	18.47	23.62
	3.5	15.41	27.46	37.78
	4.0	19.91	37.12	55.41
	4.5	24.30	46.79	76.25
	5.0	28.49	55.95	100.00
	5.5	32.41	64.25	100.00
	6.0	36.04	71.49	100.00
10	5.0	1.84	3.18	3.61
	6.0	4.31	8.39	10.13
	7.0	7.87	16.95	22.17
	8.0	12.17	28.34	40.92
	9.0	16.80	41.26	66.87
	10.0	21.46	54.21	100.00
	11.0	25.96	65.95	100.00
	12.0	30.19	75.76	100.00
20	10.0	0.19	0.35	0.37
	12.0	0.98	2.13	2.41
	14.0	3.00	7.65	9.36
	16.0	6.44	18.78	25.61
	18.0	10.92	34.91	55.08
	20.0	15.89	52.97	100.00
	22.0	20.90	69.40	100.00
	24.0	25.71	81.97	100.00

For the sake of an example, we have tabulated a small portion of the "Erlang *C* table" in Table 2. Note that once again, table lookup was a shortcut when the numerical calculations were not readily available. As is easily seen today one could easily make the calculations, even in a basic spreadsheet, so the need for a table is obsolete.

As noted earlier, the Erlang *B* and the Erlang *C* models are diametrically opposite; one assumes that the blocked user waits no time, the other model assume that such a blocked user is willing to wait an unbounded amount of time. Clearly, from a practical perspective, we'd like an "in between" modeling approach. Indeed the Poisson model, examined next, is precisely that.

3.6. The Poisson Model and Blocked Calls Held

This is the $M/M/\infty$ model. It assumes an infinite waiting room, as in the Erlang *C* model, but it assumes that arrivals that have entered the queue can depart unserved. Before we proceed, the reader should be alerted that the term "Poisson model" can be misleading. As Poisson was very prolific, many things have been named after him possibly causing confusion! In all three models, we should remind ourselves, it is assumed that arrivals occur according to a Poisson process. This assumption is no more or no less true in the Poisson model. In many ways, it would have been much simpler if this had been termed the "Erlang *D*" model, but such is not the case! Or maybe the "Erlang *H* model," *H* for *held*, as this is also referred to the situation where blocked calls are "held" until a server becomes available, or they give up and depart unserved. Had we picked such a terminology, it would be clearer that all models make the same fundamental assumptions. They differ only as to how a blocked call is handled. Be that as it may, the acceptable terminology is "Poisson model" and we are not about to try to change that.

However, we have not completely specified the model as yet. What is the statistical nature of the waiting time? In other words, what assumptions does the Poisson model make about the willingness of blocked customers to wait in the queue for service? Indeed, that is a critical component of the Poisson model. Recall that in all models, it is assumed that the service time distribution is the exponential distribution with some parameter μ. The Poisson model assumes that the waiting time has the same exponential distribution as the underlying service time. Intuitively speaking, the Poisson model assumes that if a given arrival were to keep the channel for a certain amount of time t drawn from the exponential distribution with mean $1/\mu$, the amount of time that the customer would be willing to wait in the queue for service would be drawn from the same distribution. Of course, once the service begins, it all "starts fresh" and the holding time is whatever it is.[9]

Since any customer, whether in service or not, is allowed to depart, the departure rates are not quite the same as they were in the Erlang *C* model. They are shown as part of the state-transition diagram shown in Fig. 11.

As before, one can estimate the steady-state probability distribution. From that we obtain the blocking probability,

[9] This is not as strange as one might think. Actually, it is a direct consequence of the memoryless property of the exponential distribution, which itself is somewhat counterintuitive, until one develops a deeper understanding of the exponential distribution. The interested reader is referred to good books on probability theory [e.g., 1,4].

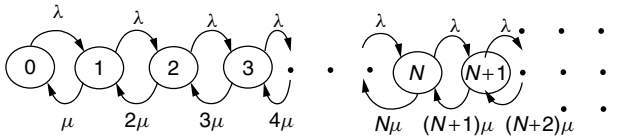


Figure 11. State-transition diagram: Poisson model.

that is, the probability that the system will be found full by an arbitrary arrival and thus it has to enter the queue:

$$P(\text{system full}) = 1 - e^{-A} \sum_{n=0}^{N-1} \frac{A^n}{n!}$$

Note that this is valid, that is, the infinite sum converges, for all values of offered load A. Unlike in the Erlang C case, we do not need to make the assumption of the load being $A < N$. Why is this so? From a mathematical perspective the answer is trivial — the infinite sum of steady-state probabilities converges, regardless of A. From a practical/intuitive perspective, the fact that departures occur with progressively higher rates as the system finds itself in the higher states, creates a stabilization force of sorts. Indeed, if we compare the state-transition diagrams of the Erlang C and the Poisson models (Fig. 10 and 11), we see that their fundamental difference was that the departure rate was the same (μN) for all states k after state N for the Erlang C model, whereas in the Poisson model the rates continue to increase indefinitely; specifically, for sate k, the rate is μk, for all values of k.[10]

As was the case with the Erlang C model, there is a semantic difficulty with respect to the term "blocking probability," but there is an even more pronounced difficulty here. Whereas in both the Erlang B and the C models there are two "classes of citizens," here there are three. Indeed, on arrival, here are the set of possibilities:

- Erlang B model
  - The call is served immediately and departs after its service is completed.
  - The call departs immediately, without being served (blocked calls).
- Erlang C model
  - The call is served immediately and departs after its service is completed.
  - The call first enters a queue, waits as long as necessary and then is served, leaving the system when service is completed (blocked calls).
- Poisson model
  - The call is served immediately and departs after its service is completed.
  - The call first enters a queue, waits for some time and departs unserved.
  - The call enters a queue, waits for some time and a server becomes available, at which time it starts its service. It departs on completion of its service.

[10] One might also note that there were no such considerations in the Erlang B model, as the state space there is finite, so there are no convergence issues.

Although it is clear that unserved calls should be counted as "blocked calls," what about those that had to wait but eventually were served? There are rational arguments that one can make for counting those calls with either the "blocked" ones or the complement, that is, the nonblocked ones. Which of these arguments may be more persuasive is irrelevant! The Poisson model allows us to (relatively easily) find the probability that the system is full and anything above and beyond that can get rather difficult. As such, the blocking probability refers to the probability that an arbitrary arrival finds the system full, thus has to enter the queue. What fraction of these "blocked" calls eventually is served is not part of the modeling question under consideration! Thus one ought to be careful, when interpreting data, that the statement that the blocking probability is 5% in fact implies that 95% of the calls were served immediately on their arrival and some fraction of the 5% were also served, but we do not know how big of small that fraction may be.

Although the term "blocking probability" has a different interpretation in each of the three models, we often find it convenient to compare the results. See Table 2 for such a comparison. Indeed, it is easy to see that the intuitive placement of the Poisson model as "in between"; the Erlang B Erlang C differentiation makes sense. From a practical perspective, we can make an even stronger statement. In practice, it is very unlikely that either extreme occurs. Namely, it is not likely that:

- Arrivals wait for no time at all. In fact, some systems provide for explicit queues, but even if no such queueing capability is explicitly provided for, customers tend to "redial," or try again in some way, thus creating somewhat of a virtual queue.[11]
- Arrivals that find the system full, "wait forever," that is, an unbounded amount of time.

Indeed, assuming some waiting time is a reasonable modeling assumption. Assuming whether the specific distribution of the Poisson model is the right one or not is almost irrelevant. If for no other reason, it is rarely the case that we have a high confidence in our assumptions as to customer behavior. Thus it is unlikely that our model will be an accurate one to start with.

Given the points outlined above, what is one to do? Very simply, if one looks at these three models as simply three data points in a continuum of possibilities, one can get a lot of insight into the problem. Indeed, in an intuitive sense, system performance is a continuous function of the waiting-time distribution. Thus by making certain assumptions about the waiting time distribution, one can check the predicted performance versus what one

[11] It is important to note that unlike in Erlang C, this virtual queue is not served in a FIFO manner. The most appropriate model for this type of situation is *random service order* (RSO); the Erlang C delay distribution needs to be modified to account for this. However, as it turns out, the probability of the system being full is the same, so we can use the Erlang C results for the blocking probability, even on this case where the service discipline is not FIFO.

is measuring in a real situation and use that information in a feedback loop to better calibrate the model of the problem at hand.

Thus it is often unreasonable to ask "Which model is the best?" More often than not, it is the case that we need to understand all these models and see which one best describes the underlying dynamics of the problem at hand.

4. OTHER RELATED TOPICS

A number of problems and applications emerge from the fundamental concepts presented thus far. They are far reaching and their fair treatment is not possible in this article. However, we will briefly discuss some of them.

4.1. Random-Access Protocols

In "conventional" traffic engineering, it is assumed that the need for a channel is made known to the system by some other means. For example, in traditional telephony, when one picks up the telephone receiver, a dial tone is allocated, which in effect activates the existing line between the telephone and the central office. The traffic engineering problem in its core concentrates on whether a line between your central office and that of the called party exists. The lines to/from one's telephone are "always" there, waiting to be activated by the dial tone or by the incoming call. The need for a channel between points A and B is known to the system by the dialed digits. In many other applications, however, the need for a channel is not known by any such external mechanism. A good example is found in the exploding area of mobile communications, even though the problem had been identified much earlier within the domain of computer networks. The terms used might be different but the underlying concepts are identical. For instance, in Fig. 12, a "central entity," labeled node A, is "in charge" of the system. It may be a base station in a mobile communications network or a host computer in a data network. It allocates idle channels to entities that request one. These entities may be mobile or peripheral devices, such as terminals, printers, or card readers — this does go back to the days of punchcard readers! The problem is very simple to state: how does the base station (node A) become aware of the need for a channel by the mobile units?[12] Note that any "central control" is precluded as they are impractical and are made possible only with unacceptably high overhead. Thus we are forced into what are called *random-access protocols*. Each mobile transmits whenever it feels like! As a moment's reflection makes clear, such transmissions are subject to collisions, as they must use the same mechanism for accessing the base station. These protocols were first analyzed in 1970 by N. Abramson of the University of Hawaii, and the simplest of a family of such protocols is called ALOHA, appropriately enough. The interested reader is referred to Rom and Sidi's study [9] for more detailed analysis and discussion of these problems. Here we present some of the most basic ideas.

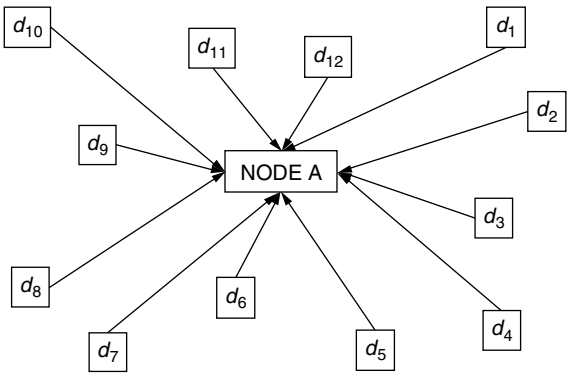


Figure 12. Abstract representation of a base station in a mobile communications network.

In the ALOHA protocol, each mobile would, quite literally, transmit whenever it feels like transmitting. The base station is assumed to receive the information successfully if and only if there is only one ongoing transmission. Any overlap of two (or more) transmissions renders all of them worthless, as even partially received data are ultimately discarded as being corrupted. The fundamental question is then *not* how many channels are available, but what is the throughput of such a system. Throughput is defined as the fraction of time that the channel is used successfully. Idle time and time spent in a colliding state count against throughput. As one can readily conjecture, the throughput of such systems is not great. If the offered load is low, most time is spent in the idle mode. As traffic increases, more collisions occur, thus more of the time is spent in the colliding state, leaving a small fraction for success! Thus the graph of Fig. 13, showing the throughput as a function of offered load, makes intuitive sense. We will return to this graph momentarily.

However, quantifying this intuitively obvious result is not trivial. We will only make some key observations here. The concept of "load" must be defined differently for this class of traffic-engineering-related problems. It should be obvious that holding time is irrelevant. What matters is the number of attempts being made per unit time. It has become common practice in these problems to normalize

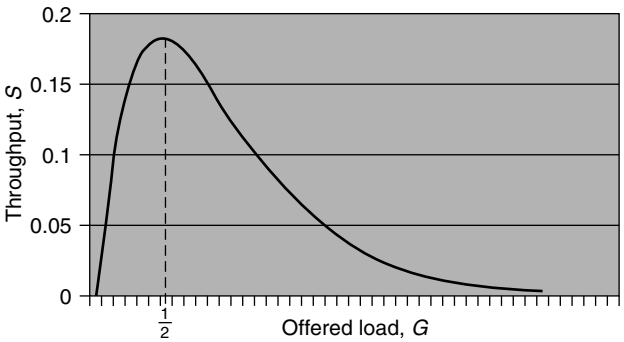


Figure 13. Throughput, S, versus offered load, G, in the ALOHA system.

[12] We will word the rest of this discussion within the framework of a hypothetical mobile communications network, as this simplifies the discussion and reflects the growing area of mobile communications.

the unit time to the average length of the transmission. Owing to the birthplace of these problems in packet switching, this is often called "packet time." Thus one talks about arrivals per packet time. Further, since the transmission is often "a packet" the arrival rate is often referred to as G packets per packet time, or G arrivals per packet time. Clearly, if the world were perfect and ultimate control and coordination were possible, the maximum value of G would be $G = 1$. In such a hypothetical case, the coordinating process would ensure that there were no collisions and that time was constantly utilized for the successful transmission of arriving packets, resulting in a throughput, S, having the ideal value of $S = 1$. Clearly this is not the case, as we realistically have to account for collisions.

Whatever the arrival rate is, collisions will occur. Furthermore, it is common to assume that arrivals form a Poisson process with rate G arrivals per packet time. What happens to colliding arrivals? Once again, the usual models assume that such arrivals are retransmitted at some future instance in time. Note that this creates an instability in the system. Colliding attempts appear at a later time, increasing the probability of future collisions, thus the need for additional transmissions, thus higher probability of collisions, and so on. This is a classic example of positive feedback, thus unstable behavior. If a system reaches such a state, then from a practical perspective it has simply collapsed. Naturally, such catastrophic events must be prevented from occurring. We will return to this subject momentarily.

Under a number of technical conditions, one can show that the throughput is expressed as a function of the offered load by

$$S = Ge^{-2G}$$

This expression is graphed in Fig. 13. However, this simple relationship can be misleading, so we need to clarify a few points:

1. As one can see by elementary calculus, it appears that the maximum throughput is $1/(2e)$, or about 18% and it occurs when $G = 0.5$ arrivals/packet time.
2. However, as G approaches 0.5 from below, the delay, D, associated with retransmissions that have been necessitated approaches infinity. The exact relationship of delay to the offered load G is too complex, requires a number of assumptions and is beyond the scope here. What is important to recognize is that regardless of the specific details, the delay does tend to infinity. Thus from a realistic perspective, one needs to back away to values of G below 0.5.
3. The region of the graph for which $G > 0.5$ is somewhat misleading. If G exceeds 0.5, the system is unstable. If G were to be larger than 0.5, it would drift to infinity as a result of retransmissions, the collisions would dominate, and the throughput would tend to 0, that is, the system would collapse. As a result, authors often refer to the region of $G > 0.5$ as being an "unstable region."
4. While this is correct, it often is interpreted as meaning that the region $G < 0.5$ as being "stable." Unfortunately, this is not the case, in the strict sense of the word "stable." Once again, under a broad set of technical assumptions, sooner or later any system will go unstable, for any value of $G > 0$.[13] However, the amount of time it will take the system to "exit" this so-called "stable region" and enter the doomed instabilities increases rapidly with the distance $0.5 - G$ [for G in the range (0,0.5)], that is, as we back away from $G = 0.5$.[14]
5. One is then entitled to pose the reasonable question of what is the highest value that G should be allowed to be. While this question has a number of interesting theoretical components [e.g., 9], no firm practical answers are agreed to by the practitioners of the trade! Apparently, it is more of a question of how much risk and what kind of safety valves have been put in place than anything else. Conservative designers want to keep G below about 0.1. More risk taking (with appropriate safety mechanisms in place, one would assume) would allow one to be at about $G = 0.2$. However, these values are debatable, subject to a number of conditions, thus should be treated as a rule of thumb rather than precise guidelines! Finally, one sees that at $G = 0.2$, $S = 0.13$, thus one often also talks about the throughout S as being no more than some value in the range of 0.1–0.2.

In conclusion, we see that the throughout of such channels operating in the ALOHA mode is quite small. One way to improve it is slotted ALOHA. In such systems, time is organized into "slots," as shown in Fig. 14. Any mobile station that at time $t = t_0$ decides that it wants to transmit, must wait until the beginning of the next slot. Clearly, collisions are still possible, though less likely, as the "vulnerability period" is cut in half, from two packet transmission times to one packet transmission time. As a result, the offered load–throughput relationship is changed to

$$S = Ge^{-G}$$

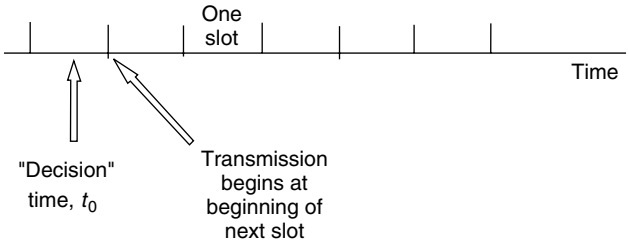


Figure 14. "Slotting" of time for Slotted ALOHA.

[13] Mathematically, the point $G = 0$ is stable, but from a practical perspective this is a useless piece of information. It simply states that a system which has no traffic is stable!

[14] Under certain conditions, this time increases exponentially with the amount by which G is less than 0.5.

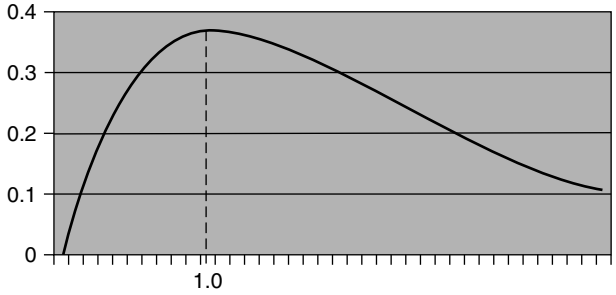


Figure 15. Throughput, S, versus offered load, G, in the Slotted ALOHA system.

This is plotted in Fig. 15. As we can see, this is very reminiscent of the behavior of the ALOHA system, except that the maximum occurs when $G = 1$ and it is $1/e$ or about 36%. While this is double of what it is in the ALOHA case, the entire discussion of the ALOHA system issues, relating to stability, delays, and other factors carries over here as well. Once again, there is no uniform agreement on what acceptable values of G (and thus of S) are, but it is certain that they are higher than in the ALOHA case. This increase in utilization comes at the expense of needing to maintain synchronization information. Whereas in some systems synchronization is necessary because of a number of other independent issues (and thus the incremental cost is virtually zero), in other systems, this additional requirement of synchronization may be prohibitive and the increased throughput may not justify the additional complexity and cost. Indeed, these are some of the problems that both theoretical and applied research addresses in specific situations.

4.2. Network Design

The concepts presented here form the basis for analyzing more realistic systems, such as large networks, consisting of nodes and links joining these nodes. These links have varying capacities, and calls can be routed trough such networks according to a number of algorithms. While space and scope limitations do not allow us to diverge into these topics, the interested reader will find the topic of network design discussed in a number of references [e.g., 3,6,7,10]. In particular, Ash has discussed the problems related to routing algorithms in circuit-switched networks in detail [3]. Similarly, Bertsekas and Gallager have discussed the problem of routing and capacity assignments in packet-switched networks [10].

In a network environment, where routing decisions are influenced by the state of the system, what might originally be Poisson traffic can be offered to various resources as non-Poisson traffic. The traffic engineering problem then increases in complexity and a number of approximate techniques can be used. Both from the theoretical and the practical perspective, these problems become quite complex and there are excellent treatments of these problems in the literature [3–5,7,8].

5. CONCLUSION

In summary, we note that traffic engineering involves a fair amount of mathematical detail that, if not accounted for, may lead to significant errors in the analysis and design of networks. Allocation of resources thus becomes a multidisciplinary problem involving many aspects of engineering, economics, and social behavior, to mention only a few.

Although the problem is quite easy to state in its most elementary form, the solutions to the real problems in a networked environment become quite more complex. Furthermore, the random-access problem, albeit originally a traffic engineering problem, presents us with quite a different set of challenges and ways of looking at a problem.

BIOGRAPHY

Apostolos K. Kakaes received his B.S. and M.S. degrees in applied mathematics with a minor in electrical engineering in 1978 and 1980, respectively, from the University of Colorado. He received a Ph.D. degree in EE from the Polytechnic University in 1987. He had joined AT&T Bell Laboratories in 1980, where he stayed until 1987 working on traffic engineering and network design in both circuit and packet switched networks. In 1987, he joined the faculty of the Department of Electrical Engineering and Computer Science of the George Washington University where his interests in both the physical layer issues and networking problems lead him to work on wireless networks. In 1996, he founded and has since been running Cosmos Communications Consulting, an independent consulting company specializing in mobile communications.

He has published, conducted IEEE tutorials, and lectured extensively around the world in all aspects of both fixed and mobile communications, including traffic engineering issues in current as well as emerging high-speed fixed and mobile communications networks.

BIBLIOGRAPHY

1. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, Wiley, New York, 1970.
2. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley, New York, 1971.
3. G. R. Ash, *Dynamic Routing in Telecommunications Networks*, McGraw-Hill, New York, 1997.
4. L. Kleinrock, *Queueing Systems*, Vol. 1, *Theory*, Wiley, New York, 1975.
5. L. Kleinrock, *Queueing Systems*, Vol. 2, Computer applications, Wiley, New York, 1976.
6. H. Akimaru and K. Kawashima, *Teletraffic Theory and Applications*, Springer-Verlag, 1992.
7. A. Girard, *Routing and Dimensioning in Circuit-Switched Networks*, Addison-Wesley, Reading, MA, 1990.
8. T. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*, Springer-Verlag, 1990.
9. R. Rom and M. Sidi, *Multiple Access Protocols; Performance and Analysis*, Springer-Verlag, 1989.
10. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

COMMUNICATIONS FOR INTELLIGENT TRANSPORTATION SYSTEMS

ORESTE ANDRISANO
GIANNI PASOLINI
ROBERTO VERDONE
University of Bologna
DEIS, Italy

MASAO NAKAGAWA
Keio University
Japan

1. INTRODUCTION

Intelligent transportation systems (ITSs) have been investigated for many years in Europe, North America, and Japan, with the aim to provide new technologies capable of improving the safety and efficiency of road transport [1–5]. In this context several technologies and applications have been investigated and demonstrated with reference to onboard communication networks, and short- and long-range communication systems [6].

Going back to the evolution of guidance support, the first efforts date back several decades (from the time of writing) and were essentially aimed at diffusing traffic information by means of road signs and FM broadcasting. An amazingly significant step in the direction of improving transport safety, featured the adoption of a millimeter-wave radar sensor installed in the front grille of vehicles in order to promptly detect obstacles and to warn drivers when they reach an unsafe distance. This technology has been improved and adopted by several car manufacturers that integrated the radar with an onboard cruise control [7]. This solution offers the possibility, for instance, to automatically maintain an optimal following distance behind cars traveling ahead, thus reducing the risk of collisions as well as the burden on the driver during long trips.

However, as soon as the first ITS systems were implemented, it was clear that their effectiveness would be greatly improved by the possibility to establish a bidirectional communication link between the different entities (cars, pedestrians, ITS service providers, etc.) acting in the road transport scenario. In this regard, the communication-based systems can play a fundamental role in the context of ITS, since they can overcome the main limitations of self-sufficient systems (those relying on passive sensors, radars, videocameras, etc.) that are solely based on the unilateral perception of the environment surrounding the vehicle. To provide advanced ITS services, localization systems will also be suitably exploited in conjunction with mobile communications.

Different technologies were investigated in Europe at the beginning of the ITS research activity, during the 1980s: infrared communications, millimeter-wave communications, and mobile radio. The first two appeared to be the most interesting at that time, and were oriented to the implementation of short-range mobile communication networks, whereas since the early 1990s, some projects have developed the concept of using cellular radio for providing ITS services [e.g., 8]; it is to be pointed out that, at that time, only second-generation (2G) systems were available, such as the Global System for Mobile communications (GSM), characterized by a limited ability to provide advanced data services with high quality levels and flexibility.

In the near future third-generation (3G) systems should be available, enabling advanced data services to be provided to mobile users in the 2000-MHz frequency band, in both Japan, where they have been under development, and Europe, where their introduction is planned for the year 2003 [9]; these systems will be able to provide different bit rates [$\leq$2 Mbps (megabits per second)], different levels of quality of service (QoS), and much more flexibility than that offered by 2G networks.

However, the evolution from 2G to 3G systems will not be sharp, and is based on the implementation of some 2.5G solutions, such as the GPRS (General Packet Radio Service) in Europe, providing intermediate bit rates (up to $\sim$100 kbps) in the GSM band (900 and 1800 MHz), based on packet-oriented techniques and flexible operating modes. Therefore, GPRS is thought to be a suitable candidate to offer services to the ITS context, too.

The increasing interest on wireless personal-area networks (WPANs) and the consequent penetration of low-cost portable devices equipped with the emerging Bluetooth technology [10], has suggested the possibility of adopting WPAN devices for the provision of short-range ITS services as well.

In this article, following a brief introduction to the typical ITS services, we first show some of the results of the research carried out in Italy and Japan concerning the role of communications in ITS; then we discuss the suitability of 2.5G and 3G systems and Bluetooth for ITS applications.

2. ITS SERVICES

The ITS services based on communication devices rely, essentially, on three kinds of techniques: *roadside-to-vehicle communication* (RVC), based on the use of an infrastructure network covering the service area, intervehicle communication (IVC), establishing a direct communication link among automobiles, and target-to-vehicle communication (TVC), based on proper active devices mounted on vehicles with the aim of detecting the presence of the unprotected road user (the target, hereafter) through suitable communication interfaces. Hereafter we provide a synthetic overview of the services that can be provided by the systems described above.

Although sometimes at an early stage, RVC based services have already been introduced in Europe, Japan, and the United States; we can mention, for example, the following applications:

- *Automatic tolling*, probably nowadays the communication-based ITS service most widely used by road users; it proved to be amazingly effective in reducing traffic congestion and improving driver convenience by cashless payment. It is based on RVC techniques.
- The *services for guidance support*, providing the user with information related to traffic, weather

conditions, and so on; they are still at a very early stage of development, and in many cases are based on traditional techniques such as road signs or broadcasting through FM stations (RVC techniques).

- The *services for traffic control*, based on the joint concepts of navigation and communication (RVC techniques).

Still based on RVC systems and not yet introduced, the *services for driving safety* should play a very important role in the future, providing the driver with real-time information concerning possible emergency situations such as those due to fog (banks), traffic accidents, or road hazards. These services are the most difficult to provide because of the stringent requirements of the application; this is the case, for instance, of the "emergency warning" service, which should be based on the possibility to inform all the vehicles in the vicinity of a dangerous situation within a short amount of time from its occurrence, and this requires a prompt system response. Consequently the introduction of such services is still seen as a long-term goal; nevertheless, we show the suitability of GPRS for this application, even if with limitations on the system response promptness.

A significant step in the evolution of ITS should be the introduction of the *cooperative driving* service, which belongs to the class of *services for driving safety* and is based on IVC systems. It consists in forming groups (platoons) of vehicles exchanging data on their status (speed, acceleration, position, etc.) in order to improve the safety and efficiency of the vehicles flow by keeping the intervehicle distances under control in all situations (sudden change of speed, position, etc.). The data exchanged among the cooperating vehicles can be processed by automatic agents and used to control the onboard actuators (brakes, etc.), or presented to the driver through suitable (e.g., vocal) interfaces. Each vehicle of the platoon must be equipped with a communication device, and direct intervehicle communication (IVC) is the only way to provide the requested promptness.

TVC-based services are finally a challenging issue for researchers since the communication link involves pedestrians, cyclists, and others whose communication devices have to be lightweight, small, low-power-consuming, and, possibly, cheap; by means of TVC systems it will be possible to prevent car accidents involving pedestrians and cyclists by means of onboard warning devices. In Europe PROTECTOR, an ITS project, developed research in this area.

3. WIRELESS STANDARDS FOR ITS: DEDICATED SOLUTIONS

3.1. Past Activities

Since the early 1980s remarkable research activity has been carried out in the United States within the context of ITS. One of the most important actors in this scenario is the Californian PATH [11], a joint venture of the University of California, the California Department of Transportation (CALTRANS), and private industry, established in 1986 to develop more efficient transit and highway systems. The goal of PATH is to increase the capacity of the busiest highways and to decrease traffic congestion, air pollution, accident rates, and fuel consumption and to perform field operational tests.

PATH participation in U.S. DoT (Federal Department of Transportation) ITS programs includes several projects within the Intelligent Vehicle Initiative (IVI) program [12], developing research in the fields of

- Sensor-friendly vehicle and roadway systems
- Forward-collision warning systems
- Rear-collision warning systems
- Automotive collision avoidance systems (ACASs)

In Europe and Japan some major research programs (e.g. DRIVE, PROMETHEUS, VASCO) in Europe, AHS and JSK programs in Japan [13–23] have been involved in the definition and testing of telecommunication systems for the field of road transportation, such as systems for automatic tolling, fleet management, traffic control, and cooperative driving; most of them were based on dedicated solutions, explicitly designed for ITS applications. Suitable frequency allocations have been proposed in order to avoid more congested bands. On the other hand, in some cases this represents an obstacle to the introduction of these systems, due to the initial high costs of dedicated solutions when entering the mass market.

For both IVC and the RVC systems, the 60–64-GHz band was chosen among the possible candidates within the DRIVE and PROMETHEUS projects in Europe for reasons related to the size of RF devices, the amount of available bandwidth, and the advantages provided by oxygen absorption, which causes spatial filtering and frequent channel reuse, suitable for short-range systems such as those based on IVC.

The results of all these programs were based on dedicated radio interfaces; we leave the interested reader to refer to the literature [13–23]. In the following we simply provide, as a synthetic view, an example of the results obtained within the TELCO project (1995–1997), which was funded in Italy by Consiglio Nazionale delle Ricerche, after the conclusion of PROMETHEUS with the aim of investigating on the possible use of the millimeter-wave band, and we mention some of the experimental test beds realized in Japan.

3.1.1. RVC at 60–64 GHz: Research in Italy. The 60–64-GHz band is characterized by a peak of oxygen absorption. To highlight the main features of the use of this band, let's consider a simple monodimensional scenario, in urban context, with beacons serving the area and separated by a distance $2R$ [18]. The multiple-access method employed is TDMA. Let's denote by N and Br the channel reuse factor and the overall bit rate, respectively. Figure 1 shows the outage probability, for a given transmission system, as a function of the cell radius with the reuse factor and the bit rate as parameters. The *outage probability* is defined as the probability that the signal-to-noise and the signal-to-interference ratios are

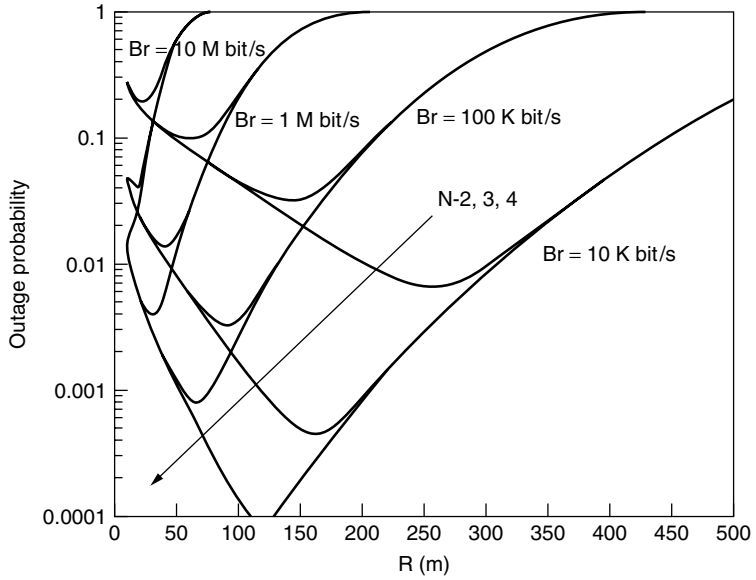


Figure 1. The outage probability as a function of the cell radius, with the reuse factor and the bit rate as parameters (© IEEE, 2000).

below specified thresholds that are defined as functions of the transmission techniques implemented. A noncoherent MSK unprotected system is considered here, as a simple reference, and omnidirectional antennas are assumed. The figure shows that a minimum outage probability is obtained for a given value of R, as a compromise between the effects of noise and interference. This is typical of the 60–64-GHz band, due to the effects of oxygen absorption: no optimum value of R would be found in a frequency band unaffected by the additional attenuation due to oxygen, or rain. As can be expected the optimum outage probability decreases when N increases or Br decreases. It is also interesting to note that, with the system parameters chosen, the values of R are $\sim$100–250 ms, thus giving a large number of beacons per kilometer and suggesting the implementation of these systems in urban areas where the number of users can be large. Suitable scaling of the bit rates and of the optimum cell size can be obtained by employing more sophisticated transmission techniques, directional antennas, or other means. In any case, the typical cell size remains below 1 km.

3.1.2. IVC at 60–64 GHz: Research in Italy. The advantages of using the 60-GHz band for IVC have been shown in several studies [13,15,17,19,20], and the results of these studies provided the design of suitable multiple-access techniques for cooperative driving applications. We let the interested reader refer to these papers.

3.1.3. IVC Experiments in Japan. In Japan, JSK carried out some IVC experiments based on infrared technology in 1996, and some new experiments were performed during the year 2000 at 5.8 GHz. The AHS project is now extending its research studies from RVC to road sensing; the RVC frequency should be more than 5 GHz due to the lack of frequency bands at lower frequencies. Road sensing uses infrared, 60-, and 90-GHz technologies.

Infrared technology was initially selected among all candidates for IVC, for economical and practical reasons, in spite of its weather-dependent characteristics. JSK carried out an experiment in 1996 to verify the possibility of networking between traveling vehicles, as a preliminary testbed in order to plan a new experiment based on microwave technology [16].

Figure 2 shows the experimental vehicles used, equipped with two infrared radiators separated by a 1-m distance on the roofs, a photodetector, a videocamera, and a processor. The distance between forward and backward vehicles was measured by the videocamera, which could record the images of the two radiators on its CCD film. These radiators not only enabled distance measurement but also sent information to the vehicle behind. The test demonstrated the feasibility of IVC, while pointing out that some features, such as network aspects, deserve thorough consideration [16].

3.1.4. CDMA or TDMA for IVC: Research in Japan. The controversy about CDMA and TDMA in cellular communications in the early 1990s concluded CDMA to be preferred to TDMA for channel capacity reasons and its ability to counteract interference. What about both access methods in IVC? Few research works about this matter have been published. Michael and Nakagawa [21] show that analysis of the interference from oncoming vehicles reveals CDMA to be better than TDMA. Since the main information from each vehicle is related to control data of the vehicle, it is characterized by regular occurrence. Consequently each vehicle should transmit periodic data bursts to other vehicles. The interval of the data bursts should be the same for all vehicles. However, if the data bursts of an

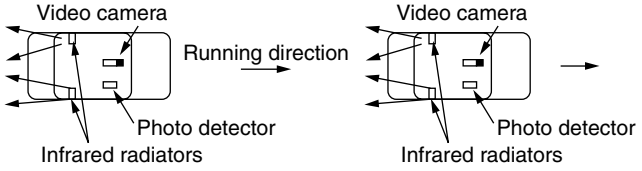


Figure 2. Top view of vehicles in the infrared experiment performed by JSK in Japan in 1996 (© IEEE, 2000).

oncoming vehicle group collide with those of the ongoing vehicle group, the collision is kept for a while and a large amount of data are lost if a TDMA scheme is used. CDMA can save the data from this type of collision; this is the main reason for the selection of CDMA for applications based on IVC. On the other hand, if different carriers were used for communications in the two directions, the problem discussed above would vanish, but this would require a centralized allocation of carriers to the vehicles.

3.1.5. Spread-Spectrum Radar and FM-CW Radar: Research in Japan. FM-CW radar was standardized for a radar system to be mounted on vehicles to detect the presence of other vehicles. When there are a number of vehicles on a road, interference between the radar signals from all the vehicles should be considered. Shiraki et al. [22] show simulation results for the FM-CW radar and spread-spectrum (SS) radar. The SS radar shows much better performance than FM-CW radar as reported in Fig. 3, in terms of signal-to-interference ratio (SIR).

This result also suggests the investigation of the possible integration of radar and communication systems, based on a common frequency band (e.g., the millimeter-wave band) and processing (e.g., SS) techniques.

3.2. Current Activities

A brief overview of the current ITS applications and testbeds is presented in order to explain to the reader which is the state of the art in this field [7].

3.2.1. Services for Guidance Support. The next generation of communication technology for guidance support is being developed behind the steering wheel and, more recently, on the wireless handset. The emerging trend in this field is to offer location-based voice and data communication tools that provide "smart" information, tailored to where the customers are and to what they are doing,

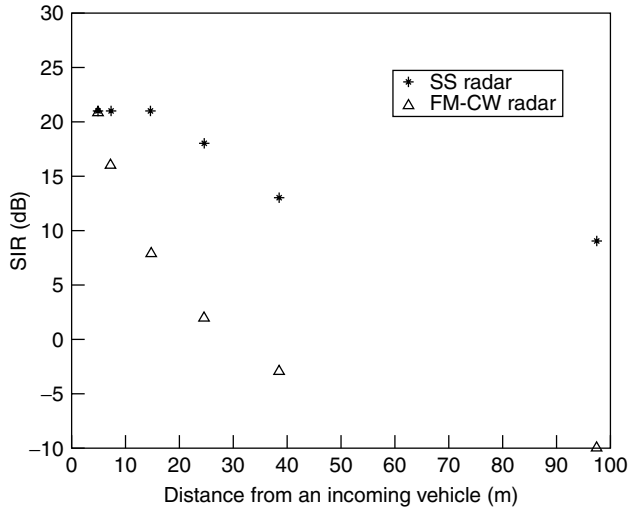


Figure 3. The ratio between the useful received power from the target and the radar interference power received from an oncoming vehicle at the same distance (© IEEE, 2000).

providing enhanced security, navigation, and convenience to mobile consumers.

The adoption of wireless terminals in this currently deeply investigated context is proving to be extremely effective, greatly improving the usefulness of ITS services; in the greater metropolitan Paris area, on the Boulevard périphérique and the Paris city roads, for instance, the time needed to travel the distance between two points is displayed by variable message signs (VMSs). Thanks to portable terminals, this information is available in the car (the terminal can be easily installed) and to people who are not on the road (they can take the terminal along). In this manner route planning is improved by conveying information on travel times before the journey commences.

3.2.2. Automatic Tolling System. The well-known system for Automatic Tolling of the Italian Autostrade S.p.A., namely, Telepass, is the first system in the world for toll collection where drivers do not need to stop at toll stations [7].

3.2.3. Services for Traffic Control. Congestion is steadily worsening on the streets of Europe's cities as more people choose to travel by car. Buses, trams and trains take less space, thus increasing their share of trips can help reduce congestion. Unfortunately, many public transport services get caught in traffic and are reputed for poor reliability. Public transport performance can be improved by better controlling and managing traffic generally. The City of Turin, as an example, has implemented a system — named 5T — that integrates nine ITS subsystems under the co-ordination of a tenth system, the "traffic and transport supervisor," which monitors and controls all the other subsystems. Among the 5T subsystems, public transport vehicle location and urban traffic control subsystems provide priority to public transport vehicles at traffic signals, particularly to those that are running late. Similar priority can be given to emergency service vehicles [7].

3.2.4. Services for Driving Safety. In the case of an accident on the motorway, significant danger is caused by vehicles that block the road (cars involved in the accident, emergency services, a tailback or bottleneck of cars that cannot pass the location). Secondary accidents, which are often more serious than the original one, can be avoided if the accident is detected immediately and the scene of the accident is cordoned off right away. In Germany, an advanced warning system that automatically detects incidents and provides immediate warning to drivers has been proposed. Beacons equipped with a light unit and additional electronics are installed along the emergency lanes.

The beacons are interconnected by cable and linked to a traffic computer center, where an operator views the accident on a computer screen, via an appropriate monitoring interface. The light units can be activated from the control centre, to flash warnings until the police arrive to cordon off the area. The light flashes, initiated within seconds of the accident, warn approaching traffic of problems downstream, and invite the drivers to slow down.

3.2.5. Dedicated Short-Range Communications. Dedicated short-range communications (DSRC) systems aim at providing one-way or two-way high-speed radio links between a fixed roadside unit and onboard equipment: ITS-related information is exchanged within the communication area, made up of the roadside antenna, therefore configuring a RVC system. DSRC systems have been deeply studied in the United States, Europe, Korea, and Japan, and several proposals have been presented so far [24–27]. The various solutions adopt different modulation formats as well as different data rates, ranging from hundreds of kilobits per second to a few megabits per second, and operate in different bands within the microwave region.

4. WIRELESS STANDARDS AND ITS

The abovementioned considerations on the difficult penetration of ITS dedicated systems in the mass market suggested that researchers investigate the possibility of providing ITS services by means of already standardized systems such as the 2.5G [28] and 3G cellular systems as well as Bluetooth [29]. The capability of these standards to offer packet-switched services will be studied in order to show how they can be exploited for the provision of services for driving safety, such as "emergency warning," and for traffic control and guidance support as well.

4.1. A GPRS-Based System for Emergency Warning, Traffic Control, and Guidance Support

GPRS supports packet switched services, based on a radio interface almost identical to that of GSM, at different rates, roughly 9–100 kbps, which can be obtained by assigning multiple GSM slots per frame to the same user. On the other hand, if the user bit rate is very low, GPRS can allocate the radio resource depending on real needs, thus preventing the inefficient use of the radio spectrum and allowing the user to pay for the amount of data actually transferred.

Let's consider a typical European highway trunk, consisting of three plus three lanes, as a reference scenario to describe the system investigated (see Fig. 4). If we assume that the whole trunk is covered by the public GPRS network, we can exploit its capability to support packet-switched services and priority-based scheduling [30].

To provide a prompt "emergency warning" (EW) service, in fact, each vehicle has to establish a connection with the network at the entrance of the trunk, in order to reserve some radio resource units to be used when needed (e.g., if a dangerous situation is monitored by the onboard computer, which also controls airbag deployment, sudden braking, etc); this event is considered to be a very infrequent but significant situation. The connection is relinquished at the output of the trunk. The packet transmission mode controlled by the network should also rely on a suitable scheduling procedure, taking the priority of services into account. In this manner, once the different vehicles have allocated their radio resource units, the system could also support other services characterized by lower priority.

Therefore, we assume that each vehicle has allocated a minimum radio resource every T_{cycle} milliseconds; the minimum radio resource should be given by a RLC (radio-link control) block, which in the GPRS is transmitted over four GSM bursts; with this choice, the packet size (at network level) is around a few hundreds of bits, but this is enough to provide an EW message.

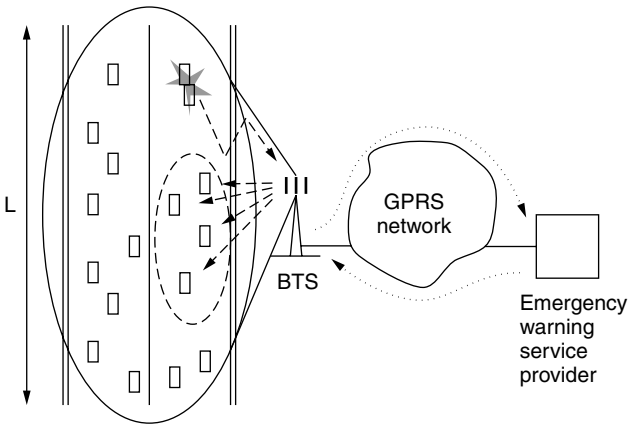


Figure 4. An ITS scenario based on GPRS: the service provider receives the Warning message from a mobile, and sends it to all the vehicles that can be involved in the information, considering the location, speed, and direction of the mobiles, the traffic conditions and the requirements of the application (© IEEE, 2000).

When a packet is generated by a mobile, it is transmitted through the GPRS network to a fixed node (representing the serving entity introduced by the service provider) that retransmits the message in multicast mode to all the vehicles in the vicinity of the source (see Fig. 4). The subset of vehicles to be addressed by the warning message has to be carefully determined taking different aspects into account, such as the speed of mobiles, the traveling direction of the source (it can be assumed that only the vehicles behind that involved in the dangerous event should be warned), and the traffic density; all these data can be known at the service provider if all vehicles periodically inform the network with their locations, speed, and other specifics. Therefore, communication, location and traffic management aspects have to be considered at the centralized node.

The implementation of a centralized control on the network side has the main advantage that the message generated on the vehicle has a fixed destination address; therefore, the determination of the subset of vehicles that should receive the warning message is controlled by the service provider, and this can be efficiently performed if all vehicles periodically inform the network of their locations. On the other hand, this choice implies a double connection, and double transfer delay, so special attention should be paid to this assessment.

The time spent in the network by the packet is subject to the delays introduced by the network entities and the time needed for radio access. The choice of T_{cycle} is of relevance, since it is a measure of the maximum delay that the message can suffer due to the radio access: in any case, the value of T_{cycle} should be small, in order to avoid a large contribution to the overall message delivery delay. On the other hand, the number of vehicles that can be served is proportional to T_{cycle}; hence, T_{cycle} should be

fixed as a tradeoff between the network capacity and the desired promptness.

It is worth noting, however, that the number of vehicles that can be served is also proportional to the number m of GPRS carriers used for the service, hence, having fixed T_{cycle}, the service coverage can be improved increasing the value of m.

Let's assume, in the six lanes of our scenario, an average intervehicle distance, d_{iv}; consequently the number of vehicles present in a trunk of length L served by the same BTS, is given by $N_{\text{veh}} = 6L/d_{\text{iv}}$; let's assume, furthermore, that the percentage of vehicles equipped with the GPRS terminal is 100 k, where k is the penetration factor.

Figure 5 shows the maximum value of the cell size L that is compatible with a traffic of kN_{veh} vehicles as a function of the penetration k; both cases of one and two GPRS carrier(s) are considered, and the intervehicle distance is fixed at 25 m.

Figure 5 shows that, for $m = 1$, at the first stage of service provision (when the penetration is low), cell sizes of around tens of kilometers; that is, the typical cell sizes of the GSM network in suburban environments, can be used, thus allowing the utilization of the existing sites. On the other hand, once the penetration becomes much larger (e.g., $k = 0.25$), new investments should be introduced, either by adding new GPRS carriers or by reducing the cell sizes; however, this would be the case of a very mature service used by millions of users, such as by a successful business.

Finally, it is worth noting that, when no packets with highest priority (i.e., EW packets) have to be transmitted, the radio resources can be dynamically reallocated to other services; therefore, the GPRS carrier(s) is (are) not dedicated to the EW service, since it (they) can be shared by other services having lower priority. For instance, the provision of information concerning traffic and weather conditions (on subscription, or on demand) or the transmission of roadmaps could be offered to the user through the same physical channels. All these services,

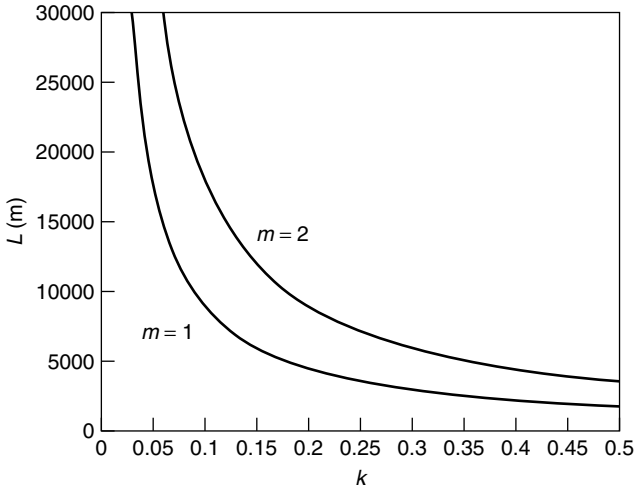


Figure 5. The maximum cell size (having fixed $T_{\text{cycle}} = 500$ ms) as a function of the penetration of users, with $m = 1, 2$ GPRS carriers and for an average intervehicle distance fixed at 25 m (© IEEE, 2000).

together with the selection of vehicles to be included in the multicast group to which the warning message should be delivered, require the knowledge (with different degrees of precision) of the vehicle locations. A mobile terminal can solve the problem of self-localization in a number of ways (through the capabilities of GSM, or through a GPS receiver, etc); in any case, the information concerning the vehicle position should be transmitted periodically to the network and this could be done by means of the allocated radio resource units.

4.2. The Role of 3G Networks

GPRS is expected to become a successful standard in the near future; however, it should only represent an evolutionary path toward more sophisticated systems such as UMTS (Universal Mobile Telecommunication System) [9], the 3G cellular network, which should replace the previous 2G/2.5G systems. It seems reasonable to assume that in many countries the 3G core network will be initially based on, or will also embrace, the GSM/GPRS core network; the main difference compared to the previous standards will be on the air interface, which in 3G systems will rely on CDMA techniques, unlike those for GSM and GPRS. However, the use of multistandard (multimode) terminals will determine a potential integration of the different air interfaces (e.g., GPRS or wideband CDMA). Therefore, the most revolutionary concept introduced by 3G solutions, that is, the possibility of having a full-coverage cellular system based on packet switching, thus allowing the extension of many services based on IP (Internet Protocol) and the adoption of a "pay what you send" billing method to the wireless context, will be introduced with GPRS. In this regard, it should be emphasized that UMTS will stimulate this concept by providing larger bandwidths.

The previous example of potential application of GPRS in the field of ITS shows some limitations; for instance, the limited bandwidth of 2.5G systems affects the promptness of the "warning message" service, as well as the delays introduced by the fixed network.

The portion of spectrum allocated to 3G standards is larger than that dedicated to 2G (and 2.5G) systems; therefore, the provision of EW services through GPRS can be seen as a first step toward a more sophisticated solution, which could be provided by means of 3G networks. In fact, the delays introduced by the GPRS network depend partly on the limited bandwidth of the GPRS system, as shown before. Within this context, the main role that 3G systems will be able to play, with respect to GPRS, will be given by the larger bandwidths that could affect the quality of the services offered.

The larger the overall bit rate, the smaller the radio access delay, the smaller the delay of message delivery, and the higher the spatial resolution of the warning message. With GPRS, taking its network delays into account, it can be expected that, at a speed of 120 km/h, a vehicle can be alerted to avoid an accident that occurred about 150–200 ms ahead. This is not enough to elude all possible events. On the other hand, this could be sufficient to avoid the catastrophic highway "chain accidents" (occurring sometimes in the presence of huge banks of fogs,

typical of different areas in Italy) that in some cases involve many cars, separated by distances much greater than 100 m. A rough EW service based on GPRS could save some human lives. The larger bandwidth of 3G standards could do even better.

However, this does not reduce the importance of the investigation of dedicated systems for ITS, which could provide even more complex services, such as "cooperative driving," which should be considered in the long term to be a very important target for increasing the safety and efficiency of road transport.

4.3. The Role of Bluetooth

The main disadvantage of the adoption of cellular communication (GSM, GPRS, or UMTS) for ITS service provision relies on the fact that direct communication between the interacting entities would not be possible; the exchange of data between the two would be performed via the cellular network. This would require suitable integration of the cellular service with localization aspects, would make the service provision prone to the network malfunction, and would increase the packet transfer delay.

An alternative solution is represented by the adoption of devices with the capability, on one hand, to automatically detect the presence of other similar devices located nearby and, on the other hand, to establish a direct communication link (hence without the need of an infrastructure network).

The Bluetooth technology, the most recent development in the field of WPANs (wireless personal-area networks), fulfills both the abovementioned requirements; furthermore, Bluetooth-based devices are expected to be lightweight, small, and economic, therefore representing a perfect candidate not only for RVC systems but also for TVC systems.

The aim of the Bluetooth technology is to allow the establishment of effortless, wireless, instant, and low-cost connections between various communication devices. The Bluetooth radio is built into a small microchip, which is estimated to cost just a few dollars, and operates in a globally available frequency band [the ISM (industrial–scientific–medical)-band at 2.4 GHz] thus ensuring communication compatibility worldwide.

Two or more units communicating one with the other(s) form a *piconet*, where one unit acts as a "master," controlling the whole traffic in the piconet and the other units act as "slaves."

The master implements a centralized control; only communications between the master and one or more slaves are possible, and there is a strict alternation between master and slave transmissions.

A simple binary, Gaussian-shaped FSK modulation scheme (GFSK) is applied in order to reduce costs and device complexity and a symbol rate of 1 Mbps can be achieved. Three different power classes are defined within Bluetooth, as reported in Table 1, where P_0 is the maximum transmitted power.

4.3.1. Estimation of the Maximum Range. An important aspect to be investigated is the maximum distance between the transmitting and receiving devices that ensures possibility of communicating.

Table 1. Bluetooth-Defined Power Classes

Power Class	Maximum Transmitter Power
1	$P_0 = 100$ mW
2	$P_0 = 2.5$ mW
3	$P_0 = 1$ mW

A precise assessment of the maximum range allowed, by the Bluetooth technology under nonideal conditions requires an experimental testbed; hence we carried out a measurement campaign with the aim to evaluate the performance of a Bluetooth link in an outdoor scenario and in different propagation conditions.

A data communication link between two commercial class 1 Bluetooth devices separated by a distance d was established and the time needed to perform a 1.57-MB (megabyte) FTP (File Transfer Protocol) transmission measured. The scope of this test was to evaluate the limit distance that determines a significant increase in the file transfer delay (T_d) due to the retransmissions requested by the ARQ strategy adopted by Bluetooth; this experiment was performed considering both a "free-space-like" [i.e., Line-of-Sight (LoS)] and a partially obstructed LoS environment.

In the first case we ensured that the volume enclosed in the ellipsoid defined by the first Fresnel zone was free from obstacles; in the second case the electromagnetic propagation was partially obstructed by the presence of the terrain (we considered the transmitter and the receiver placed at a height of 45 cm from the terrain).

In Fig. 6 the file transfer delay, T_d, is reported as a function of the distance d between the two Bluetooth devices in the case of "free-space-like" propagation conditions; as we can observe, T_d is not affected by the value of d until the distance between the two devices becomes larger than about 110 m, after which T_d rapidly increases. Ten transfers were performed.

Consequently, under the conditions considered, the maximum communication range can be estimated to be

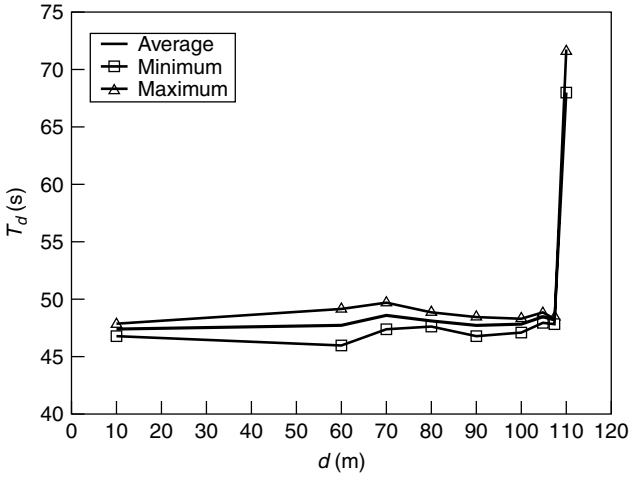


Figure 6. File transfer delay between two Bluetooth devices: "free-space-like" condition.

around 100 m for RVC systems, which are typically experiencing LOS conditions, owing to the possibility to place the beacon in suitable positions.

The experiment described above was repeated assuming partially obstructed LoS conditions: the transmitter and the receiver were placed at distances 45 cm above the terrain, made up of an asphalted surface, representing an obstruction for the first Fresnel ellipsoid when the distance d is >7 ms. This scenario could be representative of an IVC system for, example.

In Fig. 7 the result of this measurement campaign is shown. As we can observe, when the distance between the transmitting and receiving devices is lower than 60 m, the file transfer delay is not affected by the presence of the terrain, which later determines a rapid increase of T_d. Consequently, under the conditions considered, the threshold distance value that marks the transition between LoS and non-LoS conditions is significantly larger than the theoretical value of 7 ms; this observation leads to the conclusion that the reflections produced by the road can improve communication by extending the coverage distance with respect to what is predicted by simple modeling.

Analytic estimation of the maximum range, as in every wireless system, requires the assessment of a suitable propagation law, which is complex because of the presence of many vehicles and other obstacles, thus requiring the specification of the operating environment.

In this section we merely give an approximate evaluation of the maximum range, taking only free-space loss into consideration.

The minimum level of the received power, on the input to the receiver, required for a bit error rate equal to 10^{-3} is $P_r = -70$ dBm [31]; therefore, the maximum loss $L_{\max}$ (dB) $= P_0$ (dB) $- P_r$ (dB) depends on the power class and is reported in the second column of Table 2.

Knowing the $L_{\max}$ value, we can assess the maximum allowable distance $d_{\max}$; let's assume free-space loss (optimistic conditions), so that, assuming the value of carrier frequency equal to 2.4 GHz, we obtain

$$L_{\max} = 40 + 20\log(d_{\max}) - G_e - G_r$$

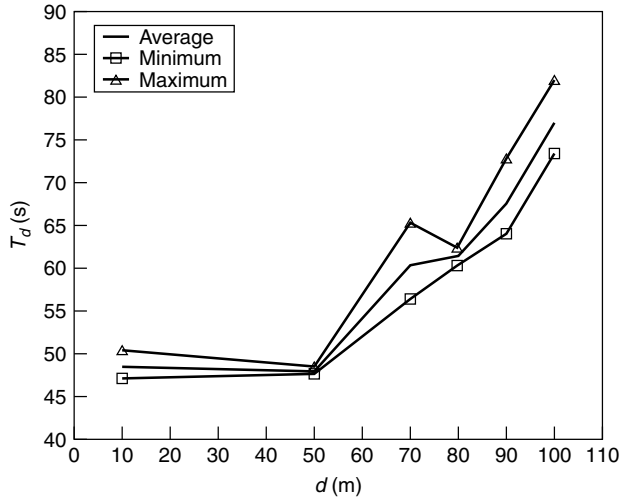


Figure 7. File transfer delay between two Bluetooth devices: "non-LoS" condition.

Table 2. Minimum Received Power Required for BER = 10^{-3}

Power Class	$L_{\max}$ (dB)	$d_{\max}$ (m) ($G_e = G_r = 0$ dB)
1	90	316
2	74	50
3	70	32

where G_e and G_r are the transmitter and receiver antenna gains, respectively.

Under the simple assumption of omnidirectional antennae ($G_e = G_r = 0$ dB), with free-space loss, we have

$$L = 40 + 20\log(d_{\max})$$

which gives the values reported in the third column of Table 2.

The coverage range calculated with our very simplified approach, which doesn't take nonideality typical of a real system into account, is in the case of class 1 devices, around a few hundreds of meters, not too different from the measured values reported above, which were obtained with commercial products, probably not transmitting at the maximum allowed level of power.

However, free-space loss is a rather optimistic assumption that in many cases could not be fulfilled in realistic conditions, especially in urban scenarios. The assessment of a loss formula taking multipath into account should be considered, but is beyond the scope of this work, as it would depend on the particular environment (urban, suburban), including the traffic conditions and the number of scatterers (cars, buildings, etc.).

4.3.2. Effects of Vehicle Speed and Multipath. We note that the Bluetooth technology was designed for static indoor environments (communication between computers, printers, etc.); the possibility of using it in an outdoor scenario, with moving terminals at different speeds, must be carefully investigated by taking the protocol and physical aspects of Bluetooth technology into account and considering the typical aspects related to the propagation of electromagnetic waves in outdoor urban environments, characterized by fading and multipath effects.

This analysis requires the consideration of the consequences of the Doppler effect (viz., the effects of the speed of terminals) on the modulation/demodulation format and the protocols.

Concerning signal dispersion, the following preliminary considerations lead to the conclusion that the dispersion due to multipath should not represent a serious drawback: Bluetooth is a short-range communication system; hence the coherence bandwidth is expected to be around tens of megahertz, that is, well beyond the signal bandwidth (1 MHz), and this renders the performance insensitive to the amount of signal dispersion. A more precise consideration of this aspect would be very complex because of the characteristics of the Bluetooth technology.

As far as the effects of speed are concerned, an analytic model has been exploited to assess the bit error rate

of a Bluetooth device under the conditions stated above (neglecting the benefits of the channel code). It is worth noting that the modulation format is nonlinear and its performance is evaluated in the presence of flat fading.

In Fig. 8 the BER (for a Bluetooth terminal neglecting the effects of channel coding, which will further improve the performance) as a function of average signal-to-noise ratio W_m, is reported where the average is made with respect to fading statistics, considered to be flat Rayleigh. The computation takes the effect of the speed of terminals into account (viz., the Doppler effect, for a relative speed equal to 110 km/h, considering a vehicle and a target running in opposite directions, one toward the other, i.e., on a "head-on collision course"). Since the implementation of the receiver characteristics is not fixed by the standard a limiter–discriminator device is assumed, as a reference, with a 4-pole Butterworth filter having normalized bandwidth equal to 1. It is expected that many implementations of the Bluetooth technology will be based on limiter–discriminator detection, which is known to be simple and efficient under hostile propagation conditions. The choice related to the filter has proved not to have a deep impact on the results.

The results show that an error floor is found, due to the speed of the terminal, at BER $= 10^{-6}$. We assume that this level of BER is sufficiently large to fulfill the performance requirement in terms of false-alarm rate. Therefore, the movement of terminals does not seem to introduce limiting effects on the performance (let's bear in mind that the effect of channel coding was not considered, and it will further lower the error floor).

In order to verify the results of our analysis, we carried out experimental activities to test the Bluetooth transmission reliability in a dynamic scenario. A communication link was established between a static class 1 Bluetooth transmitter placed on the pavement beside an urban road, at 50 cm above the terrain, and a Bluetooth receiver placed on a vehicle traveling at a constant speed and passing in front of the transmitter at a minimum distance of 3 ms.

The experiment was repeated assuming different speeds of the mobile device, and we ensured that the transmission was performed correctly even when traveling at 100 km/h, therefore confirming our expectation on the sturdiness of Bluetooth communications.

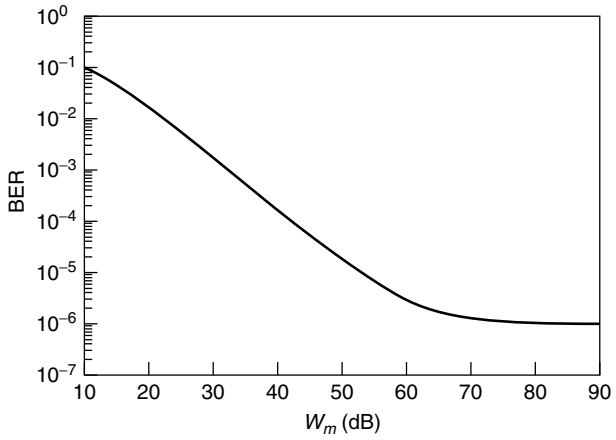


Figure 8. Bluetooth link: the effect of mobility.

As a consequence of the previous observations, we can state that link performance is not limited by the speed of the vehicle but is determined by the amount of signal-to-noise ratio and, consequently, by the transmitted power. On the other hand, it is well known that Bluetooth uses a polling technique at the MAC (media access control) level, and this means that, in the presence of many vehicles, the MAC protocol can cause a performance degradation.

5. CONCLUSIONS

This article presented an overview of the evolution of research in the field of communications for ITS. Starting from the activities oriented to dedicated solutions, the discussion then focused on the assessment of existing relevant standards in mobile communications to define and activate new ITS services. As far as dedicated solutions are concerned, this article showed some results related to radars as well as 60-GHz and infrared communication systems specifically designed for RVC and IVC. The issue of ITS service provision by means of an already standardized system was then addressed, and, as an example, the feasibility of an "emergency warning" service based on GPRS was considered. Finally, the performance of Bluetooth-based RVC and TVC links was addressed by both analytically and experimentally.

ACRONYMS

AHS	Automated highway system
ARQ	Automatic Repeat reQuest; automatic repeat request (generic)
BER	Bit error rate
BTS	Base transceiver station
CCD	Charge-coupled device
CDMA	Code-division multiple access
DSRC	Dedicated short-range communication
FM	Frequency modulation
FM-CW	Frequency-modulated continuous wave
FSK	Frequency shift keying
FTP	File Transfer Protocol
GFSK	Gaussian frequency shift keying
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
IP	Internet Protocol
ISM	Industrial–scientific–medical
ITS	Intelligent transportation system
IVC	Intervehicle communications
JSK	Association of Electronic Technology for Automobile Traffic and Driving (Japan)
LOS	Line of sight
MAC	Medium access control
MSK	Minimum shift keying
RF	Radiofrequency
RLC	Radio link control
RVC	Roadside-to-vehicle communication
SIR	Signal-to-interference ratio
SS	Spread spectrum
TDMA	Time-division multiple access

TVC	Target-to-vehicle communication
UMTS	Universal Mobile Telecommunication System
USDOT	United States Department of Transportation
VMS	Variable message sign
WPAN	Wireless personal-area network

BIOGRAPHIES

Oreste Andrisano was born in Bologna, Italy, on February 14, 1952. He received the Dr. Ing. degree in electronic engineering cum laude from the University of Bologna, Bologna, Italy, in 1975. In the same year he joined the University of Bologna, where he became a Professor of Electrical Engineering in 1985. Since 1992 he has been the Director of CSITE (Centro di studio per l'Informatica e i Sistemi di Telecomunicazioni), University of Bologna and Consiglio Nazionale delle Ricerche, Roma. Since 2000 he has been the Director of the Laboratorio Nazionale di Comunicazioni Multimediali, Napoli (Consorzio Nazionale Interuniversitario per le Telecomunicazioni, CNIT). In the period 1996-2001 he was an Editor of the IEEE Transactions on Communications (Modulation for fading channels). His research activity has been concerned with different fields in the digital communication area, such as digital signal processing, data transmission for satellite and fixed radio links applications, local wireless and mobile radio networks. He has also been active in the Intelligent Transportation Systems (ITS) area, with reference to vehicle-to-vehicle and vehicle-to-infrastructure communication systems. In 1987 and 1988 he cooperated in the definition phase of PROMETHEUS (EUREKA), as a European coordinator for the transmission systems research area. Then, he was in the Steering Committee of Project DACAR (Data Acquisition and Communication Techniques and their Assessment for Road Transport) in the framework of DRIVE I, ECC, 1988-1991. In the period 1991-1997 he was responsible for the ITS communication activities in Italy (coordination of PROCOM and TELCO). Since 1998 he is the national coordinator of the project Multimedia Systems funded at national level by MIUR and CNR (Roma). Oreste Andrisano is a member of the IEEE Communication and Vehicular Technology Societies and of the IEEE Radiocommunication Committee.

Masao Nakagawa was born in Tokyo, Japan, in 1946. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1969, 1971 and 1974 respectively. Since 1973, he has been with the Department of Electrical Engineering, Keio University, where he is now Professor. His research interests are in CDMA, consumer communications, mobile communications, ITS (Intelligent Transport Systems), wireless home networks, and optical communication with lighting. He received 1989 IEEE Consumer Electronics Society Paper Award, 1999-Fall Best Paper Award in IEEE VTC, IEICE Achievement Award in 2000, IEICE Fellow Award in 2001. He was the executive committee chairman of International Symposium on Spread Spectrum Techniques and Applications in 1992 and the technical program committee chairman of ISITA (International Symposium on Information Theory and its Applications) in 1994. He is an editor of Wireless Personal Communications and was a guest editor of the special issues on "CDMA Networks I, II, III and IV" published in IEEE JSAC in 1994(I and II) and 1996(III and IV). He chairs the Wireless Home Link subcommittee in MMAC (Multimedia Mobile Access Communication Promotion Committee).

Gianni Pasolini received his Laurea degree in telecommunications engineering from the University of Bologna, Italy, in March 1999. In May 1999 he joined CSITE-CNR (Centre for Studies in Computer Science and Telecommunication Systems of the National Research Council), and he is currently working toward his PhD. His research activity is concerned with Wireless Local Area Networks, Digital Communications and Radio Resource Management. He is a student member of IEEE.

Roberto Verdone was born in Bologna, Italy, in 1965. He received his Laurea degree in Electronic Engineering (with honours) and his Ph.D. in Electronic Engineering and Computer Science from the University of Bologna, Italy, in March 1991 and October 1995, respectively. In April 1996 he became a researcher at CSITE-CNR (Centre for Studies in Computer Science and Telecommunication Systems of the National Research Council) in Telecommunications. Since November 2001 he is a Full Professor in Telecommunications at the University of Bologna. His research activity is concerned with Digital Transmission, Cellular and Mobile Radio Systems, Wireless Local Area Networks, Digital Broadcasting and Intelligent Transportation Systems. In 1992 he was involved in the European research program PROMETHEUS. Since 1995 to 1997 he worked in the context of the National research program TELCO (TELecommunications network for COoperative Driving). He is also responsible for the activities of CSITE in projects funded by ESA and MIUR (Italy). Since 1997 to 2000 he has participated to the COST259 Action. Since 2001 he is chairman of the WG on Network Aspects within the follow-on Action COST273. He is a member of IEEE.

BIBLIOGRAPHY

1. *DRIVE Workplan*, April 26, 1988.
2. *PROMETHEUS PRO-COM White Book, Definition Phase*, Feb. 19, 1988.
3. *IEEE Commun. Mag.* (Half Special Issue on IVHS), **34**(10) (Oct. 1996).
4. *IEEE Trans. Vehic. Technol.* (Special issue on Intelligent Vehicle Highway Systems), **40**(1) (Feb. 1991).
5. *Proc. 7th World Congress on Intelligent Transportation Systems*, Turin, Italy, Nov. 6–9, 2000.
6. *Final demonstration of PROMETHEUS project*, Paris, France, Oct. 1994.
7. ERTICO homepage, *http://www.ertico.com/*.
8. I. Catling and R. Harris, SOCRATES — progress towards commercial implementation, *Proc. Vehicle Navigation Information Systems Conf.* (*VNIS'95*), Seattle, WA, July 30–Aug. 2, 1995.
9. *IEEE Trans. Vehic. Technol.* (Special Issue on the UMTS), **47**(4) (Nov. 1998).

10. Bluetooth homepage, *http://www.bluetooth.com/*.

11. PATH homepage, *http://www.path.berkeley.edu/*.

12. *California PATH Annual Report 2000*, *http://www.path.berkeley.edu/*.

13. O. Andrisano, M. Chiani, V. Tralli, and R. Verdone, Impact of cochannel interference on vehicle-to-vehicle communications at millimetre waves, *Proc. IEEE Int. Conf. Communication Systems* (*ICCS'92*), Singapore, Nov. 16–20, 1992, Vol. 2, pp. 924–928.

14. W. Kremer et al., Computer-aided design and evaluation of mobile radio local area networks in RTI/IVHS environments, *IEEE J. Select. Areas Commun.* **11**(3): 406–421 (April 1993).

15. O. Andrisano et al., Millimetre wave short range communications for advanced transport telematics, *Eur. Trans. Telecommun.* (July–Aug. 1993).

16. H. Fujii, O. Hayashi, and N. Nagakata, Experimental research on inter-vehicle communication using infrared, *Proc. IEEE Intelligent Vehicles Symp.*, 1996, pp. 266–271.

17. R. Verdone, Communication systems at millimetre waves for ITS applications, *Proc. IEEE Vehicular Technology Conf. 1997* (*VTC'97*), Phoenix, AZ, May 5–7, 1997, Vol. 2, pp. 914–918.

18. R. Verdone, Outage probability analysis for short range communication systems at 60 GHz in ATT urban environments, *IEEE Trans. Vehic. Technol.* **46**(4): 1027–1039 (Nov. 1997).

19. R. Verdone, Multi-hop R-ALOHA for inter-vehicle communications at millimetre waves, *IEEE Trans. Vehic. Technol.* **46**(4): 992–1005 (Nov. 1997).

20. M. Chiani and R. Verdone, A TDD-TCDMA radio interface at millimetre waves for ITS applications, *Proc. IEEE Vehicular Technology Conf.* (*VTC'99—Fall*), Amsterdam, The Netherlands, Sept. 19–22, 1999, Vol. 2, pp. 770–774.

21. L. Michael and M. Nakagawa, Interference characteristics in inter-vehicle communication from oncoming vehicles, *Proc. IEEE Vehicular Technology Conf.* (*VTC'99–Fall*), Amsterdam, The Netherlands, Sept. 19–22, 1999, Vol. 2, pp. 753–757.

22. Y. Shiraki et al., Evaluation of interference reduction effect of SS radar, *Proc. 1999 IEICE General Conf.*, A-17–22, 1999, p. 421.

23. CSITE homepage, *http://www-csite.deis.unibo.it/htcsite/prometheus/Indice.html*.

24. C. Cseh, Architecture of the dedicated short-range communications (DSRC) protocol, *Proc. 48th IEEE Vehicular Technology Conf.*, 1998, Vol. 3, pp. 2095–2099.

25. CEN TC 278 WG 9 homepage, *Dedicated Short-Range Communications*, *http://www.comnets.rwth-aachen.de/~ftp-wg9/*.

26. R. Yuan, North American dedicated short range communications (DSRC) standards, *Proc. IEEE Conf. Intelligent Transportation System, 1997* (*ITSC '97*), pp. 537–542.

27. O. Hyunseo, Y. Chungil, A. Donghyon, and C. Hanberg, 5.8 GHz DSRC packet communication system for ITS services, *Proc. 50th IEEE Vehicular Technology Conf., 1999* (*VTC 1999—Fall*), 1999, Vol. 4, pp. 2223–2227.

28. I. Catling, R. Harris, L. James, and N. Simmons, ITS services in Wales using the wireless application protocol (WAP), *Proc. Int. Conf. Advanced Driver Assistance Systems, 2001* (*ADAS*), pp. 73–75.

29. R. Nusser and R. M. Pelz, Bluetooth-based wireless connectivity in an automotive environment, *Proc. 52nd Vehicular Technology Conf., 2000.* (*IEEE-VTS—Fall VTC*), 2000, Vol. 4, pp. 1935–1942.

30. O. Andrisano, R. Verdone, and M. Nakagawa, Intelligent transportation systems: The role of third generation mobile radio networks, *IEEE Commun. Mag.* **38**(9): 144–151 (Sept. 2000).

31. *Specification of the Bluetooth System*, Core, version 1.0B.

COMMUNITY ANTENNA TELEVISION (CATV) (CABLE TELEVISION)

ROGER FREEMAN*
Independent Consultant
Scottsdale, Arizona

1. INTRODUCTION

The principal thrust of community antenna television (CATV) is entertainment. Lately, CATV has taken on some new dimensions. It is indeed a broadband medium, providing up to 1 GHz of bandwidth at customer premises. It was originally a unidirectional system, from the point of origin, which we call the *headend*, toward customer premises. It does, though, have the capability of being a two-way system by splitting the band, say, from 5 to 50 MHz for upstream traffic (i.e., toward the headend), and the remainder is used for downstream traffic (i.e., from the headend to customer premises). CATV is certainly a major and viable contender for *last-mile communications*.

We will briefly discuss one approach to provide capability for two-way traffic, usually voice and data. First, however, conventional CATV will be described, and it includes the concept of supertrunks and HFC (hybrid fiber–coaxial cable systems). We will involve the reader with such topics as wideband amplifiers in tandem, optimum amplifier gain, IM (intermodulation) noise, beat noise, and cross-modulation products. System layout, hubs, and last-mile or last-100-ft considerations will also be covered. There will also be a brief discussion of the conversion to a digital system using some of the compression techniques now employed on modern cable television systems. The section includes overviews of the three important competing standards for broadband hybridcoax (HFC) CATV systems. These provide the user, besides conventional TV downstream, upstream, and downstream connectivity for megabit Internet, IP intranet, various data services, still-image transmission, and POTS* telephony.

* Roger Freeman took an early retirement from the Raytheon Company, Equipment Division, in 1991 where he was principal engineer to establish *Roger Freeman Associates*, Independent Consultants in Telecommunications. He has been writing books on various telecommunication disciplines for John Wiley & Sons, since 1973. Roger has seven titles that he keeps current, including *Reference Manual, Inc. for Telecommunication Engineers* now in its third edition. His Website *www.rogerfreeman.com*, and his email address is *rogerf@pcslink.com.*

* POTS-plain old telephone service.

2. THE EVOLUTION OF CATV

2.1. The Beginnings

Broadcast television, as we know it, was in its infancy around 1948. Fringe-area problems were much more acute in that period. By "fringe area," we mean areas with poor or scanty signal coverage. A few TV users in fringe areas found that if they raised their antennas high enough and improved antenna gain characteristics, they could receive an excellent picture. Such users were the envy of the neighborhood. Several of these people who were familiar with RF signal transmission employed signal splitters so that their neighbors could share the excellent picture service. It was soon found that there was a limit on how much signal splitting could be done before signal levels got so low that they were snowy or unusable.

Remember that each time a signal splitter (even power split) is added, neglecting insertion losses, the TV signal dropped by 3 dB. Then someone got the bright idea of amplifying the signal before splitting. Now some real problems arose. One-channel amplification worked fine, but two channels from two antennas with signal combining became difficult. Now we are dealing with comparatively broadband amplifiers. Among the impairments we can expect from broadband amplifiers and their connected transmission lines (coaxial cable) are

- Poor frequency response. Some part of the received band had notably lower levels than other parts. This is particularly true as the frequency increases. In other words, there was fairly severe amplitude distortion; thus equalization became necessary.
- The mixing of two or more RF signals in the system caused intermodulation products and "beats" (harmonics), which degraded reception.
- When these TV signals carried modulation, cross-modulation (Xm) products further degraded or impaired reception.

Several small companies were formed to sell these "improved" television reception services. Some of the technicians working for these companies undertook ways of curing the ills of broadband amplifiers. What these service companies did was to install a reasonably high tower in an appropriate location. Comparatively high gain antennas were installed such that a clear, line-of-sight signal could be received from TV emitters within range. The high-tower receiving site was called a *headend*. The headend had RF amplifiers, TV line amplifiers, signal combiners, translators, and all that was necessary to process and distribute multiple TV signals to CATV subscribers. The distribution system was entirely based on coaxial cable. To keep the signal strength to a usable level, broadband amplifiers were installed at convenient intervals along the distribution line.

A subscriber's TV set was connected to the distribution system, and the signal received looked just the same as if it was taken off the air with its own antenna. In fringe areas signal quality, however, was much better than own antenna quality. The key to everything was that no changes were required in the users' TV set. This was just an extension of his/her TV set antenna. Such a simple concept is shown in Fig. 1.

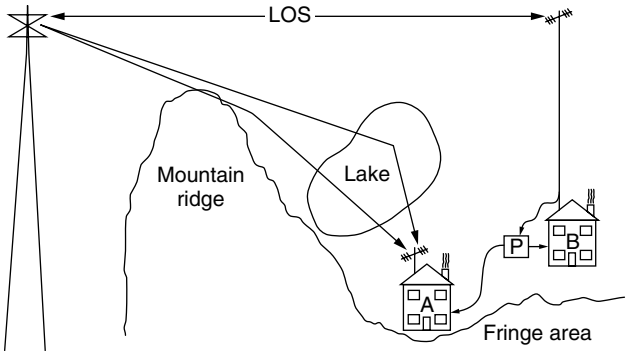


Figure 1. CATV initial concept (P = power split).

Note in Fig. 1 that home A is in the shadow of a mountain ridge and receives a weakened diffracted signal off the ridge and a reflected signal off a lake. Here is the typical multipath scenario resulting in ghosts in A's TV screen. The picture is also snowy, meaning it is noisy, as a result of a poor carrier-to-noise ratio. Home B extended the height of the antenna to be in line of sight of the TV transmitting antenna. Its antenna is of higher gain; thus it is more discriminating against unwanted reflected and diffracted signals. Home B has an excellent picture without ghosts. Home B shares its fine signal with home A by use of a 3-dB power split (P).

2.2. Early System Layouts

Figure 2 illustrates an early CATV distribution system (ca. 1968). Taps and couplers (power splits) are not shown. These systems provided 5–12 channels. A microwave system brought in channels from distant cities (50–150 mi). We had direct experience with the Atlantic City, NJ system where channels were brought in from Philadelphia and New York by microwave (MW). A 12-channel system resuited which occupied the entire assigned VHF band (i.e., channels 2–13).

As UHF TV stations began to appear, a new problem arose for the CATV operator. It was incumbent on that operator to keep the bandwidth as narrow as possible. One approach was to convert UHF channels to vacant VHF channel allocations at the headend.

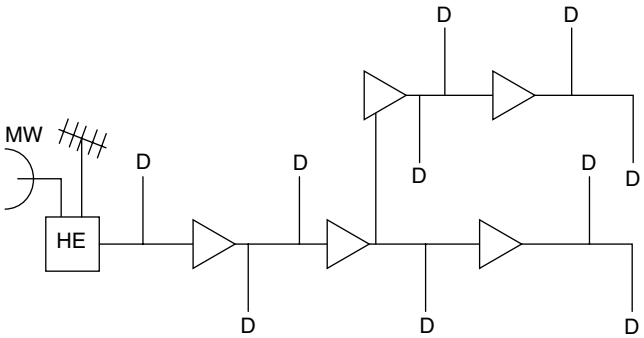


Figure 2. An early CATV distribution system (HE = headend; D = drop wire to residence; MW = microwave connectivity).

Satellite reception at the headend doubled or tripled the number channels that could be available to the CATV subscriber. Each satellite has the potential of adding 24 channels to the system. Note how the usable cable bandwidth is "broadened" as channels are added. We assume contiguous channels across the band, starting at 55 MHz. For 30 channels, we have 55–270 MHz; 35 channels, 55–300 MHz, 40 channels, 55–330 MHz; 62 channels, 55–450 MHz; and 78 channels, 55–550 MHz. These numbers of channels were beyond the capability of many TV sets of the day. Settop converters were provided that converted all channels to a common channel, an unoccupied channel, usually channel 2, 3 or 4 to which the home TV set is tuned. This approach is still very prevalent today.

In the next section we discuss impairments and measures of system performance. In Section 4, hybrid fiber–coaxial cable systems are addressed. The replacement of coaxial cable trunk by fiberoptic cable made a major stride to improved performance and system reliability/availability.

3. SYSTEM IMPAIRMENTS AND PERFORMANCE MEASURES

3.1. Overview

A CATV headend places multiple TV and FM broadcast (from 30 to 125) carriers on a broadband coaxial cable trunk and distribution system. The objective is to deliver a signal-to-noise ratio (S/N) of 42–45 dB at a subscriber's TV set. If the reader has background in the public switched telecommunications network (PSTN), he/she would expect such impairments as the accumulation of thermal and intermodulation noise. We find that CATV technicians use the term *beat* to mean intermodulation (IM) products. For example, there is triple beat distortion, defined by Bill Grant [6], as "spurious signals generated when three or more carriers are passed through a non-linear circuit (such as a wideband amplifier). The spurious signals are sum and difference products of any three carriers, sometimes referred to as 'beats.' Triple beat distortion is calculated as a voltage addition."

The wider the system bandwidth is and the more RF carriers transported on that system, the more intermodulation distortion, "triple beats" and cross-modulation we can expect. We can anticipate combinations of all the above such as *composite triple beat* (CTB), which represents the pile up of beats at or near a single frequency.

Bill Grant [6] draws a dividing line at 21 TV channels. On a system with 21 channels or less one must expect cross-modulation (Xm) to predominate. Above 21 channels, CTB will predominate.

3.2. dBmV and Its Applications

The value 0 dBmV is defined as 1 millivolt (mV) across an impedance of 75 Ω. The 75-Ω value is the standard impedance used in CATV. From the power law

$$P_w = E^2/R$$

$$P_w = 0.001^2/75 \tag{1}$$

$$0 \text{ dBmV} = 0.0133 \times 10^{-6} \text{ watts or } 0.0133 \ \mu\text{W}$$

By definition, then, 0.0133 *watts* (W) is +60 dBmV.

If $0 \text{ dBmV} = 0.0133 \times 10^{-6}$ W and 0 dBm = 0.001 W, and gain in $\text{dB} = 10\log(P_1)/(P_2)$, or in this case $10\log(0.001/0.0133 \times 10^{-6})$, then 0 dBm = +48.76 dBmV.

Remember that when working with decibels in the voltage domain, we are working with the E^2/R relationship, where $R = 75\ \Omega$. With this in mind, the definition of dBmV is

$$\text{dBmV} = 20\log(\text{voltage in millivolts})/(1 \text{ mV}) \tag{2}$$

If a signal level is 1 volt (V) at a certain point in a circuit, what is the level in dBmV?

$$\text{dBmV} = 20\log(1000)/1 = +60 \text{ dBmV}$$

If we are given a signal level of +6 dBmV, what voltage level does this correspond to?

$$+6 \text{ dBmV} = 20\log(X_{\text{mV}})/1 \text{ mV}$$

Divide through by 20:

$$\frac{6}{20} = \log(X_{\text{mV}})/1 \text{ mV}$$

$$\text{Antilog}\left(\frac{6}{20}\right) = X_{\text{mV}}$$

$$X_{\text{mV}} = 1.995 \text{ mV or } 2 \text{ mV or } 0.002 \text{ V}.$$

These signal voltages are RMS (root mean square) volts. For peak voltage, divide by 0.707. If we are given peak signal voltage and wish the RMS value, multiply by 0.707.

3.3. Thermal Noise in CATV Systems

The lowest noise levels permissible in a CATV system: at antenna output terminals, at repeater (amplifier) inputs or at a subscriber's TV set, without producing snowy pictures, are determined by thermal noise.

Consider the following, remembering that we are in the voltage domain. Any resistor or source that appears resistive over the band of interest, including antennas, amplifiers, and long runs of coaxial cable, generates thermal noise. In the case of a resistor, the noise level can be calculated on the basis Fig. 3.

To calculate the noise voltage, e_n, use the following formula:

$$e_n = (4RBk)^{1/2} \tag{3}$$

where V = an electronic voltmeter measuring the noise voltage
e_n = RMS noise voltage
R = resistance (Ω)
B = bandwidth of the measuring device (electronic voltmeter) (Hz.)

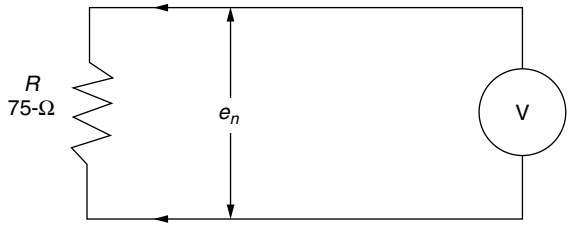


Figure 3. Resistor model for thermal noise voltage, e_n.

k = a constant equal to 40×10^{-16} at standard room temperature.

Letting the bandwidth, B, be equal to that of an NTSC TV signal be rounded to 4 MHz, the open-circuit noise voltage for a 75-Ω resistor is

$$e_n = (4 \times 75 \times 4 \times 10^{-16})^{1/2}$$
$$= 2.2\ \mu\text{V RMS}$$

Figure 4 shows a 2.2 μV noise-generating source (resistor) connected to a 75-Ω (noiseless) load. Only half of the voltage (1.1 μV) is delivered to the load. Thus the noise input to 75 Ω is 1.1 μV RMS or −59 dBmV. This is the basic noise level, the minimum that will exist in any part of a 75-Ω CATV system. The value, −59 dBmV, will be used repeatedly below.

The noise figure of typical CATV amplifiers ranges between 7 and 9 dB [4].

3.4. Signal-To-Noise Ratio (S/N) Versus Carrier-To-Noise (C/N) Ratio in CATV Systems

S/N (signal-to-noise ratio) and C/N (carrier-to-noise ratio) are familiar parameters in telecommunication transmission systems. In CATV systems S/N has a slightly different definition: *This relationship is expressed by the "signal-to-noise ratio," which is the difference between the signal level measured in dBmV, and the noise level, also measured in dBmV, both levels being measured at the same point in the system* [3].

S/N can be related to C/N on CATV systems as

$$\text{C/N} = \text{S/N} + 4.1\ \text{dB} \tag{4}$$

This is based on work by Carson [5], where the basis is "noise just perceptible" by a population of TV viewers, NTSC 4.2-MHz bandwidth TV signal. Here the S/N is 39 dB and the C/N is 43 dB.

Adding noise weighting improvement (6.8 dB), we obtain

$$\text{S/N} = \text{C/N} + 2.7\ \text{dB} \tag{5}$$

It should be noted that S/N is measured where the signal level is peak to peak and the noise level is RMS. For C/N, both the carrier and noise levels are rms. These values are based on a VSB-AM TV signal with 87.5% modulation index.

For comparison, consider another series of tests conducted by the Television Allocations Study Organization (TASO) and published in their report to the U.S. FCC in 1959. Their ratings, corrected for a 4-MHz bandwidth, instead of the 6 MHz bandwidth they used, are shown below:

TASO picture rating	S/N ratio (dB)
1. Excellent, no perceptible noise	45
2. Fine (snow just perceptible)	35
3. Passable (snow definitely perceptible, but not objectionable)	29
4. Marginal (snow somewhat objectionable)	25

Once a tolerable noise level is determined, the levels required in a CATV system can be specified. If the desired S/N has been set at 43 dB at a subscriber TV set, the minimum signal level required at the first amplifier would be −59 dBmV + 43 dB or −16 dBmV, considering thermal noise only. Actual levels would be quite a bit higher because of the noise generated by subsequent amplifiers in cascade.

It has been found that the optimum gain of a CATV amplifier is about 22 dB. When the gain is increased, intermodulation/cross-modulation products become excessive. For gains below this value, thermal noise increases, and system length is shortened or the number of amplifiers must be increased, neither of which is desirable.

There is another rule-of-thumb we should be cognizant of. Every time the gain of an amplifier is increased 1 dB, intermodulation products and "beats" increase their levels by 2 dB. And the converse is true; every time gain is decreased 1 dB, IM products and beat levels are decreased by 2 dB.

With most CATV systems, coaxial cable trunk amplifiers are identical. This, of course, eases noise calculations. We can calculate the noise level at the output of one trunk amplifier. This is

$$N_v = -59\ \text{dBmV} + \text{NF}_{\text{dB}} \tag{6}$$

where NF is the noise figure of the amplifier in decibels.

In the case of two amplifiers in cascade (tandem), the noise level (voltage) is

$$N_v = -59\ \text{dBmV} + \text{NF}_{\text{dB}} + 3\ \text{dB} \tag{7}$$

If we have M identical amplifiers in cascade, the noise level (voltage) at the output of the last amplifier is:

$$N_v = -59\ \text{dBmV} + \text{NF}_{\text{dB}} + 10 \log M \tag{8}$$

This assumes that all system noise is generated by the amplifiers, and none is generated by the intervening sections of coaxial cable (3).

Example 1. A CATV system has 30 amplifiers in tandem, and each amplifier has a noise figure of 7 dB. Assume that the input of the first amplifier is terminated in 75 Ω resistive. What is the thermal noise level (voltage) at the last amplifier output?

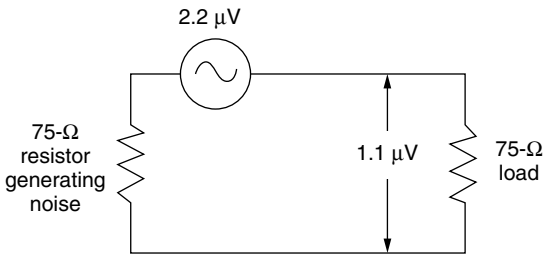


Figure 4. Minimum noise model.

Use Eq. (6):

$$N_v = -59 \text{ dBmV} + 7 \text{ dB} + 10 \log 30$$
$$= -59 \text{ dBmV} + 7 \text{ dB} + 14.77 \text{ dB}$$
$$= -37.23 \text{ dBmV}$$

For carrier-to-noise ratio (C/N) calculations, we can use the following procedures. To calculate the C/N at the output of one amplifier:

$$\text{C/N} = 59 - \text{NF}_{\text{dB}} + \text{input level (dBmV)} \quad (9)$$

Example 2. If the input level of a CATV amplifier were +5 dBmV and its noise figure were 7 dB, what would the C/N a the amplifier output be?

Use Eq. (9):

$$\text{C/N} = 59 - 7 \text{ dB} + 5 \text{ dBmV}$$
$$= 57 \text{ dB}$$

With N cascaded amplifiers, we can calculate the C/N at the output of the last amplifier, assuming all the amplifiers were identical, by the following equation:

$$\text{C/N}_L = \text{C/N(single amplifier)} - 10 \log N \quad (10)$$

Example 3. Determine the C/N at the output of the last amplifier with a cascade (in tandem) of 20 amplifiers, where the C/N of a single amplifier is 62 dB.

Use Eq. (10):

$$\text{C/N}_L = 62 \text{ dB} - 10 \log 20$$
$$= 62 \text{ dB} - 13.0 \text{ dB}$$
$$= 49 \text{ dB}$$

Another variation for calculating C/N is when we have disparate C/N ratios in different parts of a CATV system. CATV systems are usually of a tree topology where the headend is the base of the tree. There is a trunk branching to limbs and further branching out to leaf stems supporting many leaves. In the case of CATV, there is the trunk network, bridger, and line extenders with different gains, and possibly different noise figures. To solve this problem we must use a relationship of several C/N values in series:

$$\text{C/N}_{\text{sys}} = 1/[1/\text{C/N}_1) + (1/\text{C/N}_2) - - - + (1/\text{C/N}_n)] \quad (11)$$

Example 4. Out of a cascade of 20 trunk amplifiers the C/N was 49 dB; out of a bridger, 63 dB and two line extenders, 61 dB, calculate the C/N at the end of the system described.

Use Eq. (11), but first convert each decibel value to a value in decimal units and then take its inverse ($1/X$):

$$49 \text{ dB:antilog}(\tfrac{49}{10}) = 79,433\tfrac{1}{X} = 12,589 \times 10^{-9}$$
$$63 \text{ dB:antilog}(\tfrac{63}{10}) = 2,000,000\tfrac{1}{X} = 500 \times 10^{-9}$$
$$61 \text{ dB:antilog}(\tfrac{61}{10}) = 1,258,925\tfrac{1}{X} = 794 \times 10^{-9}$$

We can now sum the inverses because they all have the same exponent. Sum $= 13,883 \times 10^{-9}$. Take the inverse: 72,030. $10 \log_{(72,030)} = 48.57$ dB. Remember that the "sum" of C/N series values must be something less than the worst value of the series. That is a way of self-checking.

3.5. The Problem of Cross-Modulation (Xm)

Many specifications for TV picture quality are based on the judgment of a population of viewers. One example was the TASO ratings for picture quality given above. In the case of cross-modulation (X-mod or Xm) and CTB (composite triple beat), acceptable levels are −51 dB for Xm and −52 dB for CTB. These are good guideline values [6].

Cross-modulation is a form of third-order distortion so typical of a broadband, multicarrier system. Xm varies with the operating level of an amplifier in question and the number of TV channels being transported. Xm is derived from the amplifier manufacturer specifications. The manufacturer will specify a value for Xm (in dB) for several numbers of channels and for a particular level. The level in the specification may not be the operating level of a particular system. To calculate Xm for an amplifier to be used in given system, using manufacturer's specifications, the following formula applies:

$$\text{Xm}_a = \text{Xm}_{(\text{spec})} + 2(\text{OL}_{\text{oper}} - \text{OL}_{\text{spec}}) \quad (12)$$

where Xm_a = Xm for the amplifier in question
$\text{Xm}_{(\text{spec})}$ = Xm specified by the manufacturer of the amplifier
OL_{oper} = desired operating output signal level (dBmV)
OL_{spec} = manufacturer's specified output signal level

We spot the "2" multiplying factor and relate it to our earlier comments, namely, increase the operating level 1 dB, third-order products increase 2 dB, and the contrary for reducing signal level. And as we said, Xm is a form of third-order product.

Example 5. Suppose a manufacturer tells us that for an Xm of −57 dB for a 35-channel system, the operating level should be +50.5 dBmV. We want a longer system and use an operating level of +45 dBmV, what Xm can we expect under these conditions.

Use Eq. (12):

$$\text{Xm}_a = -57 \text{ dB} + 2(+45 \text{ dBmV} - 50.5 \text{ dBmV})$$
$$\text{Xm}_a = -68 \text{ dB}$$

CATV trunk systems have numerous identical amplifiers. To calculate Xm for N amplifiers in cascade (tandem), our approach is similar to that of thermal noise:

$$\text{Xm}_{\text{sys}} = \text{Xm}_a + 20 \log N \quad (13)$$

where N is the number of identical amplifiers in cascade, Xm_a is the Xm for one amplifier, and Xm_{sys} is the Xm value at the end of the cascade.

Example 6. A certain CATV trunk system has 23 amplifiers in cascade where Xm_a is -88 dB, what is Xm_{sys}?

Use Eq. (13):

$$\begin{aligned} Xm_{sys} &= -88 \text{ dB} + 20 \log 23 \\ &= -88 + 27 \\ &= -61 \text{ dB} \end{aligned}$$

To combine unequal Xm values, we turn to a technique similar to Eq. (11), but now, because we are in the voltage domain, we must divide through by 20 rather than 10 when converting logarithms to equivalent numerics.

Example 7. At the downstream end of our trunk system the Xm was -58 dB and taking the bridger/line extender system alone, their combined Xm is -56 dB. Now, from the headend through the trunk and bridger/line extender system, what is the Xm_{sys}? Convert -56 and -58 dB to their equivalent decimal numerics and invert ($1/X$):

$$\begin{aligned} -56 \text{ dB:antilog} - \tfrac{56}{20} &= 0.001584 \\ -58 \text{ dB:antilog} - \tfrac{58}{20} &= 0.001259 \\ \text{Sum} &: 0.002843 \end{aligned}$$

Take 20 log this value.

$$Xm_{sys} = -50.9 \text{ dB}$$

3.6. Gains and Levels for CATV Amplifiers (3)

Setting both gain and level settings for CATV broadband amplifiers is like walking a tightrope. If levels are set too low, thermal noise will limit system length (i.e., number of amplifiers in cascade). If the level is set too high, system length will be limited by excessive CTB and cross-modulation (Xm). On trunk amplifiers available gain is between 22 and 26 dB [6]. Feeder amplifiers will usually operate at higher gains, trunk systems at lower gains. Feeder amplifiers usually operate in the range of 26–32 dB gain with output levels in the range of +47 dBmV. Trunk amplifiers have gains of 21–23 dB, with output levels in the range of +32 dBmV. If we wish to extend the length of the trunk plant, we should turn to using lower loss cable. employing fiberoptics in the trunk plant is even a better alternative (see Section 4).

The gains and levels of feeder systems are purposefully higher. This is the part of the system serving customers through taps. These taps are passive and draw power. Running the feeder system at higher levels improves tap efficiency. Because feeder amplifiers run at higher gain and with higher levels, the number of these amplifiers in cascade must be severely limited to meet CTB and cross-modulation requirements at the end user.

3.7. The Underlying Coaxial Cable System

The coaxial cable employed in CATV plant is nominally 75 Ω. A typical response curve for such cable ($\frac{7}{8}$-inch, air dielectric) is illustrated in Fig. 5. This frequency response of coaxial cable is called "tilt" in the CATV industry.

For 0.5-inch cable, the loss per 100 ft at 50 MHz is 0.52 dB; for 550 MHz, 1.85 dB. Such cable systems require equalization. The objective is to have a comparatively "flat" frequency response across the entire system. An equalizer is a network that presents a mirror image of the frequency response curve, introducing more loss at the lower frequencies and less loss at the higher frequencies. These equalizers are often incorporated with an amplifier.

Equalizers are usually specified for a certain length of coaxial cable, where length is measured in dB at the highest frequency of interest. Grant [6] describes a 13-dB equalizer for a 300-MHz system, which is a corrective unit for a length of coaxial cable having 13 dB loss at 300 MHz. This would be equivalent to approximately

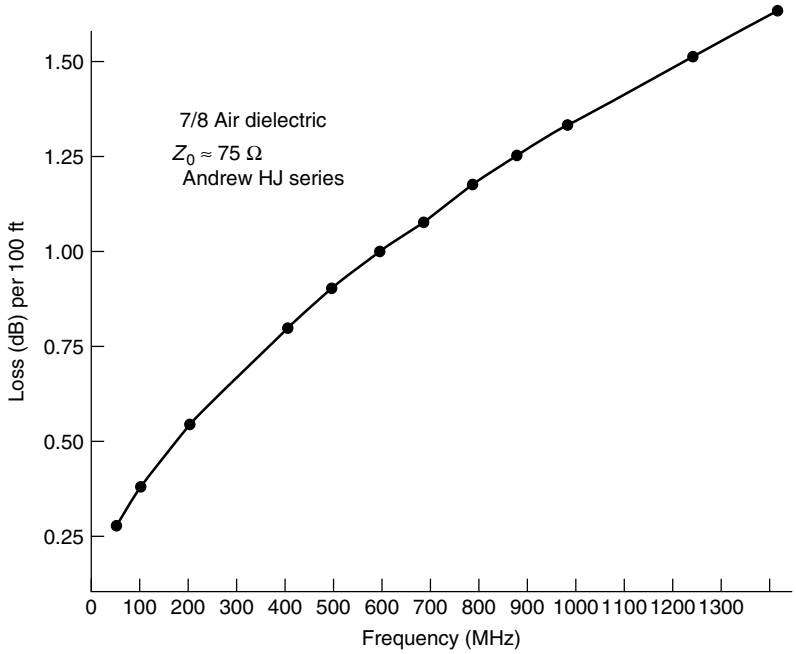


Figure 5. Attenuation–frequency response for $\frac{7}{8}$-in. coaxial cable, air dielectric, $Z_0 = 75$, Andrew HJ series heliax.

1000 ft of $\frac{1}{2}$-inch coaxial cable. Such a length of cable would have 5.45 dB loss at 54 MHz and 13 dB loss at 300 MHz. The equalizer would probably present a loss of 0.5 dB at 300 MHz and 8.1 dB at 54 MHz.

3.8. Taps

A *tap* is similar to a directional coupler. It is a device inserted into a coaxial cable that diverts a predetermined amount of its input energy to one or more tap outputs for the purpose of feeding a TV signal into subscriber drop cables. The remaining balance of the signal energy is passed on down the distribution system to the next tap or distribution amplifier. The concept of the tap and its related distribution system is shown in Fig. 6.

Taps are available to feed 2, 4, or 8 service drops from any one unit. Many different types of taps are available to serve different signal levels that appear along a CATV cable system. Commonly, taps are available in 3-dB increments. For 2-port taps, the following tap losses may be encountered: 4, 8, 11, 14, 17, 20, 23 dB. The insertion loss for the lower value tap loss may be in the order of 2.8 dB, and once the tap loss exceeds 26 dB, the insertion is 0.4 dB and remains so as tap values increase. Another important tap parameter is isolation. Generally, the higher the tap loss, the better the isolation. With 8 dB tap loss, the isolation may only be 23 dB, but with 29 dB tap loss (2-port taps), the isolation can be as high as 44 dB. Isolation in this context is the isolation between the two tap ports to minimize undesired interference from a TV set on one tap to the TV set on the other tap. For example, a line voltage signal level is +34.8 dBmV entering a tap. The tap insertion loss is 0.4 dB, so the level of the signal leaving the tap to the next tap or extender amplifier is +34.4 dBmV. The tap is 2-port. We know we want at least a +10.5 dBmV at the port output. Calculate +34.8 dBmV $-X$dB = +10.5 dBmV. Then $X = 24.3$ dB, which would be the ideal tap loss value. Taps are not available off-the-shelf at that loss value; the nearest value is 23 dB. Thus the output at each tap port will be +34.8 dBmV −23 dB = 11.8 dBmV.

4. HYBRID FIBER-COAX SYSTEMS (HFC)

The following advantages accrue by replacing the coaxial cable trunk system with optical fiber:

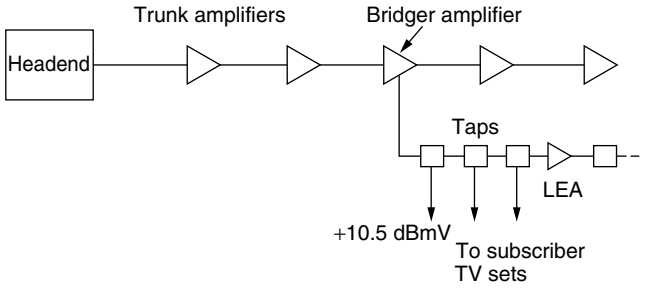


Figure 6. A simplified layout of a CATV system showing its basic elements. The objective is to provide a +10.5-dBmV signal level at the drops (tap outputs) (LEA = line extender amplifier).

- Reduces the number of amplifiers required per unit distance to reach the fartherest subscriber
- Results in improved C/N, reduced CTB and Xm levels
- Also results in improved reliability (i.e., by reducing the number of active components)
- Has the potential to greatly extend a particular CATV serving area.

Figure 7 shows the basic concept of an HFC system. One disadvantage is that a second fiber link has to be installed for reverse direction, or a form of WDM is needed, when two-way operation is required and/or for the CATV management system (used for monitoring the health of the system, amplifier degradation or failure). Figure 8 illustrates an HFC system where there are no more than three amplifiers to a subscriber tap. Note that with this system layout there cannot be a catastrophic failure. For the loss of an amplifier, only $\frac{1}{16}$ of the system is affected, worst case; with the loss of a fiber link, the worst case would be $\frac{1}{6}$ of the system.

4.1. Design of the Fiberoptic Portion of an HFC System

There are two approaches to fiberoptic transmission of analog CATV signals. Both approaches take advantage of the intensity modulation characteristics of the fiberoptic source. Instead of digital modulation of the source, amplitude modulation (analog) is employed. The most common method takes the entire CATV spectrum as it would appear on a coaxial cable and uses that as the modulating signal. The second method also uses analog amplitude modulation, but the modulating signal is a grouping of subcarriers that are each frequency-modulated. One off-the-shelf system multiplexes in a broad FDM configuration, 8 television channels, each on a separate subcarrier. Thus, a 48-channel CATV system would require six fibers, each with eight subcarriers (plus 8 or 16 audio subcarriers).

4.1.1. Link Budget for an AM System. Assume a model using a distributed feedback (DFB) laser with an output of +5 dBm coupled to the pigtail. The receiver is a PINFET, where the threshold is −5 dBm. This threshold will derive approximately 52 dB S/N in a video channel. Compared to digital operation, the C/N is around 49.3 dB, assuming that the S/N value is noise-weighted (see Section 3.4). This is a very large C/N value and leaves only 10 dB to be allocated to fiber, splice loss, and margin. If we allocated 2 dB for the link margin, only 8 dB is left for fiber/splice loss.

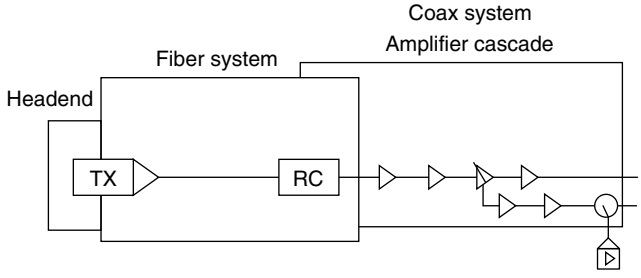


Figure 7. The concept of a hybrid fiber–coaxial cable CATV system (TX = fiberoptic transmitter, RC = fiberoptic receiver).

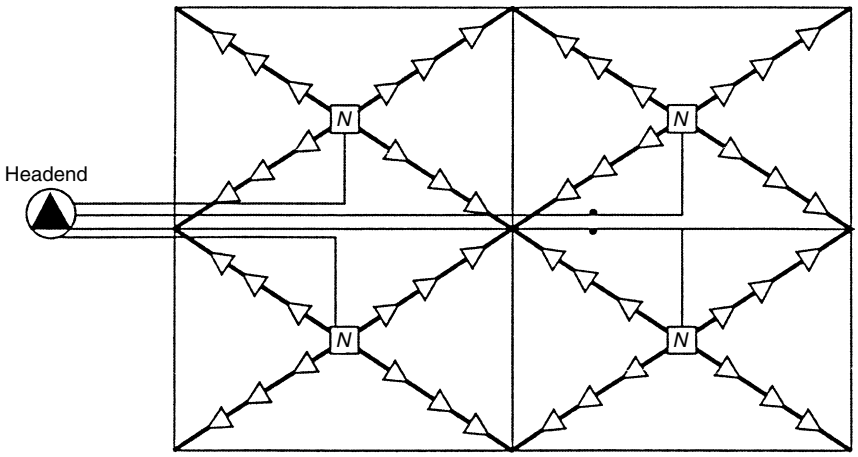


Figure 8. An HFC system layout for optimal performance (one-way). (N = node fiber) interface with coaxial cable).

At 1550-nm operation, assuming a conservative 0.4 dB/km fiber/splice loss, the maximum distance to the coax hub or fiberoptic repeater is only 8/0.4 or 20 km. Of course if we employ a EDFA (erbium-doped fiber amplifier) with, say, only a 20-dB gain, this distance can be extended by 20/0.4 or an additional 50 km. Figure 9 illustrates a typical laser diode transfer characteristic showing the amplitude-modulated input drive.

Representative design goals for the video/TV output of a fiber optics trunk are

$$\text{CNR} = 50 \text{ dB}$$

$$\text{CSO } \textit{products} = -62 \text{ dBc}$$

$$\text{CTB} = -65 \text{ dBc}$$

(where CSO = composite second order). One common technique used on HFC systems is to employ optical couplers where one fiber trunk systems feeds several hubs. A *hub* is a location where the optical signal is converted back to an electrical signal for transmission on coaxial cable. Two applications of optical couplers are illustrated in Fig. 10. Keep in mind that a signal split includes not only splitting the power but also the insertion loss of the coupler. The values shown in parentheses in Fig. 10 give the loss in the split branches (e.g., 5.7 dB and 2.0 dB).

4.1.2. FM Systems. FM systems are more expensive than their AM counterparts, but provide improved performance. EIA-250C, a well-known and respected standard for television transmission, specifies a signal-to-noise ratio of 67 dB for short-haul systems. With AM systems it is impossible to achieve this S/N, whereas a well-designed FM system can conform to EIA-250C. AM systems are degraded by dispersion on the fiber link; FM systems, much less so. FM systems can also be extended further. FM systems are available with 8, 16, or 24 channels, depending on the vendor.

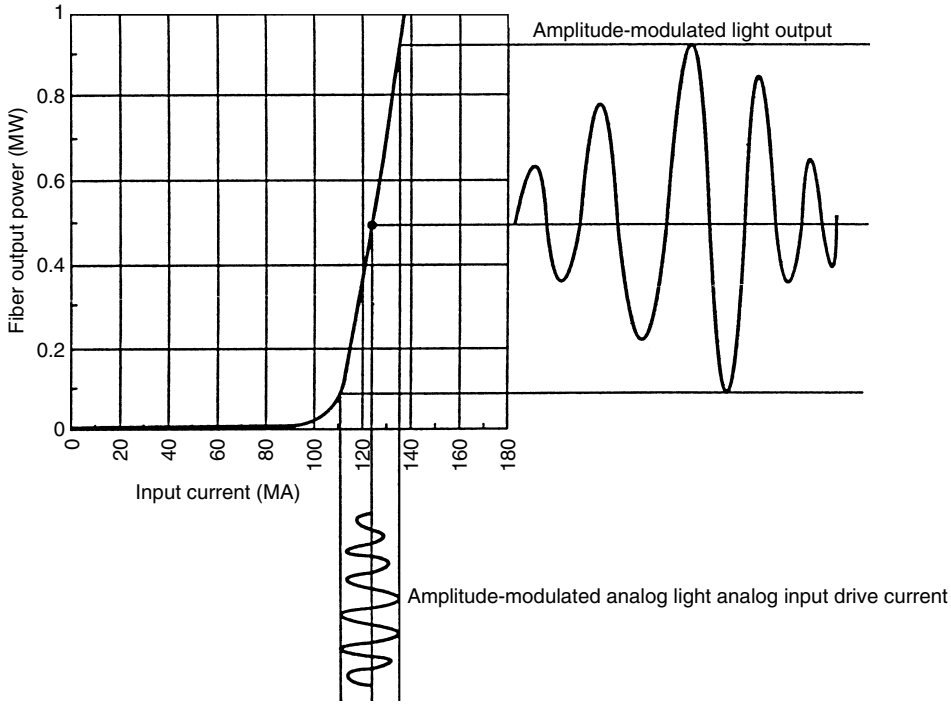


Figure 9. Laser transfer characteristics.

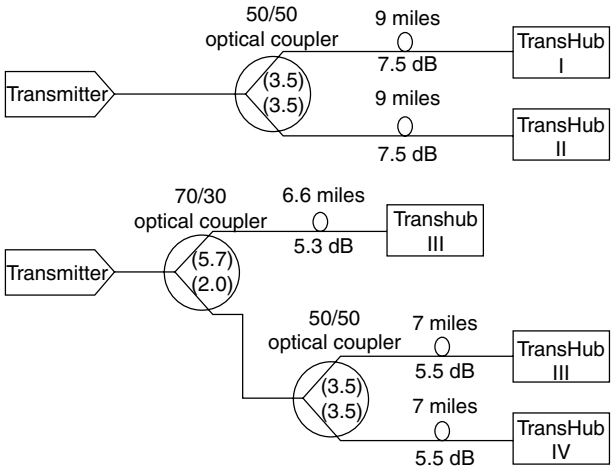


Figure 10. Two-way and three-way splits of a light signal transporting TV.

Figure 11 shows an 8-channel per fiber frequency plan, and Fig. 12 is a transmit block diagram for the video portion of the system. Figure 13 is a layout of a typical fiber hub, and Fig. 14 is a plot of optical receiver input threshold power (dBm) versus signal-to-noise ratio of individual channels derived from an FM HFC system.

Figure 12 is an FM system model. At the headend, each video and audio channel is broken out separately. And as shown in Fig. 11, each channel FM modulates a subcarrier. Note that there is a similar but separate system for the associated audio (aural) channels with 30 MHz spacing starting at 70 MHz and these audio channels may be multiplexed before transmission. Each video carrier occupies a 40 MHz slot. These RF carriers, audio and video, are combined in a passive network. The composite RF signal intensity-modulates a laser diode source. Figure 13 shows a typical fiber/FM hub.

4.1.2.1. Calculation of Video S/N for an FM System. Given the C/N (CNR) for a particular FM system, the S/N of a TV video channel may be calculated as shown in Example 8.

$$\mathrm{SNRw} = K + \mathrm{CNR} + 10\log\frac{B_{\mathrm{IF}}}{B_{\mathrm{F}}} + 20\log\frac{1.6\Delta F}{B_{\mathrm{F}}}$$

where K = a constant ($\sim$23.7 dB) made of weighting network, deemphasis, and rms to p-p conversion factors
CNR = carrier-to-noise ratio in the IF bandwidth

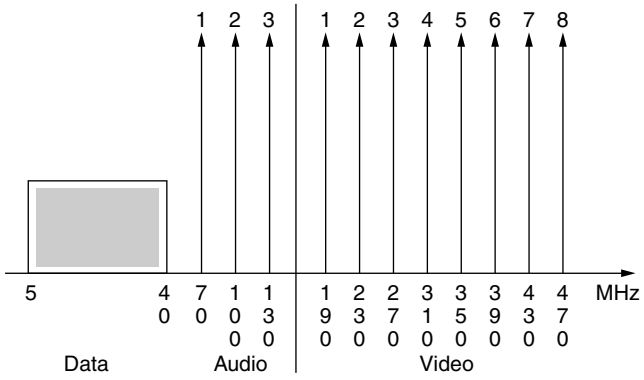


Figure 11. Eight-TV-channel frequency plan for an FM system.

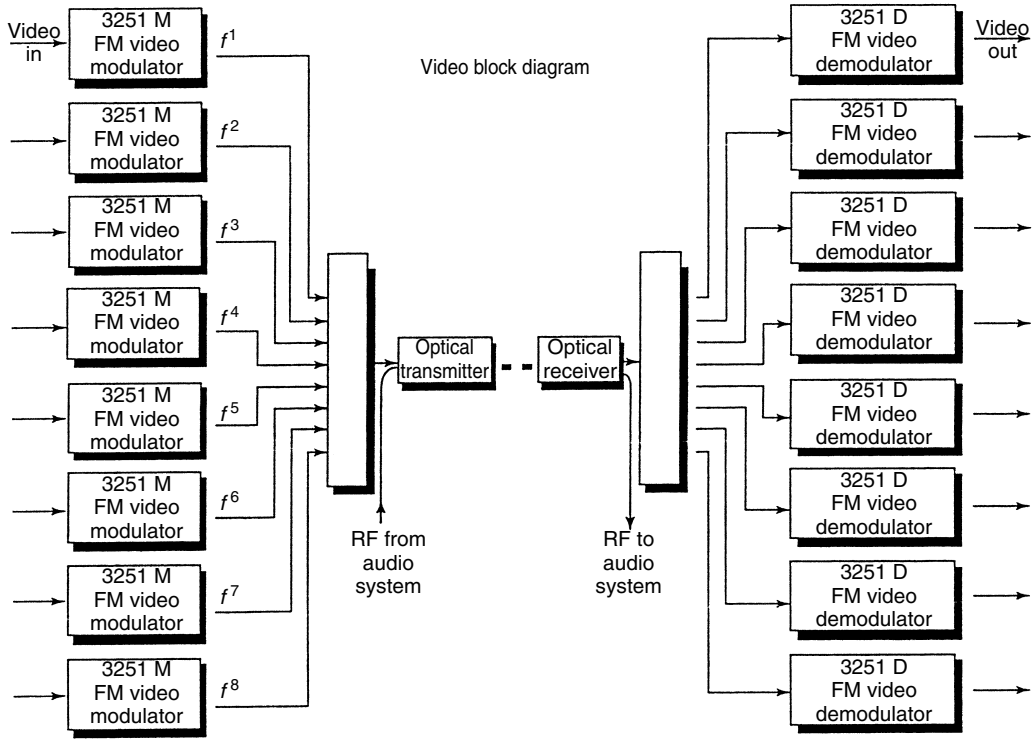


Figure 12. FM system model block diagram for video transmission subsystem. (Courtesy of Catel Corp.)

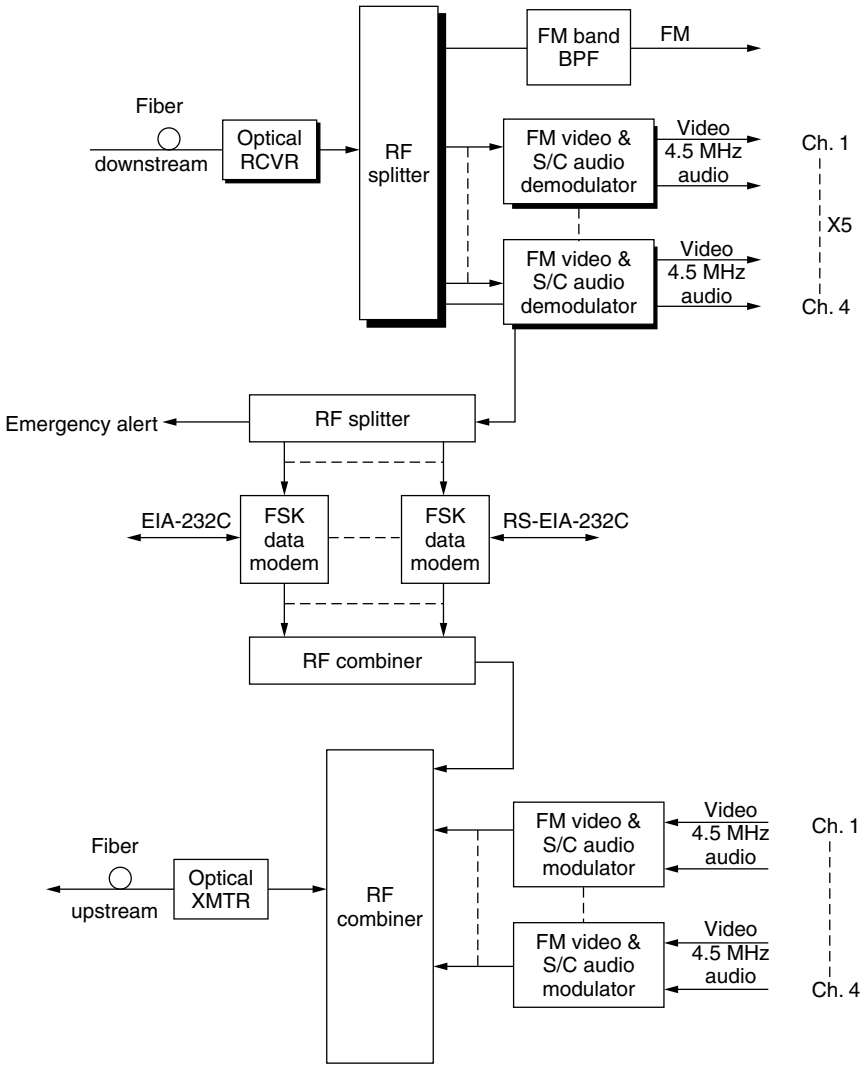


Figure 13. A typical FM/fiber hub.

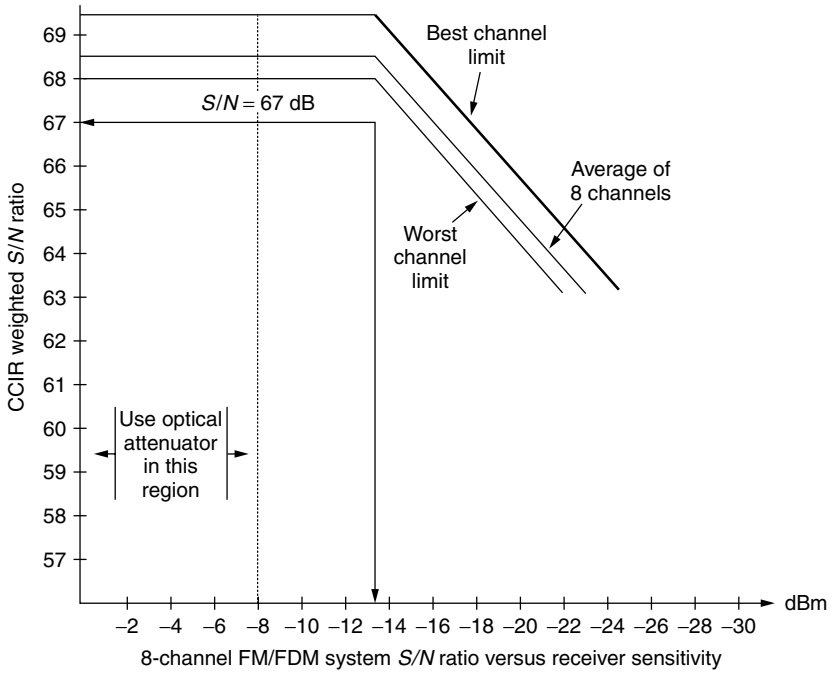


Figure 14. Link performance of an FM system. (Courtesy of Catel Corp.)

B_{IF} = IF bandwidth
B_F = baseband filter bandwidth
ΔF = sync tip-to-peak white (STPW) deviation

With $\Delta F = 4$ MHz, $B_{IF} = 30$ MHz, and $B_F = 5$ MHz, the SNRw is improved by approximately 34 dB above CNR.

Example 8. If the C/N on an FM fiber link is 32 dB, what is the S/N for a TV video channel using the values given above?

Use Eq. (14):

$$\begin{aligned} \text{S/N} &= 23.7\text{ dB} + 32\text{ dB} + 10\log(\tfrac{30}{5}) + 20\log(1.6 \times \tfrac{4}{5}) \\ &= 23.7 + 32 + 7.78 + 2.14 \\ &= 65.62\text{ dB} \end{aligned}$$

Figure 14 illustrates the link performance of an FM fiberoptic system for video channels.

Table 1 shows typical link budgets for an HFC AM system.

5. DIGITAL TRANSMISSION OF CATV

5.1. Approaches

There are two approaches to digitally transmit TV, both audio and video. The first is to transport raw, uncompressed video. The second method is to transport compressed video. Each method has advantages and disadvantages. Some advantages and disadvantages are application-driven. For example, if the objective is digital to the residence or office, compressed TV may be the most advantageous. In either case the ability to regenerate is a distinct advantage.

5.2. Transmission of Uncompressed Video on CATV Trunks

Video is an analog signal. It is converted to a digital format using techniques with some similarity to the 8-bit PCM so widely employed in the PSTN. A major difference is in the sampling. Broadcast quality TV is generally *oversampled*. Here we mean that the sampling rate is greater than the Nyquist rate. The Nyquist rate, as we remember, requires the sampling rate to be twice the highest frequency of interest. In our case this is 4.2 MHz, the bandwidth of a TV signal. Thus, the sampling rate is greater than 8.4×10^6 samples per second.

Typically, the sampling rate is based on the frequency of the color subcarrier. For NTSC television, the color subcarrier is at 3.58 MHz and we call this frequency f_{sc}. In some cases the sampling rate is set at three times this frequency ($3f_{sc}$) and in other cases four times the sampling rate ($4f_{sc}$). Thus, for NTSC color television, the sampling rate for the A/D converter is either $3 \times 3.58\text{ MHz} = 10.74 \times 10^6\text{ s}^{-1}$ or $4 \times 3.58\text{ MHz} = 14.32 \times 10^6$/s. For PAL television, the color subcarrier is 4.43 MHz and the sampling rate then may be $3 \times 4.43 = 13.29 \times 10^6$ MHz or $4 \times 4.43\text{ MHz} = 17.72 \times 10^6\text{ s}^{-1}$.

A major advantage of digital transmission is the regeneration capability just as it is in PSTN 8-bit PCM. As a result, there is no noise accumulation on the digital portion of the network. These digital trunks can be extended hundreds or more miles. The complexity is only marginally greater than an FM system. The 10-bit system can easily provide an S/N at the conversion hub of 67 dB in a video channel and an S/N of 63 dB with an 8-bit system. With uncompressed video, BER requirements are not very stringent because video contains highly redundant information.

5.3. Compressed Video

MPEG types of compression are widely used today. A common line bit rate for MPEG[2] is 1.544 Mbps. Allowing 1 bit per Hz of bandwidth, BPSK modulation and a cosine rolloff of 1.4, the 1.544 TV signal can be effectively transported in a 2 MHz bandwidth. Certainly 1000 MHz coaxial cable systems are within the state of the art. With simple division we can see that 500-channel CATV systems are technically viable. If the modulation scheme utilizes 16-QAM (4 bits per Hz theoretical), three 1.544 Mbps compressed channels can be accommodated in a 6-MHz slot. We select 6 MHz because it is the current RF bandwidth assigned for one NTSC TV channel.

6. TWO-WAY CATV SYSTEMS

6.1. Introduction

Panels (a) and (b) of Fig. 15 are two views of the CATV spectrum as it would appear on coaxial cable. Of course, with conventional CATV systems, each NTSC television channel is assigned a 6 MHz slot just as it is done with conventional broadcast television. In Fig. 15a, only 25 MHz is assigned for upstream services. Not all of this bandwidth

[2] MPEG = Motion Picture Experts' Group, a standardization agency, most know for motion picture (television) compression schemes.

Table 1. Typical Link Budgets for an AM Fiber Link

Distance (mi)	Distance (km)	Fiber Loss/km	Total Fiber Loss	Splice Loss/2 km	Total Splice Loss	Total Path Loss	Link Budget	Link Margin
Mileage, Losses, and Margins — 1310 nm								
12.40	19.96	0.5 dB	9.98	0.1 dB	1.00	10.98	13.00	2.02
15.15	24.38	0.4 dB	9.75	0.1 dB	1.22	10.97	13.00	2.03
17.00	27.36	0.35 dB	9.58	0.1 dB	1.37	10.94	13.00	2.06
Mileage, Losses, and Margins — 1550 nm								
22.75	36.61	0.25 dB	9.15	0.1 dB	1.83	10.98	13.00	2.02

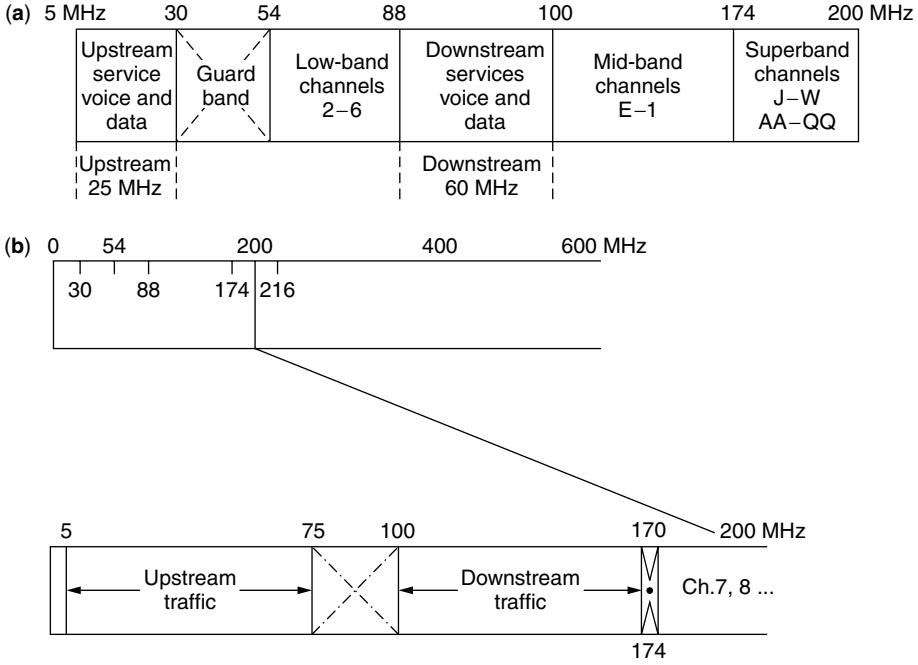


Figure 15. (**a**) CATV spectrum based on Grant [6] showing additional upstream and downstream services — note the bandwidth imbalance between upstream and downstream; (**b**) CATV spectrum with equal upstream and downstream bandwidths for other services [part (a) based on Ref. 6; part (b) prepared by the author]. On the other hand, downstream has 60 MHz assigned. In this day of the Internet, this would be providential, for the majority of the traffic would be downstream.

may be used for voice and data. A small portion should be set aside for upstream telemetry to monitor active CATV equipment. On the other hand, downstream has 60 MHz assigned. In this day of the internet, this would be providential, for the majority of the traffic would be downstream.

6.2. Comments on Fig. 15

Large guard bands isolate upstream from downstream TV and other services, with 24 MHz in A and 25 MHz in B. A small guardband was placed in the slot from 170 to 174 MHz to isolate downstream data and voice signals from conventional CATV television. We assume that the voice service will be "POTS" (plain old telephone service) and that both the data and voice would be digital.

In another approach, downstream voice, data and special video are assigned the band 550 to 750 MHz, which is the highest frequency segment portion of this system. (Ref. 12) In this case, we are dealing with a 750-MHz system.

The optical fiber trunk terminates in a node or hub. This is where the conversion from optical to the standard CATV coaxial cable format occurs. Let a node serve four groupings of subscribers, each with a coaxial cable with the necessary amplifiers, line extenders, and taps. Such subscriber groups consist of 200–500 terminations (TV sets). Assume that each termination has upstream service using the band 5–30 MHz (Fig. 15a). In our example, the node has four incoming 5–30-MHz bands, one for each coaxial cable termination. It then converts each of these bands to a higher-frequency slot 25 MHz wide in a frequency-division configuration for backhaul on a return fiber. In one scheme, at the headend, each 25-MHz slot is demultiplexed and the data and voice traffic are segregated for switching and processing.

An interesting exercise is to divide 25 MHz by 500. This tells us that we can allot each user 50 kHz full period. By taking advantage of the statistics of calling (usage), we could achieve 4–10 times bandwidth multiplier by using forms of concentration. However, upstream video, depending on the type of compression, might consume a large portion of this spare bandwidth.

There are many other ways for a subscriber can gain access, such as by TDMA, FDMA, and contention. Several protocols use combinations of time division and frequency division based on the concept of the minislot.

6.3. Impairments Peculiar to Upstream Service

6.3.1. More Thermal Noise Upstream than Downstream. Figure 16 shows a hypothetical layout of amplifiers in a CATV distribution system for two-way operation. In the downstream direction, broadband amplifiers point outward, down trunks and out distribution cables. In the upstream direction, the broadband amplifiers point inward

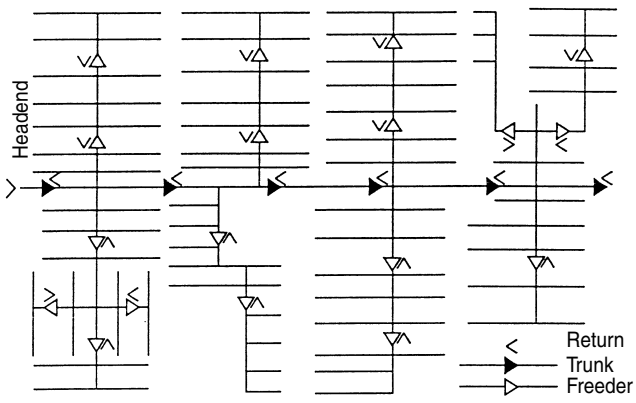


Figure 16. Trunk/feeder system layout for two-way operation. (From Grant [6]. Reprinted with permission.)

toward the headend, and all their thermal noise accumulates and concentrates at the headend. This can account for 3–20 dB additional noise upstream at the headend where the upstream demodulation of voice and data signals takes place. Fortunately, the signal-to-noise ratio requirements for good performance for voice and data are much less stringent than for video, which compensates to a certain extent for this additional noise.

6.3.2. Ingress Noise. This noise source is peculiar to CATV system. It basically derives from the residence/office TV sets that terminate the system. Parts 15.31 and 15.35 of the *FCC Rules and Regulations* govern such unintentional radiators. These rules have not been rigidly enforced.

One problem that the 75-Ω impedance match between the coaxial cable and the TV set is poor. Thus not only all radiating devices in the TV set but other radiating devices nearby in residences and office buildings couple back through the TV set into the CATV system in the upstream direction. This type of noise is predominant in the lower frequencies, that band from 5 to 30 MHz that carries the upstream signals. As frequency increases, ingress noise intensity decreases. Fiberoptic links in an HFC configuration provide some isolation, but it still can be a major problem.

6.4. Data Over Cable Service Interface Specification (DOCSIS)

DOCSIS is a complete specification for transmitting data across a cable television system. The intended service allows transparent bidirectional transfer of IP[5] traffic between the headend and customer facilities, over an all-coaxial or hybrid fiber/coax (HFC) cable network. The concept is illustrated in Fig. 17.

The specification for the cable modem (CMTS) is described in detail in DOCSIS [9,10]. A brief overview of the modulation and coding is given in the following section, which treats digital video transmission.

7. DIGITAL VIDEO TRANSMISSION STANDARD FOR CABLE TELEVISION

Based on document ANSI/SCTE 07 2000, issued Oct. 25, 1996 [9] Courtesy of Dr. Ted Woo, Technical Director, SCTE.

7.1. Introduction

This section describes the framing structure, channel coding, and channel modulation for a digital multiservice television distribution system that is specific to a cable channel. The system can be used transparently with the distribution from a satellite channel, in that many cable systems are fed directly from satellite links. The specification covers both 64- and 256-QAM waveforms. Most features of the two modulation schemes are the same. Where there are differences, the specific details of each modulation scheme are covered in the DOCSIS specification.

The design of the modulation, interleaving and coding is based on test and characterization of cable systems in North America. The modulation is quadrature amplitude modulation (QAM) with a 64-point signal constellation (64-QAM) or with a 256-point signal constellation (256-QAM), transmitter selectable. The forward error correction (FEC) is based on a concatenated coding approach that produces high coding gain at moderate complexity and overhead. Concatenated coding offers improved performance over a block code, at a similar overall complexity. The system FEC is optimized for quasi-error-free operation at a threshold output error event rate of one error event per 15 minutes.

The data format input to the modulation and coding is assumed to be MPEG-2[3] transport. However, the method used for MPEG synchronization is decoupled from the FEC synchronization. For example, this enables the system to carry asynchronous transfer mode (ATM) packets without interfering with ATM synchronization. In fact, ATM synchronization may be performed by defined ATM synchronization mechanisms.

Two modes are supported by this standard. Mode 1 has a symbol rate of 5.056 Msps, and mode 2 has a symbol rate of 5.361 Msps (megasymbols per second). Typically, mode 1 is used for 64-QAM and mode 2 is used for 256-QAM. The system is compatible with future implementations of higher-data-rate schemes employing higher order QAM extensions.

7.2. Cable System Concept

Channel coding and transmission are specific to a particular medium or communication channel. The expected channel error statistics and distortion characteristics are

[5] IP-Internet Protocol

[3] MPEG — Motion Picture Experts Group.

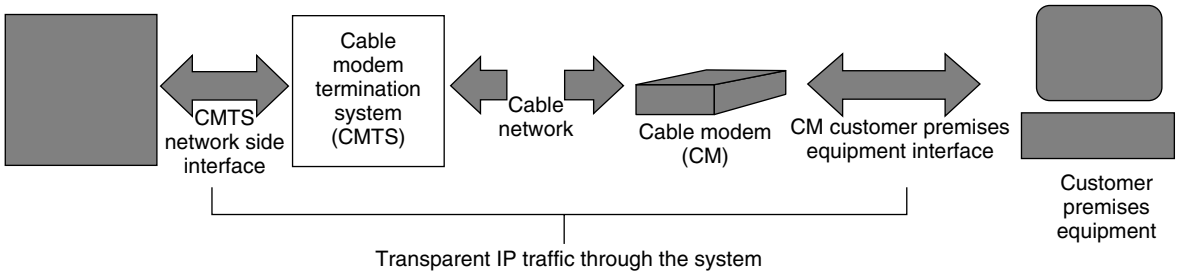


Figure 17. Transparent IP traffic through the data-over-cable system (from DOCSIS [10] Fig. 1.1, p. 2).

critical in determining the appropriate error detection and demodulation. The cable channel, including optical fiber, is regarded primarily as a bandwidth-limited linear channel with a balanced combination of white noise, interference, and multipath distortion. The quadrature amplitude modulation (QAM) technique, together with adaptive equalization and concatenated coding, is well suited to this application and channel.

The basic layered block diagram of cable transmission processing is shown in Fig. 18. The following subsections define these layers from the "outside" in, and from the perspective of the transmit side.

7.3. MPEG-2 Transport Framing

The transport layer for MPEG-2 data is composed of packets having 188 bytes, with 1 byte for synchronization purposes, 3 bytes of header containing service identification, and scrambling and control information, followed by 184 bytes of MPEG-2 or auxiliary data.

The MPEG transport framing is the outermost layer of processing. It is provided as a robust means of delivering MPEG packet synchronization to the receiver output. This processing block receives an MPEG-2 transport datastream consisting of a continuous stream of fixed-length 188-byte packets. This datastream is transmitted in serial fashion, MSB (most significant bit) first. The first byte of a packet is specified to be a sync byte having a constant value of 47_{HEX}.

7.4. Forward Error Correction

The forward error correction (FEC) definition is composed of four processing layers, as illustrated in Fig. 19. There are no dependencies on input data protocol in any of the FEC layers. FEC synchronization is fully internal and

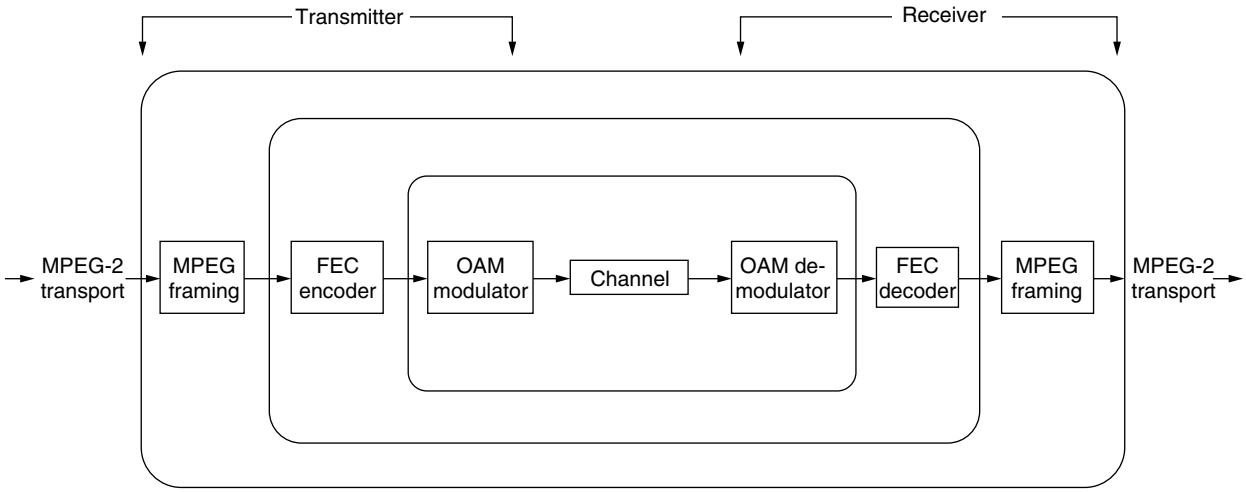


Figure 18. Cable transmission block diagram.

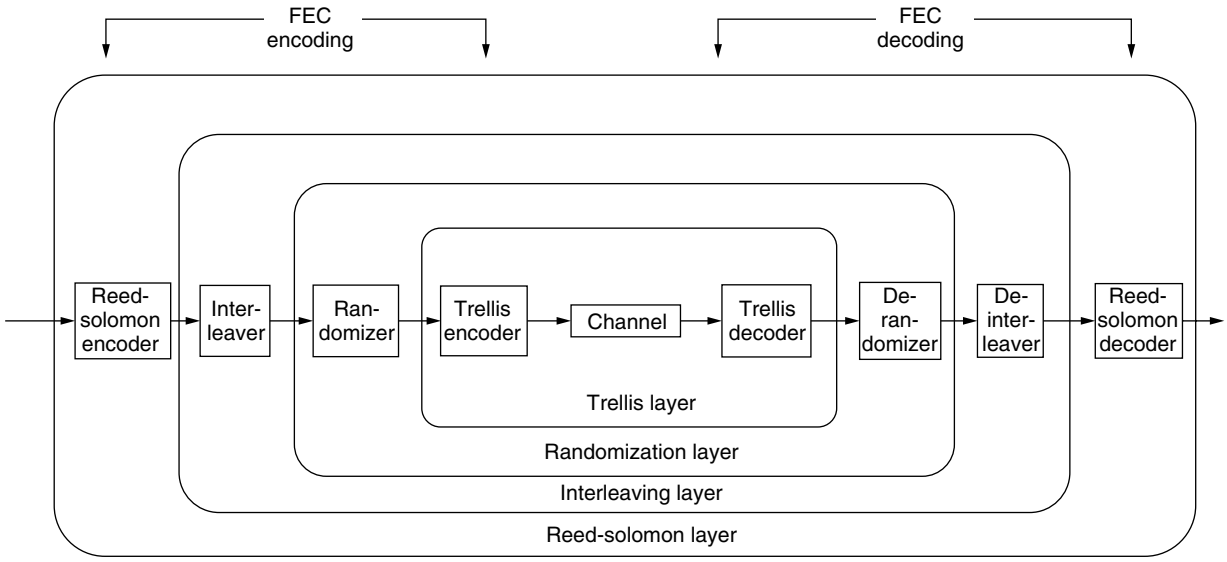


Figure 19. Layers of processing in the FEC.

transparent. Any data sequence will be delivered from the encoder input to the decoder output.

The FEC section uses various types of error-correcting algorithms and interleaving techniques to transport data reliably over the cable channel.

Reed–Solomon (RS) coding — provides block encoding and decoding to correct up to three symbols within an RS block

Interleaving — evenly disperses the symbols, protecting against a burst of symbol errors from being sent to the RS decoder

Randomization — randomizes the data on the channel to allow effective QAM demodulator synchronization

Trellis coding — provides convolutional encoding and with the possibility of using soft-decision trellis decoding of random channel errors

The following subsections define these four layers.

7.4.1. Reed–Solomon Coding. The MPEG-2 transport stream is Reed–Solomon (RS) encoded using a (128, 122) code over GF(128). This code has the capability of correcting up to $t = 3$ symbol errors per R–S block. The same R–S code is used for both 64-QAM and 256-QAM.

7.4.2. Interleaving. Interleaving is included in the modem between the RS block coding and the randomizer to enable the correction of burst noise induced errors. In both 64-QAM and 256-QAM, a convolutional interleaver is employed. The interleaver consists of a single fixed structure for 64-QAM, along with a programmable structure for 256-QAM.

7.4.3. Randomization. The randomizer is the third layer of processing in the FEC block diagram. The randomizer provides for even distribution of the symbols in the constellation, which enables the demmodulator to maintain proper lock. The randomizer adds a pseudorandom noise (PN) sequence of 7 bits over Galois Field GF(128) (i.e., bitwise exclusive-OR) to the symbols within the FEC frame to assure a random transmitted sequence.

7.4.4. Trellis-Coded Modulation. As part of the concatenated coding scheme, trellis coding is employed for the inner code. It allows introduction of redundancy to improve the threshold signal-to-noise ratio (SNR) by increasing the symbol constellation without increasing the symbol rate. As such, it is more properly termed *trellis-coded modulation*.

7.4.5. 64-QAM Modulation. For 64-QAM, the input to the trellis-coded modulator is a 28-bit sequence of four, 7-bit RS symbols, which are labeled in pairs of "A" and "B" symbols. A block diagram of a 64-QAM trellis-coded modulator is shown in Fig. 20. All 28 bits are assigned to a trellis group, where each trellis group forms 5 QAM symbols.

Of the 28 input bits that form a trellis group, each of two groups of 4 bits of the differentially precoded bitstreams in a trellis group are separately encoded by a binary convolutional coder.

The differential precoder allows the information to be carried by the change in phase, rather than by the absolute phase. For 64-QAM the third and sixth bits of the 6-bit symbols are differentially encoded, and for 256-QAM the fourth and eighth bits are differentially encoded.

7.4.6. Binary Convolutional Coder. The trellis-coded modulator includes a punctured rate-$\frac{1}{2}$ binary convolutional encode that is used to introduce the redundancy into the LSBs (least significant bits) of the trellis group. The convolutional encoder is a 16-state nonsystematic rate-$\frac{1}{2}$ encoder with the generator: G1 =

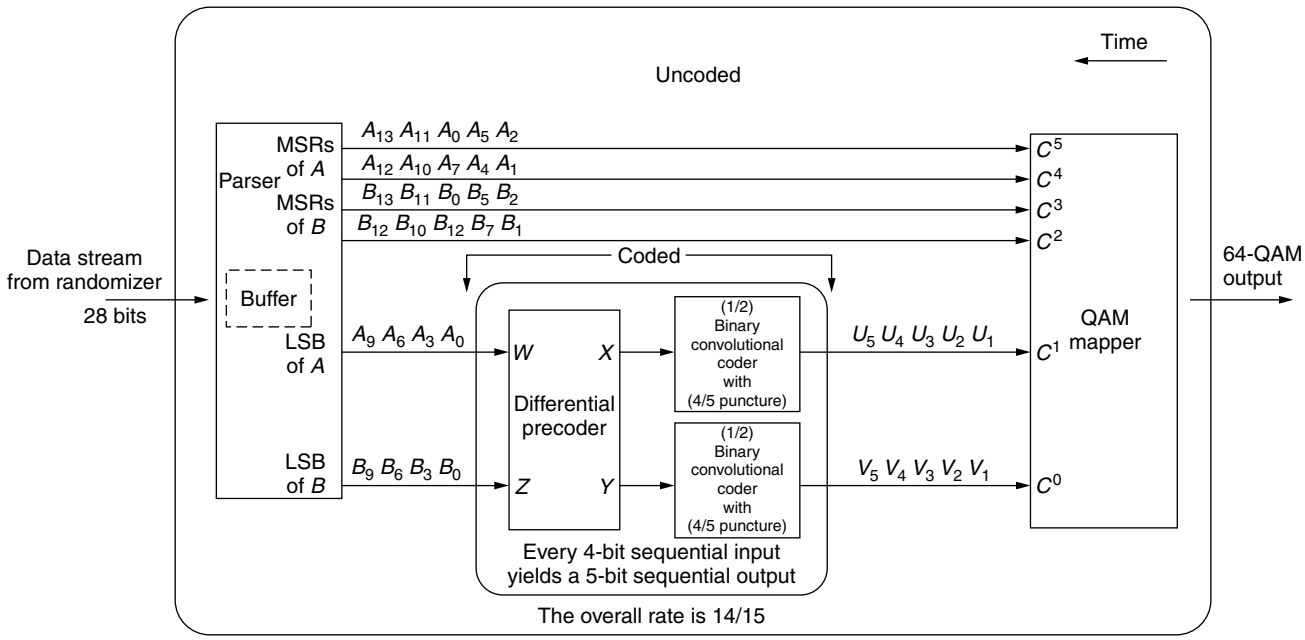


Figure 20. 64-QAM trellis-coded modulator block diagram.

010101, G2 = 011111(15, 37_{octal}), or equivalently the generator matrix[$1 \oplus D^2 \oplus D^4$, $1 \oplus D \oplus D^2 \oplus D^3 \oplus D^4$].

8. CONCLUSION

This concludes our brief description of the digital video transmission standard for CATV. A detailed description of this standard is contained in Ref. 9.

BIOGRAPHY

Roger Freeman has over 50 years experience in telecommunications including a stint in the US Navy and radio officer on merchant vessels. He attended Middlebury College and has two degrees from New York University. He has had assignments with the Bendix Corporation in Spain and North Africa which was followed by five years as a member technical staff for ITT Communications Systems. Roger then became manager of microwave systems for CATV extension at Jerrold Electronics Corporation followed by assignments at Page Communications Engineers in Washington, DC where he was a project engineer on earth stations and on various data communication programs. During this period he was assigned by the ITU as Regional Planning Expert for northern Sourth America based in Quito, Ecuador. From Quito he took a position with ITT at their subsidiary in Madrid, Spain where he did consulting in telecommunication planning. In 1978 he joined the Raytheon Company as principal engineer in their Communication Systems Directorate where he held design positions on military communications such as on the AN/TRC-170, MILSTAR, AN/ASC-30 and on wideband HF. At the same time he taught various telecommunication courses in the evenings at Northeastern University and 4-day seminars at the University of Wisconsin. These seminars were based on his several textbooks on telecommunications published by John Wiley & Sons, New York. He also gives telecommunication seminars (in Spanish) in Monterrey, Mexico City and Caracas. Roger is a contributor and guest editor (Desert Storm edition) of the IEEE Communications magazine and was advanced by the IEEE to senior life member in 1994. He served on the board of directors of the Spain Section of the IEEE and was its secretary for four years. In 1991 Roger took early retirement from the Raytheon Company and organized Roger Freeman Associates, Independent Consultants in Telecommunications. The group has undertaken over 50 assignments from Alaska to South America.

Roger may be reached at *rogerf67@cox.net*; his website is *www.rogerfreeman.com*. Also of interest would be *www.telecommunicationbooks.com* where the reader may subscribe to the on-line Reference Manual for Telecommunication Engineering, 3rd ed, updated quarterly.

BIBLIOGRAPHY

1. *How to Characterize CATV Amplifiers Effectively*, Application Note 1288–4, Hewlett-Packard Co., Palo Alto, CA, 1997.
2. R. L. Freeman, *Telecommunication Transmission Handbook*, 4th ed., Wiley, New York, 1998.
3. K. Simons, *Technical Handbook for CATV Systems*, 3rd ed, Jerrold Electronics Corp., Hatboro, PA, 1968.
4. E. R. Bartlett, *Cable Television Technology and Operations*, McGraw-Hill, New York, 1990.
5. D. N. Carson, CATV amplifiers: figure of merit and coefficient system, 1966 *IEEE International Convention Record,* Part I, *Wire and Data Communications*, IEEE, New York, March 1966, pp. 87–97.
6. W. O. Grant, *Cable Television*, 3rd ed., GWG Associates, Schoharie, NY, 1994.
7. *System Planning, Product Specifications and Services*, Catalog no. 36, Andrew Corp., Orland Park, IL, 1994.
8. *Cable Television Channel Identification Plan*, EIA-542, Electronic Industries Alliance, Washington, DC, April 1997.
9. *Digital video transmission standard for cable television*, ANSI/SCTE 07 2000, Society of Cable Telecommunications Engineers, Exton, PA, Oct. 1996.
10. *Data over Cable Service Interface Specification*, subtitled *Radio Frequency Interface Specification*, SP-RFI-103-980202, SCTE 22, Part 1, SCTE, Exton, PA, 1998.
11. R. L. Freeman, *Reference Manual for Telecommunications Engineering*, 3rd ed., Wiley, New York, 2002.
12. *Buyers Guide*, Lightwave, CMP Publications, Nashua, NY, Oct. 2000.

COMPANDERS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

The word *compander* comes from the words *comp*ressor and exp*ander*. Companders are widely used in communication systems to compress the dynamic range of an information-bearing signal, such as a speech signal, prior to digitizing the signal. In the transmission of analog signal waveforms digitally, the analog signal is usually sampled at some nominal rate that exceeds the Nyquist rate in order to avoid aliasing of frequency components. For example, in digital transmission of speech signals, the analog speech waveform at the transmitter is lowpass filtered to some nominal bandwidth, say, 3.5 kHz, and then sampled at a rate of 8 kHz. Each sample is quantized to one of a set of quantization levels and, then, represented by a sequence of bits that are transmitted via binary modulation to the receiver. The receiver reconstructs the quantized values of the samples from the received binary sequence and synthesizes the analog signal by passing the sampled values through a digital-to-analog converter. In such a system, the compressor is used at the transmitter to compress the dynamic range of the signal samples being quantized and the reverse process is performed at the receiver, where the dynamic range of the compressed signal values is inversely expanded.

Companding will be described below in the context of pulse code modulation (PCM), which is widely used to digitally encode analog signal waveforms.

2. PULSE CODE MODULATION

Pulse code modulation (PCM) is a very simple method for converting an analog signal waveform into a sequence of binary digits. The operations performed at the encoder of a PCM system are illustrated by the functional block diagram shown in Fig. 1.

The sampling is performed at a rate higher than the Nyquist rate to avoid aliasing of high-frequency signal components. The compression of the signal dynamic range is embedded in the quantizer. If the quantizer is selected to be uniform, then no signal compression is performed. On the other hand, if a nonuniform quantizer is selected that maps a signal sample x_n into a value $g(x_n)$, where $g(x_n)$ is some nonlinear function of x_n, then $g(x)$ determines the characteristics of the compressor. It is instructive to consider a uniform quantizer whose input are samples in the range$[-x_{\max}, +x_{\max}]$ and the number of quantization levels N is a power of 2, $N = 2^v$. From this, the length of each quantization region is given by

$$\Delta = \frac{2x_{\max}}{N} = \frac{x_{\max}}{2^{v-1}} \tag{1}$$

The quantized values are chosen to be midpoints of the quantization regions and, therefore, the error $\tilde{x} = x - Q(x)$ is a random variable taking values in the interval $\left(-\frac{\Delta}{2}, +\frac{\Delta}{2}\right]$. In ordinary PCM applications, the number of levels (N) is usually high and the range of variations of the input signal (amplitude variations $x_{\max}$) is small. This means that the length of each quantization region (Δ) is small and, under these assumptions, in each quantization region the error $\tilde{X} = X - Q(X)$ can be well approximated by a uniformly distributed random variable on $\left(-\frac{\Delta}{2} + \frac{\Delta}{2}\right]$. The distortion introduced by quantization (quantization noise) is therefore

$$E[\tilde{X}^2] = \int_{-(\Delta/2)}^{+(\Delta/2)} \frac{1}{\Delta}\tilde{x}^2\, d\tilde{x} = \frac{\Delta^2}{12} = \frac{x_{\max}^2}{3N^2} = \frac{x_{\max}^2}{3 \times 4^v} \tag{2}$$

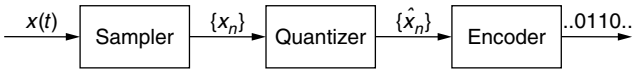


Figure 1. Block diagram of a PCM system.

where v is the number of bits per source sample. The SQNR ratio then becomes

$$\text{SQNR} = \frac{\overline{X^2}}{\overline{\tilde{X}^2}} = \frac{3 \times N^2\overline{X^2}}{x_{\max}^2} = \frac{3 \times 4^v\overline{X^2}}{x_{\max}^2} \tag{3}$$

If we denote the normalized X by $\breve{X}$, that is, $\breve{X} = \dfrac{X}{x_{\max}}$, then

$$\text{SQNR} = 3 \times N^2\overline{\breve{X}^2} = 3 \times 4^v\overline{\breve{X}2} \tag{4}$$

Note that by definition $|\breve{X}| \leq 1$ and, therefore, $\overline{\breve{X}^2} \leq 1$. This means that $3N^2 = 3 \text{ x } 4^v$ is an upperbound to the SQNR for a uniform quantizer. This also means that SQNR in a uniform quantizer deteriorates as the dynamic range of the source increases because an increase in the dynamic range of the source results in a decrease in $\overline{\breve{X}^2}$.

Expressing SQNR in decibels, one obtains

$$\text{SQNR}|_{\text{dB}} = P_{\breve{X}}|_{\text{dB}} + 6v + 4.8 \tag{5}$$

It is seen that each extra bit (increase in v by one) increases the SQNR by 6 dB.

3. COMPANDING

As long as the statistics of the input signal are close to the uniform distribution, a uniform quantizer works fine. However, in coding of certain signals such as speech, the input distribution is far from being uniformly distributed. For a speech waveform, in particular, there exists a higher probability for smaller amplitudes and lower probability for larger amplitudes. Therefore, it makes sense to design a quantizer with more quantization regions at lower amplitudes and less quantization regions at larger amplitudes. The resulting quantizer will be a nonuniform quantizer having quantization regions of various sizes.

The usual method for performing nonuniform quantization is to first pass the samples through a nonlinear element that compresses the large amplitudes (reduces dynamic range of the signal) and then perform a uniform quantization on the output. At the receiving end, the inverse (expansion) of this nonlinear operation is applied to obtain the sampled value. This techniques is called *companding* (compressing–expanding). A block diagram of this system is shown in Fig. 2.

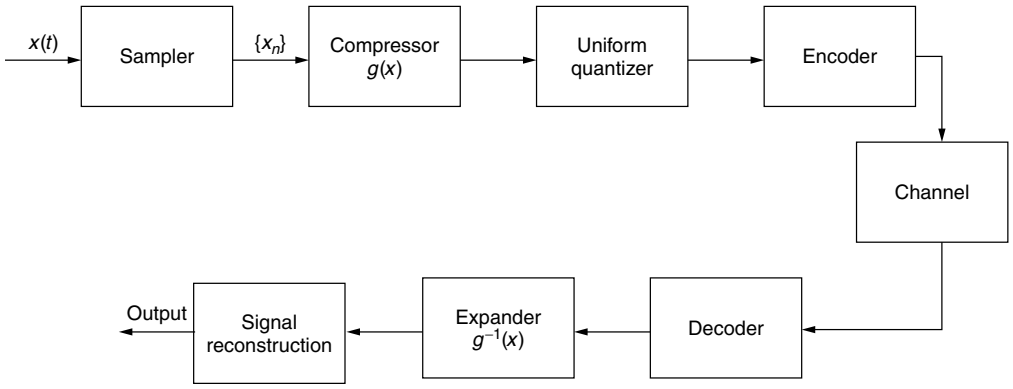


Figure 2. Block diagram of a PCM system employing a compander.

Two types of companders are widely used for speech coding. The μ-law compander used in the United States and Canada employs the logarithmic function at the transmitting side, where $|x| \leq 1$:

$$g(x) = \frac{\log(1 + \mu|x|)}{\log(1 + \mu)} \operatorname{sgn}(x) \tag{6}$$

The parameter μ controls the amount of compression and expansion. The standard PCM system in the United States and Canada employs a compressor with $\mu = 225$, followed by a uniform quantizer with 128 levels (7 bits per sample). Use of a compander in this system improves the performance of the system by ~30 dB. This means that the compander has implicitly provided an additional 5 bits of precision to that obtained by a uniform quantizer. A plot of the μ-law compander characteristics is shown in Fig. 3.

The second widely used logarithmic compressor is the A-law compander. The characteristic of this compander is given by

$$g(x) = \frac{1 + \log A|x|}{1 + \log A} \operatorname{sgn}(x) \tag{7}$$

where A is chosen to be 87.56. The performance of this compander is comparable to the performance of the μ-law compander. The characteristics of this compander are shown in Fig. 4.

4. CONCLUDING REMARKS

In digital coding of analog signals, such as speech signals, that have a nonuniform amplitude distribution, a nonuniform quantizer should be employed in order to reduce the number of bits per sample for a specified signal fidelity. Companding used in conjunction with a uniform quantizer results in a nonuniform quantization of the signal. In the case of speech signals, the μ-law or the A-law compander used in conjugation with a 7-bit/sample uniform quantizer result in an SQNR that is comparable to that obtained with a 12-bit/sample uniform quantizer without a compander. Thus, the use of a compander in the coding of speech signals via PCM has resulted in a significant reduction in the bit rate (by 5 bits per sample) that is to be transmitted over the channel.

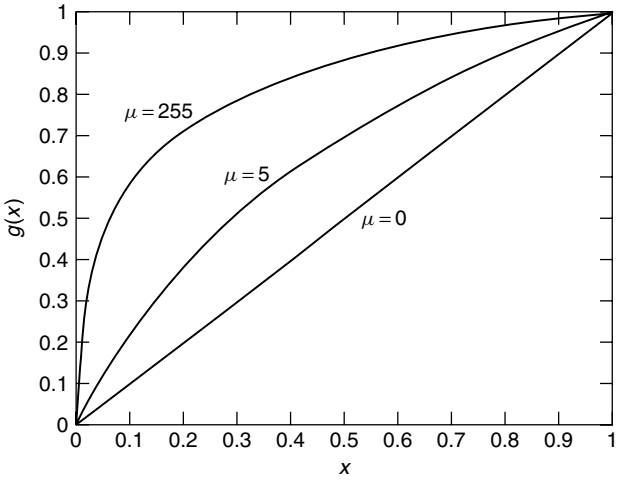


Figure 3. μ-law compander characteristics.

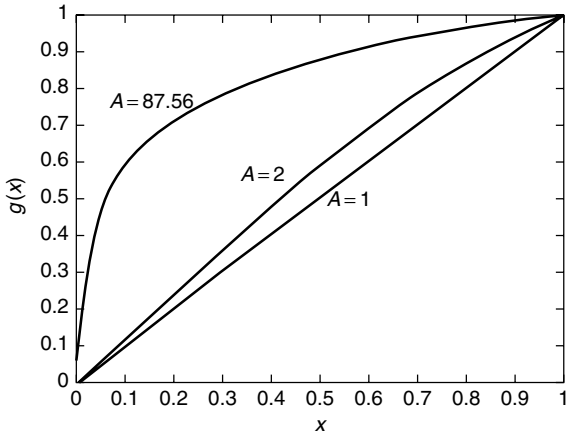


Figure 4. A-law compander characteristics.

For a detailed treatment of digital coding of signal waveforms, the reader may refer to the treatise by Jayant and Noll [1].

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing*

Laboratory (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. N. S. Jayant and P. Noll, *Digital Coding of Waveforms–Principles and Applications of Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

COMPENSATION OF NONLINEAR DISTORTION IN RF POWER AMPLIFIERS

SEKCHIN CHANG
EDWARD J. POWERS
University of Texas at Austin
Austin, Texas

1. INTRODUCTION

In all wireless communication systems, whether terrestrial or satellite, an RF power amplifier (PA) is required at the transmitter to ensure the signal will be received at the receiver with sufficient signal-to-noise ratio (SNR). For small signal level inputs, the PA input–output relationship is linear. However, as the input signal level is increased, the output power eventually saturates, causing the input–output relationship to become nonlinear [1].

On one hand, it is desirable to operate a PA near saturation because of increased RF power output (important in, e.g., satellite communications) and increased DC-to-RF power conversion efficiency (important in battery-powered RF devices, e.g., mobile phones). However, operating the PA near saturation results in a number of undesirable nonlinear effects, including amplitude and phase distortion and the generation of in-band and out-of-band intermodulation frequencies. These phenomena associated with nonlinearity lead to an increased bit error rate (BER), in-band interference, and adjacent-channel interference (ACI).

One approach to mitigating the undesirable effects associated with saturation is to back off from saturation into a more linear region, but at the expense of reducing the available RF power output and efficiency. For example, if the RF power output is backed off 10 dB (which is not unreasonable), the available RF output power will be reduced to one-tenth of the maximum output power. For these reasons there is a significant need to mitigate the effects of nonlinearities and, thereby, reduce the amount of backoff required.

To compensate for the deleterious effects of PA nonlinearities, one might try to compensate the distorted signal received at the receiver. Even when this can be done in an efficient and practical matter with some type of equalizer, it does not eliminate the undesirable effects occurring at the transmitter associated with the generation of intermodulation frequencies and the resultant in-band and out-of-band interference. Thus we focus on nonlinear compensation at the transmitter.

To mitigate the efforts of nonlinearities at the transmitter, it is customary to predistort the signal to be transmitted in a way that is equal and opposite to the distortion introduced by the PA. In such a case, the PA output should ideally correspond to the output of an ideal linear amplifier (i.e., one that is linear right up to the saturation point). Of course, no predistortion technique is perfect, so some backoff is required. The next question is how much backoff is required and how one addresses the tradeoff between large backoff (linear operation, but low RF power output) and small backoff (nonlinear operation, but high RF power output). In this article we will address the tradeoff issue via the well-known concept of total degradation versus output backoff in order to determine the optimum output power backoff.

Generally speaking, the deleterious effects of nonlinearities are less severe for constant-amplitude modulation schemes such as various forms of PSK (phase shift keying) (BPSK, QPSK, etc.) versus more bandwidth-efficient multiamplitude modulation schemes such as QAM (quadrature amplitude modulation). Thus in this article we will demonstrate the performance of predistorters using the more challenging case of QAM modulation.

Another factor to be considered involves the use of single-carrier (frequency) versus multicarrier systems. Multicarrier systems, such as OFDM (orthogonal frequency-division multiplexing) are increasingly being used in such systems as digital audiobroadcasting and future-generation personal communication systems because of their robustness to impulse noise and multipath fading. However, such multicarrier systems are characterized by high peak-to-average power ratios (PAR) because the multicarrier signals will occasionally constructively interfere. This constructive interference leads to high peak power levels that drive the PA into the saturation region with all its negative nonlinear consequences. This suggests that, in general, multicarrier systems require more output power backoff than does a single-carrier system. Thus the utilization of predistorters in multicarrier systems is the more challenging case and is, therefore, why we choose to demonstrate the use of predistorters using OFDM in this article.

In the next section we overview some of the nonlinear characteristics of power amplifiers. In Section 3 we consider the sensitivity of OFDM systems to nonlinear distortion, and in Section 4 provide an overview of various approaches to predistorters with emphasis on Volterra-based predistorters. In particular, the ability of the Volterra-based predistorter to reduce BER, total degradation, and output backoff is demonstrated via a simulation experiment using 16-QAM-based OFDM system. Conclusions are stated in Section 5.

2. NONLINEAR CHARACTERISTICS OF POWER AMPLIFIERS

The characteristics of PAs are usually expressed by amplitude modulation–amplitude modulation (AM/AM) and amplitude modulation–phase modulation (AM/PM) conversions, which represent the amplitude and phase distortions, respectively, depending on the input signal magnitude. If $x(n)$ is an input signal to the PA, it can be defined by

$$x(n) = r(n)e^{j\theta(n)} \tag{1}$$

where n denotes discrete time and $r(n)$ and $\theta(n)$ are the amplitude and the phase of the input signal $x(n)$, respectively. Let $A[\cdot]$ and $\Phi[\cdot]$ be AM/AM and AM/PM conversions, respectively. Therefore, if $s(n)$ is the output signal of the PA, it can be expressed as

$$s(n) = A[r(n)]e^{j\{\Phi[r(n)]+\theta(n)\}} \tag{2}$$

After the input signal $x(n)$ is amplified as in Eq. (2), the amplified signal $s(n)$ is transmitted into the wireless channel.

PAs can be broadly classified into traveling-wave tube amplifiers (TWTAs) and solid-state power amplifiers (SSPAs). TWTA PAs usually exhibit higher output power, but lower reliability and more severe nonlinearities than do SSPAs. Therefore, TWTAs are still dominant in applications requiring high levels of RF output power [2]. For TWTAs, the AM/AM and AM/PM conversions are modeled by the following equations [3]

$$A[(r(n)] = \frac{\alpha_a r(n)}{1+\beta_a r^2(n)} \qquad \text{(AM/AM conversion)} \tag{3}$$

$$\Phi[(r(n)] = \frac{\alpha_p r^2(n)}{1+\beta_p r^2(n)} \qquad \text{(AM/PM conversion)} \tag{4}$$

where α_a, β_a, α_p, and β_p are constants, and the subscripts a and p denote amplitude and phase, respectively. Figure 1 illustrates the curves of the AM/AM and AM/PM conversions for the typical values of constants: $\alpha_a = 2.0$, $\beta_a = 1.0$, $\alpha_p = \pi/3$, and $\beta_p = 1.0$ [3]. Note that the input and output amplitudes are normalized to their respective saturation values.

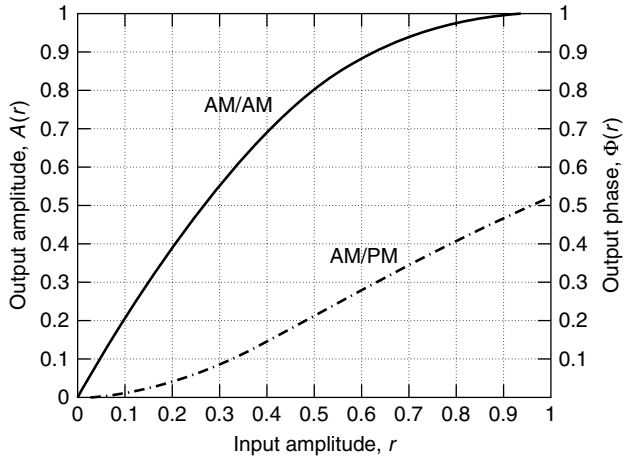


Figure 1. The AM/AM and AM/PM conversions of a traveling-wave tube amplifier (TWTA).

Unlike phase distortions in TWTAs, those in SSPAs are usually smaller [4]. In an SSPA AM/AM conversion is modeled as follows [4]:

$$A[(r(n)] = \frac{v_k r(n)}{\left(1+\left(\frac{v_k r(n)}{A_0}\right)^{2p_k}\right)^{1/2p_k}} \qquad \text{(AM/AM conversion)} \tag{5}$$

where v_k, A_0, and p_k are constants. TWTAs exhibit higher power gain and more severe nonlinearities than do SSPAs [5]. In the ideal case of either TWTAs or SSPAs, the AM/AM conversion curve should be a straight line with a constant slope, and the AM/PM conversion curve should be a constant.

Regardless of the kind of PA, the transmitted signal experiences some nonlinear distortion as indicated in Eq. (2). If the input signal $x(n)$ utilizes a simple modulation scheme such as binary phase shift keying (BPSK) or quadrature phase shift keying (QPSK), the nonlinear distortion has mild effects on the transmitted signal since these modulation schemes exhibit constant amplitude. However, if the input signal corresponds to a channel-efficient modulation scheme such as 16- or 64-QAM (quadrature amplitude modulation) where the amplitude varies, the nonlinear distortion will degrade the system performance severely. Moreover, multicarrier systems such as OFDM systems are more sensitive to nonlinear distortion as explained in Sections 1 and 3.

The degree of PA nonlinear distortion depends on the backoff amount. Backoff may be divided into input backoff (IBO) and output backoff (OBO) which are defined (in decibels) as Eqs. (6) and (7), respectively:

$$\text{IBO} = 10\log_{10}\frac{P_{i,\max}}{P_i} \tag{6}$$

$$\text{OBO} = 10\log_{10}\frac{P_{0,\max}}{P_0} \tag{7}$$

In Eq. (6), $P_{i,\max}$ represents the maximum input power of the PA at saturation and P_i denotes the mean input power of the signal at the PA input. Similarly, in Eq. (7) $P_{0,\max}$ represents the maximum output power of the PA at saturation and P_0 denotes the mean output power of the signal at the PA output. Figure 2 depicts the relationship between IBO and OBO using the AM/AM characteristics of a TWTA PA. Therefore, the operating point can be placed in the linear region of PA by backing off the PA from saturation, which leads to an effective reduction of nonlinear distortion. However, this scheme reduces the power efficiency of the PA and its output power. Therefore, a tradeoff between backoff and nonlinear distortion must be considered.

3. SENSITIVITY OF OFDM SYSTEMS TO PA NONLINEAR DISTORTION

Currently, OFDM is utilized for asymmetric digital subscriber line (ADSL), digital audiobroadcasting (DAB), and wireless local-area networks (WLANs) [6,7]. In

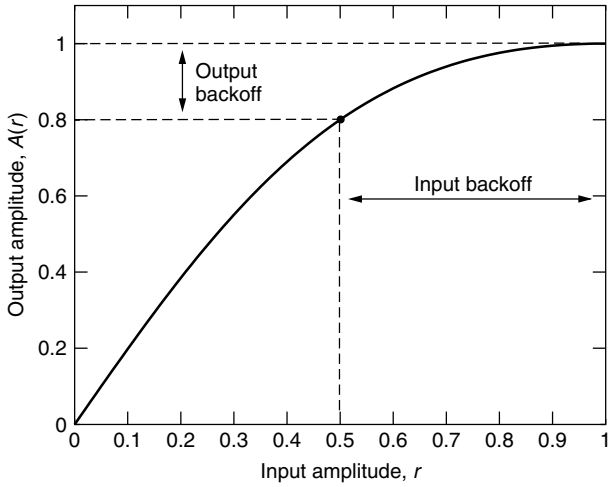


Figure 2. Input and output backoffs for a TWTA.

addition, OFDM is expected to be a strong candidate for wireless multimedia systems. Therefore, OFDM is believed to be a promising technique in wireless as well as wireline systems for high-rate data transmission [8]. OFDM utilizes a multicarrier modulation scheme, and exhibits many advantages over single-carrier modulation. Because of the increased symbol length, OFDM is robust to the effects of impulse noise and severe multipath fading [8,9]. In addition, the effects of the frequency selective fading can be greatly reduced without a high-cost equalizer by using a cyclic prefix.

However, multicarrier systems such as OFDM show great sensitivity to nonlinear distortion introduced by the PA. OFDM and other systems usually have a PA at the RF stage on the transmitter side as shown in Fig. 3 to increase the link fading margin of OFDM signals in a wireless channel. In Fig. 3, the incoming symbol input is a serial stream such as M-ary quadrature amplitude modulation (QAM) signals. After the input symbols are converted from serial to parallel (S/P) to form a vector of N M-ary symbols, the N symbols are modulated onto N subcarriers by the inverse fast Fourier transform (IFFT) and subsequently converted back from parallel to serial (P/S) data symbols. As mentioned by Karam and Sari [10], digital radio systems usually employ baseband pulseshaping at the transmitter. Therefore, the linear filter in Fig. 3 indicates the pulse shaping filter.

Since the input data to the IFFT are generally independent and identically distributed (i.i.d.) in OFDM systems, the modulated signals can be considered random signals with a zero-mean Gaussian distribution according to central limit theorem. Therefore, the signals of OFDM systems exhibit higher peak-to-average power ratio (PAR) than do those of single-carrier systems in the time domain. Therefore, OFDM systems are more sensitive to PA nonlinear distortion than are single-carrier systems because of their larger PAR. In other words, more frequent utilization of the PA's high-power region is inevitable in OFDM systems in order to produce an average output power comparable to that of single-carrier systems, thus resulting in severe performance degradation due to nonlinear distortion introduced by the PA.

4. COMPENSATION METHODS OF NONLINEAR DISTORTION

Nonlinear distortion can be reduced by backing off the PA from saturation. However, as stated earlier, back-off reduces the PA output power. Thus, some form of nonlinear compensation is desirable to mitigate nonlinear distortion while at the same time reducing the amount of backoff required. For the efficient compensation of nonlinear distortion, predistorters have been widely utilized, where the predistorter is placed in front of the PA. The predistorter distorts the input signal in such a way as to compensate for the nonlinear distortion introduced by the PA. Moreover, predistorters may be designed to be adaptive. The adaptation property is very desirable because the characteristics of PAs are time variant due to temperature variation and aging [3].

4.1. Some General Predistortion Schemes

In general, predistorters can be classified into several types according to the predistorter structure and position. Usually, predistorters belong to a minimum mean-square error (MMSE) type or amplitude/phase (AP) type based on predistorter structure. According to the location where the predistorter is placed, it is commonly called an analog or digital (or data) predistorter.

Figure 4 shows an MMSE predistorter. As depicted in this figure, the coefficients of the predistorter are trained to minimize the mean square error between output $y(n)$ of PA and input $x(n)$ to the predistorter. An MMSE predistorter has been proposed [11] and its performance analyzed in OFDM systems [12]. In these articles, the input and output responses were related with third-order polynomials. Generally, the MMSE predistorter can achieve a global compensation of the nonlinear distortion

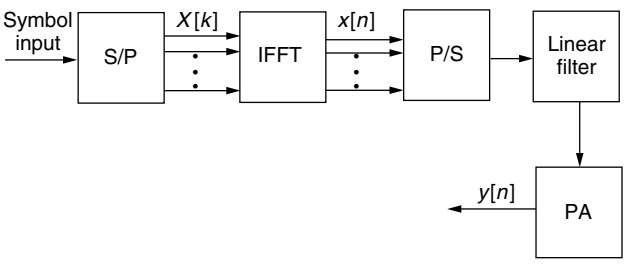


Figure 3. An OFDM system with PA.

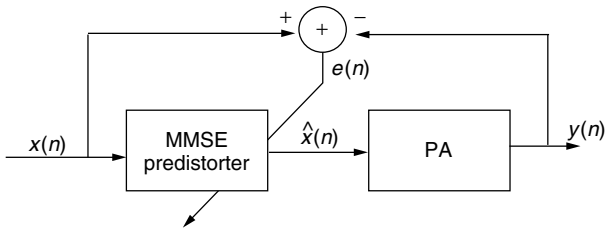


Figure 4. The general structure of a MMSE predistorter.

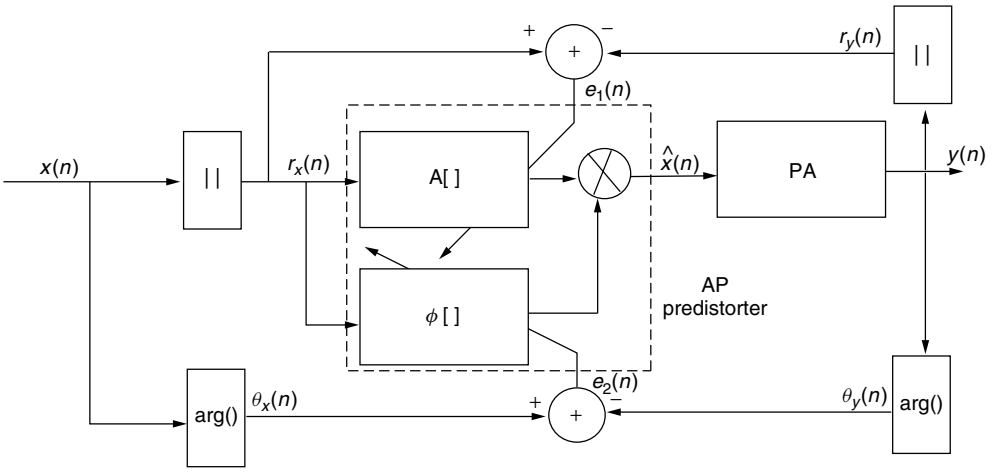


Figure 5. The general structure of an AP predistorter.

of the PA if a proper learning architecture is employed. On the other hand, the MMSE approach requires complex-valued coefficients in the predistorter to compensate for amplitude and phase distortion.

The amplitude/phase (AP) predistorter is illustrated in Fig. 5, which shows that the predistorter consists of amplitude $A[\cdot]$ and phase $\Phi[\cdot]$ coefficients. In other words, the coefficients of each type are separately trained so as to minimize the amplitude and the phase errors between output $y(n)$ of the PA and input $x(n)$ of the predistorter. The AP predistorter was introduced by D'Andrea et al. [13] and its performance was measured in OFDM systems [14]. In these articles, the linearizing functions $A[\cdot]$ and $\Phi[\cdot]$ of the predistorter were separately approximated as polynomials. Usually, each polynomial has real-valued coefficients. Therefore, the AP predistorter can exhibit a faster convergence rate than does the MMSE predistorter in training the predistorter coefficients. However, the predistorter requires additional separation and combination modules for the amplitude/phase of the input signal $x(n)$ because it must obtain separate optimal solutions to compensate for the amplitude and phase distortions of the PA.

Figure 6 depicts general structures of analog and digital (or data) predistorters. As shown in this figure, the analog predistorter is placed after the pulseshaping filter in the RF (or possibly IF) band. On the other hand, the digital predistorter is placed before the pulseshaping filter at baseband. Therefore, while the analog predistorter just compensates the memoryless nonlinearity of the PA, the digital predistorter is required to compensate a nonlinearity with memory due to the series combination of the linear pulseshaping filter, which provides memory, and the PA, which is essentially a memoryless nonlinearity. However, the digital predistorter exhibits more flexibility in determining the predistorter coefficients than the analog predistorter, since the learning algorithm is programmable in the digital predistorter. This suggests that the digital predistorter can be more adaptive to time-variant PAs.

In the following sections, we describe several predistorters that have shown good performance in compensating nonlinear distortion introduced by the PAs. We will emphasize the Volterra-based predistorter, which has been utilized for compensation of nonlinear distortion in OFDM systems. The Volterra-based predistorter belongs to the MMSE and digital class of predistorters. As indicated above, such a predistorter type has a global solution and training flexibility, but also demands a lot of learning time. Therefore, we will also indicate an efficient learning architecture for the predistorter.

4.2. Predistortion Using a Lookup Table

This efficient predistortion scheme has been applied to QAM radio systems with good results [10]. The scheme utilized a lookup table or random-access memory (RAM), which included the predistorter candidates. If an input sequence to the predistorter is given, the candidate that minimizes the error between the input data to the predistorter and the output data of the PA is selected as the output of the predistorter. Each candidate is updated until the error converges to a desired value. Since the input signals exhibit a finite number of data levels regardless of the modulation scheme used in single-carrier systems, the lookup table predistortion scheme can be easily adapted to various PAs. However, it is probably not feasible to

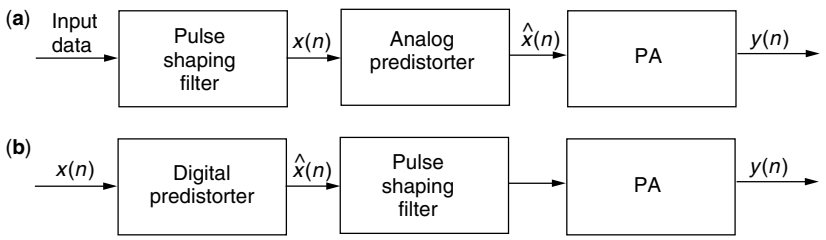


Figure 6. The general structures of analog and digital predistorters for (**a**) analog predistorter and (**b**) digital predistorter.

utilize this scheme in multicarrier systems because these systems usually show many different data levels due to the constructive and destructive interference of the multicarrier components.

4.3. Volterra-Based Predistorter

This adaptive predistorter essentially consists of a Volterra series because the Volterra series has shown excellent performance in modeling and compensating nonlinear phenomena with memory. Thus, it is particularly suitable for use as a digital predistorter. In discrete time, a third-order Volterra series for a causal, finite memory system becomes

$$y[n] = \sum_{k=0}^{N_1-1} h_k^{(1)} x[n-k] + \sum_{k=0}^{N_2-1}\sum_{l=0}^{N_2-1} h_{k,l}^{(2)} x[n-k]x^*[n-l] + \sum_{k=0}^{N_3-1}\sum_{l=0}^{N_3-1}\sum_{m=0}^{N_3-1} h_{k,l,m}^{(3)} x[n-k]x[n-l]x^*[n-m] + e[n] \tag{8}$$

where N_1, N_2, and N_3 respectively denote the memory duration of the first-order, the second-order, and the third-order terms; $x[n]$ and $y[n]$ are the complex input and output, respectively; $h_k^{(1)}$, $h_{k,l}^{(2)}$, and $h_{k,l,m}^{(3)}$ are the complex discrete time domain Volterra kernels of order 1, 2, 3, respectively; and $*$ and $e[n]$ denote the complex conjugate and the modeling error, respectively. In Fig. 6b, the PA is preceded by a linear filter. In digital communication systems, the combination of the PA and linear filter may be regarded as a nonlinear system with memory that is to be compensated [15]. Since the Volterra series may be regarded as a Taylor series with memory, a Volterra-based predistorter exhibits a structure suitable for compensation of such a system.

We may represent the third-order Volterra series of Eq. (8) in matrix form as

$$\hat{d}[n] = \mathbf{h}\mathbf{x}^T[n] \tag{9}$$

where $\hat{d}[n]$ is the estimated output of the Volterra predistorter, the superscript T denotes the transpose of the matrix, $\mathbf{h}$ is the Volterra kernel vector, and $\mathbf{x}[n]$ is the input vector, which are defined by

$$\mathbf{h} = [h_0^{(1)}, h_1^{(1)}, \ldots, h_{N_1-1}^{(1)}, h_{000}^{(3)}, h_{001}^{(3)}, h_{002}^{(3)}, \ldots, \times h_{klm}^{(3)}, \ldots, h_{(N_3-1)(N_3-1)(N_3-1)}^{(3)}] \tag{10}$$

$$\mathbf{x}[n] = [x[n], x[n-1], \ldots, x[n-N_1+1], |x[n]|^2x^*[n], \times |x[n]|^2x^*[n-1], |x[n]|^2x^*[n-2], \ldots, \times |x[n-N_3+1]|^2x^*[n-N_3+1]] \tag{11}$$

where N_1 and N_3 are the memory durations of the first-order term and the third-order term, respectively. In Eqs. (10) and (11), the absence of the second-order term is due to the fact that even-order intermodulation components do not interfere with the in-band signal for a bandpass nonlinear channel [16], since they lie out of band. However, odd-order nonlinearities generate intermodulation frequency components that lie both out of band and in band. Because of this latter fact, odd-order terms are retained in the Volterra series.

Figure 7 shows why only the odd-terms contribute to the bandpass channel. In communication systems, the PA is placed at the RF stage. Therefore, the signal $x(t)$ is upconverted to the carrier frequency f_0. The upconverted signal $r(t)$ is inputted to the nonlinear PA, thus various harmonics of the carrier frequency f_0 appear in the output. As Fig. 7 indicates, only the odd-order nonlinearities contribute to in-band components centered at f_0.

It is well known that Volterra kernels can be assumed symmetric without any loss of generality [17]. Therefore, the third term in Eq. (8) can be rewritten as follows:

$$y_3[n] = \sum_{k=0}^{N_3-1}\sum_{l=k}^{N_3-1}\sum_{m=0}^{N_3-1} h_{k,l,m}^{(3)} x[n-k]x[n-l]x^*[n-m] \tag{12}$$

Taking into account symmetry, the number of kernel coefficients for each order term will be

$$K_1 = N_1 \tag{13}$$

$$K_3 = \frac{N_3^2(N_3+1)}{2} \tag{14}$$

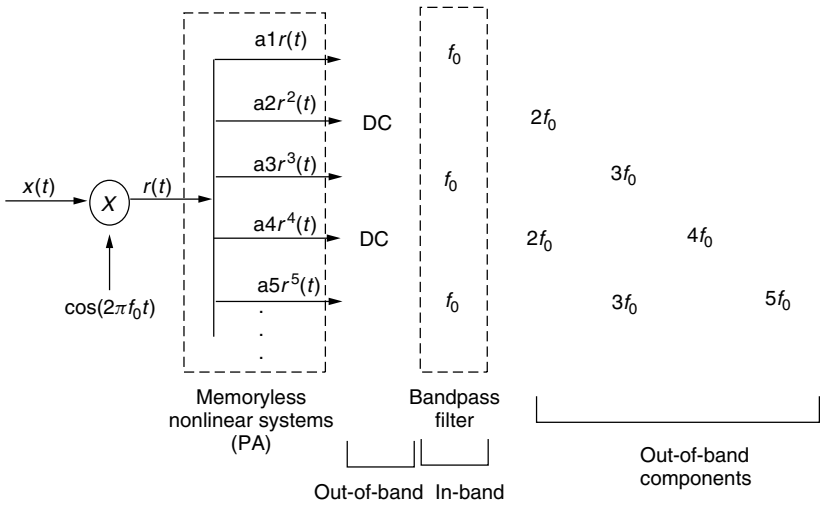


Figure 7. Abscence of even-order terms in a bandpass channel.

and the total number of Volterra coefficients for the predistorter can be represented as

$$K_T = K_1 + K_3 \tag{15}$$

The next challenge is to determine the numerical values of the Volterra coefficients for a given PA.

4.4. Learning Architecture for Volterra-Based Predistorter

The Volterra-based predistorter belongs to the MMSE and digital classes of predistorters. A particular challenge associated with this predistorter (and indeed most predistorters) is to determine the Volterra coefficients, since the desired output of the predistorter is not readily known beforehand. Thus, the predistorter design requires an efficient learning architecture to train the predistorter coefficients rapidly.

A relatively new and efficient training architecture [18] composed of both the indirect and direct learning algorithms is depicted in Fig. 8. In earlier work [15], only the indirect learning scheme was utilized. As seen in Fig. 8, there are two identical models: the actual predistorter and another for training. The two models share the same predistorter coefficient vector **h**. As stated previously, the "PA with memory" in Fig. 8 represents the PA preceded by a linear filter. In the learning structure of Fig. 8, at each iteration the predistorter coefficients are first updated using the indirect learning algorithm, which makes $\alpha[n]$ approach zero, and are then updated using the direct learning algorithm, which makes $\beta[n]$ approach zero. As $\alpha[n]$ and $\beta[n]$ approach zero, $y[n]$ approaches $x[n]$, which implies complete compensation of the PA's nonlinear distortion. These updates continue until the coefficients converge. Since the coefficients are updated twice in each iteration, this training scheme exhibits rapid convergence. Both the indirect and direct learning algorithms utilize the recursive least-squares (RLS) technique [19] for fast updates of the predistorter coefficients. A particular advantage of the approach described here is that one does not first require a Volterra model of the system to be compensated.

Figure 9 shows the learning curves (mean squared error (MSE) vs the number of iterations) for two Volterra-based adaptive predistorters, one of which was trained by the new learning scheme, direct and indirect architecture (DIA), and the other by the old learning scheme, indirect architecture-only (IAO) methods. From this figure, it is seen that the DIA scheme achieves an improvement of up to 5 dB or more in the MSE performance over the IAO scheme for 2000 or more iterations. This result signifies that the new learning algorithm is effective in increasing the convergence rate of the Volterra-based adaptive predistorter.

4.5. Simulation Experiments

In this section we carry out a simulation experiment to give the reader some feeling for the type of advantages one can achieve using predistorters.

An OFDM system with 128 subcarriers and 16-QAM symbols is considered. In this simulation, it is assumed that the nonlinear degree is 3 and the memory length of the linear filter shown in Fig. 3 is 3. Therefore, a third-order Volterra-based predistorter is used, and the first-order memory length (N_1) and the third-order memory length (N_3) for the Volterra predistorter are set to 3 in this simulation. From Eq. (15), the total number of the predistorter coefficients is 21. As mentioned in Section 2, the TWTA is known to be more nonlinear than the SSPA. Thus, we utilize a TWTA in this simulation because it poses a more severe challenge for the predistorter.

Received 16-QAM constellations of the OFDM system with and without the Volterra predistorter are shown in Fig. 10, where E_b/N_0 (energy per bit divided by the

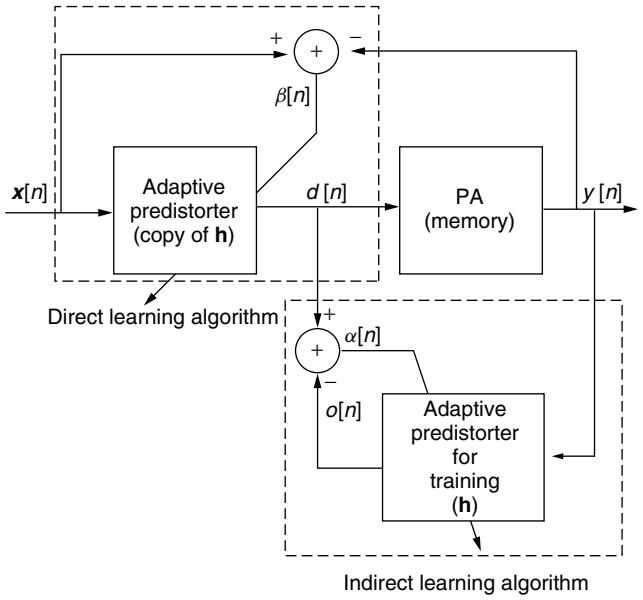


Figure 8. The learning architecture for a Volterra-based predistorter. ([18], © 2000 IEEE).

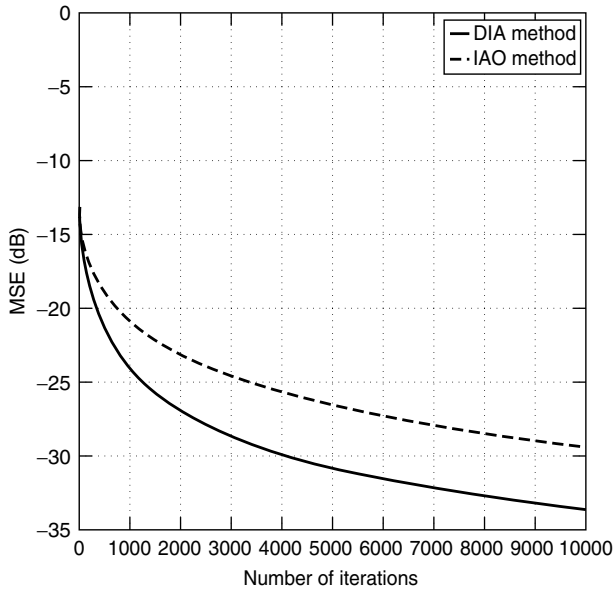


Figure 9. Learning curves of predistorters based on DIA and IAO ([18], © 2000 IEEE).

noise power spectral density) is assumed to be ∞. In the figure captions, the normalized mean-squared error (NMSE) values are also given for numerical comparisons of distorted or compensated constellations. Figures 10a,c show the distorted received signal constellations (no predistortion compensation) at output backoffs (OBOs) of 8.5 and 4.7 dB, respectively. Figures 10b,d show the received signal constellations when compensated by the proposed Volterra predistorter for OBOs of 8.5 and 4.7 dB, respectively. Comparison of Figs. 10a and 10b indicates that the signal distortion is almost perfectly compensated by the Volterra predistorter when the OBO is 8.5 dB. Figure 10d also shows some compensation of the signal distortion seen in Fig. 10c for an OBO of 4.7 dB. The relatively large clustering of Fig. 10d is due to the fact that the predistorter is driven into the saturation region by the large envelope fluctuations characteristic of multicarrier systems since the OBO of 4.7 dB is relatively small.

Therefore, we can see from this simulation result that the performance of the Volterra-based predistorter in compensating nonlinear distortion of PA is dependent on the OBO values. As a performance measure to investigate the tradeoff between nonlinear distortion and OBO in the PA, the total degradation (TD) is usually utilized and defined in decibels by

$$\mathrm{TD} = \left[\frac{E_b}{N_{0_{(\mathrm{NLPA})}}} - \frac{E_b}{N_{0_{(\mathrm{LPA})}}}\right] + \mathrm{OBO} \quad (\mathrm{dB}) \tag{16}$$

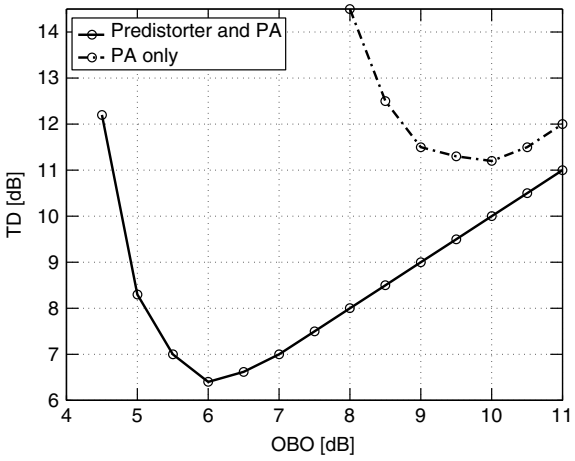


Figure 11. Total degradation versus OBO at a BER of 10^{-4} for the PA with predistorter and the PA-only ([18], © 2000 IEEE).

where $E_b/N_{0_{(\mathrm{NLPA})}}$ represents the required E_b/N_0 to obtain a specific BER when the nonlinear PA is used, and $E_b/N_{0_{(\mathrm{LPA})}}$ denotes the required E_b/N_0 to maintain the same BER, assuming that the amplifier exhibits no nonlinearities. In Fig. 11, the TD is shown for various OBO values for a BER of 10^{-4}. This figure illustrates the cases for the PA with predistorter and for the PA-only. The OBO which minimizes TD is called the *optimum OBO*. As seen in the figure, the minimum TD for the PA-only is 11.2 dB at an OBO of 10 dB, and the minimum TD of the PA with

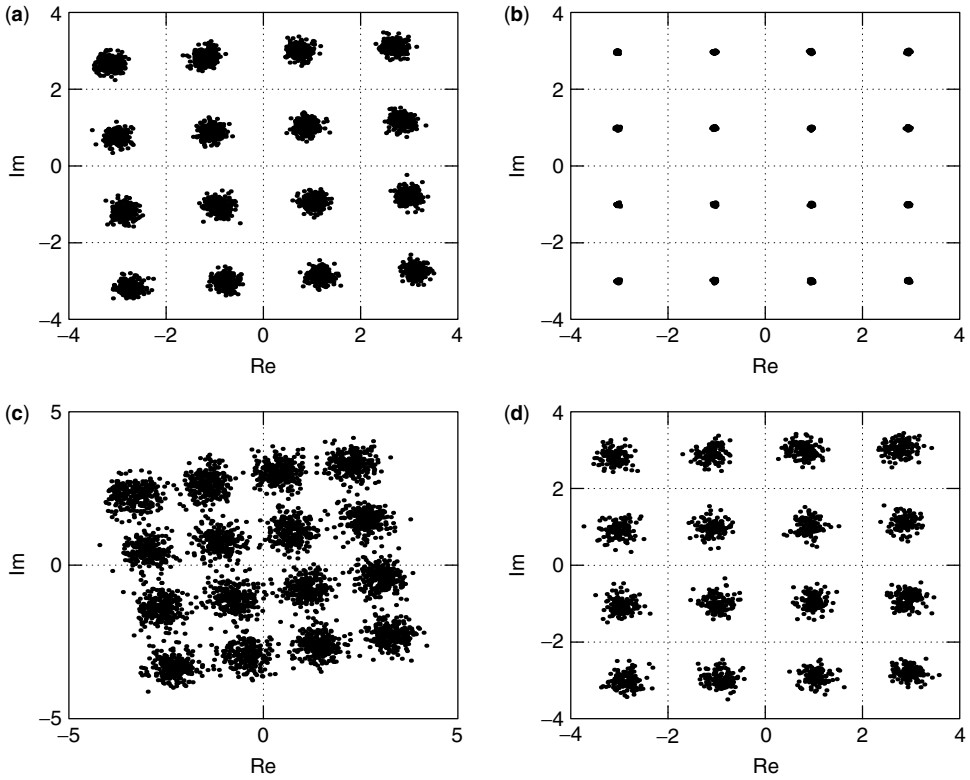


Figure 10. Received 16-QAM constellations for (**a**) PA-only (NMSE = 0.00913) at OBO = 8.5 dB, (**b**) predistorter and PA (NMSE = 0.000071) at OBO = 8.5 dB, (**c**) PA-only (NMSE = 0.0533) at OBO = 4.7 dB, and (**d**) predistorter and PA (NMSE = 0.00825) at OBO = 4.7 dB, respectively. ([18], © 2000 IEEE).

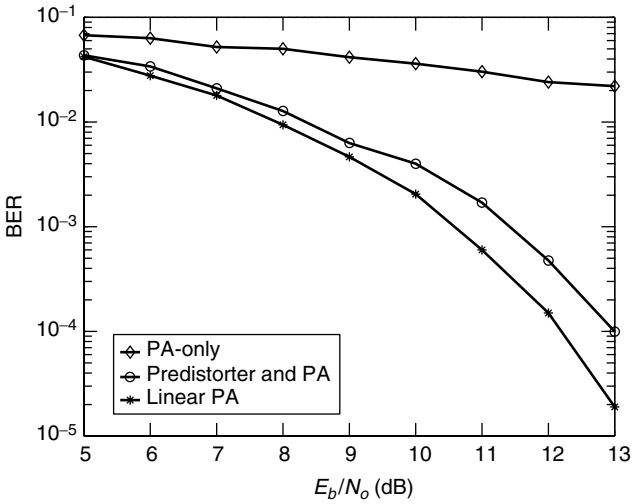


Figure 12. BER performance versus E_b/N_0 at OBO of 6.0 dB for the PA-only, the PA with Volterra-based adaptive predistorter, and the linear PA.

predistorter is 6.4 dB at an OBO of 6.0 dB. Therefore, the PA with predistorter achieves an output power increase of about 4 dB (10 dB − 6 dB) over PA-only case. Furthermore, the TD is reduced by 4.8 dB (11.2 dB − 6.4 dB).

Figure 12 compares the bit error rate (BER) performances for the cases of the PA without any predistortion and with the Volterra-based predistorter, and the ideal case without any nonlinear distortion of the PA (linear PA) at an OBO of 6.0 dB. In Fig. 12, it is shown that the BER performance is severely degraded as a result of nonlinear distortion in the case without predistortion; that is, increasing E_b/N_0 has a relatively small effect on reducing the BER. On the other hand, the BER performance of the combination of the Volterra-based predistorter and PA is fairly close to that of an ideal linear PA, thereby demonstrating the efficacy of the predistorter.

5. CONCLUSION

In this brief article the advantages of using predistorters to mitigate the effects of nonlinear distortion introduced by PAs and to reduce the amount of output backoff have been stressed and demonstrated via a predistorter example using a Volterra-based predistorter. In conclusion, it should be reiterated that there are many approaches to predistorters, some of which were mentioned in this article, and others of which are cited in the bibliography.

BIOGRAPHIES

Edward J. Powers received his B.S., M.S., and Ph.D. degrees from Tufts University, Massachusetts Institute of Technology, and Standford University in 1957, 1959, and 1965, respectively. All degrees were in electrical engineering. From 1959 to 1965 Dr. Powers was employed by Lockheed Missiles and Space Company in Sunnyvale and Palo Alto, California. In 1965, he joined the University of Texas at Austin, where he subsequently became Chair of the Department of Electrical and Computer Engineering from 1981 to 1989. He is currently the Texas Atomic Energy Research Foundation Professor in Engineering and Director of the Telecommunications and Signal Processing Research Center. His current professional interests include applications of higher-order statistical signal processing to detect, analyze, and model time series data associated with nonlinear physical phenomena; and applications of wavelets and time-frequency techniques to detect and classify various transient events in physical systems. He was elected a Fellow of IEEE in 1983.

Sekchin Chang received the B.S. and the M.S. degrees in electronics engineering from Korea University, Seoul, Korea in 1991 and 1993, respectively, and the Ph.D. degree in electrical engineering from the University of Texas at Austin in 2001. From 1993 to 1998, he was with Electronics and Telecommunications Research Institute (ETRI), Taejon, Korea, where he worked on the design and development of IS95 CDMA systems. In 2000, he joined Motorola, Austin, Texas, where he contributed to the design of modems for WCDMA systems, and is currently involved in the development of WLAN systems. His research interests include OFDM, WCDMA, adaptive predistortion, carrier and timing recovery, RAKE receivers, blind equalizers, and MIMO systems.

BIBLIOGRAPHY

1. S. Benedetto, E. Biglieri, and V. Castellani, *Digital Transmission Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
2. N. Escalera, W. Boger, P. Denisuk, and J. Dobosz, Ka-band, 30 watts solid state power amplifier, *Proc. IEEE MTT-S Int. Microwave Symp.*, Boston, June 2000, Vol. 2, pp. 561–563.
3. A. A. M. Saleh, Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers, *IEEE Trans. Commun.* **29**(11): 1715–1720 (Nov. 1981).
4. E. Bogenfeld, R. Valentin, K. Metzger, and W. Sauer-Greff, Influence of nonlinear HPA on trellis-coded OFDM for terrestrial broadcasting of digital HDTV, *Proc. 1993 IEEE Global Telecommun. Conf.*, Houston, TX, Nov. 1993, pp. 1433–1438.
5. M. Obermier and E. J. Powers, The effects of nonlinear high power amplifiers on space based phased array antenna patterns, *Proc. IEEE Int. Conf. Phased Array Systems and Technology*, Dana Point, CA, May 2000, pp. 45–48.
6. *Asymmetric Digital Subscriber Line* (*ADSL*) *Metallic Interface*, ANSI T1.413, 1995.
7. M. Alard and R. Lassalle, Principles of modulation and channel coding for digital broadcasting for mobile receivers, *EBU Rev. Techn.* **224**: 168–190 (Aug. 1987).
8. L. J. Cimini, Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing, *IEEE Trans. Commun.* **33**(7): 665–675 (July 1985).
9. J. A. C. Bingham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* **28**: 5–14 (May 1990).
10. G. Karam and H. Sari, A data predistortion technique with memory for QAM radio systems, *IEEE Trans. Commun.* **39**(2): 336–344 (Feb. 1991).

11. M. G. Di Benedetto and P. Mandarini, A new analog predistortion criterion with application to high efficiency digital radio links, *IEEE Trans. Commun.* **43**: 2966–2974 (Dec. 1995).

12. M. G. Di Benedetto and P. Mandarini, An application of MMSE predistortion to OFDM systems, *IEEE Trans. Commun.* **44**: 1417–1420 (Nov. 1996).

13. N. A. D'Andrea, V. Lottici, and R. Reggiannini, RF power amplifier linearization through amplitude and phase predistortion, *IEEE Trans. Commun.* **44**: 1477–1484 (Nov. 1996).

14. N. A. D'Andrea, V. Lottici, and R. Reggiannini, Nonlinear predistortion of OFDM signals over frequency-selective fading channels, *IEEE Trans. Commun.* **49**: 837–843 (May 2001).

15. C. Eun and E. J. Powers, A predistorter design for memoryless nonlinearity preceded by a dynamic linear system, *Proc. 1995 IEEE Global Telecommunications Conf.*, Singapore, Nov. 1995, pp. 152–156.

16. S. Benedetto and E. Biglieri, Nonlinear equalization of digital satellite channels, *IEEE J. Select. Areas Commun.* **SAC-1**(1): 57–62 (Jan. 1983).

17. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, Wiley, New York, 1980.

18. S. Chang and E. J. Powers, A compensation scheme for nonlinear distortion in OFDM systems, *Proc. 2000 IEEE Global Telecommunications Conf.*, San Francisco, Nov. 27–Dec. 1, 2000, pp. 736–740.

19. S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.

COMPUTER COMMUNICATIONS PROTOCOLS

EMMANOUEL VARVARIGOS
University of Patras
Patras, Greece

THEODORA VARVARIGOU
National Technical University
Patras, Greece

1. WHAT IS A COMMUNICATION PROTOCOL?

In the "Information Age," concepts such as "communications," "information sharing," and "networking" sometimes seem to monopolize the attention of not only the information technology (IT) experts but also the ever-growing IT end-user community. In this context, "networking" denotes the information exchange between computers or like systems, which can be very different from one another and may be scattered over wide geographic areas.

Outside the network are the computers, databases, terminals, and other devices that need to exchange information. Messages originate at these external hosts, pass into the network, pass from node to node on the communication links, and finally pass out to the external recipients. The nodes of the network, usually computers in their own right, serve primarily to forward the messages through the network.

In order for such external systems to communicate, a common "language" has to be established among them. At the human level, when two people want to communicate (effectively), they have to agree first on a language that both understand, and then on a set of rules that both have to follow; for instance, they must not speak simultaneously (or nobody will be heard), and they have to adjust their voice volume to the environmental conditions (so that voice will not fade away before it reaches the other's ears). All these rules, conventions, and distributed algorithms that communicating parts have to follow are referred to, in total, as a "protocol." Without an agreed-on protocol, communication may be hard or impossible to accomplish.

Similarly, every computer communication is governed by a certain set of rules that have to be agreed on by all the partners involved, prior to communication establishment. These rules make up the protocol of a computer communication and, typically, include items such as the data format to be used, the order in which messages are exchanged, the actions to be taken on receipt of a message, the error detection and handling techniques to be employed, and so on.

From the real world, we already know that a communication task may be rather complex. For example, imagine the case where the commander of the "BLUE" army in a battlefield has to communicate an "attack plan" to the commander of the allied "GREEN" army, who has a different nationality, speaks a different language, probably has a different perception of operations, and is located at the other side of the battlefield, with their common enemy (the "RED" army) operating in between them. To ensure that the "attack order" is correctly received and perceived by the GREEN commander, the BLUE commander has to undertake a number of subtasks. For example:

1. Use a language that the recipient of the message will understand. This can include the use of a mutually accepted language (e.g., English) and the use of a standard terminology to reflect a certain military concept of operations.
2. Use an agreed-on format to compose the message (i.e., use a certain military messaging structured format).
3. Encrypt the message with an appropriate code, so that nobody can read the original contents except for the recipient of the message, who has the proper code to decrypt it.
4. Print the encrypted message on a piece of paper and then put the paper in a stiff envelope, to protect it from environmental conditions, and probably write the recipient's name on it.
5. Give the envelope to a messenger who will carry it to its destination. The messenger must find the appropriate route to follow to arrive at the destination and avoid ambushes while crossing the hostile (RED) ground. En route to the destination, the messenger may have to modify the route, due to an unpredicted roadblock, and thus must know how to use a map, compass, or other device to reorientate.
6. The messenger, on reaching the allied campus, must find the qualified person to hand over the envelope. This person may be the GREEN commander in person, or the commander's authorized secretariat.
7. The GREEN commander or this secretariat signs an appropriate document (a "receipt"), acknowledging

receipt of the envelope, and hands it over (enclosed in a suitable envelope) to the messenger soldier.

8. The messenger then must carry the receipt back to the home campus, again crossing the RED ground, probably following a different return path.
9. The messenger, on reaching the home campus, hands over the receipt envelope to the commander (or to qualified personnel and is then dismissed).

To make things a little closer to reality, suppose now that the messenger is killed while crossing the enemy lines to carry the attack plan, and that if the two allied armies attack simultaneously, the allies win, but if they attack separately, the enemy wins. The two allied army commanders would therefore like to synchronize their attack at some given time, but each of them is unwilling to attack unless assured with certainty that the other will also attack. Thus, the first commander might send a message saying, "Let's attack on Saturday at noon; please acknowledge if you agree." The second commander, hearing such a message, might send a return message saying, "We agree; send an acknowledgment if you receive our message." It is not hard to see that this strategy leads to an infinite sequence of messages, where the last commander who sends a message is unwilling to attack until obtaining a commitment from the other side. What is more surprising is that it can be shown that no strategy (protocol) exists for allowing the two armies to synchronize with certainty. If the conditions are relaxed somewhat so as to require a high probability of simultaneous attack, the problem is solved (e.g., the first army decides on the time of the attack and sends many messengers simultaneously to the other army).

This scenario illustrates the difficulties that arise in the design of protocols, where distributed decisions based on distributed information must be made. We see how complex a communication task may get, how many subtasks have to be considered, and that protocols can be proved to be correct only in a specific, well-defined sense. Computer communications follow, in general terms, principles similar to those described in the previous real-life communication example. Those principles are enforced by certain rules and conventions that make up the various computer communications protocols and are implemented by suitably designed hardware and software components located anywhere between the communicating partners (including themselves).

2. THE OSI MODEL

Since a communication task across a computer network can be too complicated to be efficiently controlled as a whole, it is reasonable to split it up into several simpler, manageable, and cohesive subtasks and associate each subtask with a number of protocols. In this direction, the International Organization for Standardization (ISO) has developed a reference model, the *Open Systems Interconnection* (*OSI*) model, which defines seven communication subtasks arranged in a layered (hierarchical) structure, as shown in Fig. 1. Therefore, each building-block layer of

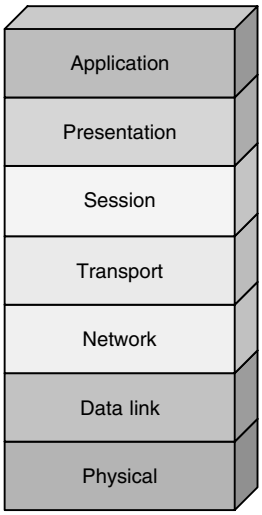


Figure 1. The seven layers of the OSI model.

the OSI model logically groups functions associated with the implementation of a specific communication subtask.

At the bottom of the OSI model, the *physical layer* is responsible for the transmission of the raw bit stream it receives from the next higher layer over the physical transmission medium. This function is usually performed by a *modem* (*mo*dulator–*dem*odulator), which maps a bit stream into a signal suitable for transmission over the underlying physical medium according to the electrical, mechanical, functional, and procedural characteristics of the medium and, at the receiving end, it does the inverse mapping and delivers the bit stream to the higher layer, via an appropriate interface. The physical layer thus provides its upper layers with a *virtual point-to-point bit pipe* and hides from them all the physical medium complexities. The design of the physical layer is usually considered as part of the job done by communication engineers as opposed to network engineers and is outside the scope of this article.

The next layer in the hierarchy, the *data-link control* (*DLC*) *layer*, provides the higher layers a *reliable point-to-point packet pipe over a single link*. It does so by organizing the exchanged bit stream into *frames* (data packets with header/trailer overhead control bits) and providing functions such as error control, retransmissions, and speed matching between sender and receiver.

The *network layer,* one layer above, is very important in internetworking, since it provides the means for end systems to communicate across a collection of communication networks. It is present at every intermediate network node (e.g., router) that interconnects communication networks, and its main purpose is to hide from its upper layers the underlying network technology and topology, providing a *virtual end-to-end packet pipe* between end systems. In order to transfer packets from the source to the destination, the network layer has to implement functions such as routing, addressing, and flow control. These functions convert the link packet pipe provided by the DLC layer to the network layer into an end-to-end packet pipe provided by the network layer to the higher layers. The network layer is considered the most complex layer of

the OSI model, since it requires the cooperation of many nodes (instead of the two end nodes, as is the case for the physical and the DLC layers of a point-to-point link).

The next layer in the hierarchy is the *transport layer,* which ensures end-to-end, reliable message transfer in a transparent way to its upper layers. Its services include message fragmentation and reassembly, and error recovery and flow control between endpoints; however, its functions and complexity vary among different network implementations and depend primarily on the reliability of the underlying network and of the lower-layer services.

The *session layer* establishes, manages, and ends the dialog between applications in end systems, thus providing a virtual session service to the higher layer. Typical functions at this layer include authentication, data grouping, and billing. It is generally considered as a rather "thin" layer.

The next higher layer is the *presentation layer, w*hich deals with the syntax of exchanged data and provides a common data representation to the user applications. Important services here include data encryption, data compression, and code conversion.

The uppermost layer of the OSI model is the *application layer*, which provides the interface of the OSI environment to user applications. It does all the remaining work that is not done by other layer. This layer provides end-application protocols like email, WWW applications (browsers), teleconferencing, FTP, Telnet, and chat services.

The layers of the OSI model are organized hierarchically so that each layer provides services to its next upper layer and is based on services provided to it by its next lower layer. Entities at the same layer between different systems ("peer" entities) communicate on the basis of a mutually agreed-on protocol. Peer entities do not communicate directly; instead, they pass their messages to their next lower layer, which then forwards them to the next lower layer, and so on, until they reach the physical layer, where they are transmitted to the interconnected system. When the messages are received at the other end of the communication medium, they travel all the way up the recipient's lower layers (*demultiplexing*), until they reach the recipient peer entity. This procedure is controlled by headers that are attached successively to each packet (*encapsulation*) by each layer as the packet crosses on its way downward (toward the physical layer) and then stripped off successively by each layer as the packet crosses on its way upward (toward the peer entity of the communication originator), as outlined in Fig. 2. Each layer has to perform its function based only on information included in the header of that layer and cannot look in the body of the message or on the headers added by other layers. This is helpful because if we need to replace the implementation of one layer, this will not affect the operations of other layers.

It is worth noting that the OSI model is not really a communication standard, but rather a framework, or guideline, for developing standards to enable the interconnection of heterogeneous computing systems. Two systems that adhere to some standard developed according to the OSI model will have the same number of layers implementing the same communication functions (although maybe in a different way) and follow a common protocol.

3. THE TCP/IP PROTOCOL SUITE

While the OSI architecture can be considered a guideline for developing communications standards, another firmly

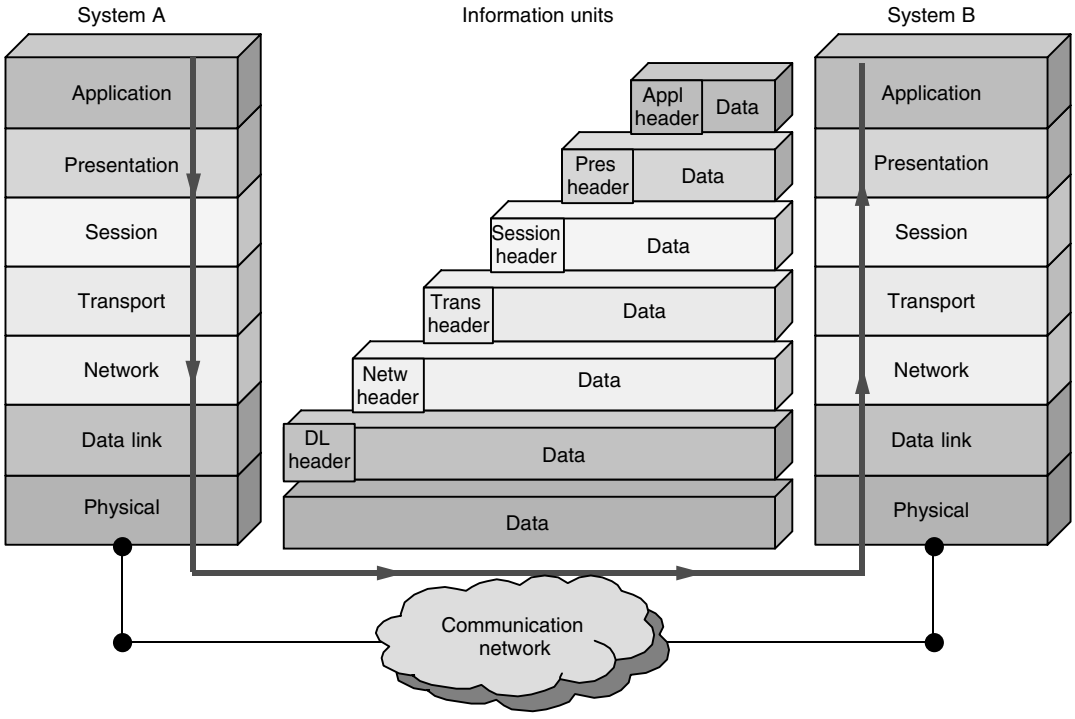


Figure 2. Computer communication through the OSI model.

established architecture has been extensively employed in computer communications. This architecture originates in the U.S. Department of Defense (DoD) efforts to support communications transparently among heterogeneous computer systems over heterogeneous physical communication networks, thus forming a "network of networks" that has nowadays evolved to what is best known as the "Internet." This architecture and its associated standards are collectively known as "TCP/IP."

The acronym *"TCP/IP"* (Transmission Control Protocol/Internet Protocol) does not imply just a pair of communications protocols, but rather refers to a large collection of interrelated protocols and applications (a *protocol suite*); the TCP and IP protocols are the more significant (and widely used) constituent parts of it. Because of its effectiveness and simplicity, the TCP/IP protocol suite has now become the dominant computer communications architecture and is almost synonymous with the terms "Internet" and "computer network."

3.1. TCP/IP Layering

The TCP/IP protocol suite architecture follows a reasoning similar to the one reflected in the OSI model — the whole communication task is divided into a number of smaller and more manageable subtasks, each subtask implemented by one or more protocols. These protocols can be organized conceptually as a hierarchical set of layers (a *protocol stack*), wherein each layer builds on its lower layer, adding new functionality to it. The TCP/IP protocol suite involves four such layers:

1. The application layer
2. The transport layer
3. The internetwork layer
4. The network interface layer

While the top layer (application layer) deals only with application details, the next three lower layers deal with the communication details of an application. Figure 3 illustrates the four layers of the TCP/IP protocol suite, as well as their functionally equivalent layers, in rough terms, in the context of the OSI Reference Model. The

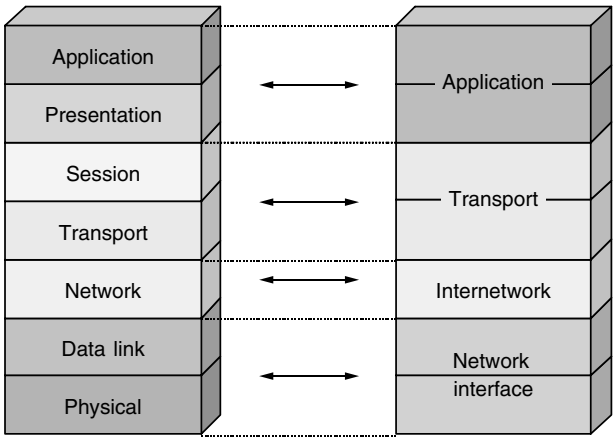


Figure 3. The TCP/IP protocol suite and its correspondence with the OSI Reference Model.

strict use of all layers is not mandated, and the layers are not explicitly stated in the standards.

The lowest layer in the hierarchy (*network interface layer* or *link layer*) forms the interface to the underlying network hardware and hides the network implementation details from the layers above. Protocols at this layer depend on the actual communication network to which the end system is attached and may provide services such as error control.

The purpose of the *internetwork* (or *network*) *layer* is to route data appropriately to the destination host, hiding from its higher layers the internetwork architecture layout between the hosts. The most important protocol at this layer is the Internet Protocol (IP), which provides connectionless services for end systems, without assuming reliability from lower layers.

The *transport layer* provides end-to-end data transfer between peer processes on different hosts. The most important protocol at this layer is the Transmission Control Protocol (TCP), which provides reliable connection-oriented services between peer processes.

Finally, the *application layer*, on top, contains network applications and services, as well as protocols for resource sharing and remote access that are needed to support the user applications. Important protocols at this layer are the Telnet protocol, the File Transfer Protocol (FTP), and the Simple Mail Transfer Protocol (SMTP).

Figure 4 shows a typical TCP/IP communications example between two application processes located at different hosts (end systems) and over two different networks interconnected with a router (an intermediate system).[1] Note in this figure that, while application- and transport-layer protocols are end-to-end, network as well as network interface-layer protocols are hop-by-hop protocols; that is, they are used between end systems and intermediate systems, or between intermediate systems across a collection of communication networks.

TCP/IP employs a two-level addressing scheme; at the low level, every host on an internetwork is assigned a unique (global) address (*IP address*), so that data can be routed to the correct host, while at the high level, every process residing on a host is assigned a unique (within the host) address (*port number*), so that data can be routed to the correct process.[2] Moreover, TCP/IP implements encapsulation and demultiplexing techniques similar to the ones used in the OSI model. On the transmit side, TCP accepts the bytestream from an application, segments it into small data blocks, and appends to each of them a header containing items such as the destination port, sequence number, and checksum for error detection. The resulting data unit, known as a *TCP segment*, is sent down to IP, which appends additional control information relevant to IP functionality (e.g., the destination host address). The resulting

[1] A *router* is a system that attaches to two or more (usually) different physical networks and forwards data packets between the networks. It consists of a network-layer protocol as well as a number of network interface protocols, depending on the actual communication networks it has to interconnect.

[2] The combination of an IP address and a port number is called a "socket" and can uniquely identify a service running on a host.

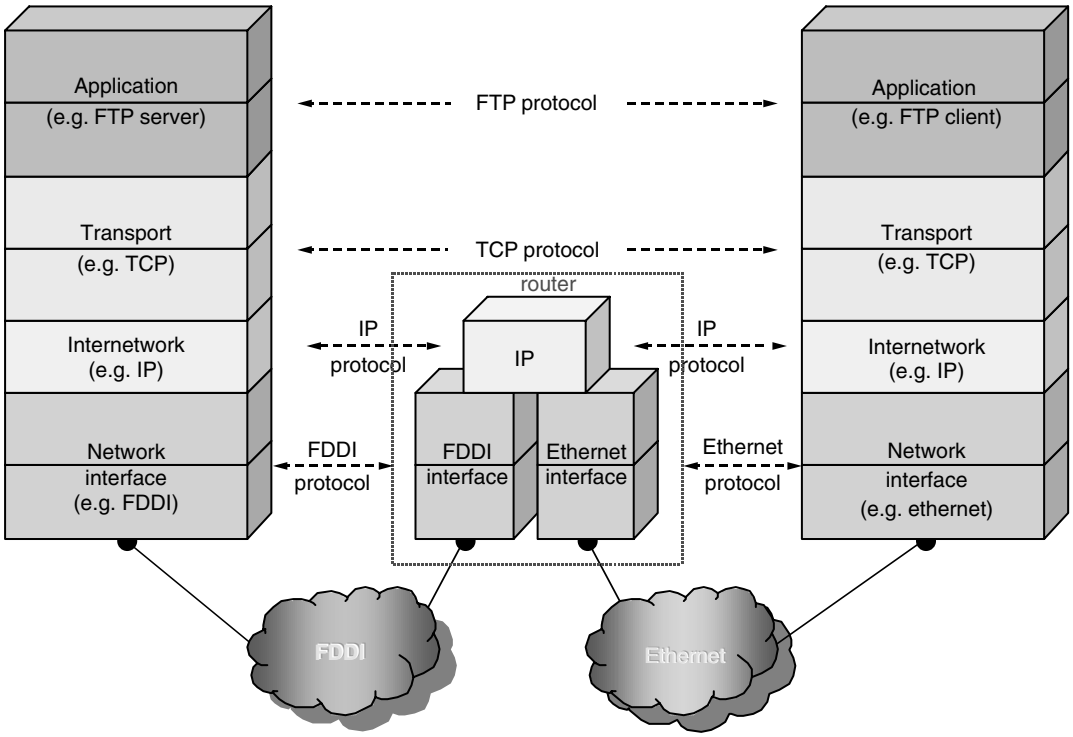


Figure 4. TCP/IP communications over two networks.

data unit, known as an *IP datagram*, is handed over to the network interface layer, which in turn adds its own specific header and transmits the resulting *frame* (which has now the structure shown in Fig. 5) across its attached communication network to the appropriate router. On the receive side, the reverse process takes place; headers are removed and used by the appropriate layers to control the communication procedure within the scope of their layer, until the original bytestream is delivered intact to the recipient application.

3.2. The Internet Protocol

The *Internet Protocol* (IP) is the most widely used internetworking protocol and provides connectionless and unreliable packet delivery services. "Connectionless" indicates that there is no predetermined route through the network between the endpoints but, rather, each packet works its way through the network independently and without any prior coordination. As a result, each packet of a session may follow a different path to reach the same destination, and therefore packets may be delivered out of order; additional services such as sequencing, if required, are dealt with at higher layers (e.g., by TCP). "Unreliable" suggests that IP makes a best effort to get a packet to its destination, but without providing any guarantee of actual delivery. Again, if reliability in the communication is required, it has to be addressed properly at higher layers.

An IP datagram consists of the IP header and the data. Its format is shown in Fig. 6, where the fields have the following meaning:

- *Version* (4 bits)—indicates the IP version number.
- *Header length* (4 bits)—indicates the length of the header (including options) in 32-bit words (i.e., rows in Fig. 6). Its normal value is 5, corresponding to a header length of 20 bytes (no options are used).
- *Type of service* (8 bits)—specifies delay, throughput, reliability, and precedence parameters to request a particular quality of service for the datagram.
- *Total length* (16 bits)—indicates the length of the total IP datagram in bytes. Thus, the maximum IP datagram size is 65,535 bytes (although, in practice, much smaller IP datagrams are used).
- *Identification* (16 bits)—a sequence number that uniquely identifies (along with the source address,

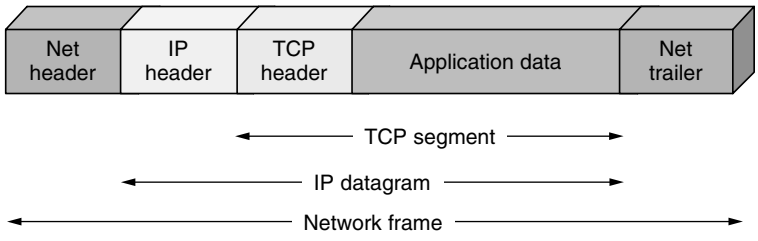


Figure 5. Encapsulation in a TCP/IP frame.

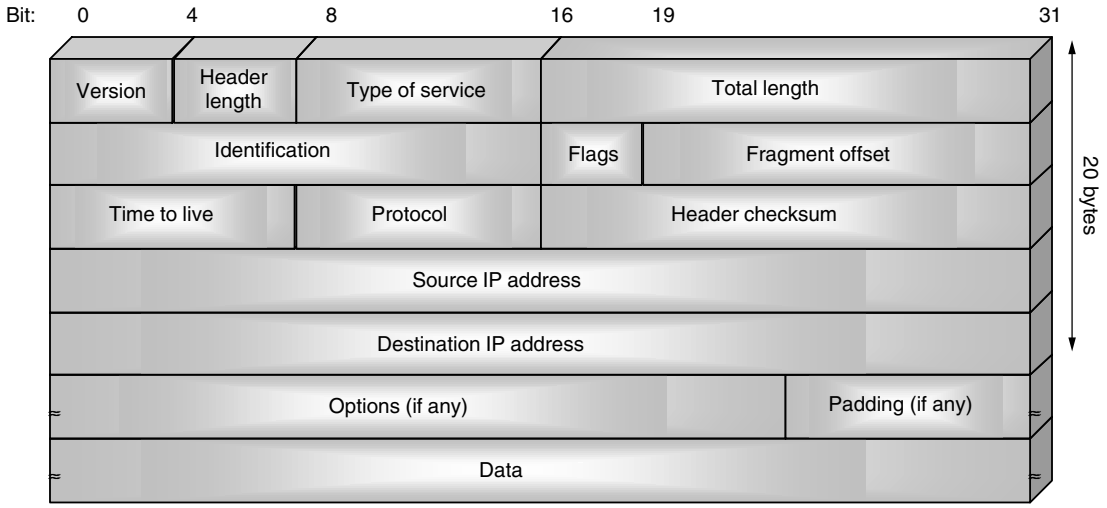


Figure 6. The format of an IP datagram.

the destination address and the protocol used) each datagram transmitted by a sender. It is used for the fragmentation and reassembly of an IP datagram.

- *Flags* (3 bits)—only 2 bits are defined. When an IP datagram is fragmented, the "more" bit is set for each fragment (except the last one). When the "don't fragment" bit is set, the IP datagram will not be fragmented.
- *Fragment offset* (13 bits)—when fragmentation is used, it indicates the offset (in 8-byte units) of this fragment from the beginning of the original IP datagram.
- *Time to live* (8 bits)—is used to prevent IP datagrams from getting caught in routing loops, by specifying the amount of time that a datagram can stay in an internetwork. Practically it has a value of 32 or 64 and is equivalent to a hop count (see Section 5.2 below).
- *Protocol* (8 bits)—indicates the higher-level protocol that has to receive the IP datagram after demultiplexing at the receiver.
- *Header checksum* (16 bits)—is used for error detection and covers only the IP header. It is maintained at each router the datagram comes through.
- *Source address* (32 bits) and *destination address* (32 bits)—indicate the IP addresses of the originator and the recipient(s) of the IP datagram (see Section 5.1.2).
- *Options* (variable length)—indicates the options requested by the sender (e.g., security restrictions, timestamping, route recording). This value is padded by 0s as necessary, to ensure that the header length is a multiple of 32 bits.

3.3. The Transmission Control Protocol

The next widely used protocol of the TCP/IP suite is the *Transmission Control Protocol* (TCP), which builds on the services provided by IP and provides additional functionality to application processes, such as reliable data transfer, error control, and flow control. TCP is implemented in edge systems, and its task is to transform the unreliable delivery service provided by IP into a reliable "connection-oriented" data transmission system, suitable for building network applications. "Reliable" suggests that packets are guaranteed to reach their destination, error-free and in the correct order, while "connection-oriented" indicates that a logical connection is established between the endpoints prior to any data transfer between them and lasts for the duration of a session. Since the actual data packets are transferred by the underlying IP protocol in a connectionless mode, TCP does not in fact set up predefined paths across which all data flow during a given session but, instead, establishes a tight relationship between the end hosts, which can be logically considered as a "virtual circuit." When communication is desired, the initiating TCP first sends a special connection request segment, and awaits a connection response. When it arrives, the initiating TCP confirms connection establishment and begins the reliable communications described earlier.

A TCP segment consists of the TCP header and the data. Its format is shown in Fig. 7, where the fields have the following meaning:

- *Source port* (16 bits) and *destination port* (16 bits)—indicate the TCP port numbers of the source and destination applications.
- *Sequence number* (32 bits)—since two applications exchange a stream of bytes across a TCP connection, the byte in the stream the first data byte in this segment represents is identified by the *sequence number* field.
- *Acknowledgment number* (32 bits)—indicates the sequence number of the next data byte that the sender application is ready to receive (implying that all data bytes with previous sequence numbers were successfully received).
- *Header length* (4 bits)—indicates the length of the header (including options) in 32-bit words (rows in Fig. 7).

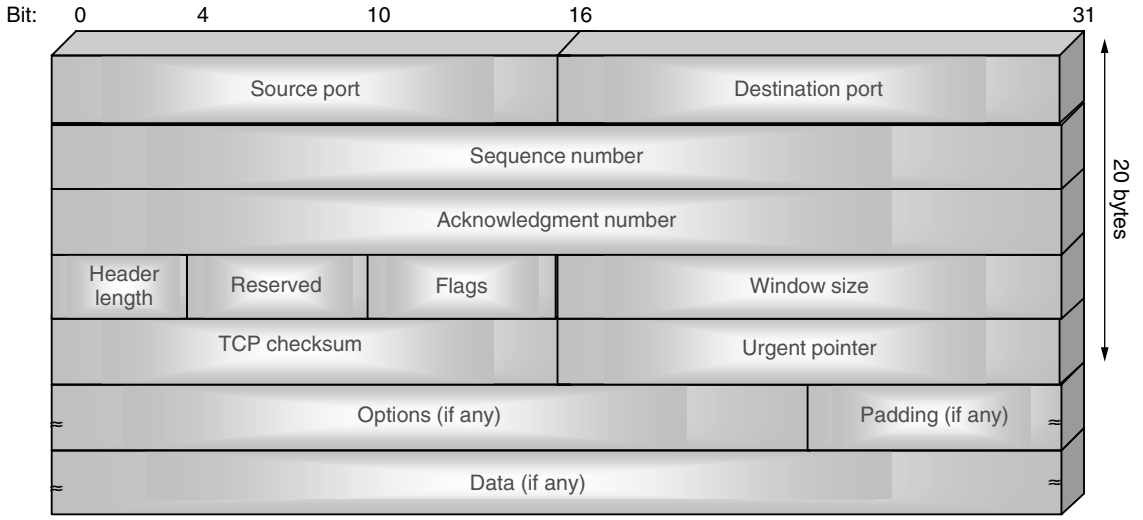


Figure 7. The format of a TCP segment.

- *Reserved* (6 bits) — these bits are reserved for future use.
- *Flags* (6 bits) — there are 6 flag bits (viz., URG, ACK, PSH, RST, SYN, FIN) that are used to control various functions in a TCP connection.
- *Window size* (16 bits) — used for the *flow control* function and indicates the number of data bytes that the sender is willing to accept (starting with the one specified by the acknowledgment number field).
- *TCP checksum* (16 bits) — used for error detection and covers both the TCP header and the TCP data. This checksum is calculated by the sender and verified by the receiver.
- *Urgent pointer* (16 bits) — the sender can transmit urgent data to the receiver, and this field points to the last byte of such data, delimiting in this way the urgent bytes in the bytestream of the segment.
- *Options* (variable length) — indicates the options requested by the sender (e.g., maximum segment size). This value is padded by 0s as necessary, to ensure that the header length is a multiple of 32 bits.

In TCP/IP, the network is simple, and doesn't guarantee anything, except a high probability of packet delivery. The complexity is in TCP, which exists only in edge systems. The edge systems themselves are powerful computers — sufficiently powerful to run TCP. We can say that the end user provides the complexity, while the Internet provides a basic service. This is in contrast to the telephone network, which was designed to provide an expensive (circuit-switched) and reliable service, but whose end systems are extremely simple (telephones).

3.4. The User Datagram Protocol and Some Basic Applications

Another important protocol of the TCP/IP suite is the *User Datagram Protocol* (UDP) at the transport layer. UDP is an alternative to TCP host-to-host protocol that adds no reliability, sequencing, error control, or flow control to IP. These functions, instead, are accommodated by the applications on top of UDP. UDP is connectionless, and its header consists of four 16-bit fields: the source and destination ports, the total length of the UDP segment, and the checksum applied to the entire UDP segment. UDP is used mainly for the exchange of self-contained data packets (or datagrams) between application processes and is extensively employed in real-time applications, like VoIP (Voice over IP) and Net audio.

The following are some basic applications that are part of the TCP/IP suite and exploit the characteristics of TCP and UDP:

1. Telnet (Telecommunications Network), a client/server application that supports terminal emulation and remote logon capability over a TCP connection.
2. FTP (File Transfer Protocol), a client/server application that enables the exchange of text and binary files between host computers, under user command. It uses two TCP connections: one for control messages and the other for the actual file transfer.
3. SMTP (Simple Mail Transfer Protocol), which supports basic electronic mail in text form over a TCP connection. MIME (Multipurpose Internet Mail Extension) is an SMTP extension that provides support for the attachment of other file forms, including audio, graphics, and video.
4. SNMP (Simple Network Management Protocol), which enables the exchange of network management information between host computers (agents) and an SNMP manager, over UDP.

4. DATA-LINK CONTROL PROTOCOLS

The physical layer of the OSI model, as we have already discussed, is responsible for the transmission of a raw bit stream over the physical communication medium. However, to manage this raw bit stream and render it useful, some form of control has to be exercised on top of the physical layer, to account for issues such as link

frame synchronization, link flow control, error control, and retransmissions over a link. This level of control is referred to as *data-link control* and is implemented by protocols residing functionally in the *data-link* layer of the OSI model.

4.1. Flow Control

Since computers that communicate across a network can vastly vary in terms of capabilities, a mechanism has to be set up to account for such inequalities. For example, a fast (high-speed) sender should be prevented from overwhelming with data a slow receiver; otherwise the latter (and maybe the network between them) will become congested. *Flow control* is the function that assumes responsibility for such congestion occurrences and there are two common mechanisms in this direction, known as *stop-and-wait* and *sliding-window*. Both mechanisms are based on the principle that the sender may transmit only when the receiver permits so.

Stop-and-wait is a simple mechanism, according to which a sender, after transmitting a frame, stops transmission and waits until the receiver "acknowledges reception" of the frame. In this way, a slow receiver can avoid getting flooded with frames by sending back reception acknowledgments only after having processed the received frame. In cases, however, where the propagation time over the communication medium is high (e.g., a satellite link), an extended version of the stop-and-wait mechanism is used, the sliding window mechanism, according to which the sender requires an acknowledgment from the receiver after a certain number of frames (determined by the *window size*) have been transmitted. This way, several frames can be in transit at the same time, increasing the utilization of the communication link.

4.2. Error Control

The error control function deals with the detection and, in some cases, correction of errors that occur during the transmission of frames (resulting in lost or damaged frames). Errors are detected at the cost of increasing the frame length by adding some extra information at transmission, which is used by the receiver to either correct the error by itself (*forward error correction*), or ask for retransmission of the frames found in error (*backward error correction*). In most cases, backward error correction is more efficient and there are three common mechanisms in this direction (collectively known as *automatic repeat request* (*ARQ*)): *stop-and-wait ARQ*, *go-back-N ARQ*, and *selective-reject ARQ*. All these mechanisms are based on the flow control techniques described previously.

Stop-and-wait ARQ is based on the principle of the stop-and-wait flow control technique, where the sender transmits a single frame and then waits to get back an acknowledgment for this frame. Only a single frame can be in transit at any one time. If the sender receives either a negative acknowledgment from the receiver or no acknowledgment at all after a certain time period, it retransmits the same frame.

The go-back-*N* ARQ improves the efficiency of the stop-and-wait ARQ but increases its complexity. It is based on the sliding-window technique, where the sender buffers the transmitted frames (up to a number determined by the window size) until an acknowledgment is received. The acknowledgment is cumulative; that is, an ACK for frame *n* acknowledges reception of all previously transmitted frames. If the sender receives a negative acknowledgment for one frame, or no answer at all after a certain time period, then it retransmits all buffered frames from the not-acknowledged frame on.

In the selective-reject ARQ, the go-back-*N* ARQ principle is further refined, by limiting retransmission to only the not-acknowledged (i.e., rejected or timed out) frames. In this way, the amount of retransmission is decreased, at the cost, however, of increasing the receiver's buffer size and both sender's and receiver's complexity.

4.3. Synchronization and Framing

For successful data exchange between computer systems, some form of control has to be imposed on the sequence of bits transmitted over a communication medium, in terms of determining the boundaries between successive frames (*framing*) and maintaining common timing parameters (duration, rate, spacing) for the frame bits between the sender and the receiver (*synchronization*). Synchronization at the data-link layer can be achieved by either asynchronous or synchronous transmission schemes.

In *asynchronous transmission*, data are transmitted one character (5–8 bits) at a time. The beginning and the end of a character are indicated by a start and a stop bit, respectively, while a parity bit before the stop bit is often added for error detection purposes (see Fig. 8). When no character is being transmitted, the communication link is in idle state and, thus, the receiver can resynchronize at each start bit (the beginning of a new character). Asynchronous transmission is widely used (e.g., at PC interfaces with low-volume and low-speed transmissions) because it is simple to implement and the relative equipment is inexpensive. Nevertheless, by using many extra bits per character, it increases the communication overhead, wasting about 20–30% of the available medium bandwidth.

Synchronous transmission, on the other hand, deals with blocks of data, avoiding the overhead of the start/stop bits around each character and providing alternate means (*pre-* and *postamble* bit patterns) to delimit a frame at both ends. In this case, the frame is treated as either a

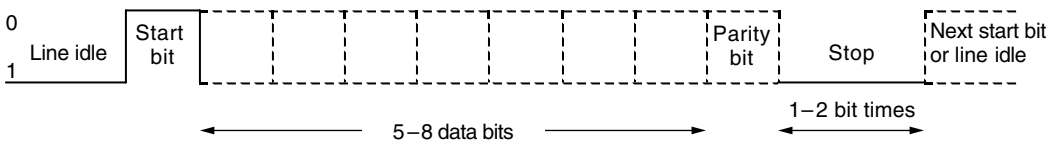


Figure 8. Character format in asynchronous transmission.

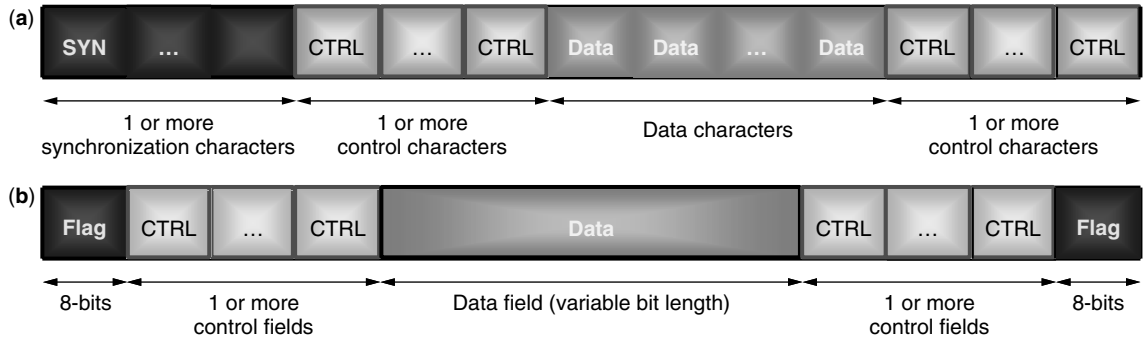


Figure 9. (**a**) Character-oriented and (**b**) bit-oriented frame format in synchronous transmission.

sequence of characters (*character-oriented* transmission) or a sequence of bits (*bit-oriented* transmission), as shown in Fig. 9. Bit-oriented protocols do not depend on specific codes (e.g., ASCII) for line control — as it is required with the character-oriented protocols — and hence they are nowadays widespread. Because of its low overhead (typically ~5%), synchronous transmission is generally used for large-volume and high-speed transmissions.

4.4. The High-Level Data Link Control (HDLC) Protocol

We pause for a moment to examine the most important Data Link Control protocol, the *High-level Data Link Control protocol* (HDLC), which has been standardized by ISO and constitutes the basis for almost all the bit-oriented protocols at the data-link layer that are in common use today: LAP-B (in the X.25 packet-switching network protocol suite), LAP-D (in the ISDN protocol suite), SDLC (in the SNA protocol suite), and IEEE 802.2 LLC (in the IEEE 802 LAN protocol suite). HDLC uses synchronous transmission and supports both full- and half-duplex communications in point-to-point and multipoint link configurations.

To organize data communications between hosts, HDLC defines two basic types of stations: the primary and the secondary stations. The *primary* station is responsible for controlling the data link (e.g., error/flow control) and transmits frames called *commands*. The secondary station responds to commands from a primary station by transmitting frames called *responses*. In a conversation, there is one primary and one or more secondary stations involved. A third type of station, the *combined* station, can also be defined, which combines the features of the primary and the secondary station.

HDLC operates in three data transfer modes:

- *Normal response mode* (NRM), in which secondary stations can transmit only in response to a poll from the primary station. This mode is used in point-to-point or multipoint link configurations (e.g., one host with many terminals attached).
- *Asynchronous response mode* (ARM), in which a secondary station may transmit without explicit permission from the primary station. This mode is used in special cases.
- *Asynchronous balanced mode* (ABM), in which all stations have equal status and may initiate transmission without receiving permission from the other station. This mode is used mainly on full-duplex point-to-point links.

Communication between primary and secondary stations in HDLC is achieved by exchanging frames (commands and responses) that fall in one of the following categories:

- *Information frames*, which are sequentially numbered and contain the user data and, sometimes, piggybacked error and flow control data (using the ARQ mechanism)
- *Supervisory frames*, which contain error and flow control data (not piggybacked as before)
- *Unnumbered frames*, which do not have sequence numbers, but are used for miscellaneous link control and management functions

An HDLC frame consists of a three-field header, the payload, and a two-field trailer, and has the general structure shown in Fig. 10. The *flag* is a special 8-bit sequence (binary 01111110) that identifies the start and the end of the frame (the end flag of one frame can also constitute the start flag of the next consecutive frame). Since the flag is essential to achieve frame synchronization, it is necessary to ensure that the flag sequence does not accidentally appear anywhere in the frame between start and

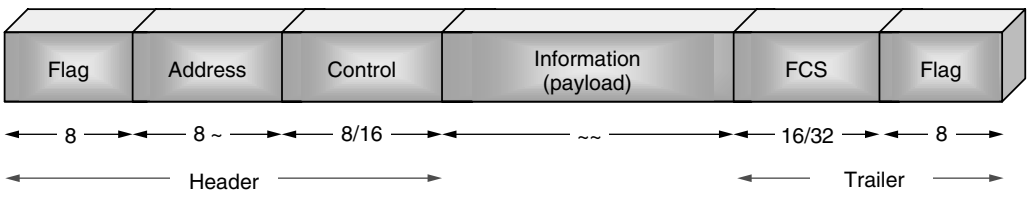


Figure 10. HDLC frame structure and fields' length (in bits).

end flags. To that end, a technique known as *bit stuffing* is used that inserts at transmitter (and deletes at receiver), an additional 0 (zero) bit, after each occurrence of five consecutive 1s in the data between the start and end flags of a frame. The *address* field is usually 8 bits long, but an extended format may also be possible. Depending on the station type of the issuer of the frame, the address field may indicate the sender (in responses) or intended receiver (in commands) of the frame. The *control* field is usually 8 bits long, but an extended format may also be possible. It is used to indicate the type of the frame (information, supervisory, or unnumbered) and it also contains some control information, according to the frame type (e.g., send sequence number for information frames).

The *information* field is the "payload" of the frame that contains the actual user data (any sequence of bits) coming from the higher layers. Its length, though undefined, is usually a multiple of 8 bits. Finally, the *frame-check sequence* (FCS) field goes after the information field and contains a 16- or 32-bit cyclic redundancy check (CRC) value that is used by the receiver to check the address, control, and information fields of the frame for errors. Practically, 16-bit CRC (2 bytes) is used for frames up to 4 KB (kilobytes).

4.5. Local-Area-Network Protocols

Due to the widespread use of local-area-networks (LANs),[3] a great number of standards and protocols have emerged that relate to LAN technology and belong functionally to the data-link layer of the OSI Reference Model. In this context, the data-link layer is usually divided into two sublayers: the *logical link control* and the *medium access control* sublayers.

Protocols in the logical link control sublayer are similar to typical data-link control protocols, like the HDLC. The most commonly used protocol at this level is LLC, described in the IEEE 802.2 standard. LLC supports both connectionless and connection-oriented services and provides mechanisms for addressing and data-link control. Moreover, it is independent of the underlying network topology, medium access control technique, and transmission medium.

Since all devices interconnected via a LAN share a common physical communication medium with a fixed capacity, some procedure has to be established to control access to the underlying medium and provide for an efficient usage of its capacity. Such procedures and techniques are provided by protocols at the medium-access control (MAC) sublayer, which, apparently, are closely dependent to the transmission medium and the network topology used. Common MAC protocols are the *carrier sense multiple access with collision detection* (CSMA/CD), the *token bus* and the *token ring,* each specified in a number of widely used standards, such as IEEE 802.3, IEEE 802.4, and FDDI.

[3] A LAN is a high-speed data network, optimized for use over small geographic areas (like buildings), that usually interconnects personal computers, peripheral devices, and other equipment over a common (shared) physical communication medium.

CSMA/CD falls under the category of contention techniques, in which all communicating parts "contend" for access to network resources. The precursor of CSMA/CD is ALOHA, which was developed in the 1970s for packet radio networks. According to ALOHA, every station transmits its packet on the air whenever it is ready to do so, without any prior coordination with the other stations in the network. In this fashion, collisions between packets of different stations that are transmitted simultaneously are quite common, especially as the number of stations increases. In case of a packet collision, the transmitting station waits for a time interval and then attempts retransmission of the packet. ALOHA is a very simple to implement technique, with a limited efficiency, however. CSMA improves this situation by requiring that each station has to listen to the medium first ("sense the carrier"), and if no other station is transmitting at that time, it can go on transmitting its own packet, or else it has to wait. CSMA can be further improved by having the transmitting station listen to the medium while transmitting, to notice early any possible collision ("collision detection") and stop its transmission (and thus free network resources). The CSMA/CD medium-access technique is used in Ethernet/IEEE 802.3 LANs, over coaxial cable or unshielded twisted pair as transmission media and on bus, tree, or star network topologies.

In the token-passing schemes, network devices access the physical medium based on the possession of a token that circulates when stations are idle. A token is like a permission to transmit, and whoever has the token can transmit or pass it to the next node. Examples of LANs that use such techniques are the *Token Bus* (IEEE 802.4), over coaxial cable or optical fiber and on bus, tree, or star network topologies; the *Token Ring* (IEEE 802.5), over shielded/unshielded twisted-pair cable and on ring topology; and the *Fiber Distributed Data Interface* (FDDI), over optical fiber and on ring topology.

5. INTERNETWORKING PROTOCOLS

In an internetworking environment, where a number of communication networks are interconnected to provide data transfer among the hosts attached to them, an addressing scheme has to be established so that all communicating entities are uniquely identified in the internetwork, to allow data to be directed (or *routed*) to the intended destination. In this section we focus on the protocols and mechanisms behind these two fundamental functions in internetworking: *addressing* and *routing*.

5.1. Addressing and Naming

In the global postal system, a postal carrier knows exactly where to deliver a letter, based on an agreed-on addressing scheme, according to which a house location can be uniquely determined, by specifying some positional parameters (country, city, street, number, etc.) that altogether constitute the recipient's "home address," which is hierarchical, and is written on the envelope of the letter. Addressing in an internetworking environment follows a similar concept; each entity in an internetwork can be reached by means of a unique (global) address, which is usually a number like a telephone number in telephony. Since, however,

humans cannot easily remember numbers, especially in the binary format that computers understand, translation schemes between symbolic names and addresses are practically used that match high-level human-intelligible names to machine-intelligible internetwork addresses and vice versa. This translation process, called *address resolution*, is usually performed in a distributed fashion by special network servers, called *name servers*. In the Internet context, address resolution is performed by the *Domain Name System* (DNS), which is a global network of name servers implementing a distributed database.

5.1.1. The Domain Name System. According to DNS, symbolic names are grouped in domains that are hierarchically organized, like the chain of command reflected in the organization chart of a big company. Within each domain there is one primary (*authoritative*) name server that has the responsibility of performing address resolution through the maintenance of a local database. It is, however, common for domain name servers to delegate authority for their subdomains to other name servers, therefore defining subareas of responsibility. The domain (and subdomains, if any) for which a name server is authoritative is referred to as "zone of authority." Consequently, DNS logically interconnects all name servers into a hierarchical tree of domains and uses a hierarchical naming structure, like the one used in telephone numbers, in which names consist of a sequence of fields [typically a top-level domain,[4] a domain, subdomain(s), and the host name] that jointly identify the entity. For example, *talos.telecom.ntua.gr* refers to the host named *talos* in the Telecommunications Laboratory (*telecom*) subdomain of the National Technical University of Athens (*ntua*) domain, in Greece (*gr*).[5]

Since DNS is a distributed database, a name server doesn't need to know all the names and addresses of hosts in the Internet; its area of knowledge and responsibility is usually confined to its own zone of authority. It does need to know, though, the name servers who are responsible for other domains. When, for instance, we type *www.cis.ohio-state.edu* on our Internet browser, a module called *resolver* tries to look up the requested address locally, if a local name cache is kept. Otherwise, it sends a DNS query to the local name server. The local name server will then follow these steps:

1. Check to see if it already knows (from its database or from a previous query) the address of *www.cis.ohio-state.edu*. If so, it finds the address in its database or cache and replies to the query. The browser then uses this (IP) address to request a connection to the host containing the desired Webpage.
2. If the local name server cannot resolve the address by itself, it will query one of the "root" servers, at the top of the DNS hierarchy, whose addresses are definitely known to all name servers, for the address of *www.cis.ohio-state.edu*.
3. Today, there are 13 root servers on the Internet, and each of them knows the IP addresses of the servers for all the top-level domains (*.com*, *.edu*, *.uk*, etc.). So, ultimately, the root server contacted will refer our name server to (i.e., give the address of) a list of *.edu* name servers.
4. Our name server will query one of the *.edu* servers for the address of *www.cis.ohio-state.edu*.
5. The *.edu* server queried will refer our name server to a list of name servers for the domain *ohio-state.edu*.
6. Our name server will then query one of the *ohio-state.edu* name servers for the address of *www.cis.ohio-state.edu*.
7. If the *ohio-state.edu* name server queried knows the address of *www.cis.ohio-state.edu*, it returns that address to our name server. If there is another name server responsible for the *cis* (delegated) subdomain, the address of that name server will be returned to our name server, which will in turn query that server for the address of *www.cis.ohio-state.edu*.
8. Finally, our browser will be given the requested (IP) address or an error if the query could not be answered.

This is a worst-case scenario that can take many seconds to complete, but actually things are simpler, since, as it was mentioned in the above process, each name server caches for a certain time period all the information it retrieves this way. This way, a huge load is removed from the root and top-level servers.

5.1.2. IP Addresses. IP addresses are 32 bits in length. They are typically written in a format known as "dotted decimal notation," according to which each of the 4 address bytes is expressed in its decimal equivalent value (0–255) and the four values are separated by periods, for example, "147.102.105.18."

IP addresses generally consist of two parts:

- The *network identifier* identifying the TCP/IP subnetwork to which the host is connected
- The *host identifier* identifying a specific host within a subnetwork

To account for networks of different sizes, IP defines several *address classes,* characterized by the length of the network identifier. Class B addresses, for instance, have a 14-bit network identifier and a 16-bit host identifier and thus can address up to 65,536 (2^{16}) hosts per network; therefore, class B addresses are used for moderate-sized networks. Five address classes are defined (A through E), while only classes A, B, and C are used for host addressing.[6]

5.1.3. Address Resolution Protocol (ARP). The *address resolution protocol* is used in specific network implementations where the Internet Protocol (IP) is applied

[4] For instance, *com*, *edu*, *org*, *gov*, or two-letter country codes such as *uk* (United Kingdom) and *it* (Italy).

[5] It is not, however, necessary for a host to be physically located at the country specified by the country code.

[6] Class D addresses are used for IP multicasting, while class E addresses are reserved for future use.

over Ethernet or token ring local-area networks (LANs). Addressing in such LANs is twofold; at the IP level, every entity in the network is assigned a unique IP address (which is contained in the IP datagram) and, at the medium-access control (MAC) level, every entity in the LAN has a unique MAC (hardware) address (6 bytes in length, which is placed in the MAC frame).

An entity's IP address is different from its MAC address. Therefore the address resolution protocol is used to translate between the two types of address, so that a sender's IP process can communicate with the intended receiver's IP process on the same network, when knowing only the receiver's IP address. The process is relatively simple: (1) the sender transmits an ARP Request message using the hardware broadcast address (so that the intended receiver will surely receive the request); (2) the ARP Request advertises the destination IP address and requests the associated MAC address; (3) the intended receiver recognizes its own IP address and forms an ARP Response that contains its own MAC address; (4) since the original Request also includes the sender's MAC address, this address is used by the receiver to unicast its ARP Response to the original sender. To reduce the number of ARP requests in a LAN, hosts maintain a cache of recently resolved addresses. ARP cache entries timeout at regular intervals (typically 20 min), and new queries are made so that changes to the network topology are properly handled.

5.2. Routing

An internetwork can be considered as a collection of communication networks. Attached to each communication network we find devices (usually computers) that support end-user applications or services. These devices are called *end systems*, as opposed to the *intermediate systems* that are used to interconnect the communication networks (and thus form an internetwork), so that end systems attached to different networks can communicate across the internetwork, share information, and provide services to each other.

On their way toward their intended recipient, data packets travel across multiple networks and reach several intermediate systems, called *routers,* that forward them to the next intermediate system/communication network. The process of moving data from source to destination across an internetwork is known as *routing*. Routing includes two basic functions: *optimal path determination* and *switching*.

The optimal path toward an internetwork destination is determined by special algorithms, called *routing algorithms*. These algorithms perform calculations based on multiple metrics, such as pathlength, packet delay, and communication cost. The results of these calculations are used to populate *routing tables* that are maintained within each router and list the next "hop" (i.e., the next router) to which a data packet should be sent on the (optimal) way to its destination. Therefore, routers forward a received data packet according to the packet's destination address and the association for this address that the routing table indicates. The packet's physical address is changed to the next router's physical address, and the packet is transmitted. This way, the packet is switched hop by hop through the internetwork, until it finally reaches its destination.

This routing process, however, can be processor-intensive. In a more efficient alternative, called *multiprotocol label switching* (MPLS), optimum end-to-end paths through the network are calculated in advance (at the network edge) and appropriate routing information is appended as a *label* between the layer 2 and layer 3 packet headers;[7] then, routers along the path use the information in this label to simply switch the packet to the next hop. MPLS is also employed in traffic engineering (i.e., establishing traffic patterns that balance overall network resources utilization) and quality-of-service (QoS) routing (i.e., selecting routes that provide a desirable level of service, e.g., bandwidth, latency, priority requirements). MPLS supports many protocols of the network layer (including IPv4, IPv6, and AppleTalk), as well as the link layer (e.g., Ethernet, token ring, ATM).

5.2.1. Communication Between Routers. Routing algorithms calculate optimal paths on the basis of the routing information available at the router at a certain point in time. This routing information reflects the status of the networks (reachability, traffic delays, etc.) and is exchanged between routers by certain protocols, called *routing protocols*. We can distinguish two broad categories of routing protocols: *interior router protocols* (IRPs) and *exterior router protocols* (ERPs), if they are used between routers within the same or in different *autonomous systems*,[8] respectively. Practically, IRPs exchange a wealth of routing information to help determine optimal paths within an AS, while ERPs exchange only summary reachability information between ASs and, therefore, are simpler than IRPs. Examples of IRPs are the *Routing Information Protocol* (RIP) and the *Open Short Path First* (OSPF), while *Border Gateway Protocol* (BGP) is a typical ERP.

RIP belongs to the distance-vector class of protocols, according to which neighboring routers send all or a portion of their routing tables by exchanging routing update messages periodically and when network topology changes occur. Whenever a router receives a routing update message, it updates its routing table in light of the new information and subsequently transmits new routing update messages (using UDP/IP) to propagate the network changes across the network. RIP uses the "hop count" metric to measure the pathlength between a source and a destination. The maximum number of hops in a path is limited to 15; a value of 16 implies that the host is unreachable. RIP is still used efficiently today in small ASs.

OSPF is a nonproprietary protocol and belongs to the link-state class of protocols, according to which a router broadcasts to all routers only the portion of its routing table that describes the status of its own links. Such updates (using IP directly) produce minimum traffic and

[7] Or, in the virtual path identifier/virtual channel identifier (VPI/VCI) cell header fields, in ATM networks (see Section 6).

[8] An *autonomous system* (*AS*) is a logical portion (i.e., a collection of routers and subnetworks) of a larger internetwork that constitutes a distinct routing domain (and usually corresponds to commercial or administrative entities). An AS is managed by a single authority and may connect to other AS, as well as other public or private networks.

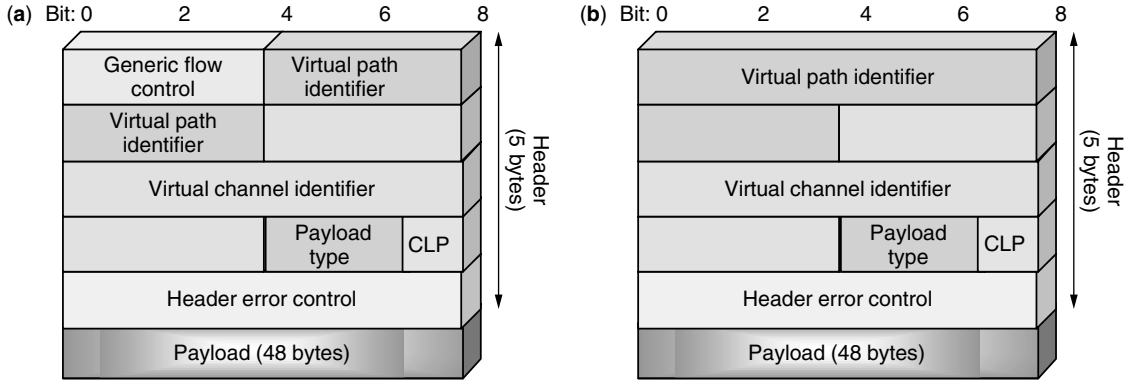


Figure 11. ATM cell format: (**a**) at the user–network interface; (**b**) at the network–network interface.

hence require less network bandwidth. Unlike RIP, OSPF is hierarchical; an AS is partitioned in a number of smaller groups of networks, called *areas*, and an area's topology is invisible to entities outside it. Routing between areas (interarea routing, as opposed to intraarea routing, which takes place within the same area) is handled by an OSPF backbone of routers. The backbone topology is invisible to all intraarea entities. OSPF uses the "shortest path first" or Dijkstra's algorithm to determine optimal paths. OSPF is more efficient and more scalable than RIP and for this reason it is widely used in TCP/IP networks and the Internet.

BGP is a scalable protocol used for the exchange of routing information between different ASs and is widely used on the Internet between the Internet service providers. BGP is a kind of distance vector protocol, with the difference that BGP neighbors don't send routing updates periodically, but exchange full routing information at the beginning (using TCP) and then send updates only about the routes that have changed. Also, a router's BGP table stores the networks that it can reach and the best route to each reachable network. BGP version 4 supports also policy-based routing based on security, legal, economic, and other issues rather than purely technical ones.

6. ASYNCHRONOUS TRANSFER MODE

Asynchronous transfer mode (ATM), also known as *cell relay*, is a high-speed packet transfer technology that tries to combine the benefits of packet switching (efficiency and flexibility for data transfer) with those of circuit switching (guaranteed bandwidth and transmission delay). ATM organizes data in packets of fixed size, called *cells*; this way the processing overhead at switching nodes is minimized. Also, since ATM takes advantage of the inherent dependability of modern communication systems, it uses minimal error control, thus further reducing the processing overhead of ATM cells and allowing for very high data transfer speeds.

The ATM cell consists of a 5-byte header and a 48-byte information (data) field, for a total cell length of 53 bytes. The cell header format is slightly different at the user–network interface and inside the ATM network, as shown in Fig. 11 parts (a) and (b), respectively. The header fields have the following meaning:

- *Generic flow control* (4 bits)—this field is present only at the user–network interface and provides some form of local cell flow control.
- *Virtual path identifier* (VPI) (8/12 bits)—identifies the virtual path a virtual channel belongs to and is used as a routing field in an ATM network. The VPI field is longer (12 bits) at the network–network interface, allowing for a large number of virtual paths to be supported inside the ATM network.
- *Virtual channel identifier* (VCI) (16 bits)—identifies a virtual channel inside a virtual path and is used as a routing field in an ATM network.
- *Payload type* (3 bits)—indicates the type of information contained in the cell's payload (control or user data) and provides some additional control information.
- *Cell loss priority* (CLP) (1 bit)—indicates the priority of the cell in case of network congestion. A cell with a CLP value of 1 is considered of low priority and subject to discard if required by network conditions.
- *Header error control* (8 bits)—used for error detection and single-bit error correction; it covers only the (first 4) bytes of the cell header.

ATM is *connection-oriented*, that is, a logical connection (called *virtual circuit* or, in ATM terminology, *virtual channel*) is established between two end stations prior to data transfer, and all packets follow the same preplanned route in the network and arrive at their destinations in sequence. ATM also defines another type of logical connection, the *virtual path*, which is essentially a bundle of virtual channels that share a large part of their path. Cell routing in an ATM network is based on both the virtual path and the virtual channel identifiers (VPI and VCI) contained in the cell header.

During a virtual channel connection (VCC) setup, a user can specify a set of parameters relating to the desired *quality of service* (QoS) and the input *traffic characteristics* of the VCC. QoS parameters include the *cell loss ratio* (CLR) (i.e., the percentage of cells that are lost in the network—due to error or congestion—to

the total transmitted cells); the *cell transfer delay* (CTD) (i.e., the delay experienced by a cell throughout the ATM network); and the *Cell Delay Variation* (CDV) (i.e., the variance of CTD). Input traffic characteristics parameters that can be negotiated between the user and the network include the *peak cell rate* (PCR), which is the maximum rate at which the user will transmit; the *sustained cell rate* (SCR), which is the average transmission rate measured over a long interval; and the *burst tolerance* (BT) and the *maximum burst size* (MBS), which define the burstiness of the sender.

On the basis of these parameters, five service categories are defined that characterize the different types of traffic that can be transferred by an ATM network:

- *Constant bit rate* (CBR), intended for applications that require a fixed data rate and an upper bound on transfer delay (e.g., videoconferencing, telephony).
- *Real-time variable bit rate* (rt-VBR), intended for applications that transmit at a variable rate and are time-sensitive (i.e., require tight upper bounds on cell transfer delay and delay variation, e.g., interactive compressed video).
- *Non-real-time variable bit rate* (nrt-VBR), intended for applications that transmit at a variable rate but are not time-sensitive (e.g., banking transactions).
- *Available bit rate* (ABR), used by normal (bursty) applications with relaxed delay and cell loss requirements (such as file transfer and email) and expected to be the most commonly used service category. ABR sources use explicit network feedback to control their cell rate.
- *Unspecified bit rate* (UBR), used by applications that are not sensitive to delay and cell loss and provide best-effort services by taking advantage of network capacity not allocated to CBR, VBR, or ABR traffic (e.g., file transfer).

6.1. ATM Congestion Avoidance and Control

ATM was designed so as to minimize the processing and transmission overhead inside the network. The only factor that can lead to cell delay variation is network congestion. ATM handles network congestion with functions that fall within two general categories: congestion avoidance and congestion control.

Congestion avoidance is concerned mainly with (1) establishing *traffic contracts* with the users, which specify the traffic parameters for a connection and the QoS parameters the network will support for that connection; and (2) enforcing the agreed-on traffic contracts, that is, watching that the agreed-on restrictions are met (*traffic policing*). Sometimes, however, congestion avoidance actions may not be effective, in which case some nodes may become congested, and congestion control functions, which are based primarily on network feedback, are brought into play.

The QoS required by CBR and VBR traffic is supported by congestion avoidance techniques based on traffic contracts and policing; cells that violate a traffic contract are discarded or tagged as low-priority cells. Traffic policing can be combined with *traffic shaping* to achieve better network efficiency, fair allocation of resources, and reduced average delays, while meeting the QoS objectives. A common mechanism used for traffic shaping is the "leaky bucket." This mechanism can smooth out a bursty cell flow by buffering arriving cells and then serving the queue (i.e., transmitting the cells) at a constant service rate of r cells per second (see Fig. 12). The leaky bucket example shown can be controlled by adjusting two parameters: the bucket capacity and the bucket leaking rate. As long as the bucket is not empty, the cells are transmitted with the constant rate of r cells per second. In case the bucket capacity is exceeded, a bucket overflow is caused and the excessive cells can be either discarded, or tagged as low priority cells.

These techniques prevent congestion buildup but get no feedback from the network concerning the congestion conditions and, therefore, are called *open-loop control techniques*. Open-loop control may be insufficient in cases where the bandwidth requirements of applications are not known at connection setup time. With ABR traffic, however, dynamic load management is possible by taking advantage of network feedback. In this case, feedback techniques are employed (*closed-loop control* techniques) that support the lossless transport of ABR traffic and the fair capacity allocation.

There are two main closed-loop control techniques: the credit-based and the rate-based. *Credit-based* schemes are based on a window flow control mechanism; each intermediate node on a session's path sends information (credit) to the previous (upstream) node and does so on a per link and per VC basis. A traffic source can transmit on a VC only when it obtains credit from the next node for this connection, which implies that the next node has adequate buffer and can accommodate the number of cells specified by the credit it grants to the source. In *rate-based* schemes, the network sends appropriate information to the user, specifying the bit rate at which the user could transmit, and the feedback control loop may extend end-to-end across the network. The rate-based approach is less expensive in terms of implementation complexity and hardware cost, but it doesn't handle bursty traffic well. The credit-based approach, on the other hand, is well suited for bursty traffic (under ideal conditions, zero cell loss can be guaranteed), but it requires complex bookkeeping at

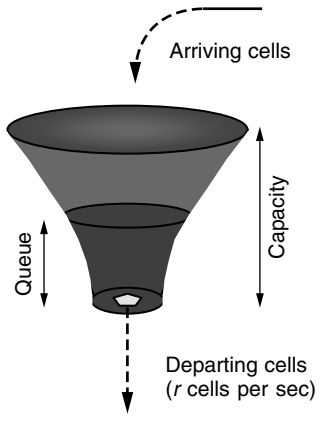


Figure 12. The leaky-bucket mechanism.

the network nodes on a per session (i.e., per VC) basis. The need for per session queuing limits the flexibility of the designer and is one of the main reasons why the ATM Forum has selected rate-based schemes for ABR traffic in ATM networks.

The rate-based scheme adopted for use in ATM uses the following mechanism. The source of an ABR flow inserts among the data cells of the flow some *resource management* (RM) cells, one RM cell after a certain number of data cells. Each RM cell contains three fields that can provide congestion feedback to the source: the *congestion indication* (CI) bit, the *no increase* (NI) bit, and the *explicit cell rate* (ER) field. The values of these fields can be changed by either the destination of the flow or an intermediate ATM switch, to reflect experienced congestion effects. Finally RM cells return to their source, where their fields are examined and corrective actions (rate increase or decrease) are taken to respond to the network or destination conditions. In particular, the source at any time is allowed to transmit cells at any rate between zero and a value called *allowed cell rate* (ACR). ACR is adjusted dynamically according to network feedback and has a lower and an upper limit, called *minimum cell rate* (MCR) and *peak cell rate* (PCR), respectively. If the source gets an RM cell with the CI bit set (signaling congestion), then it reduces ACR by an amount proportional to its current ACR (but down to MCR). If neither CI nor NI is set, then the source increases ACR by an amount proportional to the PCR (but up to PCR). Therefore, rate increases are linear, but rate decreases are exponential, so that sources respond drastically to congestion. Finally, if ACR is bigger than the value contained in the ER field (which is used to explicitly dictate a cell rate), then ACR is reduced to ER.

6.2. Connection Establishment Control Protocols for High-Speed Networks

The rapid developments in optoelectronics technology have substantially increased system transmission rates in optical communication networks since the first systems were installed, in the early–mid-1980s. Having, however, communication links of multigigabit transmission rates does not necessarily result in a communication network of the same effective capacity. An important (but not the only) issue is related to the protocols and algorithms used to perform network control. These protocols should allow full utilization of the network resources in a way that is fair to all users; they should be capable of providing delay and packet loss guarantees to the users (QoS), in the presence of node and link failures, and they should impose small processing requirements on the switches.

6.2.1. Protocols for CBR Traffic and for Traffic Consisting of Long Bursts. A sizable portion of traffic in future multigigabit-per-second networks will involve high-speed transfer of traffic at nearly constant rates (CBR traffic) and would require guaranteed lossless delivery and an explicit reservation of bandwidth. Clearly, the bandwidth–delay product being very large can result in the discarding of substantial amounts of data and retransmissions, unless bandwidth reservations are made in advance, or substantial buffer space is provided. Also, for high-speed file-transfer-type applications, long burst transmissions can easily overload the network, unless they have prenegotiated at least a minimum bandwidth with the network. Therefore, from the point of view of both transmission integrity and network efficiency, traffic of this type should be transferred only after a specific and explicit allocation precedes each data burst. This is especially true for the case of all-optical networks, where buffering has to be very limited because of technological constraints.

A key to efficiently utilizing the large bandwidth of emerging gigabit networks is to devise protocols that can overcome the problems posed by increased propagation latency of such networks. In most reservation protocols [e.g., the FRP/DT protocol, the fast bandwidth reservation schemes and the fast resource management (FRM) protocols], a setup packet is sent to the destination to make the appropriate reservations, and the capacity required by a session at an intermediate node is reserved starting at the time the setup packet arrives at that node. An obvious inefficiency in all these schemes arises because the capacity reserved for the session is not needed immediately, but it is actually needed at least one round-trip delay after arrival of the setup packet at the node. This is because the setup packet has to travel from the intermediate node to the destination, an acknowledgment has to be sent back to the source, and the first data packet of the session has to arrive from the source to the intermediate node (see Fig. 13). Over long transmission distances, the round-trip delay may be comparable to, or even larger than, the holding time of a session. In particular, if a typical session requests capacity r bps (bits per second), and transfers a total of M bits over a distance of L kilometers, the maximum percentage of time that the capacity is efficiently used is $e = (M/r)/[2Ln/c) + (M/r)]$, where c/n is the propagation speed in the fiber. Typical values of these parameters for multigigabit networks may be $r = 10$ Gbps, $M = 0.2$ Gbit, and $L = 1500$ km, which yields $e = 0.57$. This efficiency factor e decreases as r or L increase, or M decreases.

The *efficient reservation virtual circuit* (*ERVC*) protocol was designed to overcome these limitations. It is suitable for sessions that require an explicit reservation of bandwidth, and it does not suffer from the inefficiencies of the reservation protocols mentioned above. The ERVC protocol keeps track of sessions (or burst) durations and

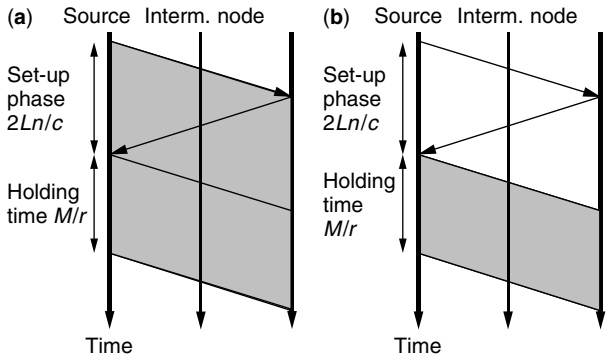


Figure 13. Comparison between (**a**) ERVC and standard reservation (SR) (**b**) protocols (shaded areas represent the time periods during which the capacity is reserved for a particular session).

reserves capacity only for the duration of a session (or burst), thus eliminating the inefficiency that results in existing schemes from holding capacity idle for a round-trip delay before it is actually used by data packets. In the ERVC protocol, session durations (or burst durations) are recorded, and each node keeps track of the utilization profile $r^l(t)$ of each outgoing link l, which describes the amount of residual capacity available on link l as a function of time t. This feature allows capacity to be reserved only for the duration of the session (or burst), starting at the time it is actually needed. Correct timing is crucial in ensuring that data transmission starts *after* all reservations are made and terminates *before* any intermediate node releases the reserved capacity. The ERVC protocol uses capacity on a demand basis, leading to more efficient utilization and a lower blocking probability than previous reservation protocols. It also has the "reservation ahead" feature that allows a node to calculate the time at which the requested capacity will become available and reserve it in advance (provided it is available within the QoS requirements of the session), avoiding in this way the wasteful repetition of the call setup phase. The protocol uses an asynchronous, distributed algorithm that allows the nodes along a session's path to collaborate when reserving capacity and to maintain timing consistency. This ensures that adequate outgoing capacity is available to service the data packets when they arrive at a link, so that the transmission is loss-free. Processing requirements at a node are minimized by using efficient update mechanisms and simple data structures that store a compact representation of the utilization profile of an outgoing link. The information required by the protocol (rates and session durations) can be recorded and processed using a simple linked list structure. The protocol is robust to link and node failures, and it allows soft recovery from processor failures. The efficiency factor e for the ERVC protocol can be as large as $e = 1$, independently of the parameters r, L, and M and efficiency is maintained even for traffic that consists of sporadic bursts of data. Also, the performance of ERVC does not depend on the round-trip delay. This is because for a single link, a different round-trip delay means only that the arrivals of sessions on link l are translated in time by a different amount; therefore, the picture in terms of load (and consequently the blocking) as seen by newly arriving sessions remains the same, irrespective of the round-trip delay.

6.2.2. The Virtual Circuit Deflection Protocol. Traffic in high-speed networks can be switched either optically or electronically. Optical switching has advantages for circuit switching, but substantial disadvantages for packet switching, because effective packet switching requires packet storage at each switch, which is difficult to achieve with current optical technology. Despite this drawback, it is believed that optical switching may open new dimensions in future networking, provided appropriate protocols that take into account its constraints are developed.

To eliminate the need for buffering (but without making advance bandwidth reservations, which requires a round-trip pretransmission delay), a variation of deflection routing, called *virtual circuit deflection* (VCD) protocol can be used. The VCD protocol is a combination of virtual circuit switching and deflection routing, and is appropriate for sessions that simultaneously require minimal pretransmission delay and lossless communication. VCD is a "tell and go" (or "immediate transmission") type of protocol, and does not therefore use end-to-end reservations. In the VCD protocol, a path (called "preferred path") is selected for a new session, based on (possibly outdated) topology and link utilization information available at its source at the time. A setup packet is sent to the destination to establish the connection, followed after a short delay (much shorter than the end-to-end round-trip delay required by reservation protocols) by the data packets. This delay should be large enough to permit the electronic processing of the setup packet, without being overpassed by the data packets. If the available capacity on a preferred link of a session is inadequate, the session may have to follow a different, longer path; we then say that the session is deflected. When the total incoming link capacity is equal to the total outgoing link capacity of a node, as is usually the case in most data networks, it can be shown that there is always adequate available capacity on the outgoing links of an intermediate node to accommodate a new session. This, however, may happen at the expense of interrupting (preempting) an existing session that originates at that node, and/or splitting the new session into two or more smaller subsessions that are routed through different paths (session splitting). Deflection or splitting of sessions at intermediate nodes is infrequent in the VCD protocol, and can happen only when the topology or link utilization information at the source is outdated and the network is congested. Resequencing of packets, which is the major drawback of conventional (datagram) deflection schemes, is much simpler to accomplish in the VCD protocol. If a session is split, a few blocks of data packets (each of which is ordered) will have to be resequenced; this is a considerably easier task to perform than the resequencing of millions of individual packets that are out of order, as is the case in conventional deflection schemes.

Even though the effective utilization of idle links is an advantage, the increase of the number of used links per call is a disadvantage of the VCD protocol. An important performance measure is the *inefficiency ratio* $\eta(\lambda)$, defined as the ratio $\eta(\lambda) = D(\lambda)/D(0)$ of the average path length $D(\lambda)$ taken by a session for a given arrival per node rate λ, over the average shortest pathlength $D(0)$ of a given network topology. Results obtained from experiments on a Manhattan street network topology, indicate that the VCD protocol can be very efficient for high-speed networks, where link capacities are big and links are shared by a large number of small sessions.

7. TCP CONGESTION CONTROL

As we discussed in Section 6.1, TCP implements a credit-based flow control mechanism, using the *window size* field of the TCP header. This way, a destination entity avoids buffer overflow by limiting the dataflow from the source entity. This mechanism, however, can be further enhanced to provide network congestion control.

A source entity using TCP maintains a queue that holds the transmitted but not yet acknowledged segments

of its datastream. If a segment is not acknowledged after a certain time period, then this segment is considered lost and is retransmitted. A critical issue here is the determination of the retransmission timeout period. This is accomplished, in most TCP implementations, on the basis of *round-trip time* (RTT) and *RTT variance* estimates for the transmitted segments. Early TCP implementations calculated RTTs (and checked for timeouts) at "clock ticks" of 500 ms. This clock coarseness, however, led to long timeout intervals and, subsequently, large delays in retransmissions. This was solved by having retransmissions occur when, aside from timeouts (a number of), duplicate ACKs are received. A receiver transmits a duplicate ACK when it cannot acknowledge a segment because an earlier segment is lost. Therefore, when a certain number of duplicate ACKs (usually 3) are received, the source is warned that the segment after the one acknowledged is lost, and so it triggers retransmission. Another parameter that can be managed by congestion control techniques is the size of the TCP window, which can be adapted dynamically to the changing network conditions. A common mechanism for doing this is the *slow-start* mechanism, according to which at the beginning of a transmission or retransmission only one segment is transmitted, and then, for each ACK received, an extra segment is transmitted (in addition to the one acknowledged in the ACK) up to a maximum value. In this way the source probes the network with a small amount of data and increases exponentially its flow up to a certain threshold, after which the flow increases linearly.[9] Increase goes on until segments are lost, which implies that the available bandwidth is exceeded; then the source responds by decreasing its window size. In other words, TCP finds the available bandwidth by congesting the network and causing own fragment losses.

The mechanisms described in the previous paragraph are found in the early TCP implementation known as *Reno*. However, as this field is the focus of extensive research, several alternative and more effective mechanisms came up (and still do!) to enhance the functionality of TCP. Several such modifications were incorporated into the well-known *TCP Vegas* implementation. Vegas is reported to achieve 37–71% better throughput, with one-fifth to one-half the losses as compared to Reno. Specifically, Vegas features a new retransmission mechanism, a congestion avoidance mechanism, and an improved slow-start mechanism. The new retransmission mechanism detects lost segments much sooner than did Reno (and without the need for a second or third duplicate ACK) by using the fine-grain system clock to calculate RTTs.[10] Consequently, when a duplicate ACK is received, the more accurate RTT estimate is compared against the timeout value, and if it is found to be larger, the source retransmits the segment without having to wait for n duplicate ACKs. Moreover, Vegas checks the first couple of nonduplicate ACKs received after a retransmission, and if the calculated (over these ACKs) RTTs exceed the timeout value, then the corresponding segments are retransmitted. In this way any other segment that was lost prior to the retransmission is detected without having to wait for a duplicate ACK. Vegas also implements a congestion avoidance (*proactive*) mechanism, in contrast to the inherently *reactive* congestion detection mechanism of Reno. This mechanism performs a comparison between the *Expected throughput* of the connection [calculated by dividing the size of the current (congestion) window by the minimum of all the measured RTTs] and the measured *Actual throughput* (calculated by dividing the actual number of bytes transmitted during the RTT of a segment by this RTT). The result of this comparison (i.e., *expected throughput* minus *actual throughput,* which is always nonnegative) is used to adjust the congestion window size, and in particular, if the difference is below a low threshold (α), the actual and expected throughput values are too close and therefore the window size is increased linearly to catch up with the available bandwidth. If, on the other hand, the difference is above a high threshold (β), the actual and expected throughput values are too distant and therefore the window size is decreased linearly to react to the network congestion observed. The window size remains constant if the difference is between the two threshold values. The two threshold values practically correspond to the number of network buffers the connection occupies; thus, Vegas detects network congestion and responds to it by trying to limit the number of occupied buffers. Finally, Vegas uses a modified slow-start mechanism, which is integrated with the congestion avoidance mechanism described previously. At the beginning of the connection, this mechanism tries to find the connection's available bandwidth, without incurring the segment losses that Reno does. This is done by allowing the (exponential) growth of the congestion window only every other RTT. In the meantime, the window size remains constant, so that a comparison between the expected and actual rates can signal congestion and trigger a switch to the linear increase/decrease mode, implying that connection's available bandwidth is reached.

[9] The exact value of the threshold is not known during the initial slow start; when, however, a retransmit timeout occurs (indicating congestion), the threshold is set to half the current window size.

[10] The RTT of a segment is calculated by subtracting the segment's transmission time from the corresponding ACK reception time (by the segment source).

BIOGRAPHIES

Emmanouel (Manos) Varvarigos was born in Athens, Greece, in 1965. He received a Diploma in Electrical and Computer Engineering from the National Technical University of Athens in 1988, and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, Cambridge, in 1990 and 1992, respectively. In 1990 he was a researcher at Bell Communications Research, Morristown, New Jersey. From 1992 to 1998 he was an Assistant and later an Associate Professor at the department of Electrical and Computer Engineering at the University of California, Santa Barbara. In 1998/99 he was an Associate Professor at the Electrical Engineering Department at Delft University of Technology, the Netherlands. In 1999 he became a Professor in the Department of Computer Engineering

and Informatics at the University of Patras, where he is currently Director of the Hardware and Computer Architecture Division and Head of the Data Transmission and Networking Lab. His research activities are in the areas of protocols and algorithms for high-speed networks, all-optical networks, high-performance switch architectures, parallel and distributed computing, interconnection networks, VLSI layout design, performance evaluation, and ad hoc networks.

Theodora A. Varvarigou received the B.Tech. degree from the National Technical University of Athens, Athens, Greece in 1988, the M.S. degrees in Electrical Engineering (1989) and in Computer Science (1991) from Stanford University, Stanford, California in 1989 and the Ph.D. degree from Stanford University as well in 1991.

She worked at AT&T Bell Labs, Holmdel, New Jersey between 1991 and 1995. Between 1995 and 1997 she worked as an Assistant Professor at the Technical University of Crete, Chania, Greece. Since 1997 she has been working as an Assistant Professor at the National Technical University of Athens.

Her research interests include parallel algorithms and architectures, fault-tolerant computation, and parallel scheduling on multiprocessor systems.

FURTHER READING

J. S. Ahn et al., Evaluation of TCP Vegas: Emulation and experiment, *Proc. SIGCOMM '95*, Aug. 1995.

D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

U. Black, *Data Link Protocols*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

F. Bonomi and K. Fendick, The rate-based flow control framework for the available bit rate ATM service, *IEEE Network* (March/April 1995).

L. Brakmo and L. Peterson, TCP Vegas: End to end congestion avoidance on a global Internet, *IEEE J. Select. Areas Commun.* (Oct. 1995).

J. Carlo et al., *Understanding Token Ring Protocols and Standards*, Artech House, Boston, 1999.

E. Carne, *Telecommunications Primer*, Prentice-Hall, Upper Saddle River, NJ, 1995.

T. Chen, S. Liu, and V. Samalam, The available bit rate service for data in ATM networks, *IEEE Commun. Mag.* (May 1996).

D. Clark, The design philosophy of the DARPA Internet protocols, *Proc. SIGCOMM '88, Computer Communication Review*, Aug. 1988.

D. Clark, S. Shenker, and L. Zhang, Supporting real-time applications in an integrated services packet network: Architecture and mechanism, *Proc. SIGCOMM '92*, Aug. 1992.

D. Comer and D. Stevens, *Internetworking with TCP/IP,* Vol. II: *Design Implementation, and Internals*, Prentice-Hall, Upper Saddle River, NJ, 1999.

D. Comer, *Internetworking with TCP/IP,* Vol. I: *Principles, Protocols, and Architecture*, Prentice-Hall, Upper Saddle River, NJ, 1995.

B. Dorling, P. Pieters, and E. Valenzuela, *IBM Frame Relay Guide*, IBM Publication SG24-4463-01, 1996.

A. Eckeberg, D. Luan, and M. Lucantoni, An approach to controlling congestion in ATM networks, *Int. J. Digital Analog Commun. Syst.* **3**(2): 1990.

A. Gersht and K. Lee, A congestion control framework for ATM networks, *IEEE J. Select. Areas Commun.* (Sept. 1991).

W. Goralski, *Introduction to ATM Networking*, McGraw-Hill, New York, 1995.

F. Halsall, *Data Communications, Computer Networks, and Open Systems*, Addison Wesley, Reading, MA, 1996.

R. Handel, N. Huber, and S. Schroder, *ATM Networks: Concepts, Protocols, Applications*, Addison Wesley, Reading, MA, 1994.

S. Haykin, *Communication Systems*, Wiley, New York, 1995.

C. Huitema, *Routing in the Internet*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

C. Huitema, *Ipv6: The New Internet Protocol*, Prentice-Hall, Upper Saddle River, NJ, 1998.

Internetworking Technologies Handbook, 3rd ed., Cisco Press, 2000.

V. Jacobson, Congestion avoidance and control, *Proc. SIGCOMM '88, Computer Communication Review*, Aug. 1988.

V. Jacobson, Berkeley TCP evolution from 4.3 Tahoe to 4.3-Reno, *Proc. 18th Internet Engineering Task Force*, Sept. 1990.

R. Jain, Congestion control in computer networks: Issues and trends, *IEEE Network Mag.* (May 1990).

R. Jain, Myths about congestion management in high-speed networks, *Internetworking: Research and Experience*, Vol. 3, 1992.

R. Jain et al., Source behavior for ATM ABR traffic management: An explanation, *IEEE Commun. Mag.* (Nov. 1996).

R. Jain, Congestion control and traffic management in ATM networks: Recent advances and a survey, *Computer Networks ISDN Syst.* **28**(13): (Oct. 1996).

P. Karn and C. Partridge, Improving round-trip estimates in reliable transport protocols, *ACM Trans. Comput. Syst.* (Nov. 1991).

N. Kavak, Data communication in ATM networks, *IEEE Network* (May/June 1995).

H. T. Kung, T. Blackwell, and A. Chapman, A credit-based flow control scheme for ATM networks: Credit update protocol, adaptive credit allocation, and statistical multiplexing, *Proc. SIGCOMM '94*, Aug./Sept. 1994.

H. T. Kung and R. Morris, Credit-based flow control for ATM networks, *IEEE Network* (March 1995).

S. Low, L. Peterson, and L. Wang, Understanding TCP Vegas: A duality model, *Proc. SIGMETRICS '01*, June 2001.

D. Mc Dysan and D. Spohn, *ATM: Theory and Applications*, McGraw-Hill, New York, 1999.

S. Miller, *IPv6: The Next Generation Internet Protocol*, Digital Press, Bedford, MA, 1998.

J. Mo et al., Analysis and comparison of TCP Reno and Vegas, *Proc. IEEE Infocom*, March 1999.

G. Moshos, *Data Communications: Principles and Problems*, West Publishing, New York, 1989.

M. Murhammer et al., *TCP/IP: Tutorial and Technical Overview*, Prentice-Hall, Upper Saddle River, NJ, 1998.

M. Naugle, *Local Area Networking*, McGraw-Hill, New York, 1996.

P. Newman, ATM local area networks, *IEEE Commun. Mag.* (March 1994).

H. Oshaki et al., Rate-based congestion control for ATM networks, *Comput. Commun. Rev.* (April 1995).

L. Peterson and B. Davie, *Computer Networks: A Systems Approach*, Morgan Kaufmann, San Francisco, 1996.

J. Pitts and J. Schormans, *Introduction to ATM Design and Performance*, Wiley, New York, 1996.

J. Proakis and M. Salehi, *Communication Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

M. Prycker, *Asynchronous Transfer Mode: Solutions for Broadband ISDN*, Ellis Horwood, New York, 1996.

A. Rodriguez et al., *TCP/IP Tutorial and Technical Overview*, 7th ed., IBM Publication GG24-3376-06, 2001.

K. Sato, S. Ohta, and I. Tokizawa, Broadband ATM network architecture based on virtual paths, *IEEE Trans. Commun.* (Aug. 1990).

M. Schwartz, *Computer-Communication Network Design and Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

A. Shah and G. Ramakrishnan, *FDDI: A High-Speed Network*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

D. Spohn, *Data Network Design*, McGraw-Hill, New York, 1994.

J. Spragins, J. Hammond, and K. Pawlikowski, *Telecommunication Protocols and Design*, Addison-Wesley, Reading, MA, 1991.

W. Stallings, *High-Speed Networks: TCP/IP and ATM Design Principles*, Prentice-Hall, Upper Saddle River, NJ, 1998.

W. Stallings, *Local and Metropolitan Area Networks*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2000.

W. Stallings, *Data and Computer Communications*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2000.

M. Steenstrup, *Routing in Communications Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

S. Steinke, IP addresses and subnet masks, *LAN Mag.* (Oct. 1995).

W. Stevens, *TCP/IP Illustrated,* Vol. 1: *The Protocols*, Addison-Wesley, Reading, MA, 1994.

A. Tanenbaum, *Computer Networks*, Prentice-Hall, Upper Saddle River, NJ, 1996.

E. Varvarigos, Control protocols for multigigabit-per-second networks, *IEICE Trans. Commun.* (Feb. 1998).

W. Weiss, QoS with differentiated services, *Bell Labs Tech. J.* (Oct.–Dec. 1998).

P. White and J. Crowcroft, The integrated services in the Internet: State of the art, *Proc. IEEE*, Dec. 1997.

G. Wright and W. Stevens, *TCP/IP Illustrated,* Vol. 2: *The Implementation*, Addison-Wesley, Reading, MA, 1995.

X. Xiao and L. Ni, Internet QoS: A big picture, *IEEE Network* (March/April 1999).

C. Yang and A. Reddy, A taxonomy for congestion control algorithms in packet switching networks, *IEEE Network* (July/Aug. 1995).

L. Zhang, Why TCP timers don't work well, *Proc., SIGCOMM '86 Symp.*, Aug. 1986.

H. Zhang, Service disciplines for guaranteed performance service in packet switching networks, *Proc. IEEE*, Oct. 1995.

CONCATENATED CONVOLUTIONAL CODES AND ITERATIVE DECODING

WILLIAM E. RYAN
University of Arizona
Tucson, Arizona

1. INTRODUCTION

Turbo codes, first presented to the coding community in 1993 [1,2], represent one of the most important breakthroughs in coding since Ungerboeck introduced trellis codes in 1982 [3]. A turbo code encoder, comprises a concatenation of two (or more) convolutional encoders, and its decoder consists of two (or more) "soft" convolutional decoders that feed probabilistic information back and forth to each other in a manner that is reminiscent of a turbo engine. This chapter presents a tutorial exposition of parallel and serial concatenated convolutional codes (PCCCs and SCCCs), which we will also call *parallel* and *serial turbo codes*. Included here are a simple derivation for the performance of these codes and a straightforward presentation of their iterative decoding algorithms. The treatment is intended to be a launching point for further study in the field and to provide sufficient information for the design of computer simulations. This article borrows from some of the most prominent publications in the field [4–12].

The article is organized as follows. Section 2 describes details of the parallel and serial turbo code encoders. Section 3 derives a truncated union bound on the error rate of these codes under the assumption of maximum-likelihood decoding. This section explains the phenomenon of interleaver gain attained by these codes. Section 4 derives in detail the iterative (turbo) decoder for both PCCCs and SCCCs. Included in this section are the BCJR (Bahl–Cocke–Jelinek–Raviv) decoding algorithm for convolutional codes and soft-in/soft-out decoding modules for use in turbo decoding. The decoding algorithms are presented explicitly to facilitate the creation of computer programs. Section 5 contains a few concluding remarks.

2. ENCODER STRUCTURES

Figure 1 depicts a parallel turbo encoder. As seen in the figure, the encoder consists of two binary rate 1/2 convolutional encoders arranged in a so-called parallel concatenation, separated by a K-bit pseudorandom interleaver or permuter. Also included is an optional puncturing mechanism to obtain high code rates [13]. Clearly, without the puncturer, the encoder is rate $\frac{1}{3}$, mapping K data bits to $3K$ code bits. With the puncturer, the code rate $R = K/(K+P)$, where P is the number of parity bits remaining after puncturing. Observe that the constituent encoders are recursive systematic convolutional (RSC) codes. As will be seen below, recursive encoders are necessary to attain the exceptional performance (attributed to "interleaver gain") provided by

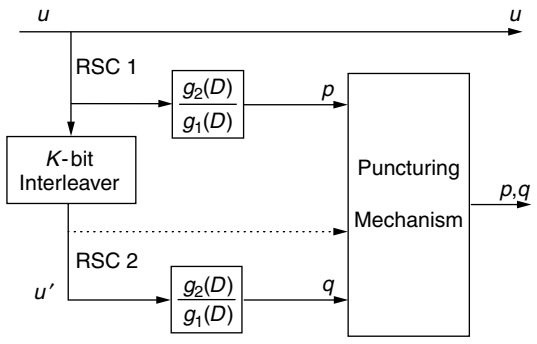


Figure 1. PCCC encoder diagram.

turbo codes. Without any essential loss of generality, we assume that the constituent codes are identical.

Figure 2 depicts a serial turbo encoder. As seen in the figure, the serially concatenated convolutional encoders are separated by an interleaver, and the inner encoder is required to be an RSC code, whereas the outer encoder need not be recursive [6]. However, RSC inner and outer encoders are often preferred since it is convenient to puncture only parity bits to obtain high code rates [14]. Further, an RSC outer code will facilitate our analysis below. The code rate for the serial turbo encoder is $R = R_o \cdot R_i$ where R_o and R_i are the code rates for the outer and inner codes, respectively.

For both parallel and serial turbo codes, the codeword length is $N = K/R$ bits, and we may consider both classes to be (N, K) block codes.

We now discuss in some detail the individual components of the turbo encoders.

2.1. The Recursive Systematic Encoders

Whereas the generator matrix for a rate $\frac{1}{2}$ nonrecursive convolutional code has the form $G_{NR}(D) = [g_1(D)\ g_2(D)]$, the equivalent recursive systematic encoder has the generator matrix

$$G_R(D) = \left[1\ \frac{g_2(D)}{g_1(D)}\right]$$

Observe that the code sequence corresponding to the encoder input $u(D)$ for the former code is $u(D)G_{NR}(D) = [u(D)g_1(D)\ \ u(D)g_2(D)]$, and that the identical code sequence is produced in the recursive code by the sequence $u'(D) = u(D)g_1(D)$, since in this case the code sequence is $u(D)g_1(D)G_R(D) = u(D)G_{NR}(D)$. Here, we loosely call the pair of polynomials $[u(D)g_1(D)\ u(D)g_2(D)]$ a *code sequence*, although the actual code sequence is derived from this polynomial pair in the usual way.

Observe that, for the recursive encoder, the code sequence will be of finite weight if and only if the input sequence is divisible by $g_1(D)$. We have the following corollaries of this fact, which we shall use later.

Fact 1. A weight 1 input will produce an infinite weight output for such an input is never divisible by a (nontrivial) polynomial $g_1(D)$. (In practice, "infinite" should be replaced by "large" since the input length is finite.)

Fact 2. For any nontrivial $g_1(D)$, there exists a family of weight 2 inputs of the form $D^j(1 + D^p)$, $j \geq 0$, which produce finite weight outputs, i.e., which are divisible by $g_1(D)$. When $g_1(D)$ is a primitive polynomial of degree m, then $p = 2^m - 1$. More generally, p is the length of the pseudorandom sequence generated by $g_1(D)$.

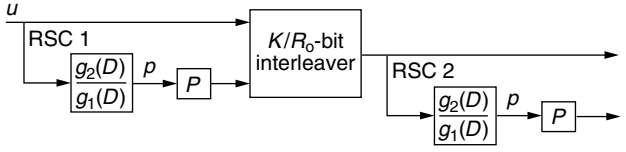


Figure 2. SCCC encoder diagram with RSC component codes. "P" signifies possible puncturing of parity bits.

Proof: Because the encoder is linear, its output due to a weight 2 input $D^j(1 + D^t)$ is equal to the sum of its outputs due to D^j and D^jD^t. The output due to D^j will be periodic with period p since the encoder is a finite-state machine (see Example 1 and Fig. 3); the state at time j must be reached again in a finite number of steps p, after which the state sequence is repeated indefinitely with period p. Now letting $t = p$, the output due to D^jD^p is just the output due to D^j shifted by p bits. Thus, the output due to $D^j(1 + D^p)$ is the sum of the outputs due to D^j and D^jD^p, which must be of finite length and weight since all but one period will cancel in the sum.

In the context of the code's trellis, fact 1 says that a weight-1 input will create a path that diverges from the all-zeros path, but never remerges. Fact 2 says that there will always exist a trellis path that diverges and remerges later, which corresponds to a weight 2 data sequence.

Example 1. Consider the code with generator matrix

$$G_R(D) = \left[1\ \frac{1 + D^2 + D^3 + D^4}{1 + D + D^4}\right].$$

Thus $g_1(D) = 1 + D + D^4$ and $g_2(D) = 1 + D^2 + D^3 + D^4$ or, in octal form, $(g_1, g_2) = (31, 27)$. Observe that $g_1(D)$ is primitive so that, for example, $u(D) = 1 + D^{15}$ produces the finite-length code sequence $(1 + D^{15}, 1 + D + D^3 + D^4 + D^7 + D^{11} + D^{12} + D^{13} + D^{14} + D^{15})$. Of course, any delayed version of this input, say, $D^7(1 + D^{15})$, will simply produce a delayed version of this code sequence. Figure 3 gives one encoder realization for this code. We remark that, in addition to elaborating on Fact 2, this example serves to demonstrate the conventions generally used in the literature for specifying such encoders.

2.2. The Interleaver

The function of the interleaver is to take each incoming block of bits and rearrange them in a pseudorandom fashion prior to encoding by the second encoder. For the PCCC, the interleaver permutes K bits and, for the SCCC, the interleaver permutes K/R_o bits. Unlike the classical interleaver (e.g., block or convolutional interleaver), which

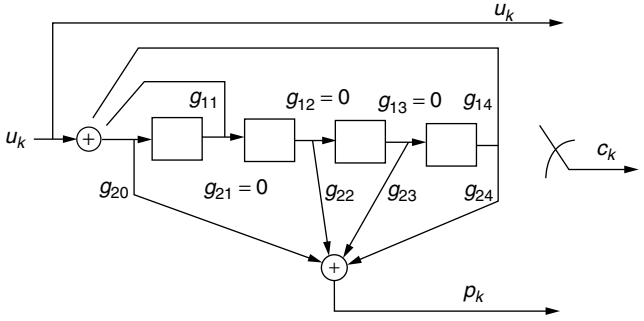


Figure 3. RSC encoder with $(g_1, g_2) = (31, 27)$.

rearranges the bits in some systematic fashion, it is crucial that this interleaver sort the bits in a manner that lacks any apparent order, although it might be tailored in a certain way for weight 2 and weight 3 inputs as will be made clearer below. The $\mathcal{S}$ random interleaver [8] is quite effective in this regard. This particular interleaver ensures that any two inputs bits whose positions are within $\mathcal{S}$ of each other are separated by an amount greater than $\mathcal{S}$ at the interleaver output. $\mathcal{S}$ should be selected to be as large as possible for a given value of K. Also, as we shall see, performance increases with K, and so $K \geq 1000$ is typical.

2.3. The Puncturer

The role of the turbo code puncturer is identical to that of its convolutional code counterpart, that is, to delete selected bits to reduce coding overhead. We have found it most convenient to delete only parity bits, but there is no guarantee that this will maximize the minimum codeword distance. For example, to achieve a rate of $\frac{1}{2}$, one might delete all even parity bits from the top encoder and all odd parity bits from the bottom one.

3. PERFORMANCE WITH MAXIMUM-LIKELIHOOD DECODING

As will be elaborated upon in the next section, a maximum-likelihood (ML) sequence decoder would be far too complex for a turbo code, due to the presence of the permuter. However, the suboptimum iterative decoding algorithm to be described there offers near-ML performance. Hence, we shall now estimate the performance of an ML decoder on a binary input AWGN channel with power spectral density $N_0/2$ (analysis of the iterative decoder is much more difficult).

Armed with the preceding descriptions of the components of the turbo encoders of Figs. 1 and 2, we can easily conclude that it is linear since its components are linear. The constituent codes are certainly linear, and the interleaver is linear since it may be modeled by a permutation matrix. Further, the puncturer does not affect linearity since all the constituent codewords share the same puncture locations. As usual, the importance of linearity in evaluating the performance of a code is that one may choose the all-zeros sequence as a reference. Thus, we shall assume that the all-zeros codeword was transmitted. The development below holds for both parallel and serial turbo codes.

Now consider the all-zeros codeword (the 0th codeword) and the kth codeword, for some $k \in \{1, 2, \ldots, 2^K - 1\}$. The ML decoder will choose the kth codeword over the 0th codeword with probability $Q(\sqrt{2d_k RE_b/N_0})$, where d_k is the weight of the kth codeword and E_b is the energy per information bit. The bit error rate for this two-codeword situation would then be

$$\begin{aligned} P_b(k|0) &= w_k \text{ (bit errors/cw error)} \\ &\quad \times \frac{1}{K} \text{ (cw/ data bits)} \\ &\quad \times Q(\sqrt{2Rd_kE_b/N_0}) \text{ (cw errors/cw)} \\ &= \frac{w_k}{K} Q\left(\sqrt{\frac{2Rd_kE_b}{N_0}}\right) \text{ (bit errors/data bit)} \end{aligned}$$

where w_k is the weight of the kth data word and "cw" = codeword. Now including all of the codewords and invoking the usual union bounding argument, we may write

$$\begin{aligned} P_b &= P_b(\text{choose any } k \in \{1, 2, \ldots, 2^K - 1\} \mid 0) \\ &\leq \sum_{k=1}^{2^K-1} P_b(k \mid 0) \\ &= \sum_{k=1}^{2^K-1} \frac{w_k}{K} Q\left(\sqrt{\frac{2Rd_kE_b}{N_0}}\right) \end{aligned}$$

Note that every nonzero codeword is included in the above summation. Let us now reorganize the summation as

$$P_b \leq \sum_{w=1}^{K} \sum_{v=1}^{\binom{K}{w}} \frac{w}{K} Q\left(\sqrt{\frac{2Rd_{wv}E_b}{N_0}}\right) \tag{1}$$

where the first sum is over the weight w inputs, the second sum is over the $\binom{K}{w}$ different weight w inputs, and d_{wv} is the weight of the v^{th} codeword produced by a weight w input. We emphasize that (1) holds for any linear code.

Consider now the first few terms in the outer summation of (1) in the context of parallel and serial turbo codes. Analogous to the fact that the top encoder in the parallel scheme is recursive, we shall assume that the outer encoder in the serial scheme is also recursive. By doing so, our arguments below will hold for both configurations. Further, an RSC outer code facilitates the design of high-rate serial turbo codes as mentioned above.

$w = 1$. From Fact 1 and associated discussion above, weight 1 inputs will produce only large weight codewords at both PCCC constituent encoder outputs since the trellis paths created never remerge with the all-zeros path. (We ignore the extreme case where the single 1 occurs at the end of the input words for both encoders for this is avoidable by proper interleaver design.) For the SCCC, the output of the outer encoder will have large weight because of Fact 1, and its inner encoder output will have large weight since its input has large weight. Thus, for both cases, each d_{1v} can be expected to be significantly greater than the minimum codeword weight so that the $w = 1$ terms in (1) will be negligible.

$w = 2$. Suppose that, of the $\binom{K}{2}$ possible weight 2 encoder inputs, $u_*(D)$ is the one of least degree that yields the minimum turbo codeword weight, $d_{2,\min}$, for weight-2 inputs. In the presence of the pseudorandom interleaver, the encoder is not time-invariant, and only a small fraction of the inputs of the form $D^j u_*(D)$ (there are approximately K of them) will also produce turbo codewords of weight $d_{2,\min}$. (This phenomenon has been called *spectral thinning* [10].) Denoting by n_2 the number of weight 2 inputs that produce weight $d_{2,\min}$ turbo codewords, we may conclude that $n_2 \ll K$. (For comparison, $n_2 \simeq K$ for a single RSC code as shifts of some worst-case input merely shifts the encoder output, thus maintaining a constant output weight.) Further, the overall minimum codeword weight, $d_{\min}$, is likely to be equal or close to $d_{2,\min}$ since low-degree, low-weight input words tend to produce low-weight

codewords. (This is easiest to see in the parallel turbo code case which is systematic.)

$w \geq 3$. When w is small (e.g., $w = 3$ or 4), an argument similar to the $w = 2$ case may be made to conclude that the number of weight w inputs, n_w, that yield the minimum turbo codeword weight for weight-w inputs, $d_{w,\min}$, is such that $n_w \ll K$. Further, we can expect $d_{w,\min}$ to be equal or close to $d_{\min}$. No such arguments can be made as w increases beyond about 5.

To summarize, by preserving only the dominant terms, the bound in (1) can be approximated as

$$P_b \simeq \sum_{w=2}^{3} \frac{w n_w}{K} Q\left(\sqrt{\frac{2Rd_{w,\min}E_b}{N_0}}\right) \tag{2}$$

where $w \geq 4$ terms may be added in the event that they are not negligible (more likely to be necessary for SCCCs). We note that n_w and $d_{w,\min}$ are functions of the particular interleaver employed. Since $w = 2$ or 3 in (1) and $n_w \ll K$ with $K \geq 1000$, the coefficients out in front of the Q function are much less than unity. (For comparison, the coefficient for a convolutional code can be much greater than unity [10].) This effect, called *interleaver gain*, demonstrates the necessity of large interleavers. Finally, we note that recursive encoders are crucial elements of a turbo code since, for nonrecursive encoders, division by $g_1(D)$ (nonremergent trellis paths) would not be an issue and (2) would not hold [although (1) still would].

When $K \simeq 1000$, it is possible to exhaustively find via computer the weight spectra $\{d_{2v}: v = 1, \ldots, \binom{K}{2}\}$ and $\{d_{3v}: v = 1, \ldots, \binom{K}{3}\}$ corresponding to the weight 2 and 3 inputs. In this case, an improved estimate of P_b, given by a truncation of (1), is

$$P_b \simeq \sum_{w=2}^{3} \sum_{v=1}^{\binom{K}{w}} \frac{w}{K} Q\left(\sqrt{\frac{2Rd_{wv}E_b}{N_0}}\right) \tag{3}$$

We remark that if codeword error rate, P_{cw}, is the preferred performance metric, then an estimate of P_{cw} may be obtained from (2) or (3) by removing the factor w/K from these expressions. That this is so may be seen by following the derivation above for P_b.

Example 2. We consider in this example a PCCC and an SCCC code, both rate $\frac{8}{9}$ with parameters $(N, K) = (1152, 1024)$. We use identical 4-state RSC encoders in the PCCC encoder whose generators polynomials are, in octal form, $(g_1, g_2) = (7, 5)$. To achieve a code rate of $\frac{8}{9}$, only one bit is saved in every 16-bit block of parity bits at each encoder output. The outer constituent encoder in the SCCC encoder is this same 4-state RSC encoder, and the inner code is a rate-1 differential encoder with transfer function $\frac{1}{1 \oplus D}$. A rate of $\frac{8}{9}$ is achieved in this case by saving one bit in every 8-bit block of parity bits. The PCCC interleaver is a 1024-bit pseudorandom interleaver with no constraints added (e.g., no $\mathcal{S}$-random constraint). The SCCC interleaver is a 1152-bit pseudorandom interleaver with no constraints added.

Figure 4 presents performance results for these codes based on computer simulation using the iterative (i.e., non-ML) decoding algorithm of the next section. Simulation results for both bit error rate P_b (BER in the figure) and

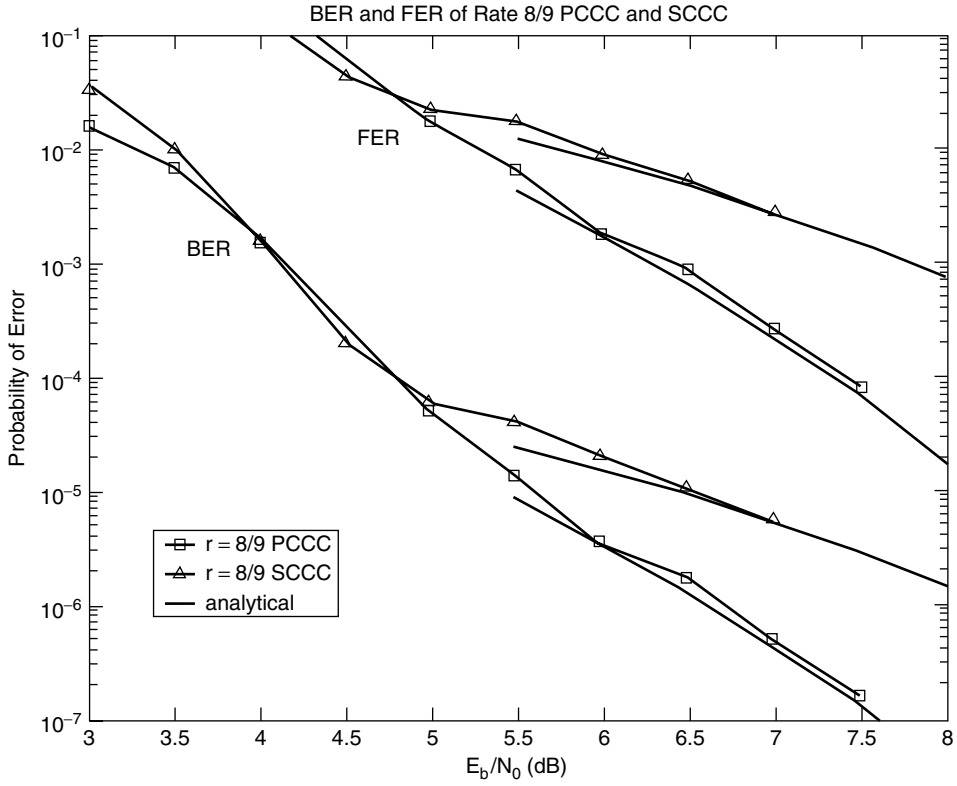


Figure 4. PCCC and SCCC bit error rate (BER) and frame error rate (FER) simulation results together with analytical result in (3).

frame or codeword error rate P_{cw} (FER in the figure) are presented. Also included in the figure are analytic performance curves for ML decoding using the truncated union bound in (3). (P_{cw} is obtained by removing the factor w/K in (3) as indicated above.) We see the close agreement between the analytical and simulated results in this figure.

In addition to illustrating the use of the estimate (3), this example helps explain the "flooring" effect of the error rate curves: it may be interpreted as the usual Q-function shape for a signaling scheme with a modest $d_{\min}$, "pushed down" by the interleaver gain $w^* n_{w^*}/K$, where w^* is the value of w corresponding to the dominant term in (2) or (3).

We comment on the fact that the PCCC in Fig. 4 is substantially better than the SCCC whereas it is known that SCCCs generally have lower floors [6]. We attribute this to the fact that the outer RSC code in the SCCC has been punctured so severely that $d_{\min} = 2$ for this outer code (although $d_{\min}$ for the SCCC is a bit larger). The RSC encoders for the PCCC is punctured only half as much, and so $d_{\min} > 2$ for each of these encoders. We also attribute this to the fact that we have not used an optimized interleaver for this example. In support of these comments, we have also simulated rate $\frac{1}{2}$ versions of this same code structure so that no puncturing occurs for the SCCC and much less occurs for the PCCC. In this case, $(N, K) = (2048, 1024)$ and $\mathcal{S}$-random interleavers were used ($\mathcal{S} = 16$ for PCCC and $\mathcal{S} = 20$ for SCCC). The results are presented in Fig. 5, where we observe that the SCCC has a much lower error rate floor, particularly for the FER curves. Finally, we remark that $w \geq 4$ terms in (2) are necessary for an accurate estimate of the floor level of the SCCC case in Fig. 5.

4. THE ITERATIVE DECODERS

4.1. Overview of the Iterative Decoder

Consider first an ML decoder for a rate $\frac{1}{2}$ convolutional code (recursive or not), and assume a data word of length $K \geq 1000$. Ignoring the structure of the code, a naive ML decoder would have to compare (correlate) 2^K code sequences to the noisy received sequence, choosing in favor of the codeword with the best correlation metric. Clearly, the complexity of such an algorithm is exorbitant. Fortunately, as we know, such a brute-force approach is simplified greatly by the Viterbi algorithm, which permits a systematic elimination of candidate code sequences.

Unfortunately, we have no such luck with turbo codes, for the presence of the interleaver immensely complicates the structure of a turbo code trellis. A near-optimal solution is an iterative decoder (also called a *turbo decoder*) involving two soft-in/soft-out (SISO) decoders that share probabilistic information cooperatively and iteratively. The goal of the iterative decoder is to iteratively estimate the *a posteriori* probabilities (APPs) $\Pr(u_k \mid \mathbf{y})$ where u_k is the kth data bit, $k = 1, 2, \ldots, K$, and $\mathbf{y}$ is the received codeword in noise, $\mathbf{y} = \mathbf{c} + \mathbf{n}$. In this equation, we assume the components of $\mathbf{c}$ take values in the set $\{\pm 1\}$ (and similarly for $\mathbf{u}$) and $\mathbf{n}$ is a noise word whose components

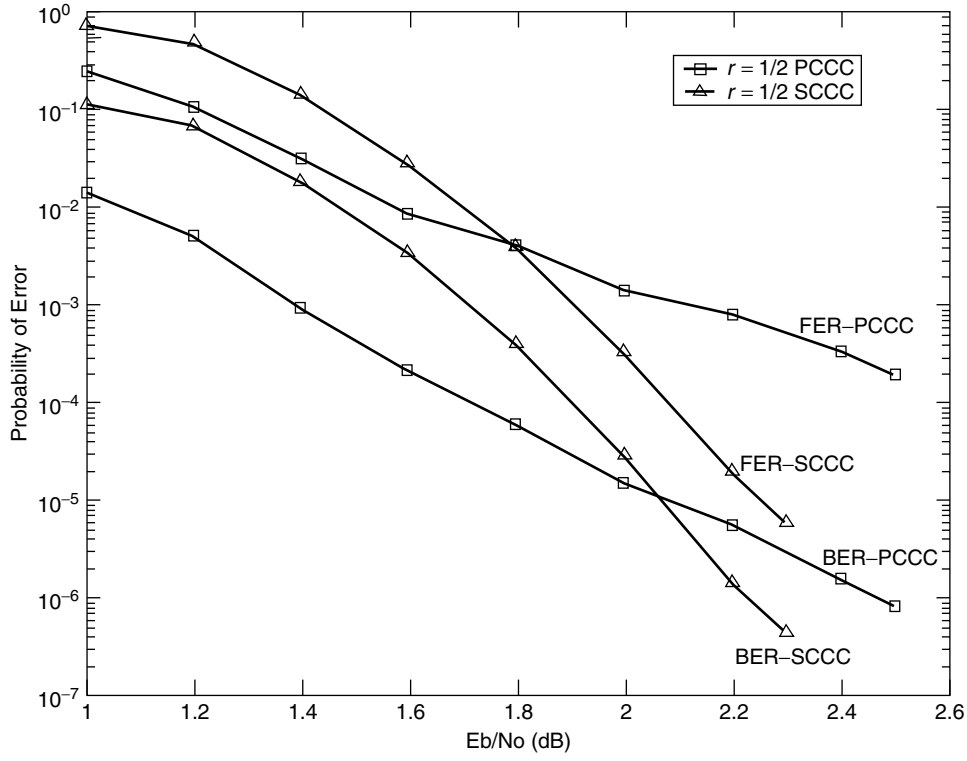


Figure 5. Rate $\frac{1}{2}$ PCCC and SCCC bit error rate (BER) and frame error rate (FER) simulation results.

are AWGN samples. Knowledge of the APPs allows for optimal decisions on the bits u_k via the maximum *a posteriori* (MAP) rule[1]

$$\frac{P(u_k = +1 \mid \mathbf{y})}{P(u_k = -1 \mid \mathbf{y})} \underset{-1}{\overset{+1}{\gtrless}} 1$$

or, more conveniently

$$\hat{u}_k = \text{sign } [L(u_k)]$$

where $L(u_k)$ is the log *a posteriori* probability (log-APP) ratio defined as

$$L(u_k) \triangleq \log\left(\frac{P(u_k = +1 \mid \mathbf{y})}{P(u_k = -1 \mid \mathbf{y})}\right) \tag{4}$$

We shall use the term *log-likelihood ratio* (LLR) in place of log-APP ratio for consistency with the literature.

The component SISO decoders that jointly estimate the LLRs $L(u_k)$ for parallel and serial turbo codes compute the LLRs for component code inputs (u_{ik}), component code outputs (c_{ik}), or both. Details on the SISO decoders will be presented below. For now, we simply introduce the convention that, for PCCCs, the top component encoder is encoder 1 (denoted E1) and the bottom component decoder is encoder 2 (denoted E2). For SCCCs, the outer encoder is encoder 1 (E1) and the inner encoder is encoder 2 (E2). The SISO component decoders matched to E1 and E2 will be denoted by D1 and D2, respectively. Because the SISO decoders D1 and D2 compute $L(u_{ik})$ and/or $L(c_{ik}), i = 1, 2$, we will temporarily use the notation $L(b_k)$ where b_k represents either u_{ik} or c_{ik}.

From Bayes' rule, the LLR for an arbitrary SISO decoder can be written as

$$L(b_k) = \log\left(\frac{P(\mathbf{y} \mid b_k = +1)}{P(\mathbf{y} \mid b_k = -1)}\right) + \log\left(\frac{P(b_k = +1)}{P(b_k = -1)}\right) \tag{5}$$

with the second term representing *a priori* information. Since $P(b_k = +1) = P(b_k = -1)$ typically, the *a priori* term is usually zero for conventional decoders. However, for *iterative* decoders, each component decoder receives *extrinsic* or *soft* information for each b_k from its companion decoder, which serves as *a priori* information. The idea behind extrinsic information is that D2 provides soft information to D1 for each b_k using only information not available to D1, and D1 does likewise for D2. For SCCCs, the iterative decoding proceeds as D2 $\rightarrow$ D1 $\rightarrow$ D2 $\rightarrow$ D1 $\rightarrow \ldots$, with the previous decoder passing soft information along to the next decoder at each half-iteration. For PCCCs, either decoder may initiate the chain of component decodings or, for hardware implementations, D1 and D2 may operate simultaneously.

This type of iterative algorithm is known to converge to the true value of the LLR $L(u_k)$ for the concatenated code provided that the graphical representation of this code contains no loops [15–17]. The graph of a turbo code does in fact contain loops [17], but the algorithm nevertheless provides near-optimal performance for virtually all turbo codes. That this is possible even in the presence of loops is not fully understood.

This section provided an overview of the turbo decoding algorithm in part to motivate the next section on SISO decoding of a single RSC code using the BCJR algorithm [18]. Following the description of the SISO decoder for a single RSC code will be sections that describe in full detail the iterative PCCC and SCCC decoders that utilize slightly modified SISO decoders.

4.2. The BCJR Algorithm and SISO Decoding

4.2.1. Probability Domain BCJR Algorithm for RSC Codes. Before we discuss the BCJR algorithm in the context of a turbo code, it is helpful to first consider the BCJR algorithm applied to a single rate $\frac{1}{2}$ RSC code on an AWGN channel. Thus, as indicated in Fig. 3, the transmitted codeword $\mathbf{c}$ will have the form $\mathbf{c} = [c_1, c_2, \ldots, c_K] = [u_1, p_1, u_2, p_2, \ldots, u_K, p_K]$ where $c_k \triangleq [u_k, p_k]$. The received word $\mathbf{y} = \mathbf{c} + \mathbf{n}$ will have the form $\mathbf{y} = [y_1, y_2, \ldots, y_K] = [y_1^u, y_1^p, y_2^u, y_2^p, \ldots, y_K^u, y_K^p]$, where $y_k \triangleq [y_k^u, y_k^p]$, and similarly for $\mathbf{n}$. As above, we assume our binary variables take values from the set $\{\pm 1\}$.

Our goal is the development of the BCJR algorithm [18] for computing the LLR

$$L(u_k) = \log\left(\frac{P(u_k = +1 \mid \mathbf{y})}{P(u_k = -1 \mid \mathbf{y})}\right)$$

given the received word $\mathbf{y}$. In order to incorporate the RSC code trellis into this computation, we rewrite $L(u_k)$ as

$$L(u_k) = \log \frac{\sum_{U+} p(s_{k-1} = s', s_k = s, \mathbf{y})}{\sum_{U-} p(s_{k-1} = s', s_k = s, \mathbf{y})} \tag{6}$$

where s_k is encoder state at time k, U^+ is set of pairs (s', s) for the state transitions $(s_{k-1} = s') \rightarrow (s_k = s)$, which correspond to the event $u_k = +1$, and U^- is similarly defined. To write (6) we used Bayes' rule, total probability, and then canceled $1/p(\mathbf{y})$ in the numerator and denominator. We see from (6) that we need only compute $p(s', s, \mathbf{y}) = p(s_{k-1} = s', s_k = s, \mathbf{y})$ for all state transitions and then sum over the appropriate transitions in the numerator and denominator. We now provide the crucial results that facilitate the computation of $p(s', s, \mathbf{y})$.

Result 1. The probability density function (pdf) $p(s', s, \mathbf{y})$ may be factored as

$$p(s', s, \mathbf{y}) = \alpha_{k-1}(s') \cdot \gamma_k(s', s) \cdot \beta_k(s) \tag{7}$$

where

$$\alpha_k(s) \triangleq p(s_k = s, \mathbf{y}_1^k)$$

$$\gamma_k(s', s) \triangleq p(s_k = s, y_k \mid s_{k-1} = s')$$

$$\beta_k(s) \triangleq p(\mathbf{y}_{k+1}^K \mid s_k = s)$$

and where $\mathbf{y}_a^b \triangleq [y_a, y_{a+1}, \cdots, y_b]$.

[1] It is well known that the MAP rule minimizes the probability of bit error. For comparison, the ML rule, which maximizes the likelihoods $P(\mathbf{y} \mid \mathbf{c})$ over the codewords $\mathbf{c}$, minimizes the probability of codeword error.

Proof: By several applications of Bayes' rule, we have

$$\begin{aligned} p(s', s, \mathbf{y}) &= p(s', s, \mathbf{y}_1^{k-1}, y_k, \mathbf{y}_{k+1}^K) \\ &= p(\mathbf{y}_{k+1}^K \mid s', s, \mathbf{y}_1^{k-1}, y_k) p(s', s, \mathbf{y}_1^{k-1}, y_k) \\ &= p(\mathbf{y}_{k+1}^K \mid s', s, \mathbf{y}_1^{k-1}, y_k) \\ &\quad \times p(s, y_k \mid s', \mathbf{y}_1^{k-1}) \cdot p(s', \mathbf{y}_1^{k-1}) \\ &= p(\mathbf{y}_{k+1}^K \mid s) \cdot p(s, y_k \mid s') \cdot p(s', \mathbf{y}_1^{k-1}) \\ &= \beta_k(s) \cdot \gamma_k(s', s) \cdot \alpha_{k-1}(s') \end{aligned}$$

where the fourth line follows from the third because the variables omitted on the fourth line are conditionally independent.

Result 2. The probability $\alpha_k(s)$ may be computed in a "forward recursion" via

$$\alpha_k(s) = \sum_{s'} \gamma_k(s', s) \alpha_{k-1}(s') \tag{8}$$

where the sum is over all possible encoder states.

Proof: By several applications of Bayes' rule and the theorem on total probability, we have

$$\begin{aligned} \alpha_k(s) &\triangleq p(s, \mathbf{y}_1^k) \\ &= \sum_{s'} p(s', s, \mathbf{y}_1^k) \\ &= \sum_{s'} p(s, y_k \mid s', \mathbf{y}_1^{k-1}) p(s', \mathbf{y}_1^{k-1}) \\ &= \sum_{s'} p(s, y_k \mid s') p(s', \mathbf{y}_1^{k-1}) \\ &= \sum_{s'} \gamma_k(s', s) \alpha_{k-1}(s') \end{aligned}$$

where the fourth line follows from the third due to conditional independence of $\mathbf{y}_1^{k-1}$.

Result 3. The probability $\beta_k(s)$ may be computed in a "backward recursion" via

$$\beta'_{k-1}(s') = \sum_{s} \beta_k(s) \gamma_k(s', s) \tag{9}$$

Proof: Applying Bayes' rule and the theorem on total probability, we have

$$\begin{aligned} \beta'_{k-1}(s') &\triangleq p(\mathbf{y}_k^K \mid s') \\ &= \sum_{s} p(\mathbf{y}_k^K, s \mid s') \\ &= \sum_{s} p(\mathbf{y}_{k+1}^K \mid s', s, y_k) p(s, y_k \mid s') \\ &= \sum_{s} p(\mathbf{y}_{k+1}^K \mid s) p(s, y_k \mid s') \\ &= \sum_{s} \beta_k(s) \gamma_k(s', s) \end{aligned}$$

where conditional independence led to the omission of variables on the fourth line.

The recursion for the $\{\alpha_k(s)\}$ is initialized according to

$$\alpha_0(s) = \begin{cases} 1, & s = 0 \\ 0, & s \neq 0 \end{cases}$$

which makes the reasonable assumption that the convolutional encoder is initialized to the zero state. The recursion for the $\{\beta_k(s)\}$ is initialized according to

$$\beta_K(s) = \begin{cases} 1, & s = 0 \\ 0, & s \neq 0 \end{cases}$$

which assumes that "termination bits" have been appended at the end of the data word so that the convolutional encoder is again in state zero at time K.

All that remains at this point is the computation of $\gamma_k(s', s) = p(s, y_k \mid s')$. Observe that $\gamma_k(s', s)$ may be written as

$$\begin{aligned} \gamma_k(s', s) &= \frac{P(s', s)}{P(s')} \cdot \frac{p(s', s, y_k)}{P(s', s)} \\ &= P(s \mid s') p(y_k \mid s', s) \\ &= P(u_k) p(y_k \mid u_k) \end{aligned} \tag{10}$$

where the event $'u_k$ corresponds to the event $s' \to s$. Note $P(s \mid s') = P(s' \to s) = 0$ if s is not a valid state from state s' and $P(s' \to s) = \frac{1}{2}$ otherwise (since we assume binary input encoders with equiprobable inputs). Hence, $\gamma_k(s', s) = 0$ if $s' \to s$ is not valid and, otherwise

$$\gamma_k(s', s) = \frac{P(u_k)}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\|y_k - c_k\|^2}{2\sigma^2} \right] \tag{11}$$

$$= \frac{1}{2\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_k^u - u_k)^2 + (y_k^p - p_k)^2}{2\sigma^2} \right] \tag{12}$$

where $\sigma^2 = N_0/2$.

In summary, we may compute $L(u_k)$ via (6) using (7), (8), (9), and (12). This "probability domain" version of the BCJR algorithm is numerically unstable for long and even moderate codeword lengths, and so we now present the stable "log domain" version of it.

4.2.2. Log-Domain BCJR Algorithm for RSC Codes. In the log-BCJR algorithm, $\alpha_k(s)$ is replaced by the *forward metric*

$$\begin{aligned} \tilde{\alpha}_k(s) &\triangleq \log(\alpha_k(s)) \\ &= \log\left(\sum_{s'} \alpha_{k-1}(s') \gamma_k(s', s) \right) \\ &= \log\left(\sum_{s'} \exp(\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s)) \right) \end{aligned} \tag{13}$$

where the *branch metric* $\tilde{\gamma}_k(s', s)$ is given by

$$\begin{aligned} \tilde{\gamma}_k(s', s) &= \log \gamma_k(s', s) \\ &= -\log(2\sqrt{2\pi}\sigma) - \frac{\|y_k - c_k\|^2}{2\sigma^2} \end{aligned} \tag{14}$$

We will see that the first term in (14) may be dropped. Note that (13) not only defines $\tilde{\alpha}_k(s)$ but also gives its recursion. These log-domain forward metrics are initialized as

$$\tilde{\alpha}_0(s) = \begin{cases} 0, & s = 0 \\ -\infty, & s \neq 0 \end{cases} \tag{15}$$

The probability $\beta_{k-1}(s')$ is replaced by the *backward metric*

$$\begin{aligned} \tilde{\beta}_{k-1}(s') &\triangleq \log(\beta_{k-1}(s')) \\ &= \log\left(\sum_s \exp \tilde{\beta}_k(s) + \tilde{\gamma}_k(s', s)\right) \end{aligned} \tag{16}$$

with initial conditions

$$\tilde{\beta}_K(s) = \begin{cases} 0, & s = 0 \\ -\infty, & s \neq 0 \end{cases} \tag{17}$$

under the assumption that the encoder has been terminated.

As before, $L(u_k)$ is computed as

$$\begin{aligned} L(u_k) &= \log \frac{\sum_{U^+} \alpha_{k-1}(s')\gamma_k(s', s)\beta_k(s)}{\sum_{U^-} \alpha_{k-1}(s')\gamma_k(s', s)\beta_k(s)} \\ &= \log\left[\sum_{U+} \exp(\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s))\right] \\ &\quad - \log\left[\sum_{U^-} \exp(\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s))\right] \end{aligned} \tag{18}$$

It is evident from (18) that the constant term in (14) may be ignored since it may be factored all the way out of both summations. At first glance, Eqs. (13)–(18) do not look any simpler than the probability domain algorithm, but we use the following results to obtain the simplification.

Result 4.

$$\max(x, y) = \log\left(\frac{e^x + e^y}{1 + e^{-|x-y|}}\right)$$

Proof: Without loss of generality, when $x > y$, the right-hand side equals x.

Now define

$$\max^*(x, y) \triangleq \log(e^x + e^y) \tag{19}$$

so that from Result 4, we obtain

$$\max^*(x, y) = \max(x, y) + \log(1 + e^{-|x-y|}) \tag{20}$$

This may be extended to more than two variables. For example

$$\max^*(x, y, z) \triangleq \log(e^x + e^y + e^z)$$

which may be computed pairwise according to the following result.

Result 5.

$$\max^*(x, y, z) = \max^*[\max^*(x, y), z]$$

Proof:

$$\begin{aligned} \text{RHS} &= \log[e^{\max^*(x,y)} + e^z] \\ &= \log[e^{\log(e^x + e^y)} + e^z] \\ &= \log[e^x + e^y + e^z] \\ &= \text{LHS} \end{aligned}$$

Given the function $\max^*(\cdot)$, we may now rewrite (13), (16), and (18) as

$$\tilde{\alpha}_k(s) = \max_{s'}{}^*[\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s)] \tag{21}$$

$$\tilde{\beta}_{k-1}(s') = \max_{s}{}^*[\tilde{\beta}_k(s) + \tilde{\gamma}_k(s', s)] \tag{22}$$

and

$$\begin{aligned} L(u_k) = &\max_{U^+}{}^*[\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s)] \\ &- \max_{U^-}{}^*[\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s)] \end{aligned} \tag{23}$$

Figure 6 illustrates pictorially the trellis-based computations that these last three equations represent.

We see from Eqs. (21)–(23) how the log-domain computation of $L(u_k)$ is vastly simplified relative to the probability-domain computation. From (20), implementation of the $\max^*(\cdot)$ function involves only a two-input $\max(\cdot)$ function plus a lookup table for the "correction term" $\log(1 + e^{-|x-y|})$. Robertson et al. [11] have shown that a table size of 8 is usually sufficient.

Note that the correction term is bounded as

$$0 < \log(1 + e^{-|x-y|}) \leq \log(2) \simeq 0.693$$

so that $\max^*(x, y) \simeq \max(x, y)$ when $|\max(x, y)| \geq 7$. When $\max^*(x, y)$ is replaced by $\max(\cdot)$ in Eqs. (21) and (22), these recursions become forward and reverse Viterbi algorithms, respectively. The performance loss associated with this approximation in turbo decoding depends on the specific turbo code, but a loss of about 0.5 dB is typical [11].

Finally, observe the sense in which the BCJR decoder is a SISO decoder: the decoder input is the "soft decision" (unquantized) word $\mathbf{y} \in \mathbb{R}^{2K}$ and its outputs are the soft outputs $L(u_k) \in \mathbb{R}$ on which final hard decisions may be made according to (4). Alternatively, in a concatenated code context, these soft outputs may be passed to a companion decoder.

4.2.3. Summary of the Log-Domain BCJR Algorithm. We assume as above a rate $\frac{1}{2}$ RSC encoder, a data block $\mathbf{u}$ of length K, and that encoder starts and terminates in the zero state (the last m bits of $\mathbf{u}$ are so selected, where m is the encoder memory size). In practice, the value $-\infty$ used in initialization is simply some large-magnitude negative number.

Initialize $\tilde{\alpha}_0(s)$ and $\tilde{\beta}_K(s)$ according to Eqs. (15) and (17).
for $k = 1{:}K$
— get $y_k = [y_k^u, y_k^p]$
— compute $\tilde{\gamma}_k(s', s) = -\|y_k - c_k\|^2/2\sigma^2$ for all allowable state transitions $s' \to s$ (note $c_k = c_k(s', s)$ here)[2]
— compute $\tilde{\alpha}_k(s)$ for all s using the recursion (21)
end

[2] We may alternatively use $\tilde{\gamma}_k(s', s) = u_k y_k^u/\sigma^2 + p_k y_k^p/\sigma^2$. (See next section.)

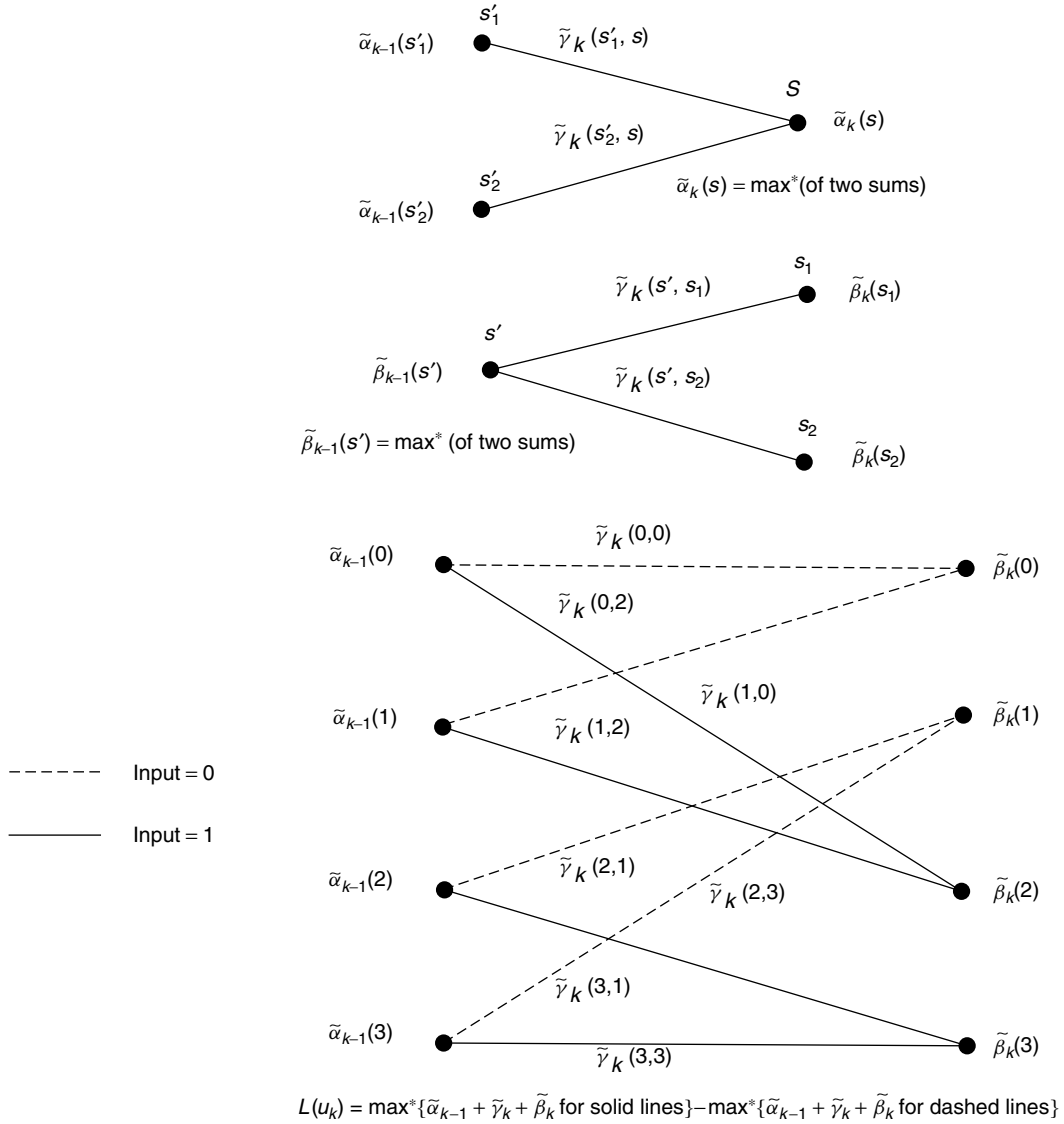


Figure 6. The top diagram depicts the forward recursion in (21), the middle diagram depicts the backward recursion in (22), and the bottom diagram depicts the computation of $L(u_k)$ via (23).

for $k = K: -1 : 2$
— compute $\tilde{\beta}_{k-1}(s')$ for all s' using (22)
end
for $k = 1{:}K$
— compute $L(u_k)$ using (23)
— compute hard decisions via $\hat{u}_k = \text{sign}\left[L(u_k)\right]$
end

4.3. The PCCC Iterative Decoder

We present in this section the iterative decoder for a PCCC consisting of two component rate $\frac{1}{2}$ RSC encoders concatenated in parallel. We assume no puncturing so that the overall code rate is $\frac{1}{3}$. Block diagrams of the PCCC encoder and its iterative decoder with component SISO decoders are presented in Fig. 7. As indicated in Fig. 7a, the transmitted codeword $\mathbf{c}$ will have the form $\mathbf{c} = [c_1, c_2, \ldots, c_K] = [u_1, p_1, q_1, \ldots, u_K, p_K, q_K]$ where $c_k \stackrel{\Delta}{=} [u_k, p_k, q_k]$. The received word $\mathbf{y} = \mathbf{c} + \mathbf{n}$ will have the form $\mathbf{y} = [y_1, y_2, \ldots, y_K] = [y_1^u, y_1^p, y_1^q, \ldots, y_K^u, y_K^p, y_K^q]$, where $y_k \stackrel{\Delta}{=} [y_k^u, y_k^p, y_k^q]$, and similarly for $\mathbf{n}$. We denote the codewords produced by E1 and E2 by, respectively, $\mathbf{c}_1 = [c_1^1, c_2^1, \ldots, c_K^1]$ where $c_k^1 \stackrel{\Delta}{=} [u_k, p_k]$ and $\mathbf{c}_2 = [c_1^2, c_2^2, \ldots, c_K^2]$ where $c_k^2 \stackrel{\Delta}{=} [u'_k, q_k]$. Note that $\{u'_k\}$ is a permuted version of $\{u_k\}$ and is not actually transmitted (see Fig. 7a). We define the noisy received versions of $\mathbf{c}_1$ and $\mathbf{c}_2$ to be $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively, having components $y_k^1 \stackrel{\Delta}{=} [y_k^u, y_k^p]$ and $y_k^2 \stackrel{\Delta}{=} [y_k^{u'}, y_k^q]$, respectively. Note that $\mathbf{y}_1$ and $\mathbf{y}_2$ can be assembled from $\mathbf{y}$ in an obvious fashion (using an interleaver to obtain $\{y_k^{u'}\}$ from $\{y_k^u\}$). By doing so, the component decoder inputs are the two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ as indicated in the Fig. 7b.

In contrast to the BCJR decoder of the previous sections whose input was $\mathbf{y} = \mathbf{c} + \mathbf{n}$ and whose output was $\{L(u_k)\}$ (or $\{\hat{u}_k\}$), the SISO decoders in Fig. 7b possess two inputs and two outputs. The SISO decoders are essentially the BCJR decoders discussed above, except that the SISO

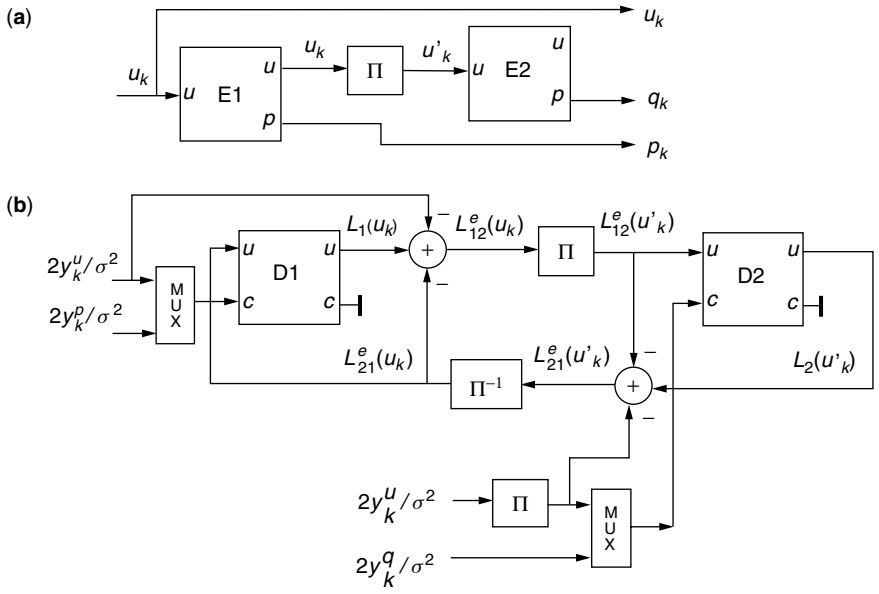


Figure 7. Block diagrams for the PCCC encoder (**a**) and iterative decoder (**b**).

decoders have the ability to accept from a companion decoder "extrinsic information" about its encoder's input (SISO input label '*u*') and/or about its encoder's output (SISO input label '*c*'). The SISO decoders also have the ability to produce likelihood information about its encoder's input (SISO output label '*u*') and/or about its encoder's output (SISO output label '*c*'). Note that the SISO decoder is to be interpreted as a decoding module not all of whose inputs or outputs need be used [7]. (Note that the RSC encoders in Fig. 7a have also been treated as modules.) As we will see, the SISO modules are connected in a slightly different fashion for the SCCC case.

Note from Fig. 7b that the extrinsic information to be passed from D1 to D2 about bit u_k, denoted $L_{12}^e(u_k)$, is equal to the LLR $L_1(u_k)$ produced by D1 minus the channel likelihood $2y_k^u/\sigma^2$ and the extrinsic information $L_{21}^e(u_k)$ that D1 had just received from D2. The idea is that $L_{12}^e(u_k)$ should indeed be extrinsic (and uncorrelated with) the probabilistic information already possessed by D2. As we will see, $L_{12}^e(u_k)$ is strictly a function of received E1 parity $\{y_k^p\}$ which is not directly sent to D2. Observe that $\{L_{12}^e(u_k)\}$ must be interleaved prior to being sent to D2 since E2 and D2 operate on the interleaved data bits u'_k. Symmetrical comments may be made about the extrinsic information to be passed from D2 to D1, $L_{21}^e(u_k)$ (e.g., it is a function of E2 parity and deinterleaving is necessary).

We already know from the previous section how the SISO decoders process the samples from the channel, $\mathbf{y}_i$ $(i = 1, 2)$, to obtain LLR's about the decoder inputs. We need now to discuss how the SISO decoders include the extrinsic information in their computations. As indicated earlier, the extrinsic information takes the role of *a priori* information in the iterative decoding algorithm [see (5) and surrounding discussion]:

$$L^e(u_k) \triangleq \log\left(\frac{P(u_k = +1)}{P(u_k = -1)}\right). \tag{24}$$

The *a priori* term $P(u_k)$ shows up in (11) in an expression for $\gamma_k(s', s)$. In the log domain, (11) becomes[3]

$$\tilde{\gamma}_k(s', s) = \log P(u_k) - \log(\sqrt{2\pi}\sigma) - \frac{\|y_k - c_k\|^2}{2\sigma^2} \tag{25}$$

Now observe that we may write

$$\begin{aligned} P(u_k) &= \left(\frac{\exp[-L^e(u_k)/2]}{1 + \exp\left[-L^e(u_k)\right]}\right) \cdot \exp\frac{u_k L^e(u_k)}{2} \\ &= A_k \exp\frac{u_k L^e(u_k)}{2} \end{aligned} \tag{26}$$

where the first equality follows since it equals

$$\left(\frac{\sqrt{P_-/P_+}}{1 + P_-/P_+}\right)\sqrt{P_+/P_-} = P_+ \text{ when } u_k = +1$$

$$\left(\frac{\sqrt{P_-/P_+}}{1 + P_-/P_+}\right)\sqrt{P_-/P_+} = P_- \text{ when } u_k = -1$$

where we have defined $P_+ \triangleq P(u_k = +1)$ and $P_- \triangleq P(u_k = -1)$ for convenience. Substitution of (26) into (25) yields

$$\tilde{\gamma}_k(s', s) = \log\left(\frac{A_k}{\sqrt{2\pi}\sigma}\right) + \frac{u_k L^e(u_k)}{2} - \frac{\|y_k - c_k\|^2}{2\sigma^2} \tag{27}$$

where we will see that the first term may be ignored.

Thus, the extrinsic information received from a companion decoder is included in the computation through the branch metric $\tilde{\gamma}_k(s', s)$. The rest of the BCJR/SISO algorithm proceeds as before using Eqs. (21)–(23).

[3] For the time being, we will discuss a generic SISO decoder so that we may avoid using cumbersome superscripts until it is necessary to do so.

Upon substitution of (27) into (23), we have

$$L(u_k) = L^e(u_k) + \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{u_k y_k^u}{\sigma^2} + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{u_k y_k^u}{\sigma^2} + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] \quad (28)$$

where we have applied the fact that

$$\|y_k - c_k\|^2 = (y_k^u - u_k)^2 + (y_k^p - p_k)^2 = (y_k^u)^2 - 2u_k y_k^u + u_k^2 + (y_k^p)^2 - 2p_k y_k^p + p_k^2$$

and only the terms dependent on U^+ or U^-, $u_k y_k^u/\sigma^2$ and $p_k y_k^p/\sigma^2$, survive after the subtraction. Now note that $u_k y_k^u/\sigma^2 = y_k^u/\sigma^2$ under the first $\max^*(\cdot)$ operation in (28) (U^+ is the set of state transitions for which $u_k = +1$) and $u_k y_k^u/\sigma^2 = -y_k^u/\sigma^2$ under the second $\max^*(\cdot)$ operation. Using the definition for $\max^*(\cdot)$, it is easy to see that these terms may be isolated out so that

$$L(u_k) = \frac{2y_k^u}{\sigma^2} + L^e(u_k) + \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] \quad (29)$$

The interpretation of this new expression for $L(u_k)$ is that the first term is likelihood information received directly from the channel, the second term is extrinsic likelihood information received from a companion decoder, and the third "term" ($\max_{U^+}{}^* - \max_{U^-}{}^*$) is extrinsic likelihood information to be passed to a companion decoder. Note that this third term is likelihood information gleaned from received parity not available to the companion decoder. Thus, specializing to decoder D1, for example, on any given iteration, D1 computes

$$L_1(u_k) = \frac{2y_k^u}{\sigma^2} + L_{21}^e(u_k) + L_{12}^e(u_k)$$

where $L_{21}^e(u_k)$ is extrinsic information received from D2, and $L_{12}^e(u_k)$ is the third term in (29) which is to be used as extrinsic information from D1 to D2.

4.3.1. Summary of the PCCC Iterative Decoder. The algorithm given below for the iterative decoding of a parallel turbo code follows directly from the development above. The constituent decoder order is D1, D2, D1, D2, and so on. Implicit is the fact that each decoder must have full knowledge of the trellis of the constituent encoders. For example, each decoder must have a table (array) containing the input bits and parity bits for all possible state transitions $s' \to s$. Also required are interleaver and de-interleaver functions (arrays) since D1 and D2 will be sharing reliability information about each u_k, but D2's information is permuted relative to D1. We denote these arrays by $P[\cdot]$ and $Pinv[\cdot]$, respectively. For example, the permuted word $\mathbf{u}'$ is obtained from the original word $\mathbf{u}$ via the pseudo-code statement: for $k = 1{:}K$, $u'_k = u_{P[k]}$, end.

We next point out that knowledge of the noise variance $\sigma^2 = N_0/2$ by each SISO decoder is necessary. Also, a simple way to obtain higher code rates via (simulated) puncturing is, in the computation of $\gamma_k(s', s)$, to set to zero the received parity samples, y_k^p or y_k^q, corresponding to the punctured parity bits, p_k or q_k. (This will set to zero the term in the branch metric corresponding to the punctured bit.) Thus, puncturing need not be performed at the encoder for computer simulations. We mention also that termination of encoder E2 to the zero state can be problematic due to the presence of the interleaver (for one solution, see Ref. 19). Fortunately, there is only a small performance loss when E2 is not terminated. In this case, $\beta_K(s)$ for D2 may be set to $\alpha_K(s)$ for all s, or it may be set to a nonzero constant (e.g., $1/S_2$, where S_2 is the number of E2 states).

Finally, we remark that some sort of iteration stopping criterion is necessary. The most straightforward criterion is to set a maximum number of iterations. However, this can be inefficient since the correct codeword is often found after only two or three iterations. Another straightforward technique is to utilize a carefully chosen outer error detection code. After each iteration, a parity check is made and the iterations stop whenever no error is detected. Other stopping criteria are presented in [9] and elsewhere in the literature.

4.3.1.1. Initialization

D1:

$\tilde{\alpha}_0^{(1)}(s) = 0$ for $s = 0$
$= -\infty$ for $\neq 0$

$\tilde{\beta}_K^{(1)}(s) = 0$ for $s = 0$
$= -\infty$ for $s \neq 0$

$L_{21}^e(u_k) = 0$ for $k = 1, 2, \ldots, K$

D2:

$\tilde{\alpha}_0^{(2)}(s) = 0$ for $s = 0$
$= -\infty$ for $s \neq 0$

$\tilde{\beta}_K^{(2)}(s) = \tilde{\alpha}_K^{(2)}(s)$ for all s (set once after computation of $\{\tilde{\alpha}_K^{(2)}(s)\}$ in the *first* iteration)

$L_{12}^e(u_k)$ is to be determined from D1 after the first half-iteration and so need not be initialized

4.3.1.2. The nth Iteration

D1:

for $k = 1{:}K$

— get $y_k^1 = [y_k^u, y_k^p]$

— compute $\tilde{\gamma}_k(s', s)$ for all allowable state transitions $s' \to s$ from (27) which simplifies to [see discussion following (27)]

$$\tilde{\gamma}_k(s', s) = \frac{u_k L_{21}^e(u_{Pinv[k]})}{2} + \frac{u_k y_k^u}{\sigma^2} + \frac{p_k y_k^p}{\sigma^2}$$

[u_k (p_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$]

— compute $\tilde{\alpha}_k^{(1)}(s)$ for all s using (21)

end

for $k = K{:}-1:2$

— compute $\tilde{\beta}_{k-1}^{(1)}(s)$ for all s using (22)

end

for $k = 1{:}K$

— compute $L_{12}^e(u_k)$ using[4]

$$L_{12}^e(u_k) = \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k^{(1)}(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k^{(1)}(s) \right]$$

end

D2:

for $k = 1{:}K$

— get $y_k^2 = [y_{P[k]}^u, y_k^q]$

— compute $\tilde{\gamma}_k(s', s)$ for all allowable state transitions $s' \to s$ from

$$\tilde{\gamma}_k(s', s) = \frac{u_k L_{12}^e(u_{P[k]})}{2} + \frac{u_k y_{P[k]}^u}{\sigma^2} + \frac{q_k y_k^q}{\sigma^2}$$

[u_k (q_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$]

— compute $\tilde{\alpha}_k^{(2)}(s)$ for all s using (21)

end

for $k = K{:}-1:2$

— compute $\tilde{\beta}_{k-1}^{(2)}(s)$ for all s using (22)

end

for $k = 1{:}K$

— compute $L_{21}^e(u_k)$ using

$$L_{21}^e(u_k) = \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}^{(2)}(s') + \frac{q_k y_k^q}{\sigma^2} + \tilde{\beta}_k^{(2)}(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}^{(2)}(s') + \frac{q_k y_k^q}{\sigma^2} + \tilde{\beta}_k^{(2)}(s) \right]$$

end

[4] Note here we are computing $L_{12}^e(u_k)$ directly rather than computing $L_1(u_k)$ and then subtracting $2y_k^u/\sigma^2 + L_{21}^e(u_k)$ from it to obtain $L_{12}^e(u_k)$ as in Fig. 5(b). We will do likewise in the analogous step for D2.

4.3.1.3. After the Last Iteration

for $k = 1{:}K$

— compute

$$L_1(u_k) = \frac{2y_k^u}{\sigma^2} + L_{21}^e(u_{p_{inv}[k]}) + L_{12}^e(u_k)$$

— $\hat{u}_k = \text{sign}\,[L(u_k)]$

end

4.4. The SCCC Iterative Decoder

We present in this section the iterative decoder for an SCCC consisting of two component rate $\frac{1}{2}$ RSC encoders concatenated in series. We assume no puncturing so that the overall code rate is $\frac{1}{4}$. Higher code rates are achievable via puncturing and/or by replacing the inner encoder with a rate 1 differential encoder with transfer function $\frac{1}{1 \oplus D}$. It is straightforward to derive the iterative decoding algorithm for other SCCC codes from the special case that we consider here.

Block diagrams of the SCCC encoder and its iterative decoder with component SISO decoders are presented in Fig. 8. We denote by $\mathbf{c}_1 = [c_1^1, c_2^1, \ldots, c_{2K}^1] = [u_1, p_1, u_2, p_2, \ldots, u_K, p_K]$ the codeword produced by E1 whose input is $\mathbf{u} = [u_1, u_2, \ldots, u_K]$. We denote by $\mathbf{c}_2 = [c_1^2, c_2^2, \ldots, c_{2K}^2] = [v_1, q_1, v_2, q_2, \ldots, v_{2K}, q_{2K}]$ (with $c_k^2 \stackrel{\Delta}{=} [v_k, q_k]$) the codeword produced by E2 whose input $\mathbf{v} = [v_1, v_2, \ldots, v_{2K}]$ is the interleaved version of $\mathbf{c}_1$, that is, $\mathbf{v} = \mathbf{c}_1'$. As indicated in Fig. 8a, the transmitted codeword $\mathbf{c}$ is the codeword $\mathbf{c}_2$. The received word $\mathbf{y} = \mathbf{c} + \mathbf{n}$ will have the form $\mathbf{y} = [y_1, y_2, \ldots, y_{2K}] = [y_1^v, y_1^q, \ldots, y_{2K}^v, y_{2K}^q]$ where $y_k \stackrel{\Delta}{=} [y_k^v, y_k^q]$, and similarly for $\mathbf{n}$.

The iterative SCCC decoder in Fig. 8b employs two SISO decoding modules (described in the previous section). Note that unlike the PCCC case, these SISO decoders share extrinsic information on the code bits $\{c_k^1\}$ (equivalently, on the input bits $\{v_k\}$) in accordance with the fact that these are the bits known to both encoders. A consequence of this is that D1 must provide likelihood information on E1 *output* bits whereas D2 produces likelihood information on E2 *input* bits as indicated in Fig. 8b. However, because LLRs must be obtained on the original data bits u_k so that final decisions may be made, D1 must also compute likelihood information on E1 input bits. Note also that, because E1 feeds no bits directly to the

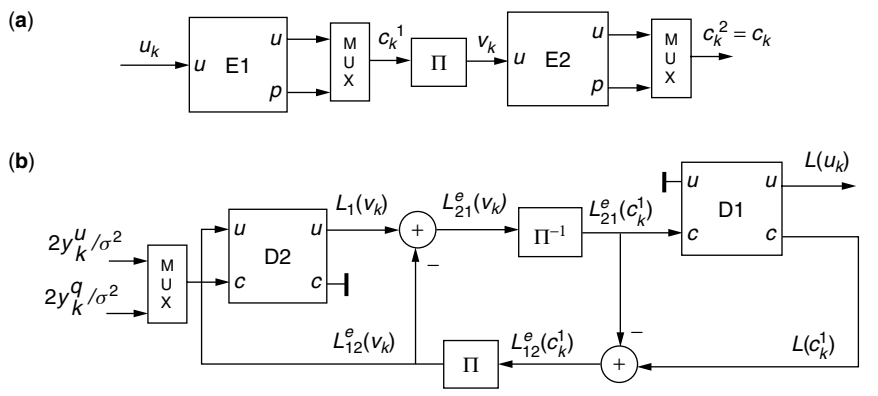


Figure 8. Block diagrams for the SCCC encoder (**a**) and iterative decoder (**b**).

channel, D1 receives no samples directly from the channel. Instead, the only input to D1 is the extrinsic information it receives from D2.

Thus, the SISO module D1 requires two features that we have not discussed in any detail to this point. The first is providing likelihood information on the encoder's input *and* output. However, since we assume the component codes are systematic, we need only compute LLRs on the encoder's output bits $[u_1, p_1, u_2, p_2, \ldots, u_K, p_K]$. Doing this is a simple matter of modifying the summation indices in (6) to those relevant to the output bit of interest. For example, the LLR $L(p_k)$ for the E1 parity bit p_k is obtained via

$$L(p_k) = \log \frac{\sum_{P^+} p(s_{k-1} = s', s_k = s, \mathbf{y})}{\sum_{P^-} p(s_{k-1} = s', s_k = s, \mathbf{y})} \tag{30}$$

where P^+ is set of state transition pairs (s', s) corresponding to the event $p_k = +1$, and P^- is similarly defined. (A trellis-based BCJR/SISO decoder is generally capable of decoding either the encoder's input or its output, whether or not the code is systematic. This is evident since the trellis branches are labeled by both inputs and outputs.)

The second feature is required by D1 is decoding with only extrinsic information as input. In this case the branch metric is simply modified as [cf. (27)]

$$\tilde{\gamma}_k(s', s) = \frac{u_k L^e_{21}(u_k)}{2} + \frac{p_k L^e_{21}(p_k)}{2} \tag{31}$$

Other than these modifications, the iterative SCCC decoder proceeds much like the PCCC iterative decoder and as indicated in Fig. 8b.

4.4.1. Summary of the SCCC Iterative Decoder. Essentially all of the comments mentioned for the PCCC decoder hold here as well and so we do not repeat them. The only difference is that the decoding order is D2, D1, D2, D1, and so on.

4.4.1.1. Initialization

D1:

$$\tilde{\alpha}^{(1)}_0(s) = 0 \text{ for } s = 0$$
$$= -\infty \text{ for } s \neq 0$$

$$\tilde{\beta}^{(1)}_K(s) = 0 \text{ for } s = 0$$
$$= -\infty \text{ for } s \neq 0$$

$L^e_{21}(c^1_k)$ is to be determined from D2 after the first half-iteration and so need not be initialized

D2:

$$\tilde{\alpha}^{(2)}_0(s) = 0 \text{ for } s = 0$$
$$= -\infty \text{ for } s \neq 0$$

$\tilde{\beta}^{(2)}_{2K}(s) = \tilde{\alpha}^{(2)}_{2K}(s)$ for all s (set after computation of $\{\tilde{\alpha}^{(2)}_{2K}(s)\}$ in the *first* iteration)

$L^e_{12}(v_k) = 0$ for $k = 1, 2, \ldots, 2K$

4.4.1.2. The nth Iteration

D2:

for $k = 1{:}\,2K$

— get $y_k = [y^v_k, y^q_k]$

— compute $\tilde{\gamma}_k(s', s)$ for all allowable state transitions $s' \to s$ from

$$\tilde{\gamma}_k(s', s) = \frac{v_k L^e_{12}(v_k)}{2} + \frac{v_k y^v_k}{\sigma^2} + \frac{q_k y^q_k}{\sigma^2}$$

[v_k (q_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$; $L^e_{12}(v_k)$ is $L^e_{12}(c^1_{P[k]})$, the interleaved extrinsic information from the previous D1 iteration.]

— compute $\tilde{\alpha}^{(2)}_k(s)$ for all s using (21)

end

for $k = 2K{:}\,-1 : 2$

— compute $\tilde{\beta}^{(2)}_{k-1}(s)$ for all s using (22)

end

for $k = 1{:}\,2K$

— compute $L^e_{21}(v_k)$ using

$$L^e_{21}(v_k) = \max^*_{V^+}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}^{(2)}_k(s)\right]$$
$$-\max^*_{V^-}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}^{(2)}_k(s)\right] - L^e_{12}(v_k)$$
$$= \max^*_{V^+}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + v_k y^v_k/\sigma^2 + q_k y^q_k/\sigma^2 + \tilde{\beta}^{(2)}_k(s)\right]$$
$$-\max^*_{V^-}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + v_k y^v_k/\sigma^2 + q_k y^q_k/\sigma^2 + \tilde{\beta}^{(2)}_k(s)\right]$$

where V^+ is set of state transition pairs (s', s) corresponding to the event $v_k = +1$, and V^- is similarly defined.

end

D1:

for $k = 1{:}\,K$

— for all allowable state transitions $s' \to s$ set $\tilde{\gamma}_k(s', s)$ via

$$\tilde{\gamma}_k(s', s) = \frac{u_k L^e_{21}(u_k)}{2} + \frac{p_k L^e_{21}(p_k)}{2}$$
$$= \frac{u_k L^e_{21}(c^1_{2k-1})}{2} + \frac{p_k L^e_{21}(c^1_{2k})}{2}$$

[u_k (p_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$; $L^e_{21}(c^1_{2k-1})$ is $L^e_{21}(v_{Pinv[2k-1]})$, the de-interleaved extrinsic information from the previous D2 iteration, and similarly for $L^e_{21}(c^1_{2k})$].

— compute $\tilde{\alpha}^{(1)}_k(s)$ for all s using (21)

end

for $k = K: -1 : 2$

— compute $\tilde{\beta}_{k-1}^{(1)}(s)$ for all s using (22)

end

for $k = 1:K$

— compute $L_{12}^e(u_k) = L_{12}^e(c_{2k-1}^1)$ using

$$L_{12}^e(u_k) = \max_{U^+}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - \max_{U^-}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - L_{21}^e(c_{2k-1}^1)$$

$$= \max_{U^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k L_{21}^e(p_k)}{2} + \tilde{\beta}_k^{(1)}(s)\right] - \max_{U^-}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k L_{21}^e(p_k)}{2} + \tilde{\beta}_k^{(1)}(s)\right]$$

— compute $L_{12}^e(p_k) = L_{12}^e(c_{2k}^1)$ using

$$L_{12}^e(p_k) = \max_{P^+}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - \max_{P^-}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - L_{21}^e(c_{2k}^1)$$

$$= \max_{P^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{u_k L_{21}^e(u_k)}{2} + \tilde{\beta}_k^{(1)}(s)\right] - \max_{P^-}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{u_k L_{21}^e(u_k)}{2} + \tilde{\beta}_k^{(-1)}(s)\right]$$

end

4.4.1.3. After the Last Iteration

for $k = 1:K$

— for all allowable state transitions $s' \to s$ set $\tilde{\gamma}_k(s', s)$ via

$$\tilde{\gamma}_k(s', s) = \frac{u_k L_{21}^e(c_{2k-1}^1)}{2} + \frac{p_k L_{21}^e(c_{2k}^1)}{2}$$

— compute $L(u_k)$ using

$$L(u_k) = \max_{U^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)\right] - \max_{U^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)\right]$$

— $\hat{u}_k = \text{sign}\left[L(u_k)\right]$

end

5. CONCLUSION

We have seen in this article the how and why of both parallel and serial turbo codes. That is, we have seen how to decode these codes using an iterative decoder, and why they should be expected to perform so well. The decoding algorithm summaries should be sufficient to decode any binary parallel and serial turbo codes, and can in fact be easily extended to the iterative decoding of any binary hybrid schemes. It is not much more work to figure out how to decode any of the turbo trellis-coded modulation (turbo TCM) schemes that appear in the literature. In any case, this article should serve well as a starting point for the study of concatenated codes (and perhaps graph-based codes) and their iterative decoders.

Acknowledgments

The author would like to thank Rajeev Ramamurthy and Bo Xia for producing Figs. 4 and 5, and Steve Wilson, Masoud Salehi, and John Proakis for helpful comments. He would also like to thank Cheryl Drier for typing the first draft.

BIOGRAPHY

William E. Ryan received his B.S. in electrical engineering degree from Case Western Reserve University, Cleveland, Ohio, in 1981, and his M.S. and Ph.D. degrees in electrical engineering from the University of Virginia, Charlottesville, in 1984 and 1988, respectively. He has been with The Analytic Sciences Corporation, Ampex Corporation, and Applied Signal Technology prior to his positions in academia. From 1993 to 1998 he was with the Electrical and Computer Engineering Department faculty at New Mexico State University, Las Cruces. Since August 1998, he has been with the Electrical and Computer Engineering Department at the University of Arizona, Tucson, where he is an associate professor. He is an associate editor for the *IEEE Transactions on Communications for Coding, Modulation, and Equalization*. His research interests are in coding and signal processing for data transmission and storage.

BIBLIOGRAPHY

1. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: turbo codes, *Proc. 1993 Int. Conf. Communications*, pp. 1064–1070.
2. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: turbo-codes, *IEEE Trans. Commun.* **44**: 1261–1271 (Oct. 1996).
3. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**: 55–67 (Jan. 1982).
4. S. Benedetto and G. Montorsi, Unveiling turbo codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* **42**: 409–428 (March 1996).
5. S. Benedetto and G. Montorsi, Design of parallel concatenated codes, *IEEE Trans. Commun.* **44**: 591–600 (May 1996).
6. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* **44**: 909–926 (May 1998).
7. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, *A Soft-Input Soft-Output Maximum a Posteriori (MAP) Module to Decode Parallel and Serial Concatenated Codes*, TDA Progress Report 42-127, Nov. 15, 1996.
8. D. Divsalar and F. Pollara, *Multiple turbo codes for Deep-Space Communications*, JPL TDA Progress Report, 42-121, May 15, 1995.
9. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**: 429–445 (March 1996).

10. L. Perez, J. Seghers, and D. Costello, A distance spectrum interpretation of turbo codes, *IEEE Trans. Inform. Theory* **42**: 1698–1709 (Nov. 1996).

11. P. Robertson, E. Villebrun, and P. Hoeher, A comparison of optimal and suboptimal MAP decoding algorithms operating in the log domain, *Proc. 1995 Int. Conf. Communications*, pp. 1009–1013.

12. A. Viterbi, An intuitive justification and a simplified implementation of the MAP decoder for convolutional codes, *IEEE JSAC* **16**: 260–264 (Feb. 1998).

13. O. Acikel and W. Ryan, Punctured turbo codes for BPSK /QPSK channels, *IEEE Trans. Commun.* **47**: 1315–1323 (Sept. 1999).

14. O. Acikel and W. Ryan, Punctured high rate SCCCs for BPSK/QPSK channels, *Proc. 2000 IEEE Int. Conf. Communications*, Vol. 1, pp. 434–439.

15. J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

16. B. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.

17. N. Wiberg, *Codes and Decoding on General Graphs*, Ph.D. dissertation, Univ. Linkoping, Sweden, 1996.

18. L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).

19. D. Divsalar and F. Pollara, turbo codes for PCS applications, *Proc. 1995 Int. Conf. Communications*, pp. 54–59.

CONSTRAINED CODING TECHNIQUES FOR DATA STORAGE

WIM M. J. COENE
HENK D. L. HOLLMANN
Philips Research Laboratories
Eindhoven, The Netherlands

1. INTRODUCTION

Modulation codes are one of the key elements in digital communication or storage systems such as a CD or DVD recorder, a hard-disk drive (HDD) in a computer, a modem, or a fax. A schematical form of a storage system is shown in Fig. 1. Here, two parts can be distinguished: the transmitting part, including the write channel in which one user stores data on the recording medium; and the receiving part of the system, including the read-channel in which the same or another user tries to restore the original information by reading out the data written on the medium.

In order to realize a sufficiently high level of reliability, the data are first encoded before being stored. This *channel encoding* typically comprises an error-correcting code (ECC) and a modulation code (MC). The channel encoder at the transmitting end consists of the error correction encoder and the modulation encoder, usually cascaded one after the other in that order.

Located at the receiving end of the channel are the physical signal detection with the read head scanning the information on the medium, followed by the bit detection module, which attempts to derive the written bits (also called *channel* bits) from the measured signals as reliably as possible. These blocks precede the channel decoding, which consists of the respective counterparts of the modules at the transmitting end, with first the MC decoder, followed by the ECC decoder.

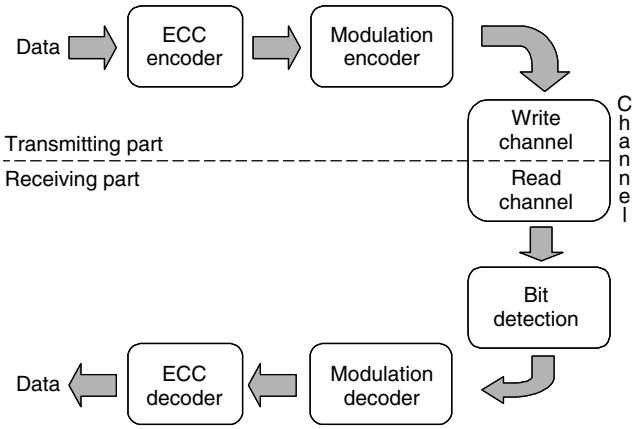


Figure 1. Schematic form of a digital storage system.

The ECC adds redundancy in the form of parity symbols, which makes it possible to restore the correct information in the presence of channel imperfections such as random errors and/or burst errors that may occur during readout from the medium. The modulation code serves to transform arbitrary (binary) sequences into sequences that possess certain "desirable" properties. Note the difference from error-correcting codes, which are used to ensure that *pairs* of encoded sequences have certain properties (i.e., being "very different"), while a modulation code serves to ensure certain properties of *each individual* encoded sequence. Which properties are desirable strongly depends on the particular storage or communication system for which the code is designed. For example, in most digital magnetic or optical recording systems, the stored sequences preferably contain neither very short nor very long runs of successive zeros or ones. The reason for this originates in how a stored sequence is read from the storage medium. The explanation is as follows.

In a storage system, the modulation of the physical signals is concerned with two physical conditions or states of the medium: (1) for magnetic recording, it is the magnetisation of magnetic domains in one of two opposite directions; (2) for optical recording, as shown in Fig. 2, it is the level of reflectivity (high and low) of the marks (or *pits*) and spaces (or *lands*) on the medium. One physical state can be associated with a channel bit (binary) 1, the other state with a channel bit (binary) 0. This representation, where the value of the channel bit represents the physical state (mark or space), is commonly known as *non-return-to-zero inverse* (NRZI), an old term originating from telegraphy. An equivalent representation of a channel bit stream is the *non-return-to-zero* (NRZ) notation, where a 1 bit indicates the start of a new mark or space on the medium, that is, a change of the physical state, and a 0 bit indicates the continuation of a mark or space.

A channel bit stream in NRZI notation can be partitioned into a sequence of phrases or *runs*, where each run consists of a number of consecutive channel bits of the

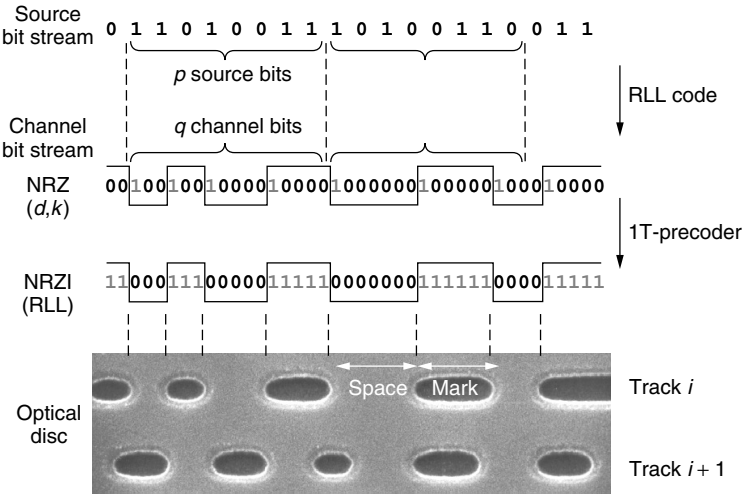


Figure 2. RLL coding in optical recording.

same type. So each run is associated with the physical area of a mark or space on the medium. The number of bits in a run is called the *run length*. As an example of a storage medium, we shall consider an optical disc in a little more detail. On the disk, data are organized in a single spiral; a small part of two successive revolutions of the spiral can be seen in Fig. 2. Along the track, physical marks and the spaces between them alternate. The marks and spaces have lengths that are multiples of the channel bit length T, and a mark or space of length nT represents a run with a run length of n bits.

Very short runs lead to small-signal amplitudes in the readout by the physical detection, and are therefore more prone to errors in the bit detection module (which is positioned directly after the read channel in Fig. 1). This is explained in more detail in Section 2. Moreover, very long runs lead to inaccuracies in the *timing recovery*, which is dealt with by a device called a *phase-locked loop* (PLL). The PLL regenerates the internal "clock" that is matched to the length of the bits on the medium; the bit clock is adjusted at each occurrence of a transition. Areas on the medium with too few transitions may cause "clock drift."

Avoiding very short and/or very long runs is achieved by using a run-length-limited (RLL) code, which typically constrains the allowable minimum and maximum run lengths that occur in the channel bit stream. The RLL constraints are described in terms of two parameters, d and k, and stipulate that the minimum and maximum run lengths are equal to $d+1$ and $k+1$, respectively. Note that the uncoded case corresponds to $d=0$ and $k=\infty$. In NRZ notation, a run of length $m+1$ is represented by a 1 bit followed by m 0 bits, that is, by 10^m in shorthand notation. Hence the (d,k) constraint in NRZ notation requires that the number of 0 bits between two successive 1 bits be at least d and at most k. Sequences that obey this constraint are called (d,k) *sequences*; a code for this constraint is referred to as a (d,k) *code*. Most RLL codes are constructed for a bit stream in NRZ notation. Subsequent transformation from NRZ to NRZI is required to obtain the channel bits that are actually written on the medium; such a transformation is formally carried out by a so-called 1T precoder, which is essentially an integrator modulo 2 (see Fig. 2).

A run-length constraint forms an example of a constraint that is specified by the absence of a finite number of "forbidden patterns." For example, a (d,k) constraint with $d>0$ can be specified in terms of the forbidden patterns $11, 101, \ldots, 10^{d-1}1$, and 0^{k+1}. Such a constraint is commonly referred to as a *constraint of finite type*. Many constraints that occur in practice are of this kind. Forbidding certain specific patterns implies that a sequence of source bits must be translated into a longer sequence of channel bits; the ratio of the length of the original sequence and the length of the encoded sequence is called the *rate* of the code.

Run-length-limited codes originated already in the 1960s through the work of Franaszek [1], Tang and Bahl [2], and others. Since then, various mechanisms for the construction of RLL and other modulation codes have been devised. A very elegant method is the ACH state-splitting algorithm [3], which will be discussed in Section 5. A detailed overview of RLL codes and their construction is given in Immink's book [4]; for other review articles, see, for example, Refs. 5 and 6.

The remainder of this survey is organised as follows. In Section 2, some practical aspects of the use of an RLL code are considered; Section 3 discusses the maximal coding rate (the capacity) of a modulation code; encoder and decoder structures are considered in Section 4; various code construction methods are described in Section 5; and finally Section 6 presents a collection of research topics and more recent trends in modulation coding.

2. PRACTICAL CONSIDERATIONS FOR THE USE OF AN RLL CODE

High-capacity storage applications employ such small bit sizes that the readout signal waveform generated by the physical detection for a given bit location does depend not only on that single bit but also on a limited number of neighboring bits. This bit-smearing effect is better known as *intersymbol interference* (ISI). For a simple read channel with linear readout characteristics, the ISI is

characterized by the impulse response of the channel, or, equivalently, by its Fourier transform, which yields the modulation transfer function (MTF) of the channel [7, Chap. 3.2]. The MTF indicates the (amplitude) response of the channel for each frequency in the system.

For storage systems, the MTF typically has a lowpass character: for instance, in optical recording, the MTF has an almost linear rolloff up to the cutoff frequency of the channel (see Fig. 3). Therefore, short run lengths in the channel bit stream, which lead to high-frequent signals, suffer most from ISI and are thus more prone to errors during read-out. One of the purposes of runlength-limited coding is to impose constraints that do not allow these high-frequency bit sequences; by doing so, the spectral content of the RLL-coded sequences is shaped to have a more lowpass character.

To illustrate this principle, we discuss the effect of employing three different d constraints, for $d = 0$ (uncoded), $d = 1$, and $d = 2$, while maintaining the same density of source bits on the storage medium in all three cases for fair comparison. So let T denote the common physical size of a source bit. Using a d-constrained code at a rate R_d, the physical channel bit size T_d will necessarily satisfy $T_d = R_d T$. Figure 4 shows the respective channel bit lengths and the highest frequency in the system (which correspond to an alternation of runs of minimum run length). Here, we of course have $R_0 = 1$ in the uncoded case. Furthermore, we assume that practical codes are

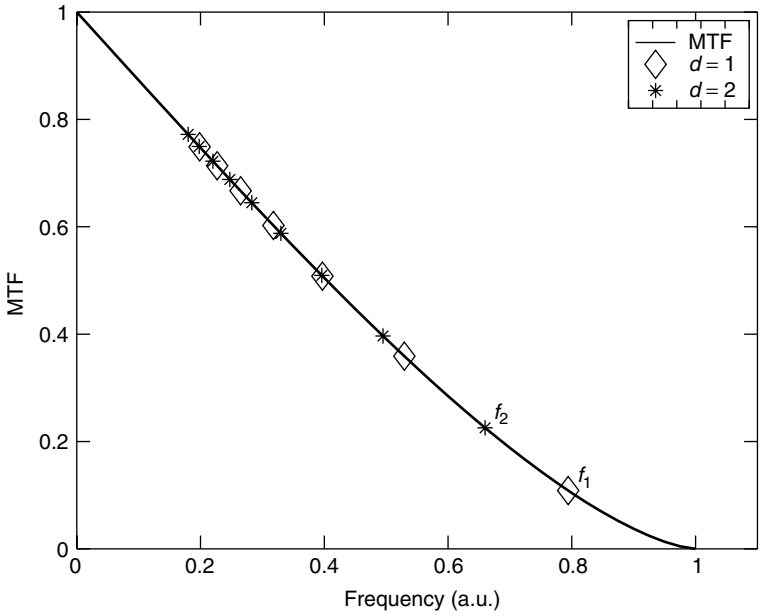


Figure 3. MTF for the optical recording channel as a function of frequency (in arbitrary units) with the frequencies of the pure tones $\ldots |nT_d \mid nT_d| \ldots$ superimposed.

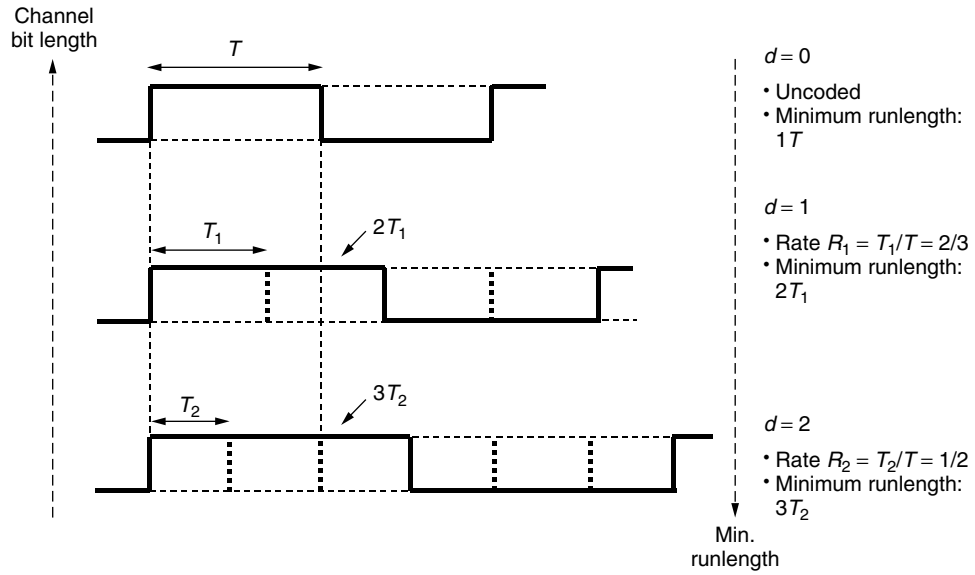


Figure 4. Channel bit length and minimum runlength for different d constraints at the same recording capacity in a storage channel.

used that have rates $R_1 = \frac{2}{3}$ and $R_2 = \frac{1}{2}$. These are reasonably close to the maximal achievable code rates of 0.6942 and 0.5515, respectively (see Section 3). For these rates, the minimum run length for $d = 1$ amounts to $2T_1 = 4/3T$, which is larger than the minimum run length T for $d = 0$; also, the minimum run length for $d = 2$ amounts to $3T_2 = 3/2T$, which is larger than the minimum run length for $d = 1$. Consequently, the highest frequencies f_d in the system for $d = 0, d = 1$, and $d = 2$ are

$$f_0 = \frac{1}{2T} > f_1 = \frac{1}{4R_1T} = \frac{3}{8T} > f_2 = \frac{1}{6R_2T} = \frac{1}{3T}$$

This relation reveals the increasing lowpass character of the code for increasing d constraint, which is the major attractiveness of RLL coding. This can also be observed from Fig. 3, where we have drawn the MTF (as a function of frequency) for the optical recording channel, with the frequencies of the pure tones $\ldots |nT_d \mid nT_d \mid nT_d| \ldots$ for $n = d + 1, d + 2, \ldots$ superimposed.

However, note that the channel bit length (or *timing window*, also known as *detection window*) decreases for increasing d constraint, which leads to a greater sensitivity with respect to *jitter* or *mark-edge noise* in the system. This counteracting effect favours the use of a *lower* d constraint.

The practical choice for a certain d constraint is a compromise between all pros and cons, and depends on many aspects, such as the actual bit detector used in the receiver (e.g., threshold detection, or some form of partial-response maximum-likelihood bit detection [7, Chap. 6]), the characteristics of the write channel, and the characteristics of the various noise sources in the system such as media noise and electronic noise.

3. CODING RATE AND CAPACITY

In the foregoing we have discussed various technical reasons for putting constraints on the sequences that we want to store or to send over a channel. Sequences that satisfy the constraints at hand are called *constrained sequences* or *codewords*, and the collection of all these sequences will be called the *constrained system*. A *modulation code* for a given constrained system consists of an *encoder* to translate arbitrary sequences into constrained sequences, and a *decoder* to recover the original sequence from the encoded sequence. The bits making up the original sequence and the encoded sequence are usually referred to as *source bits* and *channel bits*, respectively.

To achieve encoding, there is a price to be paid. Indeed, since there are more arbitrary sequences of a given length than there are constrained sequences of the same length, the encoding process will necessarily lead to an *increase* in the number of bits in the channel bit stream. This increase is measured by a number called the *rate* of the code. If, on the average, p source bits are translated into q channel bits, then the rate R of the code is $R = p/q$.

All other things being equal, we would, of course, like our code to have the highest possible rate. However, for all practical constraints there is a natural barrier for the code rate, called the *capacity* of the constraint, beyond which no encoding is possible. This discovery by Shannon [8] has been of great theoretical and practical importance. Indeed, once we know the capacity C of a given constrained system, then, on the one hand, we know that the best encoding rate that could possibly be achieved is bounded from above by C; on the other hand, once we have actually constructed a code with a rate R, the number R/C, called the *efficiency* of the code, serves as a benchmark for our engineering achievement.

In what follows, we shall sketch a derivation of Shannon's results. Consider a constrained system $\mathcal{L}$, and let N_n denote the number of constrained sequences of length n. Suppose that we can encode at a rate R. By definition, this means that, for large n, approximately Rn source bits are translated into n channel bits. There are 2^{Rn} distinct source sequences of length Rn, all of which need to be translated into distinct constrained sequences of length n, of which there are only N_n. Therefore, we necessarily have that $2^{Rn} \leq N_n$, or, equivalently, that $R \leq n^{-1} \log N_n$. (Here and in the sequel, all logarithms will be to the base 2.)

We now follow Shannon and *define* the capacity $C(\mathcal{L})$ of the constrained system $\mathcal{L}$ as

$$C(\mathcal{L}) = \lim_{n\to\infty} \frac{\log N_n}{n} \tag{1}$$

provided this limit exists.[1]

Intuitively, the capacity represents the *average amount of information* (measured in bits) *per bit of the constrained sequence*. The reasoning described above shows that encoding at a rate R is possible only if $R \leq C(\mathcal{L})$ and, moreover, suggests the possibility of encoding at rates arbitrarily close to $C(\mathcal{L})$, for example, by using long codewords glued together by a few suitably chosen merging bits.

It turns out that the kind of systems encountered in practice can typically be described in terms of a finite labeled directed *graph*. Here, a graph $G = (V, A)$ consists of a collection V of vertices or *states*, and a collection A of labeled arcs or *transitions* between states. Some authors refer to a "labeled directed graph" as defined above as a *finite-state transition diagram* (FSTD). The constrained system $\mathcal{L}(G)$ *presented* by G consists of all sequences that can be obtained by reading off the labels along a (directed) *path* in the graph, or, as we shall see, the sequences that are *generated* by paths in G. We shall refer to the graph G as a *presentation* of the system $\mathcal{L}(G)$.

Constrained systems that can be presented by some finite labeled directed graph as explained above are called *sofic systems*. Sofic systems are of great theoretical and practical importance. It turns out that about every constrained system encountered in practical applications is in fact a sofic system. (This is less remarkable than it

[1] For *subword-closed* systems, the fact that this limit exists immediately follows from Fekete's lemma [9, p. 233; 10]; namely, if the non-negative numbers a_n are such that $a_{m+n} \leq a_m + a_n$ for all $n, m \geq 0$, then $a^* = \lim_{n\to\infty} a_n/n$ exists and $a^* \leq a_n/n$ for all n. Indeed, if N_n is the number of sequences of length n contained in a subword-closed system, then obviously $N_{m+n} \leq N_m N_n$; hence an application of this lemma with $a_n = \log N_n$ shows that subword-closed systems have a well-defined capacity.

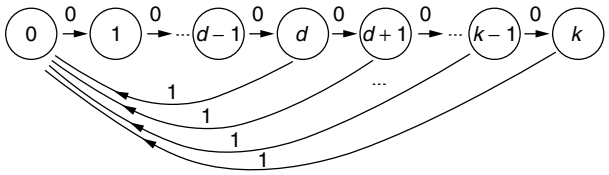


Figure 5. Presentation of a (d, k)-constrained system.

seems once we realize that about any digital device that we build is actually a *finite-state device*, whose possible output sequences necessarily constitute some sofic system.)

For example, a (d, k)-constrained system can be presented by the graph in Fig. 5. Here, the state label indicates the number of terminal zeros of a sequence generated by a path ending in that state. In the case where $k = \infty$, the constraint can be presented by a similar $(d + 1)$-state graph, where state d now indicates *at least* d terminal zeros.

We say that a presentation G is *irreducible* if there is a path in G from any state to any other state; G is said to be *primitive* if moreover all these paths may be chosen to have a fixed length h for some positive integer h.

It can be shown that each presentation can be broken down into irreducible parts. For coding purposes, it is then sufficient to consider only the "richest" component, that is, the component that presents the system with the largest capacity (which is then the capacity of the entire system). An irreducible presentation that is not primitive is in fact s-partite for some integer $s > 1$. That is, the set of states V can be partitioned into sets $V_0, \ldots, V_{s-1}$ such that each arc that starts in some V_i terminates in V_{i+1} (or in V_0 if $i = s - 1$). In that case, the sth power of the presentation (the presentation that generates s labels per arc; see Section 5) consists of s disjoint primitive components, which all have the same capacity sC, where C is the capacity of the original system.

So, for coding purposes we may in general assume that the presentation of the constrained system at hand is primitive. Note that primitivity of the presentation is precisely the property needed to assure that a *fixed* number of h merging bits, for some h, can always be used to glue any two codewords together; the merging bits can be read off from a path of length h connecting the terminal state of a path generating the first codeword to the initial state of the path generating the second codeword [see the discussion following Eq. (1) above]. For further details and proofs, the interested reader is referred, for example, to Ref. 10.

The (labeling of a) presentation is called *lossless* if any two paths with the same initial state and terminal state generate different sequences. This is an important property in connection with capacity computations. Indeed, let $G = (V, A)$ be a presentation, with $m = |V|$ states, say. If G is lossless, then at most m^2 paths can generate a given sequence, so the number P_n of paths of length n in G and the number N_n of sequences of length n presented by G are related by

$$P_n/m^2 \leq N_n \leq P_n \tag{2}$$

Hence the numbers P_n and N_n exhibit the same growth rate.

The growth rate of the numbers P_n can be computed from a matrix describing the underlying graph of the presentation, defined as follows. The adjacency matrix D of G is an $m \times m$ matrix where the entry $D(s, t)$ counts the number of arcs going from state s to state t. Note that the (s, t) entry of the nth power D^n of D counts the number of paths of length n in the graph from state s to state t. So the sum of the entries in D^n equals the number P_n of paths of length n in the graph:

$$P_n = \sum_{s,t \in V} D^n(s, t) \tag{3}$$

Now we can use a result from the classical Perron–Frobenius theory for nonnegative matrices (see e.g., Refs. 11 and 12; note that the adjacency matrix D is of this type), which states that the largest real "Perron–Frobenius" eigenvalue λ_D of a nonnegative matrix D equals its spectral radius and determines the growth rate of the entries of D^n; in fact we have [e.g., 10]

$$\lim_{n \to \infty} \frac{1}{n} \log \sum_{s,t \in V} D^n(s, t) = \log \lambda_D \tag{4}$$

Combining Eqs. (1)–(4), we conclude that the capacity $C(\mathcal{L})$ of a sofic system $\mathcal{L}$ with lossless presentation G is given by

$$C(\mathcal{L}) = \log \lambda_D \tag{5}$$

where λ_D is the largest real eigenvalue of the adjacency matrix D of G. This eigenvalue can be computed, for instance, as the largest real root of the *characteristic equation*

$$\det(\lambda I - D) = 0 \tag{6}$$

Alternatively, the capacity can be obtained by setting up linear recurrence relations for the numbers $P_n(s)$ of paths of length n starting in state s. The theory of such recurrences implies that these numbers grow as $c_s \lambda^n$, where λ is the largest real zero of the characteristic equation associated with this recurrence relation, which turns out to be the same as (6). For further details, see, for example, the book by Immink [13].

For example, the presentation of the (d, k)-constrained system in Fig. 5 is lossless. So we can apply the abovementioned method, and it turns out that the maximum code rate (the capacity) of a (d, k) code is given by the logarithm of the largest real root of the equation

$$x^{k+2} - x^{k+1} - x^{k+1-d} + 1 = 0 \tag{7}$$

if k is finite or

$$x^{d+1} - x^d - 1 = 0 \tag{8}$$

if $k = \infty$.

There are various properties of (the labeling of) a presentation that imply losslessness. All these properties are in fact concerned with the number of paths that generate a given sequence and with their localization in the graph. For example, a presentation has *finite local anticipation* if there is a number a such that any two paths of length

$a+1$ with the same initial state and generating the same sequence have the same initial arc. Similarly, the presentation has *finite local coanticipation* if the presentation obtained by reversing the direction of all arcs has finite local anticipation.

A presentation is *of finite type* if there exists a pair of numbers (m, a) (referred to as the *memory* and *anticipation* of the type) such that all paths that generate a given constrained sequence are equal, with the possible exception of at most m initial and a terminal arcs. Thus, given a (long) constrained sequence, we can reconstruct the path used to generate the sequence, up to a few arcs at the beginning and the end of the path. It turns out that a constrained system is of finite type (i.e., it can be described in terms of a finite number of "forbidden patterns") if and only if it has *some* presentation of finite type [14].

For future reference, we also introduce a slightly weaker property. A labeling is of almost finite type if it has both finite local anticipation and finite local coanticipation. As suggested by this terminology, it can be shown that a labeling of finite type is of almost finite type. A constrained system is said to be of almost finite type if it can be presented by some system of almost finite type.

A labeling with local anticipation 0 has the property that for any state, the outgoing arcs carry distinct labels. Such a labeling is called *deterministic*. This is an important property, for several reasons. Most "natural" presentations of constrained systems [e.g., the presentation of (d, k)-constrained systems in Fig. 5] are already deterministic. Moreover, any presentation G can be transformed into a deterministic presentation G^* by using a variant of the well-known *subset construction* for finite automata (see, e.g., Ref. 15 or 16). Here, G^* has as states all nonempty subsets of states of G, and for each such subset S and each labeling symbol a, the presentation G^* will have an arc $S \xrightarrow{a} T$, where T consists of all states t for which G has an arc $s \xrightarrow{a} t$ for some s in S. It is not difficult to see that indeed G and G^* present *the same* sofic system.

At this point, it is important to realize that a given sofic system can have (infinitely) many different presentations. Fortunately, every *irreducible* sofic system (i.e., one that has an irreducible presentation) has a unique *minimal* (in terms of the number of states) *deterministic* presentation, called the *Shannon cover*. It can be constructed from any irreducible deterministic presentation by an operation called *state merging*, in the following way. The set of sequences that can be generated by paths that start in a given state is called the *follower set* of that state. Now, if two states have the same follower set, then for sequence generation purposes these states are equal and can therefore be combined or *merged* into one state. Now, we repeatedly apply state merging until the follower sets in the states are all distinct. For further details about this construction, we refer to the book by Lind and Marcus [10].

The Shannon cover is important because it is "small" (hence very suitable for capacity computations), and also because many properties of a sofic system can be determined directly from its Shannon cover. For example, a sofic system is of finite type (respectively, of almost finite type) if and only if the labeling of its Shannon cover is of finite type (respectively, is of almost finite type). As a consequence, the Shannon cover is the natural starting point of many code construction methods.

4. ENCODERS AND DECODERS

An *encoder* for a given constrained system $\mathcal{L}$ is a device that transforms arbitrary binary sequences of *source bits* into sequences contained in $\mathcal{L}$. Commonly, the encoder is realized as a *synchronous finite-state device*. Such a device takes *source symbols*, groups of p consecutive source bits, as its input, and translates these into q-bit codewords, where the actual output depends on the input and possibly on the *internal state*, the (necessarily finite) content of an internal memory, of the device. The rate of such an encoder is then $R = p/q$. (To stress the individual values of p and q, we sometimes speak of a "rate $p \to q$" encoder.) Obviously, each codeword must itself satisfy the given constraint. Moreover, the encoder needs to ensure that the bit stream made up of successive codewords produced by the encoder also satisfies the constraint.

The use of the word "code" implies that it should be possible to recover or *decode* the original sequence of p-bit source symbols from the corresponding sequence of q-bit codewords. In practice, a stronger requirement is needed. Since modulation codes are typically used in the context of a noisy channel, it is important that the decoder limits the propagation of input errors. Therefore, we commonly require that the modulation code can be decoded by a *sliding-block decoder*.

A sliding-block decoder for a rate $p \to q$ encoder takes a sequence of q-bit words y_n as its input, and produces a sequence of p-bit symbols x_n as its output, where each output symbol x_n depends only on a corresponding sequence $y_{n-m}, \ldots, y_{n+a}$ of $w = m + 1 + a$ consecutive inputs, for some fixed numbers m and a, $m \le a$. We will refer to the number w as the *window size* of the decoder. (The numbers m and a are referred to as the *memory* and *anticipation* of the decoder.) The name "sliding-block decoder" refers to the image of the decoder sliding a "window" of width w over the sequence to be decoded. In Fig. 6 we depict a sliding-block decoder with $m = 2$, $a = 1$, and

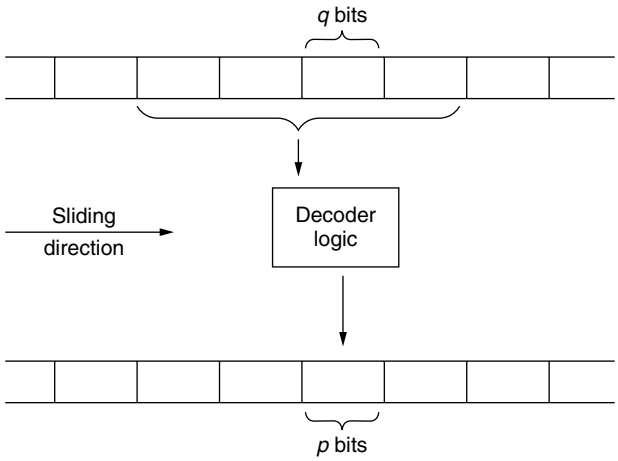


Figure 6. A sliding-block decoder.

window size $w = 4$. Note that an error in the encoded sequence will cause at most w symbol errors in the original sequence. So the *error propagation*, the amount of erroneous bits in the decoder output caused by a single input error, is limited to at most wp bits.

The window size of the decoder is an important parameter of a code: it provides an upper bound on the error propagation of the code, and also gives a good indication of the *size* of the decoder, the amount of hardware necessary to implement the decoder (i.e., it provides an upper bound on the *complexity* of the decoding operation).

Codes with a window size of one codeword (i.e., for which $w = 1$) are called *block-decodable*. For such codes, decoding a given codeword or *block* can be achieved without any knowledge of preceding or succeeding codewords. In many present-day applications, modulation codes are used in combination with symbol-error-correcting codes based on Reed–Solomon codes over some finite (Galois) field GF(2^p). In that situation, the use of a rate $p \to q$ block-decodable modulation code becomes especially attractive, since then a channel-bit error affects at most one p-bit symbol of a Reed–Solomon codeword. Note that this cannot be guaranteed if a non-block-decodable code is used, even if its decoding window size is smaller than q bits.

Note that decoding of a code requires a form of *synchronisation* between encoder and decoder in order to enable the decoder to identify the codeword boundaries in the encoded sequence. In practice this is usually achieved by using *frame-based* codes. Here, codewords are grouped into *frames* and unique *frame headers* are used to signal the beginning of a new frame. For further details, we again refer to the treatise by Immink [13].

5. CODE CONSTRUCTION METHODS

In the simplest case, the encoder translates each source symbol into a unique corresponding codeword, according to some code table. Of course, this will produce an admissible sequence only if any concatenation of these codewords also satisfies the given constraint. Codes of this type are called *block codes* or *block-encodable codes*. Decoding then consists of simply translating the codewords back into their corresponding source symbol, so these codes are block-decodable.

For example, consider the $(1, \infty)$-constrained system, in which the pattern 11 is not allowed to occur in the sequence. It is easily seen that the code where the source symbol 0 is encoded into the codeword 00 and 1 is encoded into 01 is a rate-$\frac{1}{2}$ block-decodable $(1, \infty)$ code. This simple code is called the *frequency modulation* (FM) or *biphase* code.

A slight improvement of this code is the *modified frequency modulation* (MFM) or *Miller* code. This is a rate $1 \to 2$ $(1, 3)$ code that encodes by inserting a *merging bit* between each two consecutive source bits. Here, the merging bit is set to 0 except when both surrounding bits are 0, in which case it is set to 1. Note that the FM code is obtained if the merging bit is always set to 0. Although the MFM code is not a block code, it is still block-decodable. In fact, decoding consists of simple deletion of the merging bits.

The use of merging bits is a well-established technique for constructing block-decodable (d, k) codes. They are often employed in combination with $(dklr)$ sequences, that is, (d, k)-constrained sequences in which the number of leading and trailing zeros is restricted to be at most l and r, respectively [17]. Here, p-bit source symbols are uniquely translated into $(dklr)$ sequences of a fixed length q'; in addition, a fixed number of $q - q'$ merging bits, chosen according to a set of *merging rules*, is employed to ensure that the resulting bit stream is a valid (d, k) sequence. Possible freedom in choosing these merging bits can then be used, for instance, to limit the DC content of the resulting sequence. These ideas can be found, for example, in the EFM recording code for the compact disk (see also Section 6). For some additional information on these methods, we refer to the paper by Weber and Abdel–Ghaffar [18]. A similar idea has been applied [19] to construct *almost-block-decodable* (d, k) codes.

Next, we shall discuss a number of code construction methods that employ an *approximate eigenvector* to guide the construction. We shall first explain this concept and motivate its importance for code construction.

Consider a given constrained system $\mathcal{L}$, presented by some (irreducible deterministic) graph G. Suppose that we wish to design a code with rate $p \to q$, where we assume that $p/q \leq C(\mathcal{L})$. Since we are interested in admissible sequences of q-bit codewords, it is convenient to consider the qth power graph G^q of G. This graph has the same states as G, and has an arc for each path of length q in G, labeled with the q-bit sequence generated by this path. Obviously, G^q essentially generates the same sequences as G, but does so with q bits or one codeword at the time. Note that if D is the adjacency matrix of G, then the adjacency matrix of G^q is given by D^q.

The desired encoder encodes arbitrary sequences of p-bit symbols into sequences of codewords; we shall refer to the collection of all these codeword sequences as the *code system* of the code. Note that each such sequence corresponds to one or more *encoding paths* in G^q.

Intuitively, we would expect that the number $\phi_s^{(n)}$ of encoding paths of length n in G^q that start in state s grows exponentially in 2^p, the number of source symbols. Indeed, under certain assumptions, it can be shown [20] that for each state s the limit

$$\phi_s = \lim_{n \to \infty} \frac{\phi_s^{(n)}}{2^{pn}} \tag{9}$$

exists and takes on an *integer value*. We will write ϕ for the vector with as entries the numbers ϕ_s.

Since an encoding path of length n starting in s consists of a transition in G^q from s to some state t, say, followed by an encoding path of length $n - 1$ starting in t, the vector $\phi^{(n)}$ with the numbers $\phi_s^{(n)}$ as entries satisfies

$$\phi^{(n)} \leq D^q \phi^{(n-1)} \tag{10}$$

(Note that we have an inequality here since not each path of length n obtained in this way needs to be an encoding path.) If we now combine Eqs. (9) and (10), we conclude that

$$2^p \phi \leq D^q \phi \tag{11}$$

A nonnegative integer vector ϕ that satisfies the inequality (11) is called a $[D^q, 2^p]$-*approximate eigenvector*, or a $[G^q, 2^p]$-*approximate eigenvector* if we wish to stress the connection with the presentation. We think of the numbers ϕ_s as weights associated with the states. In terms of these weights, the inequalities state that the sum of the weights of the (terminal states of) arcs leaving a state is at least 2^p times the weight of this state. The discussion above makes precise the idea that in code construction, the weight of a state is an indicator for the *relative encoding power* of that state.

It can be shown [12] that a $[G^q, 2^p]$-approximate eigenvector exists if and only if $p/q \leq C$, the capacity of the system presented by G. The following simple algorithm, first introduced by Franaszek [21], produces such an approximate eigenvector with components not larger than a given number M, provided such a vector exists. In this algorithm, we initially set $\phi_s = M$ for all states s, and then repeatedly perform the operation

$$\phi \leftarrow \min\{\phi, \lfloor 2^{-p} D^q \phi \rfloor\} \tag{12}$$

(where both the rounding operation $\lfloor\ \rfloor$ and taking the minimum are performed componentwise), until either $\phi = 0$ (in which case no such approximate eigenvector exists) or the vector ϕ remains unchanged (in which case ϕ is the desired approximate eigenvector).

Approximate eigenvectors were first introduced by Franaszek. In a series of pioneering papers [1,21–24], he developed a number of code construction methods that all employ approximate eigenvector in an essential way.

For example, the *principal-state method* is based on the observation that the existence of a *binary* $[D^q, 2^p]$-approximate eigenvector is equivalent to the existence of a set of *principal states* with the property that from each principal state there are (at least) 2^p distinct paths of length q in G (i.e., 2^p arcs in G^q) ending in another principal state. (The principal states are the states with weight one.) If it exists, a binary approximate eigenvector can be obtained by the recursive elimination procedure (12) by taking $M = 1$.

Given such a collection of paths in G, we can immediately construct a rate $p \rightarrow q$ encoder as follows. In each principal state, assign each of the 2^p possible source symbols to a q-bit codeword label of a path leaving this state. Now, the encoder moves from one principal state to another, using, for example, an encoding table in each principal state to translate input symbols into codewords. Usually, the encoding tables are implemented by using *enumerative methods* [13,25]. Such codes are called *fixed-length principal-state codes*.

If the constraint is of finite type, then any assignment of source symbols to codewords will lead to a sliding-block decodable code: in that case, the sequence of principal states traversed by the encoder can be reconstructed from the sequence of codewords.

The principal-state method may lead to prohibitively large values of p and q. (Note that these values have a large impact both on the complexity of the source-to-codeword assignment and on the size of the encoding tables.) For example, the minimum codeword lengths of a fixed-length principal-state rate-$\frac{2}{3}(1, 7)$ code and a rate-$\frac{1}{2}(2, 7)$-code are 33 and 34, respectively. Sometimes this problem can be avoided if we allow the coding paths between principal states to have varying lengths. As an example, we consider the design of a rate-$\frac{1}{2}(2, \infty)$ code. (The forbidden patterns are 11 and 101.) The minimum codeword length of a *fixed-length* rate-$\frac{1}{2}$ code for this constraint is 14 [e.g., 13]. The constraint can be presented by a three-state graph G with states named 0, 1, and 2, (see Fig. 7). Here, sequences ending in state 0 or 1 have precisely 0 or 1 terminal zeros, respectively, and a sequence ending in state 2 has at least 2 terminal zeros. The second-power graph G^2 of G needed in the construction of a rate $1 \rightarrow 2$ code is depicted in Fig. 8.

Let the set of principal states consist of the single state 2. Now consider the possible encoding paths, that is, paths in G^2 starting and ending in state 2. By inspection of G^2, we obtain the three paths

$$2 \xrightarrow{00} 2, \quad 2 \xrightarrow{10} 1 \xrightarrow{00} 2, \quad 2 \xrightarrow{01} 0 \xrightarrow{00} 2$$

from which a code can now be designed. The encoding rules are specified as

$$0 \rightarrow 00, \quad 10 \rightarrow 10.00, \quad 11 \rightarrow 01.00$$

The code can be decoded with a sliding-block decoder that has a window size of two codewords.

In general, for this method to succeed we need a (finite) collection $\mathcal{P}$ of paths π in G^q between principal states such that in each principal state s, we have that

$$\sum 2^{-|\pi|} \geq 1 \tag{13}$$

where the sum is over all paths π starting in s (and ending in the same or another principal state) and where $|\pi|$ denotes the *length* (number of transitions in G^q) of π. This condition, the Kraft–McMillan inequality for prefix codes [e.g., 26] is necessary and sufficient for the existence of a

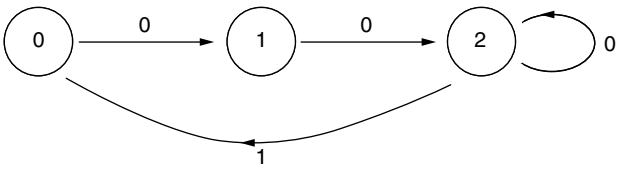


Figure 7. Presentation of the $(2, \infty)$ constraint.

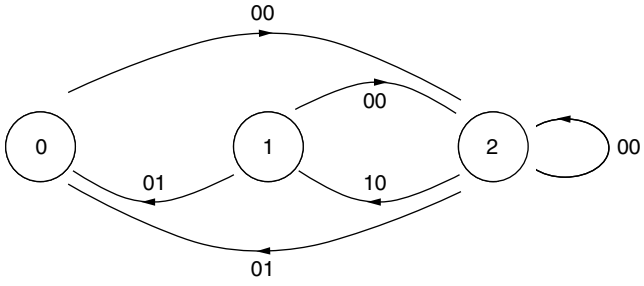


Figure 8. Presentation of G^2.

prefix code (such as the source words 0, 10, and 11 in the preceding example) with wordlengths equal to the path lengths. A similar but more complicated example of the above method is the rate-$\frac{1}{2}(2, 7)$ code [27].

Codes for constraints of finite type found by this method are always sliding-block-decodable. Such codes are called *variable-length principal-state codes* or (synchronous) *variable-length codes*. A special case of this method is the *substitution method*. Here, we set up some simple encoding rules, and then try to remove violations of the constraints by suitable substitutions. For example, consider again the design of a rate-$\frac{1}{2}(2, \infty)$ code. Suppose that we try to use the simple encoding rules $0 \to 00$ and $1 \to 10$. This works fine except when the source sequence contains two consecutive 1s, which produces 1010. These violations can be removed by replacing, from left to right, each occurrence of 1010 in the encoded stream by 0100. Note that we have obtained the same code as the one constructed earlier. A similar but more complicated example is the rate-$\frac{2}{3}(1, 7)$ code ("Jacoby code") [28].

Franaszek [23,24] and Lempel and Cohn [29] also introduced the class of *bounded-delay encodable codes* and the *bounded-delay method* for constructing such codes. Bounded-delay encodable codes involve a finite-state encoder where the encoding of the current source symbol may depend on the present state, on a bounded number of upcoming source symbols (so *lookahead* may be employed), and on a bounded number of previous states (the *history*) in the encoding path. The bounded-delay construction method for such codes employs an approximate eigenvector in an essential way. The idea is to construct in each state a suitable collection of *independent path sets*, each consisting of a set of encoding paths for this state, by exhaustive search up to a fixed maximum pathlength. Although this is a powerful construction method (in fact, much later it was shown [30] to be as powerful as the ACH method to be mentioned next, see also Refs. 31–33, and Ref. 34, Chaps. 4 and 5), there is no easy way to turn this method into an algorithm.

A technique called *state combination* [35] uses an approximate eigenvector with all components equal to 0, 1, or 2 to construct block-decodable codes that can be encoded by employing one-symbol lookahead, a special case of bounded-delay encodable codes. State combination is especially suited to the construction of (d, k) RLL codes. The method has been further developed and extended [32,36].

A breakthrough in code construction was achieved by the discovery of the state-splitting method or *ACH algorithm* [3]. This method employs an approximate eigenvector to construct an encoder, and does so in a number of steps bounded by the sum of the components of the approximate eigenvector. It can be used to prove the following theorem.

Theorem 1. For any given constraint of finite type and for any given rate $p \to q$ for which $p/q \leq C$, the capacity of the constraint, there exists a sliding-block decodable code for that constraint, with a synchronous finite-state encoder that encodes binary data at a constant rate $p \to q$.

Starting point in this method is a pair (H, ϕ), where $H = G^q$ is the qth power of an (irreducible) graph G presenting our constraint and ϕ is a $[H, 2^p]$-approximate eigenvector. (Here we may assume without loss of generality that H is irreducible and that $\phi > 0$.) The algorithm repeatedly transforms the current pair (H, ϕ) into another such pair (H', ϕ') by an operation called *weighted state splitting*. (This operation is called *ϕ-consistent state splitting* in Ref. 37, to which we refer for an excellent overview of the method.) This transformation is guaranteed to succeed except when all nonzero weights are equal. It has the property that the new weights are in some sense "smaller" than the original weights.

So with each transformation the approximate eigenvector gets "smaller" until finally a pair (G, ϕ) is reached where all nonzero components of ϕ are equal. Then the approximate eigenvector inequalities for ϕ show that in each state with a nonzero weight there are at least 2^p transitions leading to other such states, that is, we have obtained an encoder for our constraint.

State splitting, on which the transformation is based, can be understood intuitively as follows. While generating a sequence in the graph, each state encountered stands for a collection of future opportunities (represented by the arcs leaving the state) from which we may choose one. Now subdivide or *split* the state, this collection of opportunities, into *nonempty* parts called *substates* (to each of which we assign the corresponding part of the original arcs). Before this splitting operation, we could move from another state to this state without worrying about which opportunity (which arc) to utilize later, but now we have to choose first, before moving, from *which* of the parts we wish to pick our next opportunity. We do not lose opportunities, but we have to choose earlier.

So if s is a state and A_s is the collection of arcs leaving this state, then to split this state, we proceed as follows. First we partition the set A_s into *nonempty* parts $A_{s_1}, \ldots, A_{s_r}$. Then, in the graph G we replace the state s by states $s_1, \ldots, s_r$ (the substates of s), we replace each arc α in part A_{s_i}, $i = 1, \ldots, r$, by an arc from s_i with the same label and terminal state as α, and then we replace each arc β ending in s by r arcs $\beta_1, \ldots, \beta_r$, with the same label and initial state as β, with β_i ending in s_i, $i = 1, \ldots, r$.

It should be evident from the preceding discussion that the new graph obtained by splitting a state presents the same sequences as the original graph. Moreover, it is not difficult to see that if the original graph is of finite type, then the new graph is again of finite type, with the same memory and with an anticipation that has increased by at most one. Note that if the final graph has memory m and anticipation a, then the encoder obtained from this graph will have a decoding window of size at most $m + 1 + a$.

Weighted state splitting is another type of simple state splitting where we also distribute the weight of the state that is split over the substates (where each substate should receive a *nonzero* amount), and in such a way that the resulting weights constitute an approximate eigenvector for the new graph. A state that allows weighted state splitting can always be found among the states with maximum weight [3], provided not all nonzero weights are equal, and an algorithm is given to find such a state.

Example 1. Consider again the presentation $H = G^2$ in Fig. 8. Take $\phi = (1, 1, 2)$. Obviously, ϕ is a $[G^2, 2]$-approximate eigenvector. We shall split state 2 into two states, 2_1 and 2_2, according to the partition $A_{2_1} = \{2 \xrightarrow{00} 2\}, A_{2_2} = \{2 \xrightarrow{01} 0, 2 \xrightarrow{10} 1\}$. We also distribute the weight of state 2 over the two descendents of state 2 by assigning weight 1 to both 2_1 and 2_2. Since state 2_1 has successor state 2 (i.e., both 2_1 and 2_2) of total weight 2, and since state 2_2 has as successors the states 0 and 1, also of total weight 2, the new weights again form an approximate eigenvector, so the state split is ϕ-consistent. The new approximate eigenvector $\phi' = (1, 1, 1, 1)$ is constant, hence the resulting graph (depicted in Fig. 9) includes an encoder graph. By choosing a suitable assignment of 1-bit source symbols to the arcs (also shown in Fig. 9), we (essentially) obtain the rate-$\frac{1}{2}(2, \infty)$ code constructed earlier.

Ashley and Marcus [31], have shown that the ACH algorithm is *universal* in the sense that, given a sliding-block decoder, a matching finite-state encoder can be obtained by weighted state splitting (see also Refs. 32 and 33 and Ref. 34, Chaps. 4 and 5). However, precisely for that reason, the algorithm offers a large amount of freedom, first in the choice of the approximate eigenvector, then both in the choice of the states to be split and in the actual splits themselves, and it is not clear how to choose in order to obtain a good code.

It is usually best to choose the *smallest possible* approximate eigenvector. This works well in practice; however, pathological cases are known where the best possible code can be produced only by arbitrarily large approximate eigenvectors [20].

Also, some heuristics have been developed to guide the state-splitting process [e.g., 37] in order to minimize the number of encoder states of the resulting encoder. However, in practical applications we are often especially interested in codes with a small *decoding window*. This problem is addressed in Ref. 32, where weighted state splitting is combined with ideas from the bounded-delay method of Franaszek to obtain heuristics to guide the choices of which states to split. For completeness, we mention that Theorem 1 holds even for constraints of almost finite type [14].

Which code construction method to use strongly depends on the structure of the power graph G^q presenting the constraint at the desired rate, and in particular on the maximum number of arcs leaving a state. If this number is relatively *small*, then use an ACH-type method. (An attractive choice is the variant described by Hollmann [32].) On the other hand, if this number is relatively *large*, then such methods are less suitable because of the increasing number of choices for the state-splitting steps. In that case, use a method based on merging bits, or a method for (almost) block-decodable codes such as described in Ref. 35 or 33, or any other more heuristic method that does the job.

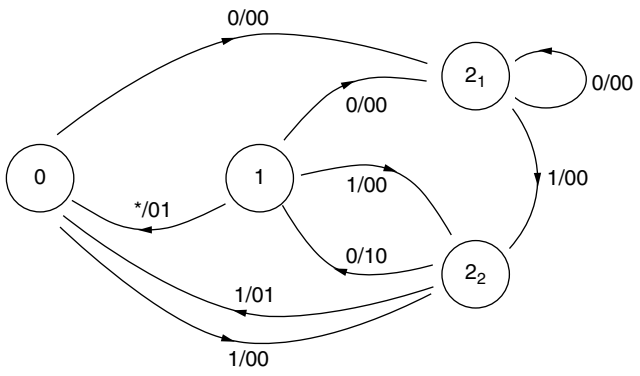


Figure 9. Presentation after state splitting.

Extremely large values of p and q are encountered in the design of high-rate codes, or when the efficiency of the code should be very high. In such cases, methods as *guided scrambling* [38], or a promising variant of enumerative encoding (see Refs. 39 and 40, and Ref. 34, Chap. 1) should be considered. Here, the extremely large block-length imposes a special system architecture to limit error propagation. For further details, we refer to the paper by Immink [40].

6. SELECTION OF RELATED TRENDS AND TOPICS

6.1. DC Control in RLL Codes for Optical Recording

Run-length-limited (RLL) codes have been successfully applied in optical storage, as witnessed by the well-known examples of EFM [41] for CD and EFMPlus [42] for DVD. The EFM and EFMPlus codes both employ a $(2, 10)$ constraint. EFM stands for *8-to-14 modulation*: the EFM code uses a single code table that maps a byte to a 14-bit channel word. Successive channel words are concatenated via the insertion of three merging bits, so the EFM code has rate $\frac{8}{17}$. The EFMPlus code maps a byte to a 16-bit channel word. EFMPlus is an ACH-type sliding-block code (see Section 5) based on a 4-state encoder graph. Both EFM and EFMPlus are byte-oriented, which is desirable for formats with the error correction coding (ECC) based on 8-bit symbols. An overview of the EFM and EFMPlus codes is given by Immink [43]. More recently, the *Blu-Ray Disc* (sic) or BD (formerly known as the DVR system [44]), which is the third generation of optical recording after CD and DVD, employs a code called "17PP" that uses a $(1, 7)$ constraint.

All RLL codes used in optical recording are *DC-free*; that is, they have almost no content at low frequencies. This property is an example of a *frequency-domain constraint*. Here, restrictions are enforced on the energy content per time unit of the sequence at certain frequencies, that is, on the *power spectral density function* of the sequence. (Constraints like run-length constraints are called *time-domain* constraints.) Most of these constraints belong to the family of *spectral null constraints*, where the power density function of the sequence must have a zero of a certain order at certain specific frequencies. The constraint that specifies a zero at DC, the zero frequency, is referred to as the *DC-free* constraint. We shall represent the NRZI channel bits by the bipolar values ± 1. A sequence $x_1, x_2, \ldots$ is called *DC-free* if its *running digital sum* (RDS)

$$\mathrm{RDS}_i = x_1 + \cdots + x_i$$

takes on only finitely many different values. In that case, the power spectral density function vanishes at DC.

One common way to ensure this is to constrain the code sequence to be *N-balanced*; we allow only sequences whose RDS takes on values between $-N$ and N. Note that such a constraint cannot be specified in terms of a finite collection of forbidden patterns; that is, it is not of finite type.

The DC-free property is needed in optical recording for a number of reasons: (1) it is necessary to separate the data signal from low-frequency disk noise such as fingerprints, dust, or defects; (2) DC-free coding is needed for control of the slicer level in the case of nonlinearities in the physical signals like pit–land asymmetries [45]; and (3) servo systems used for tracking of the laser spot position typically require a DC-free data signal.

We shall now discuss a general method to achieve DC control in RLL sequences. As discussed above, DC control is performed via control of the running digital sum (RDS). A very useful concept herein is the *parity*, the number of ones modulo 2, of a sequence of bits. Recall that an NRZ 1 bit indicates the start of a new run in the (bipolar) NRZI bit stream. Hence, because of the 1T precoder between NRZ and NRZI channel bit streams (see Section 1), each 1 bit in the NRZ bit stream changes the polarity in the corresponding NRZI bit stream. Consequently, an *odd* number of ones in a segment of the NRZ bit stream *reverses* the NRZI polarity after that segment while an *even* number of ones leaves the polarity unchanged.

This observation can be used for DC control as follows (see also Fig. 10). Suppose that for a certain segment of the NRZ bit stream, we can choose between two candidate sequences, one with parity 0 and the other with parity 1. Then the part of the NRZI bit stream *after* this segment will have a contribution to the RDS where the *sign* but not the magnitude depends on which of the two sequences is chosen. The *best* choice is, of course, the one that keeps the value of the RDS as close to zero as possible. For obvious reasons, we shall refer to these segments as *DC-control segments*.

In order to realize DC control, we have to insert DC-control segments at regular positions in the bit stream. Such positions are referred to as *DC-control points*. This is the basic mechanism for DC control in RLL codes.

Two types of DC-control segments can be distinguished. One type just serves the primary purpose of parity selection; another type additionally encodes some data bits. Further, we can differentiate between two types of DC-control: one type providing *guaranteed* DC control where parity selection is possible at each DC-control point, and another type providing *stochastic* DC-control where the possibility of parity selection depends on the data that is to be encoded.

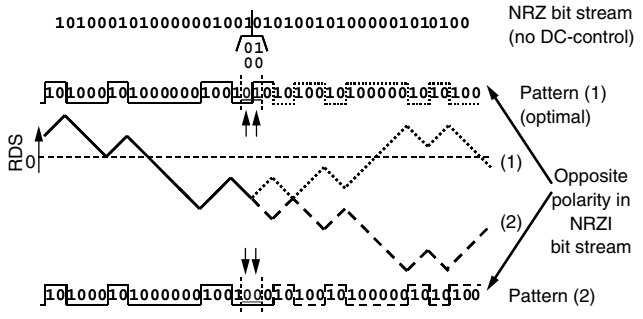


Figure 10. Principle of DC control via insertion of merging bits in a valid $(1, \infty)$ bit stream.

We shall now review several practical design methods for DC control that use the above mechanism in one form or another.

1. Insertion of a DC-control segment of N channel bits in a NRZ bit stream already satisfying a (d, k) constraint. At each DC-control point, the original bit stream is cut open, and the segment is inserted. If we require that both parities (0 and 1) can be selected without violation of the (d, k) constraint, then the minimum possible value of N is $N = 2(d + 1)$ if $2d + 1 \leq k < \infty$ and $N = d + 1$ if $k = \infty$. This can be seen as follows. At a DC-control point, the situation looks like $\ldots 10^i \cdot 0^{r-i}1\ldots$, for some i and r with $0 \leq i \leq r$ and $d \leq r \leq k$. First, let $k = \infty$. A possible segment of odd parity is $0^{d-i}10^i$ if $i \leq d$ or 10^d if $i \geq d$; if $i = 0$, no shorter segment is possible. Of course, the segment 0^{d+1}, of even parity, can always be inserted. Next, suppose that $2d + 1 \leq k < \infty$. Now, a possible segment of even parity is $0^{d-i}10^d10^i$ if $i \leq d$ or 10^d10^d if $i \geq d$; again, if $i = 0$, no shorter segment is possible. Also, a possible segment of odd parity is $0^{2d+1-i}10^i$ if $i \leq 2d + 1$ or 10^{2d+1} if $i \geq 2d + 1$. The case where $k = \infty$ and $d = 1$, which requires only two merging bits, is illustrated in Fig. 10. Here, following the 1 bit, two segments 01 and 00, of opposite parity, can be merged into the sequence; for both choices, the respective RDS traces are drawn. Since the RDS for the first pattern remains close to zero, whereas the RDS trace of the second pattern drifts away, obviously the first pattern is the best choice.

2. Insertion of merging bits between successive NRZ channel words, as is used in the EFM code. The function of the merging bits in EFM is twofold; they (a) serve to prevent violations of the $(2, 10)$ run-length constraint and (b) provide a means for DC control via the available freedom in the choice of the merging bits. Note that whether parity selection is possible depends on the two EFM channel words at hand; that is, the DC control of the EFM code is of a stochastic nature. For example, when the previous EFM word ends in a 1 and the current EFM word starts with 1 or 01, then the only valid merging bit pattern (i.e., not violating the $d = 2$ constraint) is 000, so that no DC control is possible.

3. Use of a substitution table for DC control. Here a code is used where certain source words can be encoded into two NRZ channel words (a *standard* word and a *substitute* word) of opposite parity. This mechanism is used, for example, in the EFMPlus code. Here, code construction via the ACH algorithm yields a 4-state encoder graph. In each state, the main encoding table contains 256 entries; because of the presence of *surplus* words, 88 of these entries have a substitution entry. (Additional DC control in EFMPlus is achievable via occasional swapping of encoder state [43].) Note that the DC-control mechanism via substitution tables (as in EFMPlus) is of a stochastic nature; whether a byte can be used as a DC-control point depends on whether the encoding table has a substitute entry at that location.

4. DC control via the use of a *parity-preserving* code [46]. Such a code preserves the parity on RLL

encoding; that is, the parity of a source word is identical to the parity of the corresponding channel word. The major difference with the previous methods is that single DC-control bits are inserted in the *source* bit stream. Changing a DC-control bit from 0 to 1 changes the parity in the source bit stream and hence also in the NRZ channel bit stream; this property enables the selection of the polarity of the NRZI channel bit stream, and thus allows for DC control.

5. DC control via the use of *combicodes* [47]. This is a typical example of a method where some data bits are encoded in the DC-control segment. A combicode for a given constraint consists of a set of at least two codes for that constraint, possibly with different rates, where the encoders of the various codes share a *common* set of encoder states. As a consequence, after each encoding step the encoder of the current code may be replaced by the encoder of any other code in the set, where the new encoder has to start in the ending state of the current encoder. Typically, one of the codes, called the *standard* code or *main* code, is an efficient code for standard use; the other codes serve to realize certain additional properties of the channel bit stream. Sets of sliding-block decodable codes for a combicode can be constructed via the ACH algorithm; here the codes are jointly constructed starting with suitable presentations derived from the basic presentation for the constraint and using the *same* approximate eigenvector [47].

Coene has used the combicode idea [47] for DC-control as follows. The *main* code has a single channel word for each source word. As a second code, of a lower rate, a *substitution code* is used. This code has for *every* source word a set of two possible channel words, of opposite parity and ending in the same encoder state.

In a practical format, the sequence of uses of the various codes of the combicode needs to be chosen beforehand. This choice will be the result of a tradeoff between the overall rate and the required property (e.g., the amount of DC control) to be realized. Obviously, DC control by use of these combicodes is of the *guaranteed* type.

6.2. Time-Varying Codes

The concept of time-varying codes [48] is related to a generalization of the state-splitting (ACH) algorithm. In this generalization, the starting point is a *periodic* presentation of the constraint where the codeword labels may have different lengths in different phases. Thus, the set V of vertices can be partitioned into subsets $V_0, V_1, \ldots, V_{s-1}$, say, with the property that an arc starting in some part V_i ends in V_{i+1} (or in V_0 if $i = s - 1$); moreover, labels on arcs that start in the same part have the same length. An encoder derived from such a periodic presentation operates cyclically in s phases, where the length of the codeword produced by the encoder depends only on the current phase.

The framework of time-varying codes is useful for the design of efficient codes with reduced error propagation. Note that a time-varying code may be considered as a pair of codes where the order in which the codes have to be used is fixed beforehand; this in contrast with combicodes, where the order in which the various codes can be used is completely free. The concept of multiple phases in a time-varying code [48] is strongly related to the idea of representing a bit by a number of (virtual) *fractional* bits, as used by Coene [47] to generate a highly efficient DC-free combicode for $d = 1$.

6.3. RLL Parity-Check Coding

In the standard digital storage system as sketched in Fig. 1, random channel errors that may occur in the channel bit stream after bit detection are repaired by the ECC error correction decoder situated after the MC decoder. In the early 1990s, it was realized that a *combination* of error correction coding and modulation coding may be quite advantageous in terms of overall efficiency and performance [e.g., 49–52]. RLL parity-check coding focuses on the most prominent error patterns that are left by the bit detector in the receiver. For magnetic recording, the most simple parity-check coding schemes aim at peak shift errors in the NRZ bit stream, where the 1 bits are shifted (to the left or right). For optical recording, where the bit detector regenerates the NRZI bit stream, the most prominent random errors are transition shifts that cause the runs at the left and right sides of the transition to become one or more bits longer or shorter. Because of the 1T precoder, the transition error in the NRZI bit stream corresponds to a peak shift error in the NRZ bit stream. In parity-check coding, the channel bits are partitioned into fixed-length *segments*, and a parity check is generated for each segment. (Subsequently, these parity checks can be handled in various ways.) As a consequence, inspection of the NRZ bit stream after detection will reveal the violation of the parity check in the case that a channel error occurs, thus enabling error detection; for error correction, some side information from the signal waveform is needed. One scheme [52] considers merging of parity blocks into the original NRZ bit stream in such a way that these blocks become an integral part of the corresponding parity-check segment. Another scheme [53] is concatenated parity-check coding, where parity bits that are generated on the NRZ bit stream are separately RLL-encoded. The merging scheme has a low coding efficiency, but the advantage is its simplicity and lack of error propagation. The concatenated scheme has a high efficiency but suffers from error propagation. A promising route is to use a parity-check coding scheme based on a combination of codes [54], like the *combicodes* used for DC control. Apart from a *main* RLL code, a second RLL code, the *parity-check-enabling code*, is required. The latter code is used to set the parity-check constraint of a segment of NRZ bits to a predetermined value. For DC control together with parity-check control, a third code is needed: the *substitution* code. All three codes are jointly constructed, so that the channel words of these codes can be freely concatenated.

6.4. MTR Constraint for Magnetic Recording

The *maximum transition run* (MTR) constraint [55] specifies the maximum number of consecutive 1 bits in the NRZ bit stream. Equivalently, in the NRZI bit stream, the MTR constraint limits the number of successive 1T runs. The MTR constraint can also be combined with a d constraint, in which case the MTR constraint limits the

number of consecutive *minimum runlengths*. An application for $d = 1$ can be found in the paper by Narahara et al. [44]. The basic idea behind the use of MTR codes is to eliminate the *dominant* error patterns, that is, those patterns that would cause most of the errors in the *partial-response maximum-likelihood* (PRML) sequence detectors used for high-density recording. A highly efficient rate $16 \rightarrow 17$ MTR code limiting the number of consecutive transitions to at most two has been described [56].

6.5. (0, G/I) Constraint for Magnetic Recording

In magnetic hard-disk drives using the PR4 or *class 4* partial-response maximum-likelihood (PRML) sequence detectors [57] with response equal to $1 - D^2$, it is advantageous to use modulation coding with so-called $(0, G/I)$ constraints [58,59]. The PR4 detector can be partitioned into two $1 - D$ sequence detectors, where each detector operates at half the bit rate on either the even- or odd-indexed bits. The 0 in $(0, G/I)$ stands for $d = 0$, and the G and the I stand for the *global* and *interleaved* constraint on the maximum number of consecutive zeros in the *joint* and in both *interleaved* bit streams. The global constraint is just a k constraint and is meant for the purpose of timing recovery; the interleaved constraint is introduced in order to limit the effects of truncation of the path memory depths in the separate Viterbi detectors for each of the $1 - D$ responses [7, Chaps. 6 and 7].

6.6. Two-Dimensional Constraints

Two-dimensional (2D) modulation codes have received considerable attention. This research effort is based on the recognition that 2D coding might lead to significant coding gains, especially in technologies such as holographic data storage [60] with a 2D page-based readout mechanism. Just like their 1D counterparts (as explained in Section 2), one of the aims of 2D modulation codes is to combat intersymbol interference (ISI), which is achieved by forbidding certain patterns that lead to high spatial frequencies. 2D lowpass filtering constraints are typically defined in terms of restrictions on neighbouring bits for 2D bit arrays on a rectangular lattice; some examples can be found in Ref. 61 (where they are called "checkerboard codes") and in Ref. 62. Another class of 2D codes are the *multitrack codes* [63], designed for systems where multiple tracks are encoded, written and read in parallel; the d constraint is maintained for each track independently as in the 1D case (in view of ISI), but the k constraint used for timing recovery (or clocking) is defined *jointly* for a set of tracks, since synchronisation is also carried out jointly. Other 2D constraints with potential practical interest are 2D (d, k) constraints, where the 1D (d, k) constraint has to be satisfied in both horizontal and vertical directions. In general, unlike the case for 1D RLL coding, only bounds can be derived for the capacity of 2D (d, k) constraints. This area has been the subject of intensive research (see Refs. 64 and 65 and references cited therein). A practical proposal to use 2D $d = 1$ modulation coding in optical recording can be found in Ref. 66.

Concerning the construction of 2D codes, 2D bit arrays can be encoded in one-dimensional strips containing a number of bit rows (or tracks) as proposed in Refs. 61 and 63. In this way, 2D code construction can be reduced to a 1D problem for which code construction methods such as the ACH sliding-block codes discussed in Section 5 can be used.

Acknowledgments

With great pleasure we thank our colleagues A. H. J. Immink, L. M. G. Tolhuizen, and J.H. van Lint for their many helpful comments made during the preparation of this text.

BIOGRAPHIES

Wim Coene received a Ph.D. in physics from the Center for High-Voltage Electron Microscopy (University of Antwerp) for work on computational modeling of electron diffraction and image formation in a transmission electron microscope (TEM). In 1988 he joined Philips Research Laboratories, where he first worked on signal processing for ultra-high-resolution TEM, in particular on phase retrieval methods used for digital correction of aberration artifacts. In 1996, after one year of work on MPEG-2 video coding, he started to work on signal processing for optical storage in the field of bit detection algorithms and channel modulation codes and techniques. Currently, his research activities are focused on coding and signal processing for next-generation (optical) storage channels.

Henk Hollmann was born in Utrecht, the Netherlands, on March 10, 1954. He received the master's degree in mathematics in 1982, with a thesis on association schemes, and the Ph.D. degree in 1996, with a thesis on modulation codes, both from Eindhoven University of Technology. In 1997, he was awarded the SNS bank prize for the best Ph.D. thesis in fundamental research of this university. In 1982 he joined CNET, Issy-les-Moulineaux, France, where he worked mainly on number-theoretic transforms. Since 1985 he has been with Philips Research Laboratories, Eindhoven, the Netherlands. His research interests include discrete mathematics and combinatorics, information theory, cryptography, and digital signal processing.

BIBLIOGRAPHY

1. P. A. Franaszek, Sequence-state coding for digital transmission, *Bell Syst. Tech. J.* **47**: 143–157 (1968).
2. D. T. Tang and L. R. Bahl, Block codes for a class of constrained noiseless channels, *Inform. Control* **17**: 436–461 (1970).
3. R. L. Adler, D. Coppersmith, and M. Hassner, Algorithms for sliding block codes. An application of symbolic dynamics to information theory, *IEEE Trans. Inform. Theory* **IT-29**(1): 5–22 (Jan. 1983).
4. K. A. S. Immink, *Codes for Mass Data Storage Systems*, Shannon Foundation Publishers, The Netherlands, 1999.
5. B. H. Marcus, R. M. Roth, and P. H. Siegel, Constrained systems and coding for recording channels, in V. S. Pless and W. C. Huffman, eds., *Handbook of Coding Theory II*, Elsevier, Amsterdam, 1998, 1635–1764.

6. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorders, *IEEE Trans. Inform. Theory* (Special Commemorative Issue) 2260–2299 (1998).
7. J. W. M. Bergmans, *Digital Baseband Transmission and Recording*, Kluwer, Amsterdam, 1996.
8. C. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (1948).
9. M. Fekete, Uber die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit Ganzzahligen Koeffizienten, *Math. Zeitschr.* **17**: 228–249 (1923).
10. D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press, Cambridge, MA, 1995.
11. H. Minc, *Nonnegative Matrices*, Wiley, New York, 1988.
12. E. Seneta, *Non-negative Matrices and Markov Chains*, (2nd ed.,) Springer, New York, 1981.
13. K. A. S. Immink, *Coding Techniques for Digital Recorders*, Prentice-Hall, Englewood Cliffs, NJ, 1991.
14. R. Karabed and B. H. Marcus, Sliding-block coding for input-restricted channels, *IEEE Trans. Inform. Theory* **IT-34**(1): 2–26 (1988).
15. J. E. Hopcroft and J. D. Ullmann, *Formal Languages and Their Relation to Automata*, Addison-Wesley, Reading, MA, 1969.
16. J. E. Hopcroft and J. D. Ullmann, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA, 1979.
17. G. F. M. Beenker and K. A. S. Immink, A generalized method for encoding and decoding runlength-limited binary sequences, *IEEE Trans. Inform. Theory* **IT-29**(5): 751–754 (1983).
18. J. H. Weber and K. A. Abdel-Ghaffar, Cascading runlength-limited sequences, *IEEE Trans. Inform. Theory* **IT-39**(6): 1976–1984 (1993).
19. K. A. S. Immink, Constructions of almost block-decodable runlength-limited codes, *IEEE Trans. Inform. Theory* **IT-41**(1): 284–287 (Jan. 1995).
20. H. D. L. Hollmann, On an approximate eigenvector associated with a modulation code, *IEEE Trans. Inform. Theory* **IT-43**(5): 1672–1678 (1997).
21. P. A. Franaszek, A general method for channel coding, *IBM J. Res. Dev.* **24**: 638–641 (1980).
22. P. A. Franaszek, On future-dependent block coding for input-restricted channels, *IBM J. Res. Dev.* **23**: 75–81 (1979).
23. P. A. Franaszek, Synchronous bounded delay coding for input restricted channels, *IBM J. Res. Dev.* **24**: 43–48 (1980).
24. P. A. Franaszek, Construction of bounded delay codes for discrete noiseless channels, *IBM J. Res. Dev.* **26**: 506–514 (1982).
25. T. M. Cover, Enumerative source coding, *IEEE Trans. Inform. Theory* **IT-19**: 73–77 (1973).
26. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
27. J. S. Eggenberger and P. Hodges, Sequential encoding and decoding of variable word length, fixed rate data codes, U.S. Patent 4,115,768, (1978).
28. G. Jacoby and R. Kost, Binary two-thirds rate code with full word look-ahead, *IEEE Trans. Magn.* **MAG-20**(5): 709–714 (1984).
29. A. Lempel and M. Cohn, Lookahead coding for input-restricted channels, *IEEE Trans. Inform. Theory* **IT-28**: 933–937 (1982).
30. P. A. Franaszek, Coding for constrained channels: A comparison of two approaches, *IBM J. Res. Dev.* **33**: 602–608 (1989).
31. J. Ashley and B. H. Marcus, Canonical encoders for sliding block decoders, *Siam J. Disc. Math.* **8**(4): 555–605 (1995).
32. H. D. L. Hollmann, On the construction of bounded-delay encodable codes for constrained systems, *IEEE Trans. Inform. Theory* **IT-41**(5): 1354–1378 (1995).
33. H. D. L. Hollmann, Bounded-delay encodable, block-decodable codes for constrained systems, *IEEE Trans. Inform. Theory* Special Issue on Codes and Complexity **IT-42**(6): 1957–1970 (1996).
34. H. D. L. Hollmann, *Modulation Codes*, doctoral thesis, Eindhoven Univ. Technology, Eindhoven, The Netherlands, 1996.
35. K. A. S. Immink, Block-decodable runlength-limited codes via look-ahead technique, *Philips J. Res.* **46**(6): 293–310 (1992).
36. H. D. L. Hollmann, Bounded-delay encodable codes for constrained systems from state combination and state splitting, *Proc. 14th Benelux Symp. Information Theory*, Veldhoven, 1993, pp. 80–87.
37. B. H. Marcus, P. H. Siegel, and J. K. Wolf, Finite-state modulation codes for data storage, *IEEE J. Select. Areas Commun.* **10**(1): 5–37 (1992).
38. I. J. Fair, W. D. Gover, W. A. Krzymien, and R. I. Macdonald, Guided scrambling: A new line coding technique for high bit rate fiber optic transmission systems, *IEEE Trans. Commun.* **COM-39**(2): 289–297 (1991).
39. L. Pátrovics and K. A. S. Immink, Encoding of *dklr*-sequences using one weight set, *IEEE Trans. Inform. Theory* **IT-42**(5): 1553–1554 (1996).
40. K. A. S. Immink, A practical method for approaching the channel capacity of constrained channels, *IEEE Trans. Inform. Theory* **IT-43**(5): 1389–1399 (1997).
41. K. A. S. Immink and H. Ogawa, Method for encoding binary data, U.S. Patent 4,501,000 (1985).
42. K. A. S. Immink, EFMPlus: The coding format of the multimedia compact disc, *IEEE Trans. Consum. Electron.* **41**(3): 491–497 (1995).
43. K. A. S. Immink, A survey of codes for optical disk recording, *IEEE J. Select. Areas Commun.* **19**: 756–764 (2001).
44. T. Narahara et al., Optical disc system for digital video recording, *Jpn. J. Appl. Phys.* **39**(2B)(Part 1): 912–919 (2000).
45. A. F. Stikvoort and J. A. C. van Rens, An all-digital bit detector for compact disc players, *IEEE J. Select. Areas Commun.* **10**(1): 191–200 (1992).
46. J. A. H. Kahlman and K. A. S. Immink, Device for encoding/decoding N-bit source words into corresponding M-bit channel words, and vice versa, U.S. Patent 5,477,222 (1995).
47. W. Coene, Combi-codes for DC-free runlength-limited coding, *IEEE Trans. Consum. Electron.* **46**(4): 1082–1087 (2000).

48. J. J. Ashley and B. H. Marcus, Time-varying encoders for constrained systems: an approach to limiting error propagation, *IEEE Trans. Inform. Theory* **IT-46**: 1038–1043 (2000).

49. H. M. Hilden, D. G. Howe, and E. J. Weldon, Shift error correcting modulation codes, *IEEE Trans. Magn.* **27**: 4600–4605 (1991).

50. Y. Saitoh, I. Ibe, and H. Imai, Peak-shift and bit error-correction with channel side information in runlength-limited sequences, *Proc. 10th Int. Symp. Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, (AAECC), 1993, Vol. 10, pp. 304–315.

51. A. V. Kuznetsov and A. J. H. Vinck, A coding scheme for single peak-shift correction in (d, k)-constrained channels, *IEEE Trans. Inform. Theory* **IT-39**: 1444–1450 (1993).

52. P. Perry, M.-C. Lin, and Z. Zhang, Runlength-limited codes for single error-detection with mixed type errors, *IEEE Trans. Inform. Theory* **IT-44**: 1588–1592 (1998).

53. S. Gopalaswamy and J. Bergmans, Modified target and concatenated coding for $d = 1$ constrained magnetic recording channels, *Proc. IEEE Int. Conf. Communications*, New Orleans, LA, 2000, pp. 89–93.

54. W. M. J. Coene, H. P. Pozidis, and J. W. M. Bergmans, Runlength limited parity-check coding for transition-shift errors in optical recording, *Proc. Global Telecommunications Conf. 2001*, GLOBECOM'01; *IEEE* **5**: 2982–2986 (2001).

55. J. Moon and B. Brickner, Maximum transition run codes for data storage systems, *IEEE Trans. Magn.* **32**(5): 3992–3994 (1996).

56. T. Nishiya et al., Turbo-EEPRML: An EEPRML channel with an error correcting post-processor designed for 16/17 rate quasi MTR code, *Proc. Globecom'98*, Sydney, 1998, pp. 2706–2711.

57. H. Kobayashi and D. T. Tang, Application of partial-response channel coding to magnetic recording systems, *IBM J. Res. Dev.* **14**: 368–375 (1970).

58. B. Marcus and P. Siegel, *Constrained Codes for PRML*, IBM Report RJ 4371, 1984.

59. P. H. Siegel and J. K. Wolf, Modulation and coding for information storage, *IEEE Commun. Mag.* **29**(12): 68–86 (1991).

60. J. Ashley et al., Holographic data storage, *IBM J. Res. Dev.* **44**(3): 341–368 (2000).

61. W. Weeks and R. E. Blahut, The capacity and coding gain of certain checkerboard codes, *IEEE Trans. Inform. Theory* **IT-44**(3): 1193–1203 (1998).

62. J. J. Ashley and B. M. Marcus, Two-dimensional low-pass filtering codes, *IEEE Trans. Commun.* **46**(6): 724–727 (1998).

63. M. W. Marcellin and H. J. Weber, Two-dimensional modulation codes, *IEEE J. Select. Areas Commun.* **10**(1): 254–266 (1992).

64. A. Kato and K. Zeger, On the capacity of two-dimensional runlength constrained channels, *IEEE Trans. Inform. Theory* **IT-45**(5): 1527–1540 (1999).

65. R. M. Roth, P. H. Siegel, and J. K. Wolf, Efficient coding schemes for the hard-square model, *IEEE Trans. Inform. Theory* **IT-47**(3): 1166–1176 (2001).

66. S. Taira et al., *Study of Recording Methods for Advanced Optical Disks*, Technical Report of IEICE, MR2001-117, 2002, pp. 57–64.

CONTINUOUS-PHASE-CODED MODULATION

CARL-ERIK W. SUNDBERG
iBiquity Digital Corp.
Warren, New Jersey

JOHN B. ANDERSON
Lund University
Lund, Sweden

1. INTRODUCTION

Transmission cost is often proportional to bandwidth and the radiofrequency spectrum is a limited resource. The motivation for searching for spectrally efficient modulation is thus clear. The available transmitter power is also limited for many applications. For satellite and land-mobile radio applications, modulation with a constant RF (radiofrequency) envelope is advantageous because transmitters use more efficient nonlinear amplifiers. The combination of these factors dictates a coded modulation system based on constant-amplitude sinusoids. In this article we outline the power and spectrum properties of a large class of such signals. This class, continuous-phase modulation (CPM), can be viewed as a generalization of minimum shift keying (MSK) containing such schemes as tamed frequency modulation (TFM) and Gaussian MSK (GMSK). We review the performance of the CPM class with optimum reception and show how the choice of system parameters affects its error performance and the signal bandwidth. We also show how these two trade off against each other. Transmitters and simplified receivers are also discussed.

CPM coding was the first coded modulation class to be extensively studied and marked the advent of a combined energy and bandwidth view in practical coding schemes. It had its origins in the invention of MSK [1] and a related scheme called *fast frequency shift keying* [2]. It gained impetus by several papers [3–5] that introduced continuous-phase frequency shift keying (CPFSK) and its optimal detection in the early 1970s. With Miyakawa et al. [6] and Anderson and Taylor [7] in the mid-1970s, more sophisticated codinglike ideas were introduced; the second paper introduced the modern concepts of distance calculation, trellis decoding, and simultaneous consideration of code energy and bandwidth. CPM reached its full flower as a coded modulation method in 1981 with the basic papers of Aulin, Sundberg and others [8–11].

MSK has attracted new study since the 1970s as well. We will outline methods to improve on MSK while maintaining a constant amplitude. By *improvement* is meant a narrower power spectrum, lower spectral sidelobes, cheaper implementation, better error probability, or all the above. The main part of the article considers a number of methods for constructing constant-amplitude signals that significantly outperform MSK in either energy, bandwidth, or both. An important issue is at what level of complexity these improvements are obtained.

2. THE CPM SIGNAL CLASS

A large class of constant-amplitude modulation schemes is defined by the signal

$$s(t) = \left[\frac{2E}{T}\right]^{1/2} \cos(2\pi f_0 t + \phi(t, \boldsymbol{a})) \tag{1}$$

where the transmitted information is contained in the phase

$$\phi(t, \boldsymbol{a}) = 2\pi h \sum_{i=-\infty}^{\infty} a_i q(t - iT) \tag{2}$$

with $q(t) = \int_{-\infty}^{t} g(\tau)\, d\tau$. Normally the function $g(t)$ is a smooth pulseshape over a finite time interval $0 \le t \le LT$ and zero outside. Thus the parameter L is the length of the pulse (per unit T) and T is the symbol time; E is the energy per symbol, f_0 is the carrier frequency and h is the modulation index. The M-ary data symbols a_i take values $\pm 1, \pm 3 \cdots \pm(M-1)$. M is normally selected to be a power of 2, and we will mainly consider binary, quaternary, and octal systems ($M = 2,4,8$). From the definition above we note that the pulse $g(t)$ defines an instantaneous frequency and its integral $q(t)$ is the phase response. The precise shape of $g(t)$ determines the smoothness of the transmitted information carrying phase. The rate of change of the phase (or instantaneous frequency) is proportional to the parameter h, which is the *modulation index*. The pulse $g(t)$ is normalized in such a way that $\int_{-\infty}^{\infty} g(t)\, dt$ is $\frac{1}{2}$. This means that for schemes with positive pulses of finite length, the maximum phase change over any symbol interval is $(M-1)h\pi$.

By choosing different pulses $g(t)$ and varying the parameters h and M, a great variety of CPM schemes can be obtained. Some of the more popular pulseshapes are listed in Table 1. These include CPFSK, tamed frequency modulation (TFM) [12], generalized TFM (GTFM) [13], Gaussian MSK (GMSK) [14], duobinary FSK (2REC) [15], raised cosine (LRC), and spectrally raised cosine (LSRC). In the table we use the notation LXX for a pulse of length L symbol intervals; thus 3RC is a raised cosine pulse of length $3T$. For spectral raised cosine (LSRC), the main time lobe has width LT. The pulse of length $1T$, namely, 1REC, is another name for CPFSK. The 2REC pulse is also called duobinary.

MSK is obtained as a special case of the signals defined in Eq. (1) by selecting the pulse 1REC ($L = 1$) from Table 1 and using binary ($M = 2$) data with $h = \frac{1}{2}$. The CPM signal can be viewed as phase modulation or as frequency modulation, but for understanding the optimum coherent receiver, it is advantageous to view the signal as phase modulation.

Memory is introduced into the CPM signal by means of its continuous phase. Each information-carrying phase function $\phi(t, \boldsymbol{a})$ is continuous at all times for all combinations of data symbols. Further memory can be built into the CPM signal by choosing a $g(t)$ pulse with $L > 1$. These schemes have overlapping pulse shaping and called *partial-response* techniques. CPM signals with $L \le 1$ are *full-response* schemes. In this case the memory is in the continuous phase only.

Although the CPM signals in Eq. (1) are in principle conceivable for any value of the modulation index h, a key to the design of practical maximum-likelihood detectors is to consider CPM schemes with rational values of h. For $h = 2k/p$ where k and p have no common factors, the phase

Table 1. Definition of the Frequency Pulse Function $g(t)$

LRC	$g(t) = \begin{cases} \dfrac{1}{2LT}\left[1 - \cos\left(\dfrac{2\pi t}{LT}\right)\right]; & 0 \le t \le LT \\ 0 \quad ; & \text{otherwise} \end{cases}$
	L is the pulse length, e.g., 3RC has $L = 3$
TFM[a]	$g(t) = \dfrac{1}{8}[Ag_0(t-T) + Bg_0(t) + Ag_0(t+T)]; \quad A = 1; B = 2$
	$g_0(t) \approx \sin\left(\dfrac{\pi t}{T}\right)\left[\dfrac{1}{\pi t} - \dfrac{2 - \dfrac{2\pi t}{T}\cot\left(\dfrac{\pi t}{T}\right) - \dfrac{\pi^2 t^2}{T^2}}{\dfrac{24\pi t^3}{T^2}}\right]$
LSRC	$g(t) = \dfrac{1}{LT} \cdot \dfrac{\sin\left(\dfrac{2\pi t}{LT}\right)}{\dfrac{2\pi t}{LT}} \cdot \dfrac{\cos\left(\beta \cdot \dfrac{2\pi t}{LT}\right)}{1 - \left(\dfrac{4\beta}{LT} \cdot t\right)^2}; \quad 0 \le \beta \le 1$
GMSK	$g(t) = \dfrac{1}{2T}\left[Q\left(2\pi B_b \dfrac{t - \dfrac{T}{2}}{\sqrt{\ell n 2}}\right) - Q\left(2\pi B_b \dfrac{t + \dfrac{T}{2}}{\sqrt{\ell n 2}}\right)\right]; \quad 0 \le B_b T < \infty$
	$Q(t) = \int_t^{\infty} \dfrac{1}{\sqrt{2\pi}} e^{-\tau^2/2} d\tau$
LREC	$g(t) = \begin{cases} \dfrac{1}{2LT}; & 0 \le t \le LT \\ 0 \quad ; & \text{otherwise} \end{cases}$
	$L = 1$ yields CPFSK

[a] The class of GTFM pulses is obtained by varying A, B, and $g_0(t)$.

$\phi(t, \boldsymbol{a})$ during interval $nT \le t \le (n+1)T$ can be written

$$\phi(t, \boldsymbol{a}) = 2\pi h \sum_{i=n-L+1}^{n} a_i q(t - iT) + \theta_n = \theta(t, \boldsymbol{a}) + \theta_n \qquad (3)$$

where $\theta_n = \left[h\pi \sum_{i=-\infty}^{n-L} a_i \right]$ modulo 2π has only p different values. Thus the total number of states that (at most) is needed to describe the signal in Eq. (1) is $S = pM^{(L-1)}$, where a state is defined as the vector $(\theta_n, a_{n-1}, a_{n-2}, \ldots, a_{n-L+1})$. The state vector consists of the phase state θ_n and $M^{(L-1)}$ *correlative states* for partial response systems. For a full-response system the number of states is p. The finite state description for CPM signals allows the use of a finite Viterbi decoder.

As an example of a CPM scheme, we choose binary 3RC; thus the pulse $g(t)$ is a raised-cosine pulse of length 3 symbol intervals. Figure 1a shows the pulse $g(t)$ and phase response $q(t)$ for 3RC. For comparison, the corresponding functions are also shown for 1REC (CPFSK). The information-carrying phase function $\phi(t, \boldsymbol{a})$ is illustrated both for 1REC and 3RC in Fig. 1b for a particular data sequence. Note that all changes in the 3RC phase take longer time than in the CPFSK scheme. Figure 1c shows all phase functions starting at the arbitrary phase 0° and

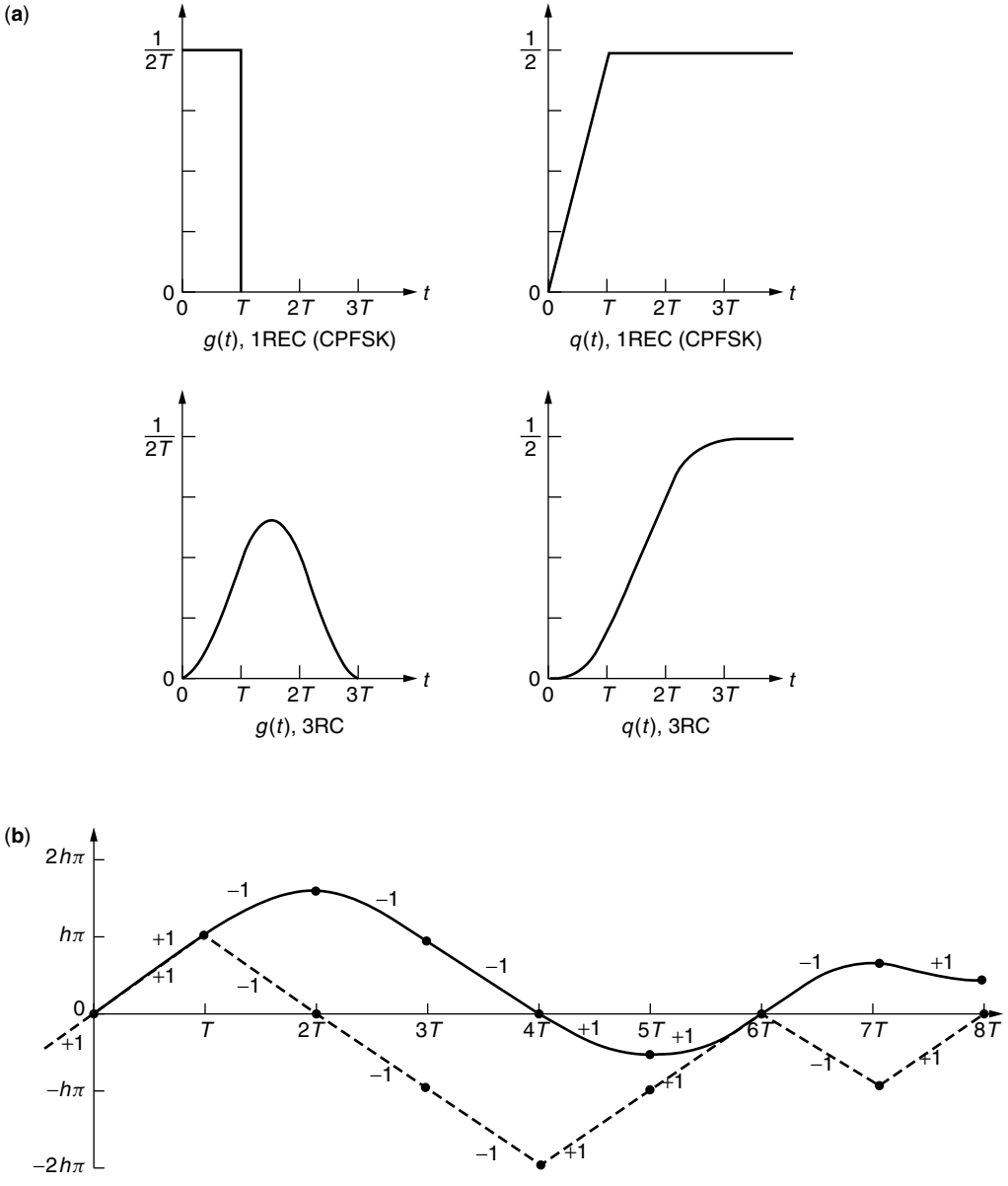


Figure 1. (**a**) Pulseshapes $g(t)$ and phase responses $q(t)$ for the full-response 1REC (CPFSK) and partial response 3RC CPM schemes; (**b**) examples of the phase function $\phi(t, \boldsymbol{a})$ for 1REC (dashed) and 3RC (solid) for the data sequence +1, −1, −1, −1, +1, +1, −1, +1; (**c**) phase tree for binary 3RC with $h = \frac{2}{3}$. The state description of the signal is also given. The transitions with arrows are also shown in Fig. 2a.

time 0, where the two previous data symbols are +1,+1. It is obvious from Fig. 1c that all phase changes are very smooth. The corresponding phase tree for MSK has linear phase changes with sharp corners when the data change. The phase tree in Fig. 1c also displays the state vector at each node.

The finite-state description of the CPM signal implies a trellis description, and this finiteness stems from the modulo-2π property of the phase. By plotting $I = \cos[\phi(t, \boldsymbol{\alpha})]$ and $Q = \sin[\phi(t, \boldsymbol{\alpha})]$ versus time in a three-dimensional plot, all signals appear on the surface of a cylinder. Such a phase cylinder is shown in Fig. 2a for the parameters used in Fig. 1c. This scheme has 12 states. Contrary to Fig. 1c, where restrictions were imposed for $t < 0$, the phase cylinder shows all signals over three symbol intervals. The phase nodes and some of the state vectors are also shown. To clarify the connection between the tree in Fig. 1c and the trellis in Fig. 2a, we have marked three identical transitions with arrows in both figures.

While the appearance of the phase tree does not change with h, the number of phase states does and so does the phase cylinder. Figure 2b shows the simplicity of the phase cylinder for $h = \frac{1}{2}$ and Fig. 2c shows how the complexity has grown for $h = \frac{3}{4}$. It is seen in Fig. 2b that with a proper phase offset, I and Q exhibit quite open-amplitude eye diagrams. We will see that this property leads to simple linear receivers and this is the reason for the popularity of the binary $h = \frac{1}{2}$ schemes.

An interesting generalization of the CPM class is to let the modulation index vary cyclically with time. This gives a so called multi-h scheme. These systems have better performance than the fixed h schemes. Typically most of the available improvement is obtained with two or three different h values.

3. CPM POWER SPECTRA

A large number of methods are available in the literature for calculating the power spectrum of CPM; for a detailed treatment, see, for instance, Ref. 16. Computer simulations can also be employed to estimate the power spectra. As a measure of signal bandwidth in what follows, we will generally take the frequency band around the carrier frequency containing 99% of the signal power. Figure 3 shows the power spectral density of some binary CPM schemes. The term *GMSK4* here means that the GMSK pulse in Table 1 is truncated symmetrically to 4 symbol intervals. The CPM schemes 3RC, GMSK4 with $B_bT = 0.25$, and 3SRC6 have comparable power spectra. The corresponding pulses $g(t)$ are also quite similar as are their detection properties — the rate of reduction of the spectral sidelobes is determined by the smoothness of the pulse $g(t)$. For most applications, the raised-cosine pulses probably have sufficient smoothness. Figure 3 also shows the MSK and 2REC (duobinary) schemes. Notice the lower spectral sidelobes of the smooth partial response schemes.

In comparison, Fig. 4 shows power spectra for some four-level schemes ($M = 4$). Figures 3 and 4 illustrate that a longer pulse $g(t)$ narrows the power spectra for fixed h and M. TFM has a power spectrum similar to binary 3.7 RC and 3.7 SRC. GMSK with $B_bT = 0.18$ corresponds approximately to 4RC and GMSK with $B_bT = 0.2$ to TFM.

The width of the main spectral lobe decreases with increasing L but increases with increasing h and M. The relationship between bandwidth or fractional out of band power and pulseshape $g(t)$ is rather complicated.

4. DETECTION AND ERROR PROBABILITY

Coherent maximum likelihood sequence detection (Viterbi detection) can be performed for all CPM schemes that

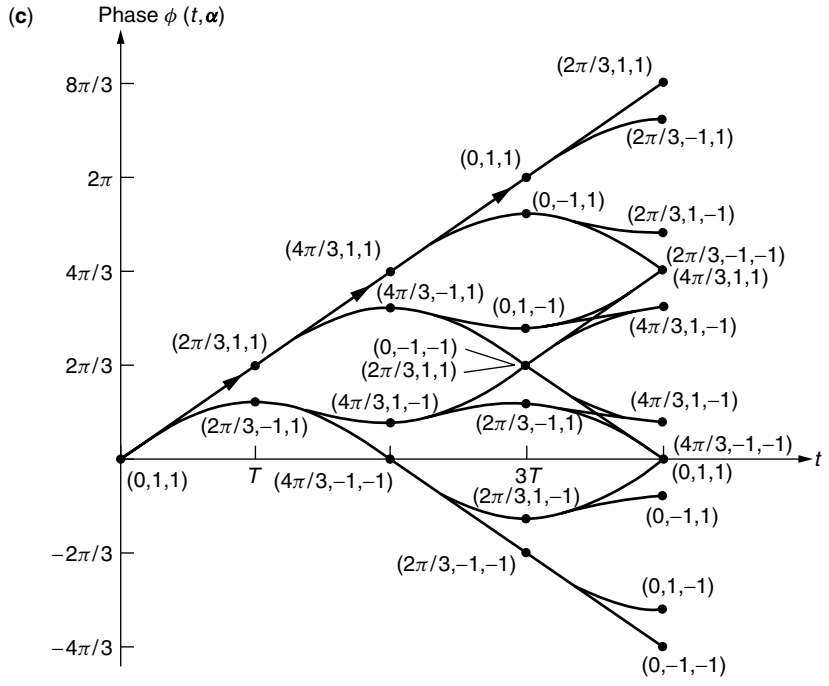


Figure 1. (*Continued*)

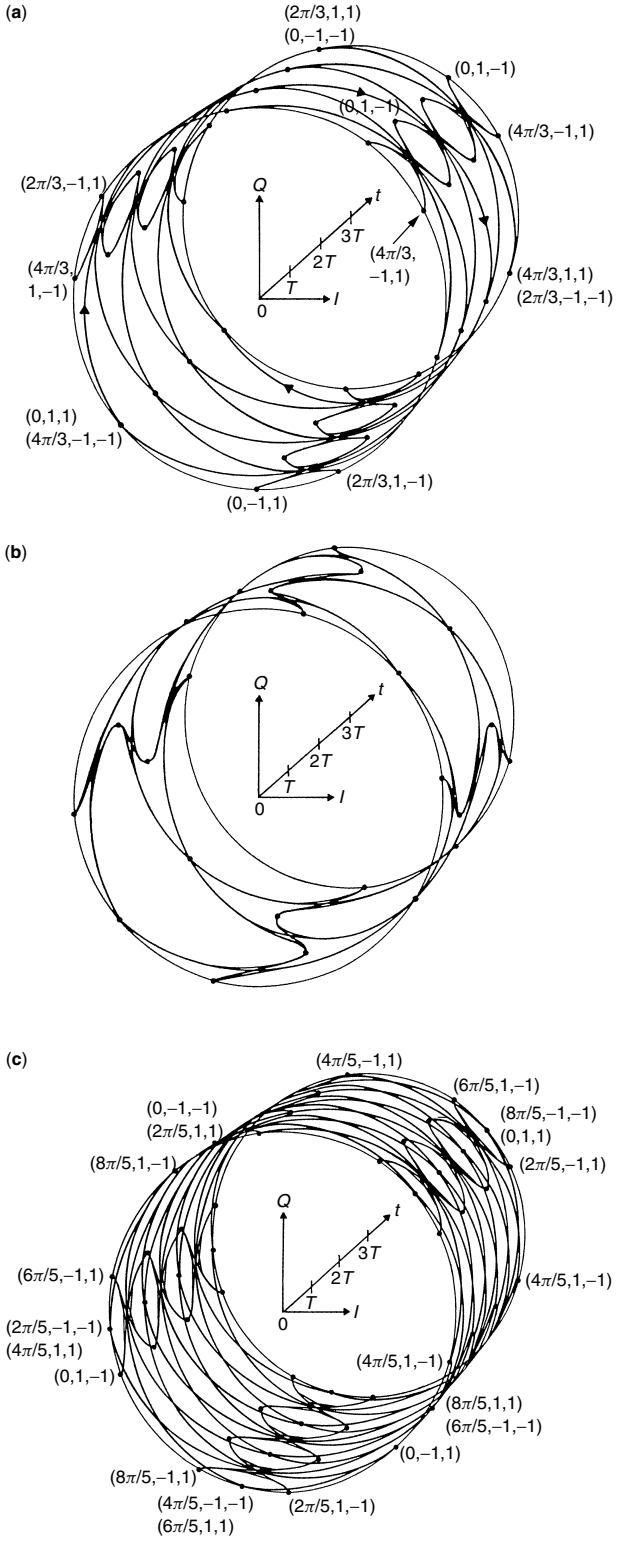


Figure 2. (**a**) Phase cylinder for 3RC with $h = \frac{2}{3}$ and $M = 2$. Compare the phase tree in Fig. 1c. Note the arbitrary phase offset between the phase in the tree in Fig. 1c and the cylinder, in Fig. 2a. Also note that the tree is plotted over T and the cylinder over $3T$. Notice the transitions with arrows; these are also shown in the tree in Fig. 1c. (**b**) phase cylinder for 3RC, $h = \frac{1}{2}$, $M = 2$. (**c**) phase cylinder for $M = 2$, 3RC with $h = \frac{4}{5}$ (10 states).

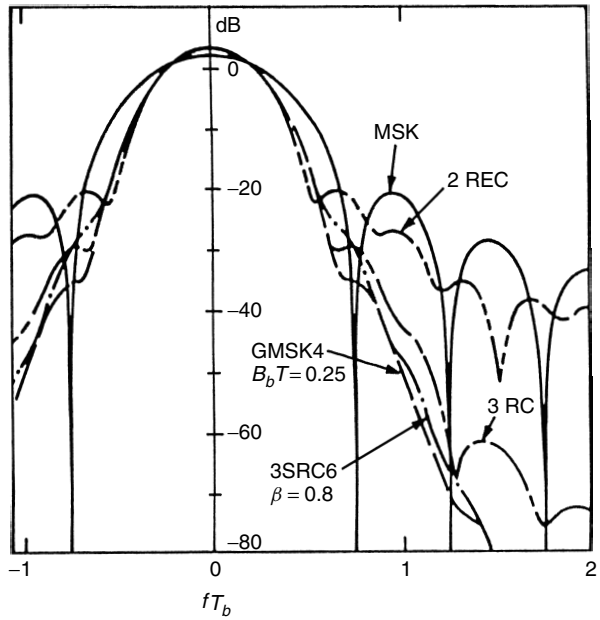


Figure 3. Average power spectrum for some CPM schemes with $h = \frac{1}{2}$ and $M = 2$. See Table 1 for details.

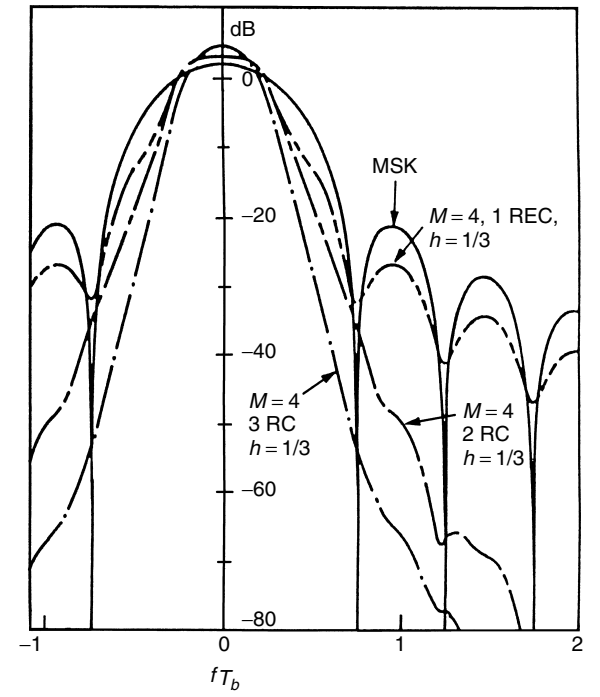


Figure 4. Average power spectra for MSK ($M = 2$, 1REC, $h = \frac{1}{2}$) and the $h = \frac{1}{3}$, $M = 4$ schemes with 1REC, 2RC, and 3RC pulses.

can be described by the finite state and trellis description given above. Although the structure of the optimum ideal coherent receiver for CPM is known [16], it is difficult to evaluate its bit error probability performance. Simulations are required for low channel signal-to-noise ratios. The most convenient and useful parameter for describing the error probability of CPM schemes with maximum-likelihood sequence detection is the minimum Euclidean

distance between all pairs of signals

$$D^2_{\min} = d^2_{\min} \cdot 2E_b = \min\left\{2E_b \log_2(M)\frac{1}{T} \times \int_0^{NT} [1 - \cos[\phi(t, \boldsymbol{a}) - \phi(t, \boldsymbol{b})]]dt\right\} \quad (4)$$

where E_b is the signal energy per bit given by $E_b \log_2 M = E$ and NT is length of the receiver observation interval. When N is sufficiently large, the largest obtainable distance, called the *free distance*, is reached.

For ideal coherent transmission over an additive white Gaussian noise (AWGN) channel, the bit error probability for high signal-to-noise ratios E_b/N_0 is approximately

$$P_b \approx Ce^{-d^2_{\min} E_b/N_0} \quad (5)$$

where C is a constant.

Efficient algorithms are available for computing the minimum distance for different $g(t)$, L, h, and M [16,17]. Figure 5 shows an upper bound d_B^2 to the free distance d_f^2 as a function of h for binary LRC codes; d_B is found by considering certain error events [16]. The term d_B^2 equals the free distance d_f^2 for almost all h values in Fig. 5. Note that the distance grows with L, at least for larger h. The vertical axis of Fig. 5 also has a decibel (dB) scale that gives the gain in E_b/N_0 relative to MSK (which has $d_f^2 = 2$).

The comparisons in Fig. 5 are somewhat artificial, since the bandwidth of a CPM scheme also changes with L and h. Another comparison of CPM schemes is given in the scatterplot in Fig. 6, where each point

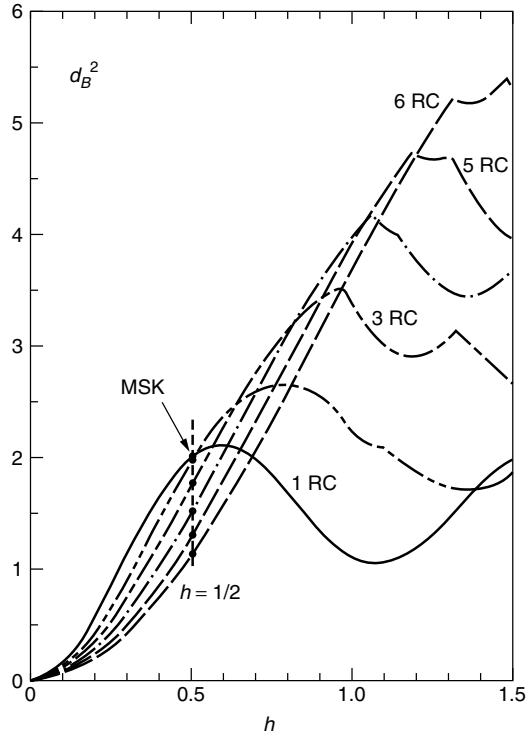


Figure 5. Upper bound d_B^2 on the distance for the binary CPM schemes 1 RC, 2 RC, ..., 6 RC.

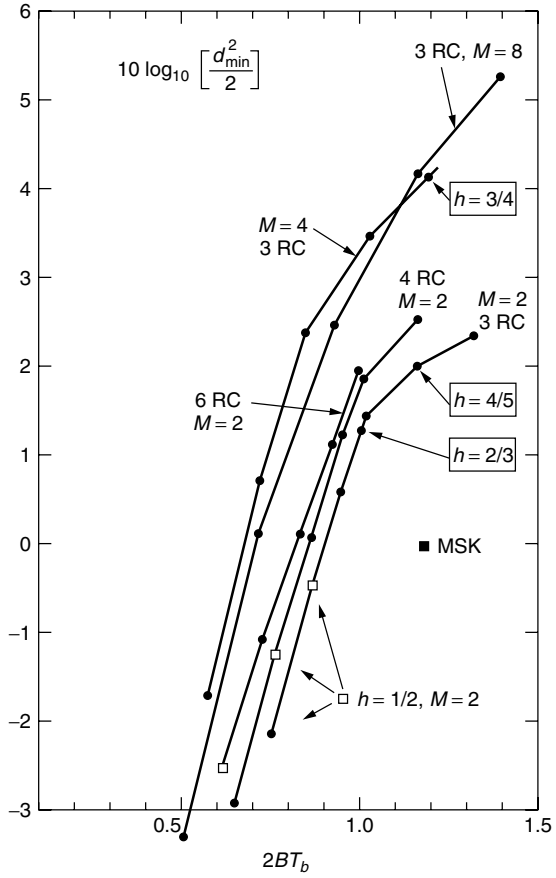


Figure 6. Power–bandwidth tradeoff for CPM schemes using raised-cosine pulses.

represents a system with its 99% power RF bandwidth $2BT_b$ (where $T_b = T/\log_2 M$) and the required signal-to-noise ratio (SNR) relative to MSK in dB at high E_b/N_0. Thus schemes on the same vertical line (for example, through the MSK point) have the same bandwidth at equal data rates; schemes on the same horizontal line have similar error probability for high SNRs. It is evident that larger L and larger M yield more energy- and bandwidth-efficient systems. Not surprisingly, the system complexity increases in the same direction. We have marked some binary $h = \frac{1}{2}$ schemes, some binary 3RC schemes and some quaternary 3RC schemes in Fig. 6. The $h = \frac{2}{3}$ and $\frac{4}{5}$ binary 3RC schemes correspond to the phase cylinders shown above in Fig. 2a,c. For these two schemes, the number of states is 12 and 20, respectively.

The bandwidth and energy of CPM signaling may also be compared to that of other coded modulations. As a class, CPM lies in a middle range of bandwidth and energy; it achieves a moderate error rate ($\sim 10^{-5}$) in 0.5–1.5 Hz·s/data bit and 4–12 dB for E_b/N_0, depending on the code parameters and complexity. Ordinary parity-check coding together with a simple modulation like QPSK work at wider bandwidth and lower energy, and this coding partially overlaps the CPM range on the wideband/low-energy side. In the area of overlap, CPM has the advantage of constant RF envelope. Trellis-coded modulation (TCM) is another coded modulation method

based on set partitioning. It works in a region of narrower bandwidth and higher energy than CPM and partially overlaps it on the other side. In the overlap region, TCM needs either half the bandwidth or 5 dB less energy than CPM. TCM and CPM, however, are not directly comparable because TCM demands a linear channel with accurate amplitude response and CPM does not. Transmitter amplifiers with sufficient linearity for TCM are 2–4 dB less efficient than nonlinear amplifiers of the sort that can be used with CPM signals [18].

5. GENERATING CPM SIGNALS

A conceptual general transmitter structure is shown in Fig. 7. This is based on Eq. (1). This structure demonstrates the fact that CPM is "digital FM"; that is, it is ordinary linear pulsetrain modulation, with pulseshaper $g(t)$, but applied to an FM modulator instead of an amplitude modulator. However, the structure is not easily converted into hardware for coherent systems. The reason is that an exact relation between the symbol rate and the modulation index is required and this requires control circuitry.

The most general and straightforward way of implementing a robust CPM transmitter is to use stored lookup tables. This is seen by rewriting the normalized CPM waveform $S_0(t, \boldsymbol{a}_n) = S(t, \boldsymbol{a}_n)/[2E/T]^{1/2}$ as

$$S_0(t, \boldsymbol{a}_n) = I(t)\cos(2\pi f_0 t) - Q(t)\sin(2\pi f_0 t) \qquad (6)$$

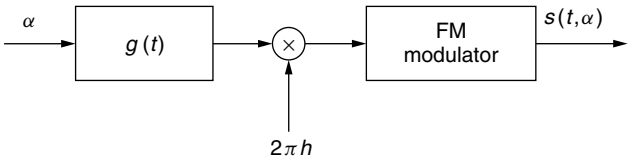


Figure 7. Conceptual modulator for CPM based on Eq. (1).

where $I(t) = \cos(\theta(t, \boldsymbol{a}_n) + \theta_n)$ and $Q(t) = \sin(\theta(t, \boldsymbol{a}_n) + \theta_n)$. The subscript n on $\boldsymbol{a}_n$ indicates that we are considering the data symbol a_n and sufficiently many of the previous symbols. Figure 8 shows a transmitter based on Eq. (6) where the two read only memories contain sampled and quantized versions of $I(t)$ and $Q(t)$ for each data symbol a_n, correlative state vector (the $L-1$ previous data symbols) and phase-state value. The address field for the ROM is roughly $L\log_2 M + [\log_2 p] + 1$ bits and the ROM size is $p \cdot M^L \cdot m \cdot m_q$ bits, where m is the number of samples per symbol time and m_q is the number of bits per quantized sample. The transmitter in Fig. 8 also contains a small sequential machine with a phase state lookup table for calculating the next phase state, given the previous one and the incoming data symbol. The transmitter also contains two D/A converters.

For binary 3RC with $h = \frac{2}{3}$, for example, the ROM (read-only memory) address length is 5-bits and the ROM size is 1024 bits. For a wide range of CPM parameters, the ROM size is manageable. Alternative transmitter structures are possible that have reduced lookup tables.

Several special modulator structures have been devised for MSK. Because MSK is a quadrature-multiplexed modulation scheme, it can be optimally detected by coherently demodulating its in-phase and quadrature components in parallel. The quadrature channels of the modulator and demodulator must be time synchronized, amplitude balanced, and in phase quadrature. The serial method [19] is an alternative approach to parallel modulation and demodulation of MSK which avoids some of these problems.

6. RECEIVERS

Receivers for coherent CPM are an active area of research. For a general CPM scheme with a rational modulation index h and a pulse of finite length L, ideal

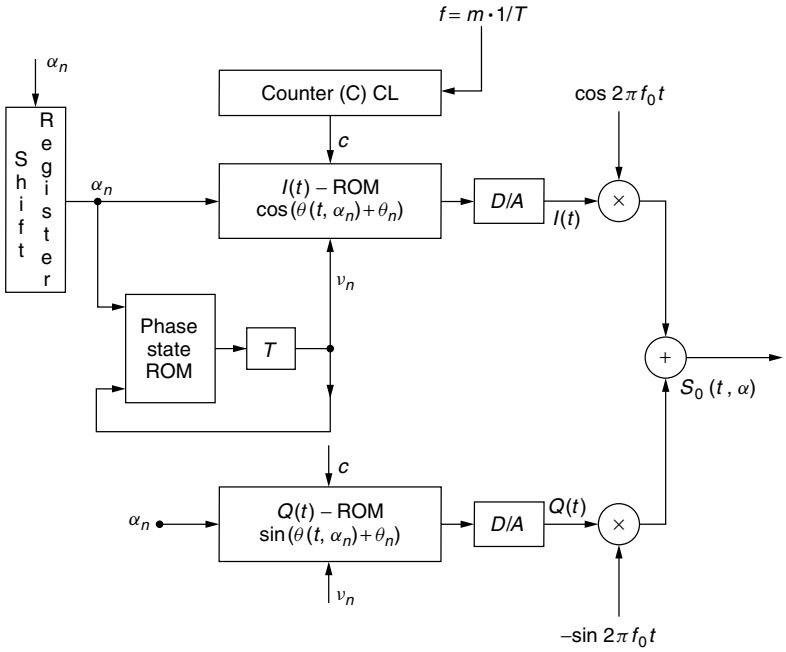


Figure 8. General CPM transmitter with the lookup table principle.

optimum coherent detection can be performed by means of the Viterbi algorithm. The state and trellis description discussed earlier is used. The metric is calculated in a bank of linear filters that are sampled every symbol interval. Figure 9 shows the conceptual diagram for this general receiver. The path memory in the trellis processor causes a delay of N_T symbol intervals. The N_T needed is related to the growth of the minimum distance with the observation interval length. It should be sufficiently large that the free distance is obtained between all paths.

Although there are no special theoretical problems in constructing a receiver based on the principles illustrated in Fig. 9 there are several practical ones. As with all Viterbi detectors, the complexity grows exponentially with signal memory. The limiting factors are the number of states $S = pM^{L-1}$ and the number of filters $F = 2M^L$ for calculating the metrics. For many cases with long smoothing pulses, the optimum receiver can be approximated by a receiver based on a shorter and simpler pulseshape $g_R(t)$ of length $L_R < L$. Thus the complexity of the suboptimum receiver is reduced by a factor of M^{L-L_R} for both the number of states and the number of filters in the filter bank. The simpler pulse shape can be optimized (for large SNRs) for a given transmitter pulseshape and modulation index. The loss in error performance can be very small.

For some cases of CPM it is not necessary to use the Viterbi detector. Much work has been devoted to the so-called MSK-type receiver, which is based on the structure given in Fig. 10. The circuit is inspired by the parallel MSK receiver; it has only two filters and just a small amount of processing. The receiver makes single

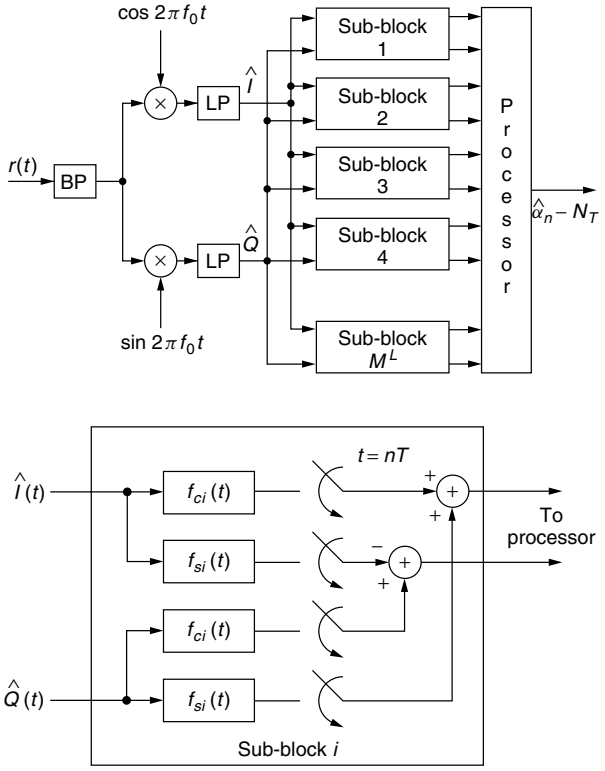


Figure 9. A general receiver structure for CPM based on the Viterbi algorithm. There are $4M^L$ linear filters.

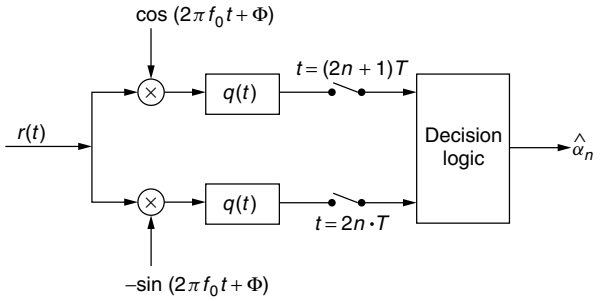


Figure 10. Receiver structure for a parallel MSK-type receiver for binary $h = \frac{1}{2}$ CPM.

symbol decisions. This simplified receiver is suboptimum but it works well for binary modulations with modulation index $h = \frac{1}{2}$. The decisions are made every $2T$ in alternate quadrature arms. Various ideas for selecting the receiver filters are analyzed in the early papers on this subject and an optimum filter has been derived for various correlative FSK schemes (i.e., ones with piecewise linear phase functions) [20] and for smooth pulses [21]. The performance for this type of receiver is almost equal to the optimum Viterbi receiver for schemes with a moderate degree of smoothing, that is, overlapping frequency pulses of length L up to three to four symbol intervals, like 3RC, 4RC, TFM, and some GMSK schemes.

A large literature exists on linear receiver simplifications for CPM. Further details are given in Ref. 16 and the survey article [22].

7. ADVANCED TOPICS

Several topics on more advanced levels are suggested:

Combinations of Convolutional Codes and CPM. The CPM-coded modulations just described may be preceded by an ordinary convolutional code with a rate such as $\frac{1}{2}$ or $\frac{2}{3}$. Pioneering work appeared in [23]. The result is properly considered a concatenated code. Modern techniques of iterative decoding and soft information transfer between the two decoders may be applied, but the joint code trellis is simple enough that outright maximum-likelihood decoding is practical. Convolutional plus CPM code constructions have a wider bandwidth than CPM coding alone, but they have very good minimum distance.

Partially Coherent Detection. A problem of practical interest is how to detect CPM signals when the phase reference is not completely known. Only the derivative of phase (i.e., frequency) may be known, or the phase reference may be stable but subject to an unknown constant offset. In the latter case the minimum distance of CPM is often unaffected.

Reduced-Search Receivers. CPM codes are trellis codes, and as such they can be detected with trellis search algorithms that view only a reduced part of the code trellis. Reduced searching for CPM in fact works well, better than it does for ordinary convolutional codes. Schemes such as the M algorithm need only extend two to four paths

through a complicated trellis. Indicative results are presented in Ref. 24.

Shannon Theory of CPM. An information-theoretic channel that models CPM is unfortunately one with memory, and the problem of computing Shannon information rates for CPM channels is difficult. However, methods exist to compute the cutoff rate, an underbound to capacity that is generally tight at moderate to high signal energy. Cutoff rate studies have been performed by a number of authors, beginning with Ref. 11.

Filtered CPM. An extensive literature exists on CPM that has undergone channel filtering. Early research is discussed in Ref. 16. In general, it can be said that removal of spectral sidelobes by filtering has little effect on CPM detection. More narrow filtering can have a severe effect unless special receivers are used, in which case detection losses are much reduced and can sometimes be removed completely [25].

8. FURTHER READING

A monograph devoted to CPM is Ref. 16. General textbooks that include an introductory chapter about CPM include, for example, Ziemer and Peterson [26] and Proakis [27]. Useful tutorial articles in the literature include Refs. 18 and 22.

BIOGRAPHY

Carl-Erik W. Sundberg received the M.S.E.E. and the Dr. Techn. degrees from the University of Lund, Lund, Sweden in 1966 and 1975 respectively. From 1977 to 1984, he was a Research Professor (Docent) in Telecommunication Theory, University of Lund. From 1984 to 2000, he was a Distinguished Member of Technical Staff (DMTS) at Bell Laboratories, Murray Hill, New Jersey, and during 2001 he was a DMTS at Agere Systems, Murray Hill. He currently is a Senior Scientist at iBiquity Digital Corp., Warren, New Jersey. His research interests include source and channel coding, digital modulation, fault-tolerant systems, digital mobile radio, digital audio broadcasting, spread-spectrum, digital satellite systems, and optical communications. He has published more than 95 journal papers and contributed over 140 conference papers. He has 67 patents, both granted and pending. He is a coauthor of *Digital Phase Modulation*, (New York: Plenum, 1986), *Topics in Coding Theory*, (New York: Springer-Verlag, 1989) and *Source-Matched Digital Communications* (New York: IEEE Press, 1996). In 1986 he received the IEEE Vehicular Technology Society's Paper of the Year Award, and in 1989 he was awarded the Marconi Premium Proc. IEE Best Paper Award. Two of his papers were selected for inclusion in the IEEE Communications Society 50th Anniversary Journal Collection, Volume 2002. He is a Fellow of the IEEE and is listed in *Marquis Who's Who in America*.

John B. Anderson was born in New York State in 1945 and received the Ph.D. degree in electrical engineering from Cornell University in 1972. He has been a faculty member at McMaster University in Canada, Rensselaer Polytechnic Institute in Troy, New york, and since 1998 has held the Ericsson Chair in Digital Communication at Lund University, Lund, Sweden. His researcg work is in coding and digital commuication, bandwidth-efficient coding, and practical application of these. Dr. Anderson has served as president of the IEEE Information Theory Society and editor-in-chief of IEEE Press. He is the author of five textbooks, including the forthcorning *Coded Modulation Systems* (Plenum, 2002). He is fellow of the IEEE (1987) and received the Humboldt Research Prize in 1991 and the IEEE Third Millenmium Medalj in 2000.

BIBLIOGRAPHY

1. U.S. Patent 2,917,417 (March 28, 1961), M. L. Doelz and E. H. Heald, Minimum shift data communication system.
2. R. de Buda, Coherent demodulation of frequency-shift keying with low deviation ratio, *IEEE Trans. Commun.* **COM-20**(3): 429–436 (June 1972).
3. M. G. Pelchat, R. C. Davis, and M. B. Luntz., Coherent demodulation of continuous phase binary FSK signals. *Proc. Int. Telemetering Conf.*, Washington, DC, Nov. 1971, pp. 181–190.
4. W. P. Osborne and M. B. Luntz, Coherent and noncoherent detection of CPFSK, *IEEE Trans. Commun.* **COM-22**(8): 1023–1036 (Aug. 1974).
5. T. A. Schonhoff, Symbol error probabilities for M-ary CPFSK: Coherent and noncoherent detection, *IEEE Trans. Commun.* **COM-24**(6): 644–652 (June 1976).
6. H. Miyakawa, H. Harashima, and Y. Tanaka, A new digital modulation scheme — multimode binary CPFSK, *Proc. 3rd Int. Conf. Digital Satellite Communications*, Kyoto, Japan, Nov. 1975, pp. 105–112.
7. J. B. Anderson and D. P. Taylor, A bandwidth-efficient class of signal space codes, *IEEE Trans. Inform. Theory* **IT-24**(6): 703–712 (Nov. 1978).
8. T. Aulin, *Three Papers on Continuous Phase Modulation* (*CPM*), Ph.D. thesis (on telecommunication theory), Univ. Lund, Lund, Sweden, Nov. 1979.
9. T. Aulin and C.-E. Sundberg, Continuous phase modulation — Part I: Full response signaling, *IEEE Trans. Commun.* **COM-29**(3): 196–209 (March 1981).
10. T. Aulin, N. Rydbeck, and C.-E. Sundberg, Continuous phase modulation — Part II: Partial response signaling, *IEEE Trans. Commun.* **COM-29**(3): 210–225 (March 1981).
11. J. B. Anderson, C.-E. Sundberg, T. Aulin, and N. Rydbeck, Power-bandwidth performance of smoothed phase modulation codes, *IEEE Trans. Commun.* **COM-39**(3): 187–195 (March 1981).
12. F. de Jager and C. B. Dekker, Tamed frequency modulation, a novel method to achieve spectrum economy in digital transmission, *IEEE Trans. Commun.* **COM-26**(5): 534–542 (May 1978).
13. K. S. Chung, General tamed frequency modulation and its application for mobile radio communication, *IEEE J. Select. Areas Commun.* **SAC-2**(4): 487–497 (July 1984).
14. K. Murota and K. Hirade, GMSK modulation for digital mobile telephony, *IEEE Trans. Commun.* **COM-29**(7): 1044–1050 (July 1981).
15. G. S. Deshpande and P. H. Wittke, Correlative encoded digital FM, *IEEE Trans. Commun.* **COM-29**(2): 156–162 (Feb. 1981).

16. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.

17. S. G. Wilson and M. G. Mulligan, An improved algorithm for evaluating trellis phase codes, *IEEE Trans. Inform. Theory* **IT-30**(6): 846–851 (Nov. 1984).

18. J. B. Anderson and C.-E. Sundberg, Advances in constant envelope coded modulation, *IEEE Commun. Mag.* **30**(12): 36–45 (Dec. 1991).

19. F. Amoroso and J. A. Kivett, Simplified MSK signal technique, *IEEE Trans. Commun.* **COM-25**(4): 433–441 (April 1977).

20. P. Galko and S. Pasupathy, Optimal linear receiver filters for binary digital signals, *Proc. Int. Conf. Communication*, Philadelphia, PA, June 1982, pp. 1H.6.1–1H.6.5.

21. A. Svensson and C.-E. Sundberg, Optimum MSK-type receivers for CPM on Gaussian and Rayleigh fading channels, *IEE Proc., Part F, Commun., Radar, Signal Process* **131**(8): 480–490 (Aug. 1984).

22. C.-E. Sundberg, Continuous phase modulation, *IEEE Commun. Mag.* **24**(4): 25–38 (April 1986).

23. G. Lindell, *On Coded Continuous Phase Modulation*, Ph.D. thesis (on telecommunication theory), Univ. Lund, Lund, Sweden, May 1985.

24. S. J. Simmons and P. H. Wittke, Low complexity decoders for constant envelope digital modulations, *IEEE Trans. Commun.* **COM-31**(12): 290–295 (Dec. 1983).

25. N. Seshadri and J. B. Anderson, Asymptotic error performance of modulation codes in the presence of severe intersymbol interference, *IEEE Trans. Inform. Theory* **IT-34**: 1203–1216 (Sept. 1988).

26. R. E. Ziemer and R. L. Peterson, *Digital Communications and Spread Spectrum Systems*, Macmillan, New York, 1985.

27. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

CONTINUOUS PHASE FREQUENCY SHIFT KEYING (CPFSK)

THOMAS A. SCHONHOFF
Titan System Corporation
Shrewsbury, Massachusetts

1. INTRODUCTION

Continuous phase frequency shift keying (CPFSK) is a modulation that, as its name implies, can be characterized as a traditional frequency shift keyed (FSK) signal constrained to maintain continuous phase at its symbol time boundaries. This constraint offers two important advantages from a communication point of view:

1. The continuous phase at the symbol boundaries essentially "smooths" the waveform, thereby offering a signal bandwidth that can be considerably smaller than conventional modulations such as FSK or phase-shift-keying (PSK). The spectral characteristics are presented in Section 3.
2. The waveform during each symbol period is dependent on the data and waveforms during previous symbol periods (i.e., the signal waveform contains memory). This memory can be used to improve error rate performance relative to more conventional modulations. The error rate performance is discussed in Section 4.

Historically, CPFSK is a generalization of minimum shift keying (MSK) [1] and, as shown below, MSK is indeed one form of CPFSK. In turn, CPFSK has been generalized to continuous phase modulation (CPM) [2]. Some specific CPM techniques and their relationship to CPFSK are given in Section 2.3.

2. DEFINITION OF CPFSK

During the ith symbol period, the transmitted CPFSK waveform can be written as

$$s(t) = \sqrt{\frac{2E}{T}} \cos\left(2\pi f_c t + \frac{d_i \pi h[t-(i-1)T]}{T} + \pi h \sum_{j=i-1} d_j + \phi\right) \quad (1)$$

where E is the transmitted signal energy during symbol period T and f_c is the carrier frequency. d_i represents the digital data during the ith symbol; for M-ary signaling, $d_i = \pm 1, \ldots, \pm(M-1)$. h is referred to as the deviation ratio. Its importance is made evident in succeeding sections. The starting phase at the beginning of the i^{th} symbol is seen to be $\pi h \sum_{j=i-1} d_j$. This term shows that previous symbols have an effect on the transmitted waveform. ϕ is the initial starting phase.[1]

2.1. Phase Trajectories of CPFSK

From the preceding equation, the phase term is $\vartheta(t) = \frac{d_i \pi h[t-(i-1)T]}{T} + \pi h \sum_{j=i-1} d_j$. This is known as the phase trajectory, and an example for quaternary CPFSK is shown in Fig. 1.

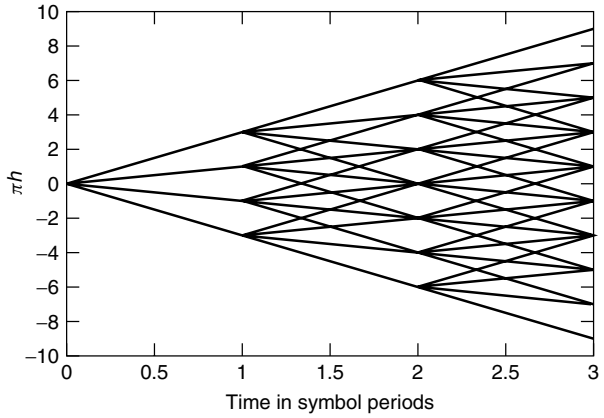


Figure 1. Quaternary CPFSK phase trajectories.

[1] It is assumed that starting phase at time zero is 0.

For the specific case of $h = \frac{1}{2}$, the signal is known as minimum shift keying (MSK) and is widely used in radio systems for which the transmitted frequency is 500 kHz or less.[2]

2.2. Frequency modulation Interpretation of CPFSK

Since the instantaneous transmitted frequency of the signal is the time derivative of the phase, it is clear from Eq. (1) that, during every symbol periods, one of M possible transmitted frequencies are sent, namely $2\pi f_c + \dfrac{d_i \pi h}{T}$, where $d_i = \pm 1, \ldots, \pm(M-1)$.

2.3. Relationship to General CPM

Continuous phase modulation (CPM) is a generalization of CPFSK where the generalization uses the two criteria of performance improvements explained in the Introduction. For example, one approach for CPM is to make the modulation even "smoother" (i.e., make not only the phase continuous but also higher derivatives of the phase continuous). The second approach is to introduce more memory into the modulation by using phase and/or frequency pulse shapes that extend over more than one symbol period. In this case, memory is introduced not only by having continuous phase, frequency, etc. at the symbol boundaries, but also by the pulse shapes themselves. This latter approach has come to be called partial-response CPM.

Adapting the notation of Ref. 4, a general CPM signal can be written in terms of its phase as

$$\vartheta(t;\vec{d}) = 2\pi \sum_{i=-\infty}^{n} d_i h_i q(t - iT), \quad nT \le t \le (n+1)T \qquad (2)$$

In this article, we simply identify the parameters in this equation. The interested reader is directed to Ref. 2 or 4, This equation shows that the phase is determined, in general, by the vector of all past data $\vec{d}$, which represent a sequence of independent M-ary symbols taken from the set $\{\pm 1, \pm 3, \ldots, \pm(M-1)\}$.

Another item of note in Eq. (2) is that, in general, a different value of the modulation index h_i can be used for each symbol period. At present, a popular digital modulation candidate for military UHF satellite systems is a form of this multi-h quaternary modulation. The normalized phase shape $q(t)$ extends over a general L symbol periods and, in general, is constrained only by its end regions. These constraints can be written as

$$q(t) = \begin{cases} 0, & t < 0 \\ \frac{1}{2}, & t \ge LT \end{cases} \qquad (3)$$

3. TRANSMITTED SPECTRAL PROPERTIES OF CPFSK

The general derivation of the power spectral density of CPFSK was first given in Salz [3], although the derivation and terminology of Proakis [4] is used herein. The power spectral density of CPFSK is shown in Ref. 4 to be

$$\Phi(f) = T\left[\frac{1}{M}\sum_{n=1}^{M} A_n^2(f) + \frac{2}{M^2}\sum_{n=1}^{M}\sum_{m=1}^{M} B_{nm}(f) A_n(f) A_m(f)\right] \qquad (4)$$

where

$$A_n(f) = \frac{\sin\pi\left[fT - \frac{1}{2}(2n-1-M)h\right]}{\pi\left[ft - \frac{1}{2}(2n-1-M)h\right]}$$

$$B_{nm}(f) = \frac{\cos(2\pi fT - \alpha_{nm}) - \varphi\cos\alpha_{nm}}{1 + \varphi^2 - 2\varphi\cos 2\pi fT}$$

$$\alpha_{nm} = \pi h(m + n - 1 - M) \qquad (5)$$

and

$$\varphi = \frac{\sin M\pi h}{M \sin \pi h}$$

Figure 2 shows a plot of the one-sided power spectral density of binary CPFSK for three values of the deviation ratio h, namely, MSK, for which h is $\frac{1}{2}$, the value of h that gives the best binary error rate, namely, $h = 0.715$, and an intermediate value of $h = 0.6$.

Figure 3 shows a comparable one-sided power spectral density for three selected values of h for quaternary CPFSK.

4. RECEIVER STRUCTURES AND ERROR RATE PERFORMANCE OF CPFSK

Much of this section is based on the original developments of Refs. 5 and 6. The initial work in Ref. 5 developed the

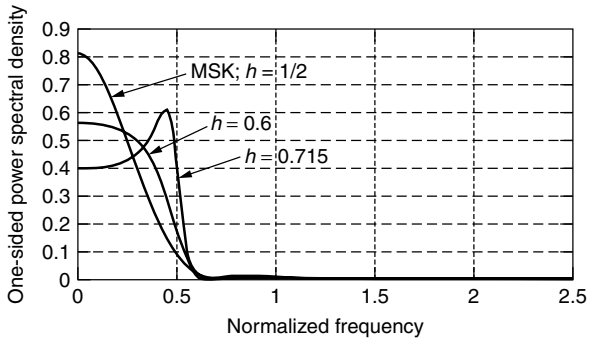


Figure 2. One-sided power spectral density of binary CPFSK.

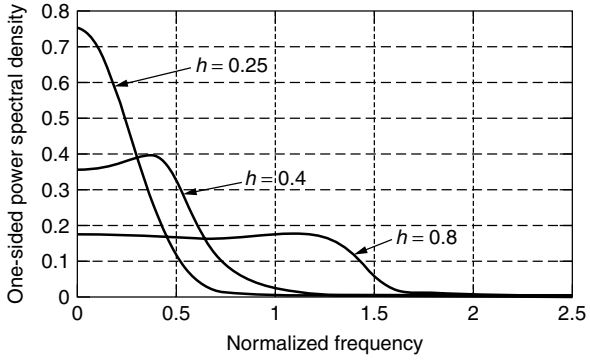


Figure 3. One-sided power spectral density of quaternary CPFSK.

[2] Unfortunately, when $h = \frac{1}{2}$, the memory benefits of CPFSK are significantly reduced so that the error rate performance is theoretically identical to antipodal PSK.

theory and structures for binary signaling, and this was generalized and expanded upon in Ref. 6. Two receiver structures associated with coherent and noncoherent processing, respectively, are presented in Sections 4.1 and 4.2, whereas a compilation of alternative structures is presented in Section 4.3.

4.1. Coherent Structures and Performance

A shorthand notation for the received CPFSK signal can be written as

$$r(t) = \alpha s(t, d_1, D_k) + n(t) \tag{6}$$

where α corresponds to the received signal attenuation, $s(t, d_1, D_k)$ is a compact representation of the transmitted signal of Eq. (1), d_1 is the first symbol on which we wish to make a decision, $D_k = \{d_2, \ldots, d_n\}$ corresponds to the next n symbols, and $n(t)$ is a narrowband zero-mean white Gaussian noise process with a double-sided power spectral density of $N_0/2$. We wish to observe n symbols and make a decision on the first.

It follows that the M likelihood parameters can be written as

$$l_1 = \iiint_{n-\text{fold}} \exp\left(\frac{2}{N_0}\right) \int_0^{nT} r(t)s(t, 1, D_k) f(D_k)\, dD_k$$

$$l_2 = \iiint_{n-\text{fold}} \exp\left(\frac{2}{N_0}\right) \int_0^{nT} r(t)s(t, -1, D_k) f(D_k)\, dD_k$$

$$\vdots$$

$$l_M = \iiint_{n-\text{fold}} \exp\left(\frac{2}{N_0}\right) \int_0^{nT} r(t)s(t, -(M-1), D_k) \times f(D_k)\, dD_k \tag{7}$$

where $f(D_k)$ is the discrete pdf of the $(n-1)$-tuple. Evaluating over these n integrals results in the n decision variables

$$U_1 = \sum_{j=1}^{m} \exp\left(\frac{2}{N_0} \int_0^{nT} r(t)s(t, 1, D_j)\, dt\right)$$

$$U_2 = \sum_{j=1}^{m} \exp\left(\frac{2}{N_0} \int_0^{nT} r(t)s(t, -1, D_j)\, dt\right)$$

$$\vdots$$

$$U_M = \sum_{j=1}^{m} \exp\left(\frac{2}{N_0} \int_0^{nT} r(t)s(t, -(M-1), D_j\, dt\right) \tag{8}$$

where $m = M^{n-1}$ and corresponds to all of the possible sequences of the $(n-1)$-tuple D_j.

Although this set of decision variables is very complicated, it allows us to develop an optimal and a suboptimal receiver structure for coherent CPFSK. From Eq. (8), we can see that an optimum receiver structure can be depicted as shown in Fig. 4. $m = M^{n-1}$ correlators (or matched filters) are used for each of the M possible received symbols. All m correlators associated with one of the symbols (e.g., d_i are added to produce the decision

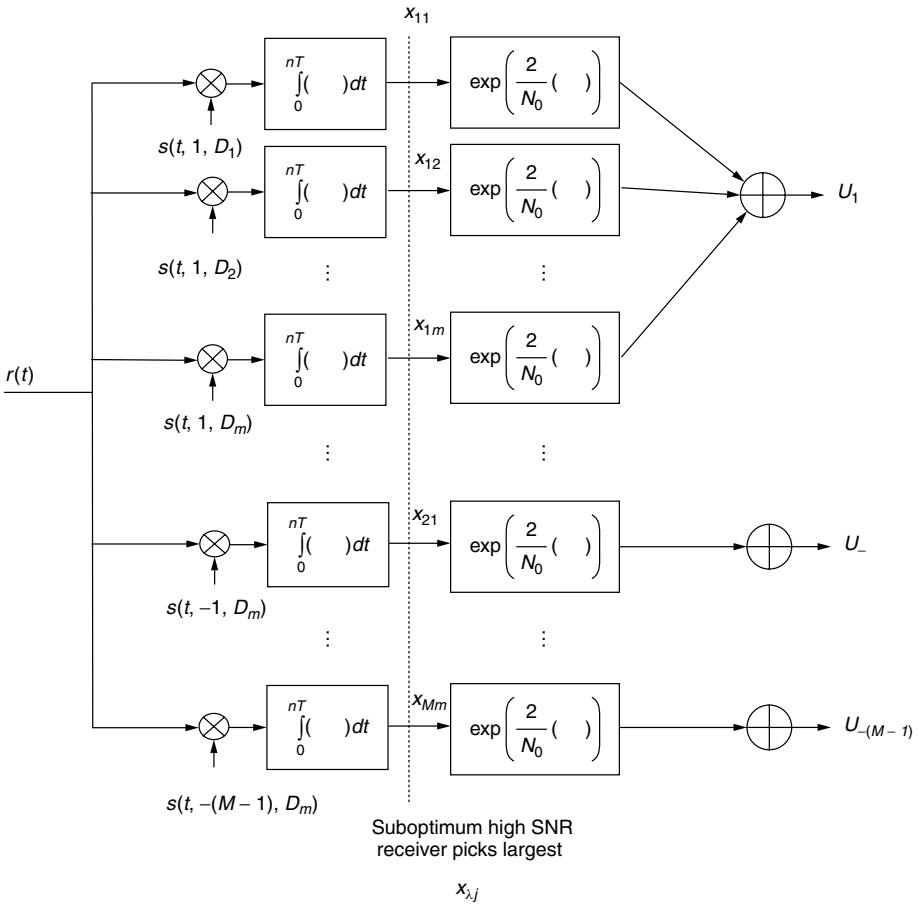


Figure 4. Optimal and suboptimal coherent CPFSK receivers.

variable U_i), and the maximum decision variable is used to estimate which symbol was transmitted during the i^{th} symbol period.

Because of the nonlinear nature of the optimum decision variables, it is analytically impossible to determine the error rate performance of this optimum receiver. Nonetheless, both low snr and high snr bounds can be used to estimate its performance [5,6]. In particular, the high snr approximation uses the fact that the exponential function is monotonic. Thus, if we truncate the receiver structure along the dotted line in Fig. 4, we determine the mM correlator outputs $x_{\lambda j}$, pick the largest, and then base our decision on the subvariable λ. That is, we still make one decision every symbol period, but we use correlators or matched filters that span the last n symbols.

Since the noise is Gaussian, it is clear that each of the correlator outputs is Gaussian and the performance can be estimated using either a union bound as indicated in Refs. 5 and 6 or the minimum distance of all the possible sequences as used in Ref. 2. At high snr, both approaches give comparable estimates. Figures 5 and 6 give binary and quaternary high snr receiver structure error rate performances respectively. These graphs were extracted from the same data as that of [6].

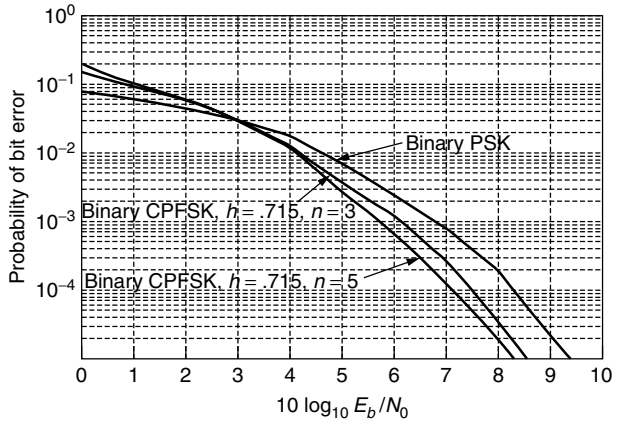


Figure 5. Probability of bit error for selected coherent binary modulations.

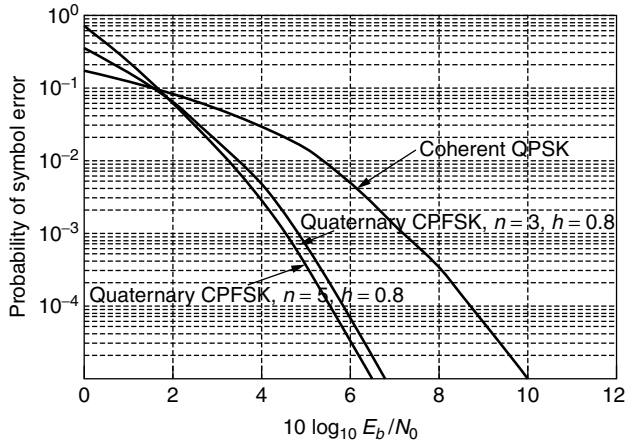


Figure 6. Probability of symbol error for selected coherent quaternary modulations.

For the binary error rate results of Fig. 5, it is known that the optimal value of the deviation ratio is $h = 0.715$. It can be seen from the figure that observing the signal for $n = 3$ or $n = 5$ bits both result in performance improvements over antipodal PSK. For $n = 5$, the improvement is approximately 0.8 dB at an error rate of 10^{-5}.

The results of Fig. 6 show that the improvements of quaternary CPFSK are much more impressive than those of binary CPFSK. Indeed, at an error rate of 10^{-5}, up to 3.5 dB improvement in SNR is possible. The values of h in the last two figures are the optimal values found; however, for quaternary CPFSK in particular, other values of h also result in improved performance. The interested reader is directed to Refs. 2 and 6.

4.2. Noncoherent Structures and Performance

The results of this section again parallel those of the developments in [6]. For noncoherent detection, the initial phase ϕ of Eq. (1) is assumed unknown with a uniform pdf from $(0, 2\pi)$. We use a slightly amended shorthand notation from [6] and model the received signal as

$$r(t) = s(t, d_{n+1}, \Delta_k, \phi) + n(t) \tag{9}$$

where we are observing $2n + 1$ symbols and making a decision on the middle symbol $d_{n+1}.\Delta_k$ is a $2n$-tuple consisting of the symbols before and after the decision symbol $d_{n+1}.\Delta_k$ can then be written as $\Delta_k = \{d_1, d_2, \ldots, d_n, d_{n+2}, \ldots, d_{2n+1}\}$. This implies that the subscript k progresses over $\mu = M^{2n}$ different values. The optimum noncoherent receiver structure is derived from the M likelihood parameters

$$l_1 = \int_\phi \iiint_\Delta \exp\left(\frac{2}{N_0}\int_0^{(2n+1)T} r(t)s(t, 1, \Delta_k, \phi)\,dt\right) \times f(\Delta)f(\phi)\,d\phi\,d\Delta$$

$$l_2 = \int_\phi \iiint_\Delta \exp\left(\frac{2}{N_0}\int_0^{(2n+1)T} r(t)s(t, -1, \Delta_k, \phi)\,dt\right) \times f(\Delta)f(\phi)\,d\phi\,d\Delta$$

$$\vdots$$

$$l_M = \int_\phi \iiint_\Delta \exp\left(\frac{2}{N_0}\int_0^{(2n+1)T} r(t)s(t, -(M-1), \Delta_k, \phi)\,dt\right) \times f(\Delta)f(\phi)\,d\phi\,d\Delta \tag{10}$$

These likelihood ratios can be seen to be similar to those of the coherent receiver structure given by Eq. (7) except for the averaging over the initial phase ϕ which results in a Bessel function of order zero $I_0(x)$. Performing the averages of Eq. (10) results in the M noncoherent decision variables

$$U_1 = \frac{1}{\mu}\sum_{k=1}^{\mu} I_0\left(\frac{2}{N_0}\chi_{1k}\right)$$

$$U_2 = \frac{1}{\mu}\sum_{k=1}^{\mu} I_0\left(\frac{2}{N_0}\chi_{2k}\right)$$

$$\vdots$$

$$U_M = \frac{1}{\mu}\sum_{k=1}^{\mu} I_0\left(\frac{2}{N_0}\chi_{Mk}\right) \tag{11}$$

where χ_{Nk} is a Rician statistical variable defined as

$$\chi_{Nk} = \sqrt{x_{Nk}^2 + y_{Nk}^2} \tag{12}$$

and x_{Nk} and y_{Nk} are the in-phase and quadrature variables, respectively, defined in terms of our shorthand notation as

$$x_{Nk} = \begin{cases} \displaystyle\int_0^{(2n+1)T} r(t)s(t, N, \Delta_k, 0)\,dt & Nodd \\ \displaystyle\int_0^{(2n+1)T} r(t)s(t, \_(N-1), \Delta_k, 0)\,dt & Neven \end{cases} \tag{13}$$

and

$$y_{Nk} = \begin{cases} \displaystyle\int_0^{(2n+1)T} r(t)s\left(t, N, \Delta_k, \frac{\pi}{2}\right)\,dt & Nodd \\ \displaystyle\int_0^{(2n+1)T} r(t)s\left(t, \_(N-1), \Delta_k, \frac{\pi}{2}\right)\,dt & Neven \end{cases} \tag{14}$$

Figure 7 shows the functional block diagram of the optimal noncoherent CPFSK receiver structure, which is developed from Eqs. (11) through (14). As can be seen, a total of M^{2n+1} noncoherent correlators or matched filters are used to make a decision on the middle symbol.

The performance of one example of noncoherently detected CPFSK is shown in Fig. 8. This figure shows that significant improvement is still possible even when the input starting phase is not estimated.

4.3. Other Receiver Structures

The phase trellis, shown in the example in Fig. 1, leads to another receiver structure based on the Viterbi algorithm (VA). This was first identified by Forney in Ref. 7 and explored specifically for CPFSK in Ref. 8. The VA is most useful when the deviation ratio h is a convenient fraction. For example, for binary CPFSK, if $h = \frac{2}{3}$, which is close to the optimum value of 0.715, the phase trellis can be reduced to a phase state diagram as shown in Fig. 9. The VA uses a traditional state history and state metric to determine the most likely path through the trellis or phase-state diagram. References 2 and 8 show that the error rate performance of the VA is virtually indistinguishable from that of the fully implemented multicorrelator receiver as derived in Sections 4.1 and 4.2.

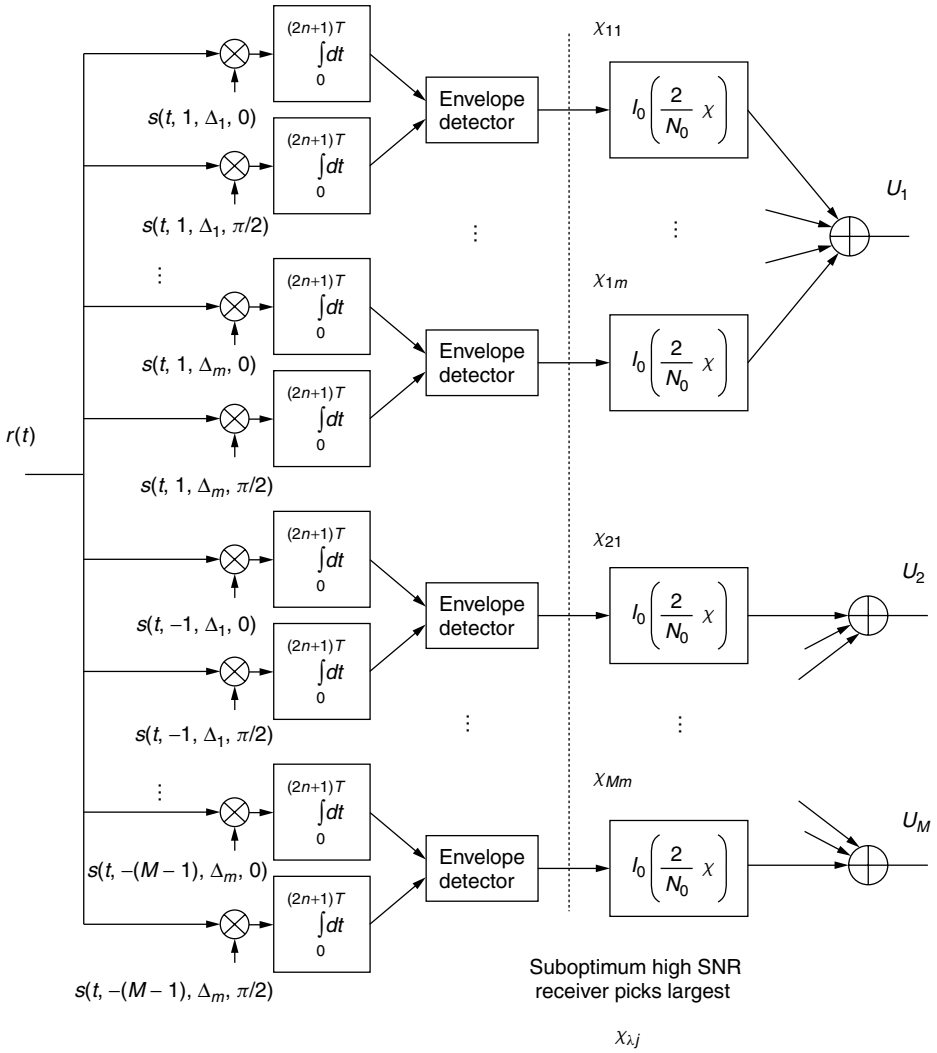


Figure 7. Optimal and high SNR noncoherent CPFSK receiver structure.

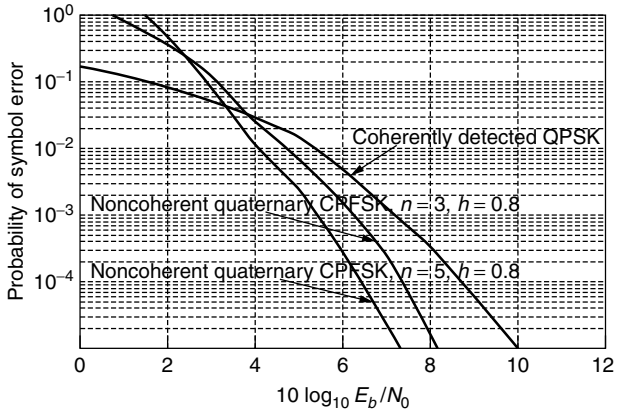


Figure 8. Probability of symbol error for quaternary noncoherent CPFSK.

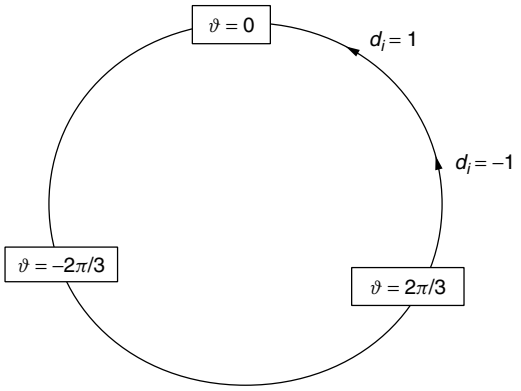


Figure 9. Phase-state diagram for binary CPFSK with deviation ratio of $\frac{2}{3}$.

5. CONCLUSION

This article has presented the concept of continuous phase frequency shift keying (CPFSK). Its transmitted spectra is given and examples are presented. Optimal receiver structures for coherent and noncoherent detection are derived, and examples of symbol error rate are presented. Finally, an alternate receiver structure, based on the maximum likelihood sequence estimator (MLSE) or Viterbi algorithm is outlined.

BIOGRAPHY

Thomas A. Schonhoff received his bachelor's degree from M.I.T., his master's degree from Johns Hopkins University, and his Ph.D. from Northeastern University. He has worked at six different corporations, although he has been with LinCom Corporation (now Titan System Corporation, Communication and Software Solutions Division) since 1985. For the past 21 years, Dr. Schonhoff has also taught graduate courses as an adjunct at Worcester Polytechnic Institute.

BIBLIOGRAPHY

1. Rudi de Buda, Coherent demodulation of frequency shift keying with low deviation ratio, *IEEE Trans. Commun.* 429–435 (1972).
2. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
3. R. R. Anderson and J. Salz, Spectra of digital FM, *Bell System Tech. J.* 1165–1189 (1965).
4. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
5. W. P. Osborne and M. B. Luntz, Coherent and noncoherent detection of CPFSK, *IEEE Trans. Commun.* 1023–1036 (1974).
6. T. A. Schonhoff, Symbol error probabilities for M-ary CPFSK: coherent and noncoherent detection, *IEEE Trans. Commun.* 644–652 (1976).
7. G. D. Forney, Jr., The Viterbi algorithm, *Proc. IEEE* March 268–278 (1973).
8. T. A. Schonhoff, H. Nichols, and H. Gibbons, Use of the MLSE algorithm to demodulate CPFSK, *1978 International Conference on Communications*, Toronto, June 1978.

CONVOLUTIONAL CODES

RICHARD D. WESEL
University of California at Los Angeles
Los Angeles, California

1. INTRODUCTION

Convolutional codes represent one technique within the general class of channel codes. Channel codes (also called *error-correction codes*) permit reliable communication of an information sequence over a channel that adds noise, introduces bit errors, or otherwise distorts the transmitted signal. Elias [1,2] introduced convolutional codes in 1955. These codes have found many applications, including deep-space communications and voiceband modems. Convolutional codes continue to play a role in low-latency applications such as speech transmission and as constituent codes in Turbo codes. Two reference books on convolutional codes are those by Lin and Costello [3] and Johannesson and Zigangirov [4].

Section 2 introduces the shift-register structure of convolutional encoders including a discussion of equivalent encoders and minimal encoders. Section 3 focuses on the decoding of convolutional codes. After a brief mention of the three primary classes of decoders, this section delves deeply into the most popular class, Viterbi decoders. This discussion introduces trellis diagrams, describes the fundamental add-compare-select computation, compares hard and soft decoding, and describes the suboptimal (but commonly employed) finite traceback version of Viterbi decoding.

Section 4 defines the free distance of a convolutional code and describes how free distance may be computed by a specialized application of the Viterbi algorithm. This procedure also yields an analytic lower bound on the decision depth that should be used for finite-traceback decoding. Catastrophic encoders are also discussed in this section. Section 5 describes the generating function that enumerates all the paths associated with error events in the decoder trellis. This section then gives union bounds

on bit error rate that are computed from the generating function. Section 6 provides some final remarks regarding the effective blocklength of convolutional codes and their role today.

2. ENCODER STRUCTURE

As any binary code, convolutional codes protect information by adding redundant bits. A rate-k/n convolutional encoder processes the input sequence of k-bit information symbols through one or more binary shift registers (possibly employing feedback). The convolutional encoder computes each n-bit symbol ($n > k$) of the output sequence from linear operations on the current input symbol and the contents of the shift register(s). Thus, a rate k/n convolutional encoder processes a k-bit input symbol and computes an n-bit output symbol with every shift register update. Figures 1 and 2 illustrate feedforward and feedback encoder implementations of a rate-$\frac{1}{2}$ code. Section 2.1 explores the similarities and differences between feedforward and feedback encoders by examining their state diagrams.

2.1. Equivalent Encoders

Convolutional encoders are finite-state machines. Hence, state diagrams provide considerable insight into their behavior. Figures 3 and 4 provide the state diagrams for the encoders of Figs. 1 and 2, respectively. The states are labeled so that the least significant bit is the one residing in the leftmost memory element of the shift register. The branches are labeled with the 1-bit (single-bit) input and the 2-bit output separated by a comma. The most significant bit (MSB) of the two-bit output is the bit labeled MSB in Figs. 1 and 2.

If one erases the state labels and the single-bit input labels, the remaining diagrams for Figs. 3 and 4 (labeled

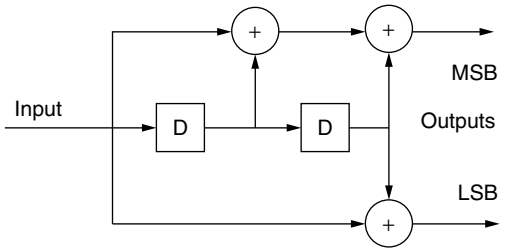


Figure 1. Rate-$\frac{1}{2}$ feedforward convolutional encoder with two memory elements (four states). MSB and LSB refer to the most and least significant bits, respectively.

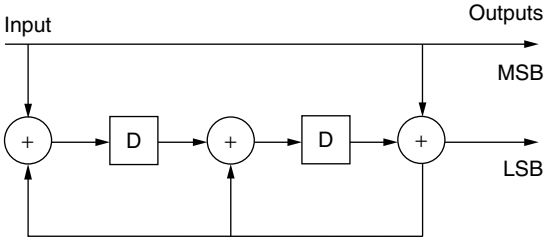


Figure 2. Rate-$\frac{1}{2}$ feedback convolutional encoder with two memory elements (four states).

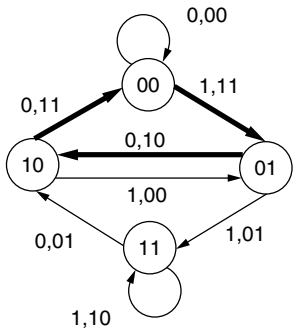


Figure 3. State diagram for rate-$\frac{1}{2}$ feedforward convolutional encoder of Fig. 1.

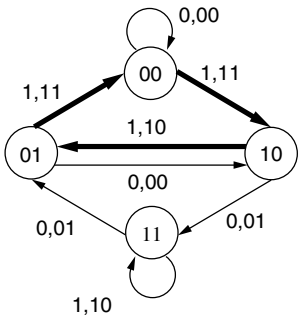


Figure 4. State diagram for rate-$\frac{1}{2}$ feedback convolutional encoder of Fig. 2.

with only the 2-bit outputs) would be identical. This illustrates that the two encoders are equivalent in the sense that both encoders produce the same set of possible output sequences (or codewords). Strictly speaking, a code refers to the list of possible output sequences without specifying the mapping of inputs sequences to output sequences. Thus, as in the above example, two equivalent encoders have the same set of possible output sequences, but may implement different mappings from input sequences to output sequences. In the standard convolutional coding application of transmission over additive white Gaussian noise (AWGN) with Viterbi decoding, such encoders give similar BER performance, but the different mappings of inputs to outputs do lead to small performance differences.

The three-branch paths emphasized with thicker arrows in Figs. 3 and 4 are each the shortest nontrivial (i.e., excluding the all-zeros self-loop) loop from the all-zeros state back to itself. Notice that for Fig. 3, the state diagram corresponding to the feedforward encoder, this loop requires only a single nonzero input. In contrast, for the state diagram corresponding to Fig. 4, this loop requires three nonzero inputs. In fact, for Fig. 4 no nontrivial loop from the all-zeros state to itself requires fewer than two nonzero inputs. Thus the feedforward shift register has a finite impulse response, and the feedback shift register has an infinite impulse response.

This difference is not particularly important for convolutional codes decoded with Viterbi, but it is extremely important to convolutional encoders used as constituents in Turbo codes, which are constructed by concatenating

convolutional codes separated by interleavers. Only feedback encoders (with infinite impulse responses) are effective constituents in Turbo codes. Thus, equivalent encoders can produce dramatically different performance as constituents in Turbo codes, depending on whether or not they meet the requirement for an infinite impulse response.

2.2. Minimal Encoders

A practical question to ask about a convolutional encoder is whether there is an equivalent encoder with fewer memory elements. This question may be answered by performing certain diagnostic computations on the encoder matrix. Furthermore, if the encoder is not minimal, it may be easily "repaired" yielding an encoder that is equivalent but requires fewer memory elements. Forney's classic paper [5] treats this fundamental area of convolutional coding theory elegantly. More recently, Johannesson and Wan [6] extended Forney's results by taking a linear algebra approach. This fascinating area of convolutional code theory is important to convolutional code designers, but less so for "users." Any code published in a table of good convolutional codes will be minimal.

3. DECODING CONVOLUTIONAL CODES

Convolutional code decoding algorithms infer the values of the input information sequence from the stream of received distorted output symbols. There are three major families of decoding algorithms for convolutional codes: sequential, Viterbi, and maximum a posteriori (MAP). Wozencraft proposed sequential decoding in 1957 [7]. Fano in 1963 [8] and Zigangirov in 1966 [9] further developed sequential decoding. See the book by Johannesson and Zigangirov [4] for a detailed treatment of sequential decoding algorithms. Viterbi originally described the decoding algorithm that bears his name in 1967 [10]. See also Forney's work [11,12] introducing the trellis structure and showing that Viterbi decoding is maximum-likelihood in the sense that it selects the sequence that makes the received sequence most likely.

In 1974, Bahl et al. [13] proposed MAP decoding, which explicitly minimizes bit (rather than sequence) error rate. Compared with Viterbi, MAP provides a negligibly smaller bit error rate (and a negligibly larger sequence error rate). These small performance differences require roughly twice the complexity of Viterbi, making MAP unattractive for practical decoding of convolutional codes. However, MAP decoding is crucial to the decoding of Turbo codes. For the application of MAP decoding to Turbo codes, see the original paper on Turbo codes by Berrou et al. [14] and Benedetto et al.'s specific discussion of the basic turbo decoding module [15].

When convolutional codes are used in the traditional way (not as constituents in Turbo codes), they are almost always decoded using some form of the Viterbi algorithm, and the rest of this section focuses on describing Viterbi. The goal of the Viterbi algorithm is to find the transmitted sequence (or codeword) that is closest to the received sequence. As long as the distortion is not too severe, this will be the correct sequence.

3.1. Trellis Diagrams

The state diagrams of Figs. 3 and 4 illustrate what transitions are possible from a particular state regardless of time. In contrast, trellis diagrams use a different branch for each different symbol time. As a result, trellis diagrams more clearly illustrate long trajectories through the states. Figure 5 shows one stage (one symbol time) of the trellis diagram associated with the rate-$\frac{1}{2}$ feedforward encoder of Figs. 1 and 3. Each column of states in the trellis diagram includes everything in the original state diagram. All the branches emanating from states in a particular column are incident on the states in the adjacent column to the right. In other words, each state transition in the trellis moves the trajectory one stage to the right.

To avoid crowding in Fig. 5, branch labels appear at the left and right of the trellis rather than on the branch itself. For each state, the top label belongs to the top branch emanating from or incident to that state. Figure 6 uses thick arrows to show the same path emphasized in the state diagram of Fig. 3. However, in Fig. 3 the beginning and end of the path were not clear. In Fig. 6 the path clearly begins in state 00 and then travels through 01 and then 10 before returning to 00.

3.2. The Basic Viterbi Algorithm

The Viterbi algorithm uses the trellis diagram to compute the accumulated distances (called the *path metrics*) from the received sequence to the possible transmitted sequences. The total number of such trellis paths grows

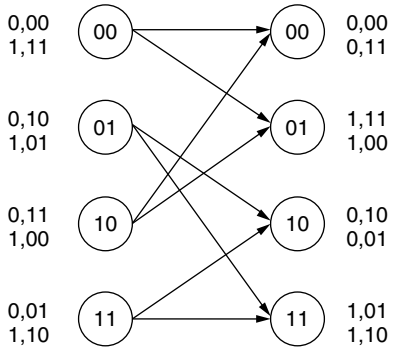


Figure 5. One stage of the trellis diagram for rate-$\frac{1}{2}$ feedforward convolutional encoder of Figs. 1 and 3.

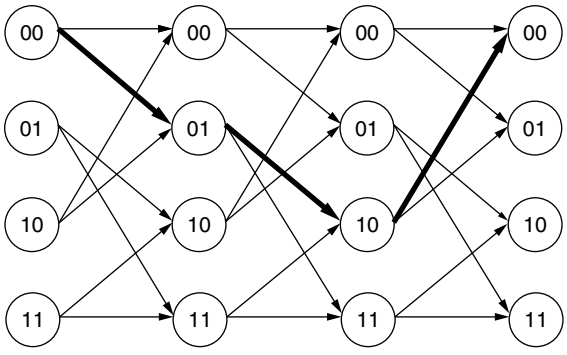


Figure 6. Trellis diagram for the path emphasized in Fig. 3.

exponentially with the number of stages in the trellis, causing potential complexity and memory problems. However, the Viterbi algorithm takes advantage of the fact that the number of paths truly in contention to have the minimum distance is limited to the number of states in a single column of the trellis, assuming that ties may be arbitrarily resolved.

As an example of the Viterbi algorithm, consider transmission over the binary symmetric channel (bit error channel) where the probability of a bit error is less than $\frac{1}{2}$. On such a channel, maximum likelihood decoding reduces to finding the output sequence that differs in the fewest bit positions (has the minimum Hamming distance) from the received sequence. For this example, assume the encoder of Fig. 1 with the state diagram of Fig. 3 and the trellis of Fig. 5. For simplicity, assume that the receiver knows that the encoder begins in state 00.

Figure 7 illustrates the basic Viterbi algorithm for the received sequence 01 01 10. Beginning at the far left column, the only active state is 00. The circle representing this state contains a path metric of zero, indicating that as yet, the received sequence differs from the possible output sequences in no bit positions. Follow the two branches leaving the first column to the active states in the second column of the trellis. Branch metrics label each branch, indicating the Hamming distance between the received symbol and the symbol transmitted by the encoder when traversing that branch. The two hypothetical transmitted symbols are 00 for the top branch and 11 for the bottom (see Fig. 5). Since both differ in exactly one bit position from the received symbol 01, both branch labels are one.

The path metric for each destination state is the sum of the branch metric for the incident branch and the path metric at the root of the incident branch. In the second column, both path metrics are one since the root path metric is zero. These equal path metrics indicate that no path is favored at this point. Now follow the branches from the second column to the third. Exactly one branch reaches each state in the third column. Once again, adding the branch metric and the associated root path metric produces the new path metric.

When following branches from the third column to the fourth, two branches are incident on each state. Only the path with the minimum path metric needs to survive. For example, state 00 (the top state) in the fourth column has a path incident from state 00 in the third column with a path metric of $2 + 1 = 3$. It also has a path incident from state 10 in the third column with a path metric of $3 + 1 = 4$. Only the path with the smaller path metric needs to survive. Figure 7 shows the incident branches of survivor paths with thicker arrows than nonsurvivor paths. Each state in the fourth column has exactly one survivor path, and the values shown indicate the path metrics of the survivor paths.

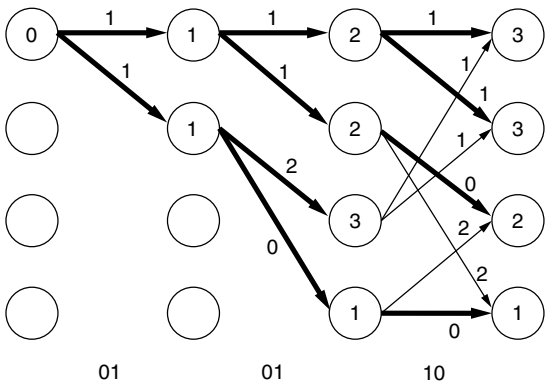


Figure 7. Illustration of the basic Viterbi algorithm on the bit error channel. This is also the trellis for hard Viterbi decoding on the AWGN channel in contrast to the soft Viterbi decoding shown in Fig. 8.

After all received symbols have been processed, the final step in decoding is to examine the last column and find the state containing the smallest path metric. This is state 11, the bottom state, in the fourth column. Following the survivor branches backward from the minimum-path-metric state identifies the trellis path of the maximum likelihood sequence. Reference to Fig. 5 reveals that the maximum likelihood path is the state trajectory $00 \rightarrow 01 \rightarrow 11 \rightarrow 11$. This state trajectory produces the output symbol sequence 11 01 10, which differs in exactly one bit position from the received sequence as indicated by its path metric. The input information sequence is decoded to be 1 1 1.

In this short example, only one trellis stage required path selection. However, once all states are active, path selection occurs with each trellis stage. In fact, if the initial encoder state is not known, path selection occurs even at the very first trellis stage. The basic computational module of the Viterbi algorithm is an add-compare-select (ACS) module. Adding the root path metric and incident branch metric produces the new path metric. Comparing the contending path metrics allows the decoder to select the one with minimum distance. When there is a tie (where two incident paths have the same path metric), a surviving path may be arbitrarily selected. In practice, ties are not uncommon. However, ties usually occur between paths that are ultimately destined to be losers.

3.3. Hard Versus Soft Decoding

The integer branch and path metrics of the binary error channel facilitate a relatively simple example of the Viterbi algorithm. However, the AWGN channel is far more common than the bit error channel. For the AWGN channel, binary phase shift keying (BPSK) represents binary 1 with 1.0 and binary 0 with -1.0. These two transmitted values are distorted by additive Gaussian noise, so that the received values will typically be neither 1.0 nor -1.0. A novice might choose to simply quantize each received value to the closest of 1.0 and -1.0 and assign the appropriate binary value. This quantization would effectively transform the AWGN channel to the bit error channel, facilitating Viterbi decoding exactly as described above. This method of decoding is called hard decoding, because the receiver makes a binary (hard) decision about each bit before Viterbi decoding.

Hard decoding performs worse by about 2 dB than a more precise form of Viterbi decoding known as *soft decoding*. Soft decoding passes the actual received values to the Viterbi decoder. These actual values are called soft values because hard decisions (binary decisions) have

not been made prior to Viterbi decoding. Soft Viterbi decoding is very similar to hard decoding, but branch and path metrics use squared Euclidean distance rather than Hamming distance. Figure 8 works an example analogous to that of Fig. 7 for the case where 1.0 and −1.0 are transmitted over the AWGN channel and soft Viterbi decoding is employed. A fixed-point implementation with only a few bits of precision captures almost all the benefit of soft decoding.

3.4. Finite Traceback Viterbi

The maximum likelihood version of Viterbi decoding processes the entire received sequence and then selects the most likely path. Applications such as speech transmission can conveniently process relatively short data packets in this manner. However, stream-oriented applications such as a modem connection cannot wait until the end of the received sequence before making any decisions about the information sequence. In such cases, a suboptimal form of Viterbi decoding is implemented in which decisions are made about transmitted bits after a fixed delay. This fixed delay is called the traceback depth or decision depth of the Viterbi decoder. The exact choice of the traceback depth is usually determined by simulation, but there is an analytic technique that identifies a good lower bound on what the traceback depth should be. We will discuss this "analytic traceback depth" in the next section, since its computation is a natural by-product of computing the free distance of a convolutional code.

Figures 9 and 10 illustrate finite traceback Viterbi decoding. Figure 9 shows the path metrics and survivor paths (indicated by thick arrows) for a soft Viterbi decoder in steady state. Actually, the selection of survivor paths and path metrics is the same for maximum-likelihood Viterbi decoding and finite-traceback Viterbi decoding. Figure 10 shows the distinguishing behavior of finite-traceback Viterbi decoding. Rather than wait until the end of the received sequence, each k-bit input symbol is decoded after a fixed delay. In Fig. 10 this delay is three symbols. After each update of the path metrics, the path with the smallest metric (identified in Fig. 10 by a thick circle) is traced back three branches and the k-bit input symbol associated with the oldest branch is decoded. Notice that the paths selected by this algorithm do not have to be consistent with each other. For example, the two paths traced back in Fig. 10 could not both be correct, but this inconsistency does not force the decoded bits to be incorrect.

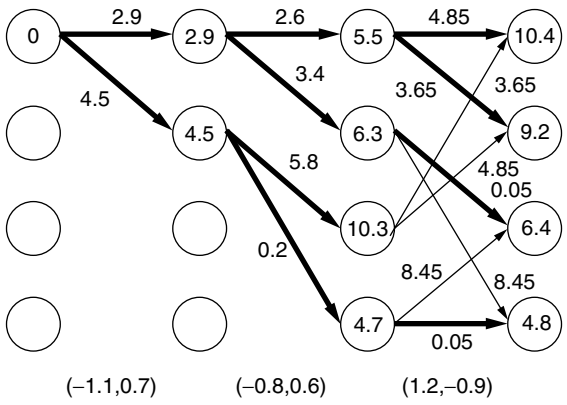


Figure 8. Illustration of soft Viterbi decoding.

4. FREE DISTANCE

The ultimate measure of a convolutional code's performance is the bit error rate (BER) or block error rate (BLER) of the code as a function of signal-to-noise ratio (SNR). However, free distance gives a good indication of convolutional code performance. The free distance of a convolutional code is the minimum distance (either Hamming or Euclidean) between two distinct valid output sequences. Unlike algebraic block codes, which are designed to have specific distance properties, good convolutional codes are identified by an exhaustive search over encoders with a given number of memory elements. Often free distance is the metric used in these searches.

For simplicity, we will restrict the discussion of free distance to free Hamming distance. For BPSK, this restriction imposes no loss of generality since the Hamming and squared Euclidean free distances are related by a constant.

4.1. Computation of Free Distance

The set of distances from a codeword to each of its neighbors is the same for all codewords. Hence, the free distance

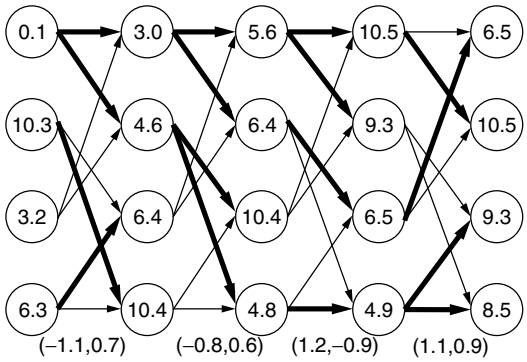


Figure 9. Steady state soft Viterbi state updates. Survivor paths are shown as thick arrows.

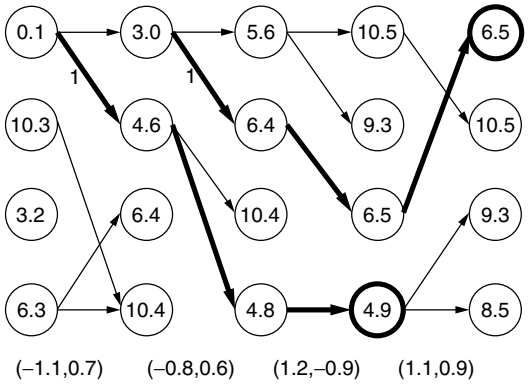


Figure 10. Finite traceback soft Viterbi decoding with a traceback depth of 3. Only the survivor paths of Fig. 9 are shown. Each traceback operation decodes only $k = 1$ bit. Thick arrows identify two such traceback paths.

is the distance from the all-zeros output sequence to its nearest-neighboring codeword. A Viterbi decoding operation with some special restrictions efficiently performs this computation. Viterbi decoding is performed on the undistorted all-zeros received sequence, but the first trellis branch associated with the correct path is disallowed. Thus prevented from decoding the correct sequence, the Viterbi algorithm identifies the nearest-neighbor sequence. Since the received sequence is noiseless, the path metric associated with the decoded sequence is the distance between that sequence and the all-zeros sequence, which is the free distance.

Figure 11 illustrates the computation of free Hamming distance using the Viterbi algorithm for the encoder described in Figs. 1, 3, and 5. The disallowed branch is shown as a dashed line. Only survivor branches are shown, and the thick branches indicate the minimum distance survivor path. Below each column is the minimum survivor path metric, which is called the *column distance*. The *free distance* is formally defined as the limit of the column distance sequence as the survivor pathlength tends to infinity. This limit is 5 in Fig. 11.

For noncatastrophic feedforward convolutional encoders, the free distance is equivalent to the minimum distance of a path that returns to the zero state. In general, the minimum distance path need not be the shortest path. For encoders with more states than the simple example of Fig. 11, there are typically several such paths having the same minimum distance. The number of minimum-distance paths is the number of nearest-neighbor output sequences. This is sometimes called the *multiplicity* of the free distance. If two codes have the same free distance, the code with the smaller multiplicity is preferred.

4.2. Analytic Decision Depth

As mentioned in Section 3.4, the specific decision depth used in finite traceback Viterbi is usually determined by simulation. However, a good estimate of the decision depth helps designers know where to begin simulations. For noncatastrophic feedforward encoders, Anderson and Balachandran [16] compute a useful lower bound on the required decision depth as a by-product of the free-distance computation.

This analytic decision depth is the pathlength at which the survivor path incident on the zero state has a path metric that is the unique minimum distance in the column. In other words, the path metric of the survivor path to the zero state is the only distance in the column equal to the column distance. In Fig. 11, this analytic decision depth is 8; after the eighth branch the path metric of the survivor path to the zero state is the only distance in the column equal to the column distance of 5. For noncatastrophic feedforward encoders, the Viterbi decoding procedure for computing free distance may be stopped when the analytic decision depth is identified. The column distance remains fixed thereafter.

When using this analytic decision depth, finite traceback decoding gives performance consistent with the first-order metrics of free distance and multiplicity. The asymptotic performance (in the limit of high SNR) is the same as maximum-likelihood Viterbi. In practice, a somewhat larger decision depth is often used to capture some additional performance at SNRs of interest by improving second-order metrics of performance (i.e., distances slightly larger than the minimum distance). For example, the analytic decision depth of the standard rate-$\frac{1}{2}$ 64-state feedforward convolutional encoder is 28, but simulation results show that a decision depth of 35 gives a noticeable performance improvement over 28. Decision depths larger than 35 give only negligible improvement.

4.3. Catastrophic Encoders

A convolutional encoder is catastrophic if a finite number of errors in the output sequence can cause an infinite number of errors in the input sequence. With such an encoder, the consequences of a decoding error can be truly catastrophic. Catastrophic encoders are certainly undesirable, and they are never used in practice. In fact, they are easily avoided because they are not minimal encoders. Hence if an encoder is catastrophic it also uses more memory elements than does a minimal equivalent encoder, which is not catastrophic.

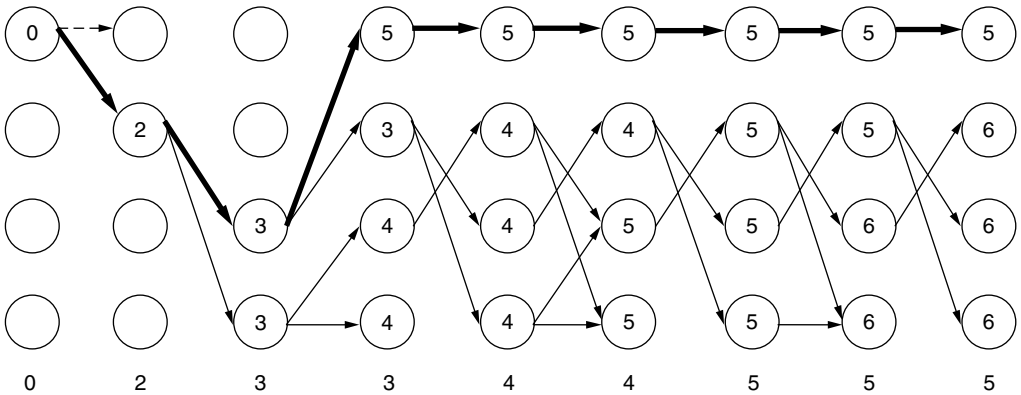


Figure 11. Application of the Viterbi algorithm to identify the free Hamming distance of the code described by Figs. 1, 3, and 5. The column distances are shown below each column. The disallowed branch is shown as a dashed line. Only survivor paths are shown, and the minimum distance path is shown with thick arrows.

An encoder is catastrophic if and only if its state diagram has a loop with zero output weight and nonzero input weight. Catastrophic encoders still have a free distance as defined by the limit of the column distance, but this free distance is seldom equal to the minimum survivor path metric to the zero state. Usually, some of the survivor path metrics for nonzero states never rise above the minimum survivor path metric to the zero state. An additional stopping rule for the Viterbi decoding computation of free distance resolves this problem: If the column distance does not change for a number of updates equal to the number of states, the free distance is equal to that column distance.

Noncatastrophic encoders may also require this additional stopping rule if they have a nontrivial zero-output-weight loop. Such a loop does not force catastrophic behavior if it is also a zero-input-weight loop. Such a situation only occurs with feedback encoders since feedforward encoders do not have loops with zero output weight and zero input weight except the trivial zero-state self-loop. In cases where this stopping criterion is required, the analytic decision depth of Anderson and Balachandran is not well defined. However, a practical place to start simulating decision depths is the pathlength at which the Viterbi computation of free distance terminates.

Because nontrivial zero-output-weight loops indicate a nonminimal encoder, their free distance is not often computed. However, there are circumstances where computation of the free distance is still interesting. As described by Fragouli et al. [17], these "encoders" arise not from poor design but indirectly when severe erasures in the channel transform a minimal encoder into a weaker, nonminimal encoder. An alternative to the additional stopping rule is simply to compute the free distance of an equivalent minimal encoder.

5. BOUNDS ON BIT ERROR RATE

As mentioned at the beginning of Section 4, BER and BLER as functions of SNR are the ultimate metrics of convolutional code performance. Monte Carlo simulation plays an important role in the characterizing BER and BLER performance. However, accurate characterization by Monte Carlo simulation at very low BER or BLER, say, less than 10^{-10}, is not computationally feasible with today's technology. However, analytic upper bounds on BER are very accurate below BER 10^{-5}. Thus, the use of bounds in conjunction with Monte Carlo simulation for high BER provides a good overall performance characterization.

5.1. The Generating Function

To facilitate the bound, a generating function or transfer function enumerates in a single closed-form expression all paths (including nonsurvivors) that return to the zero state in an infinite extension of the trellis of Fig. 11. The bound itself is analogous to the moment generating function technique for computing expectations of random variables. Figure 12 shows an altered version of the state diagram of Fig. 3 where the zero state has been split into a beginning zero state and an ending zero state. This new

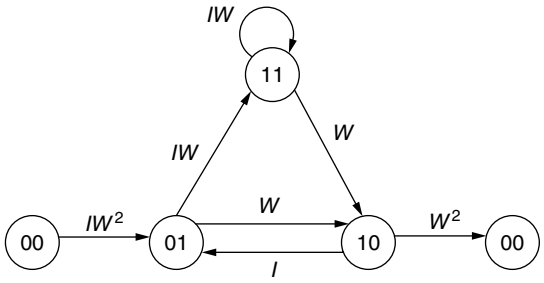


Figure 12. Split-state diagram for the encoder described by Figs. 1, 3, and 5. The exponent of W indicates the Hamming weight of the output error symbol. The exponent of I indicates the Hamming weight of the input error symbol.

diagram is called a *split-state diagram*. For the bound, only the Hamming weights of input and output symbols are needed. These Hamming weights are given as exponents of I and W, respectively. These values appear as exponents, so that when the labels along any path from the beginning zero state to the ending zero state are multiplied, the result is a single expression I^iW^w, where i is the overall input Hamming weight of the path and w is the overall output Hamming weight of the path.

Let A be the matrix of branch labels for all branches that neither begin nor end in a zero state. Each column of A represents a beginning state, and each row represents an ending state for a branch. A zero indicates no branch between the corresponding beginning and ending states. For Fig. 12

$$A = \begin{bmatrix} 0 & I & 0 \\ W & 0 & W \\ IW & 0 & IW \end{bmatrix} \tag{1}$$

Let b be the column of branch labels for all branches that begin in the zero state. For Fig. 12

$$b = \begin{bmatrix} IW^2 \\ 0 \\ 0 \end{bmatrix} \tag{2}$$

Let c be the row of branch labels for all branches that end in the zero state. For Fig. 12

$$c = \begin{bmatrix} 0 & W^2 & 0 \end{bmatrix} \tag{3}$$

The shortest path from the beginning zero state to the ending zero state has three branches; it is the path shown by thick arrows in Fig. 11. The product of the labels for this path may be computed as $cAb = IW^5$. Note that the exponent of 5 is consistent with the path metric of 5 in Fig. 11, which is also the free distance. There is one four-branch path and its label product is $cA^2b = I^2W^6$. There are two five-branch paths and their label products are $cA^3b = I^3W^7 + I^2W^6$. In general, $cA^{L-2}b$ gives the label products of all L-branch paths. Thus, the equation

$$T(W, I) = \sum_{L=3}^{\infty} cA^{L-2}b \tag{4}$$

$$= c(I - A)^{-1}b \tag{5}$$

enumerates the label products of all paths from the beginning zero state to the ending zero state. Note that I is

the input weight indeterminate in Eq. (4) but the identity matrix in Eq. (5). $T(W, I)$ is the generating function (or transfer function) of the convolutional encoder.

5.2. Union Bounds

Manipulation of $T(W, I)$ produces upper bounds on the BER for both the bit error channel and the AWGN channel. These upper bounds compute the sum over all error events e

$$\sum_e i_e P_e \tag{6}$$

where an error event is simply a path from the beginning zero state to the ending zero state. The input Hamming weight i_e associated with the error event e counts the total number of bit errors that all shifts of this error event can induce on a fixed symbol position. P_e is the probability that this path is closer to the received sequence than the transmitted (all-zeros) sequence. Note that (6) is an upper bound because P_e does not subtract probability for situations where more than one path is closer to the received sequence than the all-zeros sequence. For the bit error channel with bit error probability p,

$$\text{BER} \le \frac{1}{k} \left\{ \frac{\partial T(W, I)}{\partial I} \right\}_{I=1, W=2(p-p^2)^{1/2}} \tag{7}$$

For BPSK transmission of $\pm E_s^{1/2}$ over the AWGN channel with noise variance $N_0/2$, we obtain

$$\text{BER} \le \frac{1}{k} Q\left[\left(\frac{2 d_{\text{free}} E_s}{N_0} \right)^2 \right] e^{d_{\text{free}} E_s / N_0} \times \left\{ \frac{\partial T(W, I)}{\partial I} \right\}_{I=1, W=e^{-E_s/N_0}} \tag{8}$$

$$\le \frac{1}{2k} \left\{ \frac{\partial T(W, I)}{\partial I} \right\}_{I=1, W=e^{-E_s/N_0}} \tag{9}$$

where the tighter bound of Eq. (8) requires knowledge of the free distance d_{free}, but the looser bound of Eq. (9) does not.

6. FINAL REMARKS

Although a block code has a well-defined blocklength, convolutional codes do not. Convolutional codes are sometimes considered to have infinite blocklength, and this perspective is valuable for certain derivations, such as the derivation of union bounds on BER presented in Section 5. However, Sections 3.4 and 4.2 demonstrate that most of the useful information for decoding a particular input symbol lies within a relatively small interval of output symbols called the *decision depth*. In two important senses of blocklength, the latency required for decoding and the general strength of the code, the (properly chosen) decision depth is a good indicator the effective blocklength of a convolutional code. The decision depth of standard convolutional codes is small, certainly less than 50 for the standard rate-$\frac{1}{2}$ code with six memory elements in a single shift register.

Shannon's channel capacity theorem [18] (see also the treatise by Cover and Thomas [19]) computes the maximum rate that can be sent over a channel (or the maximum distortion that can be tolerated for a given rate). This theorem applies only as blocklength becomes infinite; in general it is not possible to achieve the performance promised by Shannon with small-blocklength codes. Indeed, convolutional code performance is hampered by their relatively small effective blocklength. For a bit error rate (BER) of 10^{-5}, they typically require about 4 dB of additional signal-to-noise ratio (SNR) beyond the Shannon requirement for error free transmission in the presence of AWGN. In contrast, for a BER of 10^{-5} Turbo codes and low-density parity-check codes, which both typically have blocklengths on the order of 10^3 or 10^4, require less than 1 dB of additional SNR beyond the Shannon requirement for error-free transmission in AWGN.

On the other hand, the performance of convolutional codes is actually quite good, given their short blocklengths. Applications such as speech transmission that require very low latency continue to employ convolutional codes because they provide excellent performance for their low latency and may be decoded with relatively low complexity. Furthermore, since Turbo codes contain convolutional encoders as constituents, a good understanding of convolutional codes remains essential even for long-blocklength applications.

BIOGRAPHY

Richard D. Wesel received both B.S. and M.S. degrees in electrical engineering from MIT in 1989 and the Ph.D. degree in Electrical Engineering from Stanford University in 1996. From 1989 to 1991 he was with AT&T Bell Laboratories, where he worked on nonintrusive measurement and adaptive correction of analog impairments in AT&T's long-distance network and the compression of facsimile transmissions in packet-switched networks. He holds patents resulting from his work in both these areas.

From July 1996 to July 2002 he was an Assistant Professor in the Electrical Engineering Department of the University of California, Los Angeles. Since July 2002 he is an Associate Professor at UCLA. His research is in communication theory with particular interests in the topics of channel coding the distributed communication. In 1998 he was awarded a National Science Foundation CAREER Award to pursue research on robust and rate-compatible coded modulation. He received an Okawa Foundation Award in 1999 for research in information and telecommunications, and he received the 2000 TRW Excellence in Teaching Award from the UCLA School of Engineering and Applied Science. Since 1999 he has been an Association Editor for the *IEEE Transactions on Communications* in the area of coding and coded modulation.

BIBLIOGRAPHY

1. P. Elias, Coding for noisy channels, *Proc. IRE Conv. Rec. part 4* 37–46 (1955) (this paper is also available in Ref. 2).
2. E. R. Berlekamp, ed., *Key Papers in the Development of Coding Theory*, IEEE Press, 1974.

3. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, 1983.

4. R. Johannesson and K. Sh. Zigangirov, *Fundamentals of Convolutional Coding*, IEEE Press, 1999.

5. G. D. Forney, Jr., Convolutional codes I: Algebraic structure, *IEEE Trans. Inform. Theory* **16**(6): 720–738 (Nov. 1970).

6. R. Johannesson and Z. Wan, A linear algebra approach to minimal convolutional encoders, *IEEE Trans. Inform. Theory* **39**(4): 1219–1233 (July 1993).

7. J. M. Wozencraft, Sequential decoding for reliable communication, *Proc. IRE Conv. Rec. part 2* 11–25 (1957).

8. R. M. Fano, A heuristic discussion of probabilistic decoding, *IEEE Trans. Inform. Theory* **9**: 64–74 (April 1963).

9. K. Sh. Zigangirov, Some sequential decoding procedures, *Probl. Peredachi Inform.* **2**: 13–25 (1966) (in Russian).

10. A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Trans. Inform. Theory* **13**: 260–269 (April 1967).

11. G. D. Forney, Jr., The Viterbi algorithm, *Proce. IEEE* **61**: 268–278 (March 1973).

12. G. D. Forney, Jr., Convolutional codes II: Maximum likelihood decoding, *Inform. Control* **25**: 222–266 (July 1974).

13. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**(2): 248–287 (March 1974).

14. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error correcting coding and decoding: Turbo-codes, *Proc. Int. Conf. Communication*, May 1993, pp. 1064–1070.

15. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, A soft-input soft-output APP module for the iterative decoding of concatenated codes, *IEEE Commun. Lett.* **1**(1): 22–24 (Jan. 1997).

16. J. B. Anderson and K. Balachandran, Decision depths of convolutional codes, *IEEE Trans. Inform. Theory* **35**(2): 455–459 (March 1989).

17. C. Fragouli, C. Komninakis, and R. D. Wesel, Minimality under periodic puncturing, *IEEE Int. Conf. Communication*, Helsinki, Finland, June 2001, pp. 300–304.

18. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (1948).

19. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

CRYPTOGRAPHY

IAN F. BLAKE
University of Toronto
Toronto, Ontario, Canada

1. INTRODUCTION

The need for secure communications has existed for centuries, and many ciphers such as substitution, transposition, and other types were in use by the Middle Ages. An excellent account of the historical development of cryptography is given in the comprehensive book of Kahn [3].

The first and most influential contribution of the modern era is the work of Claude Shannon [8], where fundamental notions of secrecy systems, including the principles of diffusion and confusion, unicity distance, and an information-theoretic approach to secrecy, were introduced. The notions of diffusion and confusion are very much in evidence in the design of the Data Encryption Standard, introduced in 1977 and used worldwide to this time. Only relatively recently has a replacement been announced, the Advanced Encryption Standard. These two block ciphers, as well as other so-called symmetric key ciphers, including stream ciphers, are considered in the next section.

It is the paper by Diffie and Hellman [1] that has ushered in the modern era of cryptography and decisively changed the landscape. It proposed the notions of asymmetric key (public key) cryptography, one-way functions, trap-door one-way functions, and digital signatures (called *one-way authentication* there) that have revolutionized secure communications in a networked world. Perhaps the most elegant incarnation of their ideas is that of RSA [6] (standing for Rivest, Shamir, and Adleman, the inventors), which includes the first realization of a digital signature. The notion of a cryptographic protocol, a sequence of steps to achieve a given purpose, arose out of these works and continues as a vital area of research.

This article overviews the subject of modern cryptography in a manner that those with a technical background will hopefully find useful, while omitting mathematical proofs.

Three excellent reference works on cryptography are those by Schneier [7], which gives a comprehensive and readable account of the subject and those by Menezes et al. [5] and Stinson [9], which give a more detailed and mathematical account that those interested in current research directions and implementation will find invaluable. Both of these last two references will be referred to liberally in this article.

The Website of the National Institute of Standards and Technology (NIST) contains an organized, authoritative, and up-to-date reference on standards, documentation and ongoing activity on the theory and practice of cryptography that is invaluable. It publishes the standards in documents referred to as *Federal Information Processing Standards* (FIPS), which invariably become de facto worldwide standards. These will be referred to throughout the article.

This article attempts to cover many of the important aspects of modern cryptography. A guide to this has been the list provided by NIST on its Website, which lists the following items in its cryptographic toolkit:

- Encryption
- Digital signatures
- Prime-number generation
- Modes of operation
- Authentication
- Random-number generation
- Secure hashing
- Key management

Each category contains references and links to standards and further materials, as appropriate. It is a reference list

for proper cryptographic practice in all of its important aspects. The list is used as a guide for this article. Authentication here is taken to include both entity authentication and data integrity.

Cryptography has among its goals to achieve confidentiality, data integrity, authentication, identification, and nonrepudiation of data transmission, storage, and transactions, and this article will outline how cryptography is used to achieve these goals. Formally, *cryptography* is the study of "secret writing," and *cryptanalysis* is the study of breaking a cryptographic technique. Together they are referred to as *cryptology*.

The following section first considers symmetric key encryption systems, including both block and stream ciphers, where the transmitter and receiver must have a common key. Before proceeding to examine public key cryptosystems, we briefly consider the mathematical problems and their complexity on which such systems are based. The main functions that public key systems are used for, including key exchange, encryption, digital signatures, and authentication techniques, are then presented. The article concludes with a brief look at prime-number and random-number generation and some indication of future directions of cryptography.

2. SYMMETRIC KEY ENCRYPTION SYSTEMS

Symmetric key encryption systems require a common key at both the transmitter and receiver. In a network environment this might mean that each pair of users requires a unique shared key, implying that each user in a community of n users is required to store $\binom{n}{2}$ keys, often a prohibitively large number. An efficient solution to this problem will be considered below.

The set of symmetric key systems is divided into block and stream systems. In a block system the input data are divided into blocks of equal length and produces equal length ciphertext blocks of output, usually of the same length as the input blocks. A stream cipher encrypts one symbol at a time using a system that produces a long sequence of symbols in some complex manner. They are typically generated by hardware, tend to be faster than block ciphers and require less complex hardware and storage. They find particular application in constrained devices such as wireless systems.

2.1. Symmetric Key Ciphers

As noted, a block cipher, represented in Fig. 1, operates on a block of plaintext P bits with a key K to produce a block of ciphertext bits, C. The decryption process uses the same key as the encryption process.

The best known and most widely used block cipher has been the Data Encryption Standard (DES). Originally introduced in 1977, it was the result of a solicitation by NIST (then the National Bureau of Standards) for an encryption method for computer data and was derived from a submission by IBM. A detailed description of its operation is given in FIPS 46 (1977) (and modifications given in FIPS 46-1, 1988 and FIPS 46-2, 1993).

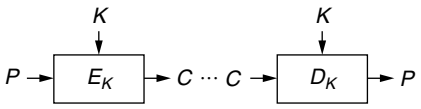


Figure 1. Basic block encryption.

The operation of DES is briefly described, with some details omitted. It operates on input plaintext data blocks of length 64 bits, producing ciphertext blocks of the same length. While its key length is 64 bits, only 56 bits are used in the algorithm itself; every 8th bit is a parity bit that is ignored in producing ciphertext. The first step of the algorithm is a fixed permutation of the block held in a register. After this step, the data are divided into left and right halves, each of 32 bits. The algorithm proceeds in 16 rounds with a typical round represented in Fig. 2.

At each of the 16 rounds, a different 48 bit subkey, K_i, is derived from the 56 bit key K in a simple and deterministic manner, not given here. The function f takes the 32 bit R_i together with the 48-bit subkey K_i to produce a 32-bit output. The block R_i is first expanded to 48 bits, by repeating certain of the bits, and XORed with the subkey K_i to produce 8 blocks of 6 bits each. Each subblock addresses one of eight so-called S boxes, which are 4×16 arrays containing integers 0–15. The first and last bits of the 6-bit subblock address the row and the middle 4 bits address a column, producing an integer representing 4 bits. Encryption concludes with the inverse of the initial permutation. The decryption process is very similar to the encryption process, with certain parts of the algorithm inverted or run backward as appropriate.

The algorithm is actually used in one of four modes (referred to as the *modes of operation*). The basic technique described is referred to as *electronic codebook* (ECB). The other modes are called *cipher feedback* (CFB), *output feedback* (OFB), and *cipher block chaining* (CBC). These modes introduce feedback and hence dependence between cipher blocks, useful in avoiding certain types of repetition attacks, among other uses. The four modes of operation can actually be used with any block cipher.

It has been observed for many years that the 56-bit key is too short to withstand a brute-force attack on current computing equipment, and indeed DES is now viewed as insecure, and its use is no longer recommended. It is possible to cascade DES ciphers, using double or triple encryption, thereby expanding the key length to 112 or

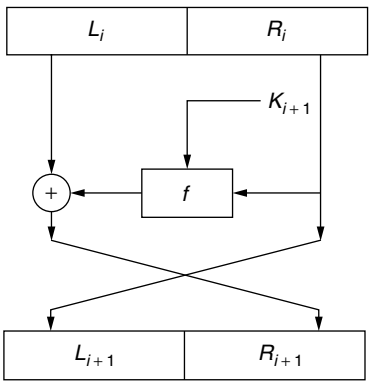


Figure 2. One round of DES.

168 bits, although here also, care must be taken. These so-called double and triple DES versions will likely be in use for many more years as they are in widespread applications in both hardware and software. DES has proved to be a remarkably successful block cipher, far outliving its planned lifetime with no inherent weaknesses discovered.

Recognizing the need for stronger encryption, NIST solicited the cryptographic community for algorithms to replace DES. Of the many submissions received, an algorithm from Belgium, Rijndael, was chosen, after careful consideration of its security, speed in both software and hardware, and suitability on a variety of platforms. The algorithm as adopted by NIST is referred to as the Advanced Encryption Standard (AES), and it is described in detail in FIPS 197 (Dec. 2001). An overview of the algorithm is given here to contrast it with the DES algorithm described above. The algorithm supports key lengths of 128, 192, and 256 bits. While the original Rijndael algorithm was designed to support block lengths 128, 192, and 256, only the data block length 128 is supported in the current standard.

The algorithm is byte- and word-oriented (4 bytes to a word) and involves two types of arithmetic. For arithmetic on bytes, the 8-bit sequence $(a_0, a_1, \ldots, a_7)$ is interpreted as the polynomial $a(x) = a_0 + a_1x + a_2x^2 + \cdots + a_7x^7$. Arithmetic in the finite field with 256 elements, $\mathbb{F}_{2^8}$, is taken as the arithmetic of polynomials of degree <8 over $\mathbb{F}_2$, modulo the irreducible polynomial $m(x) = 1 + x + x^3 + x^4 + x^8$. In particular, the inverse of a byte is defined for all nonzero bytes (the inverse of the all-zero byte, when called for in the algorithm, is taken as the all-zero byte). The second type of arithmetic involves arithmetic on polynomials with byte coefficients. The array of bytes $\{b_0, b_1, b_2, b_3\}$ is equated with the polynomial $b(x) = b_0 + b_1x + b_2x^2 + b_3x^3$. Arithmetic on such polynomials is taken modulo the (reducible) polynomial $M(x) = x^4 + 1$.

Using these arithmetics, the encryption algorithm is briefly described. Assume a key and block length of 128 bits; the algorithm for the other key sizes is easily derived from this. A *state* matrix is first defined as a 4×4 matrix with each entry a byte, and this is initially set to the input block to be encrypted, the matrix filled in by the data bytes down columns. The state matrix is continually modified at each step of the algorithm, with the final state matrix containing the output ciphertext block. The algorithm uses four basic operations (three described here and the fourth, in the next paragraph): (1) for each byte in the state matrix a *SubBytes* operation is defined as an affine transformation that replaces the byte $b = (b_0, b_1, \ldots, b_7)$ by $b' = (b'_0, b'_1, \ldots, b'_7)$, where $b' = Ab + c$, where A is a circulant matrix with first row [10001111] (as bits) and where the vector c is [11000110]; (2) a *ShiftRows* operation on the state matrix is defined as shifting row i of the matrix i positions to the left, cyclically, for rows $i = 0, 1, 2, 3$; and (3) the last operation, *MixColumns*, is defined on the columns of the state matrix by replacing a column that has a polynomial representation $a(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ (a_i a byte) and multiplying it by the (fixed) polynomial $c(x) = c_0 + c_1x + c_2x^2 + c_3x^3$ modulo $M(x)$ where, in hexadecimal notation $c_0 = 0 \times$ '02', $c_1 = 0 \times$ '01', $c_2 = 0 \times$ '01', $c_3 = 0 \times$ '03' (e.g., $c_3 = [00000011]$).

Finally the encryption process requires the derivation of a number of rounds $N_r + 1$ of a key schedule $K[i]$, $i = 0, 1, \ldots, N_r$. The number of rounds required depends on the key size and is 10, 12, and 14 depending on whether the key size is 4, 6, or 8 words in length. The precise generation of the key rounds from the original key is mechanical and is not described here. Suffice it to say that at the ith round a 4×4 matrix of bytes, $K[i]$, is generated and this matrix is added to the state matrix under byte addition (XOR), in the operation referred to as *AddRoundKey*.

With the operations described, the encryption algorithm is simply described as follows; recall that the initial state matrix is the 128 bits of the data to be encrypted fed in byte-wise down columns:

```
AddRoundKey (state, K[0])
for (i = 1 to Nr − 1) do {
   SubBytes (state);
   ShiftRows (state);
   MixColumns (state);
   AddRoundKey (state, K[i]);
   }
SubBytes (state);
ShiftRows (state);
AddRoundKey (state, K[ Nr]);
```

The encrypted output block is then the state matrix. The decryption process is similar in structure.

The AES will be the block encipherment algorithm of choice for the foreseeable future. Because of the varying block and key lengths, the modes of operation for AES are still under consideration and a *counter* mode has been added to the four standard modes mentioned previously. As with any block cipher, the last block is padded in an appropriate manner to round the data sequence out to an integral number of blocks.

There are numerous other block ciphers in use, some in the public domain and others proprietary [7]. While some may have specific advantages such as speed, use of AES will likely dominate the future of block encryption.

2.2. Stream Ciphers

From an information-theoretic point of view, the only secure cipher is a one-time pad, where purely random bits are recorded and given to the transmitter and receiver. These bits may then be XORed to the bits of the message to form the ciphertext. The plaintext can then be recovered at the receiver by XORing the ciphertexts with the one-time pad bits. Stream ciphers attempt, in some sense, to emulate this situation. More complicated operations than XORing might also be used. A general additive stream cipher is shown in Fig. 3. Here the key K may be the initial

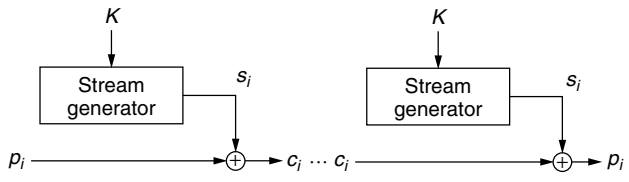


Figure 3. Configuration for an additive stream cipher.

state of the circuit used to generate the bit stream. Clearly the machines at the transmitter and receiver must be initially in the same state to generate the same sequence, allowing the plaintext to be recovered at the receiver.

There is a wealth of literature on the generation of suitable sequences. Many, if not most, of these techniques use maximum-length shift register sequences in some manner. As shown in Fig. 4, these sequences are generated by a shift register of length n, say, where the linear feedback connections are determined by a certain type of polynomial, a primitive polynomial, over the field of two elements, $\mathbb{F}_2$. Tables of such polynomials exist to quite high degrees. Clearly the sequences, coming from a deterministic circuit with a finite number of states, must be periodic. For a given shift register length, these sequences have the maximum length possible, $2^n - 1$ for a register of length n. Such sequences themselves are not secure — it is known the feedback connections of the register may be determined from knowledge of approximately $2n$ bits. However, the outputs of several such registers may be combined in a highly nonlinear Boolean function to produce a binary sequence of much greater complexity more suitable for cryptographic applications.

As noted, such stream ciphers are attractive in some applications for their high speed and relatively low circuit complexity. They have been incorporated in some standards, although many systems use proprietary generation techniques.

3. THE COMPLEXITY OF CERTAIN MATHEMATICAL PROBLEMS

Public key cryptography depends very much on the computational complexity of certain mathematical problems using the currently best known algorithms to solve them, which gives a notion of computational security. A few of the most important such problems, in terms of their use in public key cryptography, are reviewed here. Specifically these will be the problems of integer factorization, modular discrete logarithms, and modular square roots. Only the discrete logarithm problem will be considered in any detail. Informally we classify a problem as being *computationally feasible* or easy, if it is likely that a presumed attacker will have the resources to solve the system in a reasonable amount of time. Otherwise we refer to a problem as being *computationally infeasible*.

A *one-way function* f [5] is one for which it is "easy" to compute $f(x)$ for all elements in its domain X but for a randomly selected value $y \in \mathrm{Im}(f)$ it is computationally infeasible to find a value $x \in X$ such that $f(x) = y$. A *trapdoor one-way function* is a one-way function for which, given some extra information, it becomes computationally feasible to find such a value of $x \in X$, where the trapdoor information is independent of x.

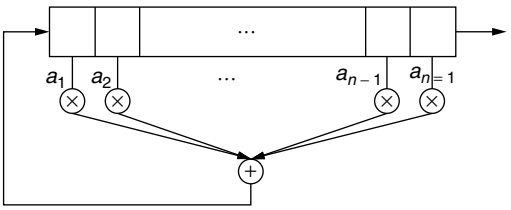


Figure 4. An LFSR for the polynomial $f(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n$.

A measure of the computational complexity of the problems that will be of interest for the problems considered, is

$$L_n(a, c) = O(\exp(c + o(1)[(\log(n))^a (\log\log(n))^{1-a}]))$$

where $O(\cdot)$ and $o(\cdot)$ are the standard complexity notation. This function is referred to as *subexponential* in $\log(n)$ since if $a = 1$ it is exponential in $\log(n)$ and if $a = 0$ it is polynomial in $\log(n)$.

The difficulty of factoring an integer plays a central role in many public key systems. It might be argued the hardest integers of a given size to factor are integers that are the product of two large primes of approximately the same size, $n = pq$, where p and q are primes on the order of $\sqrt{n}$. The most efficient factoring algorithm currently, to factor a general integer (one with no special structure), is the general number field sieve (NFS). It has a conjectured complexity [5] of $L_n\left(\frac{1}{3}, c\right)$, where $c = \left(\frac{64}{9}\right)^{1/3}$ in both time and space. The NFS continues to be developed.

The problem of determining square roots in $\mathbb{Z}_n^*$, the group of units of $\mathbb{Z}_n$, where $n = pq$ is the product of two odd primes, is formally equivalent to factoring n of such form. It can also be shown there are efficient algorithms for determining square roots modulo a prime number and for primes of the form $p \equiv 3 \pmod 4$ they are particularly so. In fact a solution to the equation $x^2 \equiv a \pmod p$ for $p \equiv 3 \pmod 4$, when a is a square $(\mathrm{mod}\ p)$, is given by [5]

$$u \equiv \pm a^{(p+1)/4} \pmod p$$

An efficient nondeterministic algorithm for finding such square roots exists for any prime. Determining square roots modulo the product of two primes is also easily accomplished if the two primes are known. For instance, if $n = pq$, p and q primes, to solve the equation $x^2 \equiv a \pmod n$, one first solves it modulo p, then modulo q and combines the solutions by use of the Chinese remainder theorem to determine the (four) solutions modulo n, assuming that the equation has solutions both modulo p and modulo q. Thus determining solutions modulo n is a simple computation if the factorization of n is known. It is surprising, then, that the problem is equivalent to factoring n when the factorization of n is not known, since the factoring problem is a known difficult problem for integers of the assumed form. This is an example of a trapdoor one-way function.

To discuss the discrete logarithm problem (DLP) in prime fields, consider the multiplicative group of the integers modulo a large prime p, $\mathbb{F}_p^*$ and let $\alpha \in \mathbb{F}_p^*$ be a primitive element, namely, an element of order $p - 1$. The discrete logarithm problem in $\mathbb{F}_p^*$ is then

$$\text{DLP: Given } \alpha,\ p,\ \text{and } y = \alpha^x \pmod p,\ \text{find } x \pmod{p-1}$$

While the most efficient algorithm currently available to solve the DLP is an adaptation of the number field sieve

algorithm mentioned previously for factoring integers, a simpler algorithm, referred to as the *index calculus method*, is briefly discussed here. In fact, similar algorithms can be applied to the discrete logarithm problem in any algebraic structure that possesses a norm. In the case of the integers modulo a prime, $\mathbb{F}_p^*$, the first phase of the algorithm considers a *factor base*, $\mathcal{D}$, consisting of all the prime numbers less than some suitably chosen bound, often several hundred thousand. It attempts to establish a sufficient number of random relations between the discrete logarithms of primes in $\mathcal{D}$. One method, for example, might be to choose random powers of the primitive element α, say, $\alpha^u \in \mathbb{F}_p^*$, for randomly chosen u, and determine whether this integer factors in $\mathcal{D}$. If it does, an equation is obtained that relates u with the discrete logs of elements in $\mathcal{D}$. Such an integer would be called *smooth* with respect to the bound on $\mathcal{D}$. If a sufficient number of such relations are found (at least $|\mathcal{D}|$), then matrix reduction techniques should determine the logarithms of all elements in $\mathcal{D}$.

In the second phase of the algorithm, one attempts to find the log of the given element y by multiplying it successively by a random power of α (to yield, say, $w = \alpha^v y \in \mathbb{F}_p^*$) and determine whether it factors entirely over the factor base. If it does, the logarithm of w and hence of y can be found.

This simple idea has turned out to be a powerful one and is the basis of most of the current most effective algorithms for both the DLP and integer factoring. It turns out that most algorithms designed to factor integers can be modified to find discrete logarithms. The current most efficient algorithm to determine logarithms is the number field sieve and it has a conjectured complexity in both time and space of $L_n\left(\frac{1}{3}, c\right)$, where $c \approx \left(\frac{64}{9}\right)^{1/3} \approx 1.923$ the same as the integer factorization problem.

The discrete logarithm problem can take place in any Abelian group and typically, for reasons of security and efficiency, it is possible to reduce it to being in a cyclic group of prime order.

There is an additive version of the DLP, which is briefly described since it is finding increasing favor for cryptography on small devices. The set of solutions to an elliptic curve can be shown to be an Abelian group under point addition. In the case of a curve over a prime field $\mathbb{F}_p$ such a curve can, without loss of generality, be taken to be of the form

$$y^2 = x^3 + ax + b$$

and the Abelian group consists of the points (x, y) satisfying this equation, together with a point at infinity. The operation of point addition is very natural and is discussed here only for the case of $\mathbb{F}_p$.

Another type of finite field, one of characteristic two, is also popular for cryptographic applications and the equations of the curve and for point addition are slightly different for that case. The addition stems from the observation that if a straight line intersects the curve in at least two points, there is a unique third point of intersection. The situation is depicted in the Fig. 5.

For the last equation above, let $E_{a,b}(\mathbb{F}_p) = \{(x, y) \mid y^2 = x^3 + ax + b\}$ and let $P_1 = (x_1, y_1)$, $-P_1 = (x_1, -y_1)$ and $P_3 = (x_3, y_3) = P_1 + P_2$. If $P_1 \neq P_2$ define $\lambda = (y_2 - y_1)/(x_2 - x_1)$,

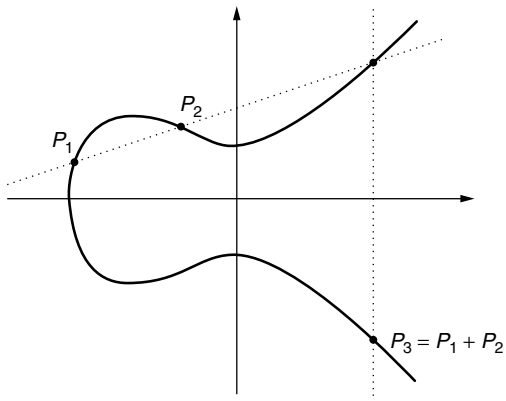


Figure 5. Elliptic curve addition.

$x_1 \neq x_2$, then $x_3 = \lambda^2 - x_1 - x_2$ and $y_3 = (x_1 - x_3)\lambda - y_1$. Similar equations hold if $P_1 = P_2$, in which case the line through the points is a tangent and the addition is referred to as a *point doubling*.

With this notion of point addition, for a given point $P = (x, y)$ one defines a general point multiple $k \cdot P = Q$ and the DLP in this setting is, given Q and P (as well as p and the curve equation), find k. This will be referred to as the ECDLP. The interesting aspect of this problem, for carefully chosen curves, is that no notion of smoothness has been found and hence we have been unable to formulate an index calculus attack on such a problem. As a consequence, the complexity of this ECDLP problem (on carefully chosen curves) is $O(\sqrt{p})$, which is considerably greater than the equivalent DLP in $\mathbb{F}_p^*$. Hence the ECDLP on an elliptic curve over $\mathbb{F}_p$ appears to be a considerably more difficult problem than the DLP in $\mathbb{F}_p^*$ for the same size prime. An implication of this is that the ECDLP on an elliptic curve over $\mathbb{F}_p$ can be used with a much smaller field size, for the same level of security as the DLP in $\mathbb{F}_p^*$.

The subject of elliptic curves has been of interest to mathematicians for over a century and is a deep and fascinating area of research. To use elliptic curves for cryptography it is necessary to know the order of the cyclic subgroup to be used. This requires determining the exact number of points on the elliptic curve and the factorization of this number, from which the order of the subgroup can be found. An important result in this direction is the Hasse–Weil theorem which says that $\#E_{a,b}(\mathbb{F}_p) \in (p + 1 \pm 2\sqrt{p})$. Once the order of the group is known, it is often straightforward to determine a generating point for the cyclic subgroup to be used.

In the following sections the three problems noted here will be used in various cryptographic protocols. The next section discusses the important problem of key exchange.

4. KEY EXCHANGE

In a network environment with a large user community, the use of a symmetric key cryptosystem such as AES described earlier, introduces the problem of establishing unique common keys between each pair of users. A solution to this problem, now in almost universal use in one form or another, is the Diffie–Hellman (DH) key exchange introduced in their 1976 paper [1], based on the discrete

logarithm problem. A version of the protocol in $\mathbb{F}_p$ is as follows. Given public information p, a prime, and $\alpha \in \mathbb{F}_p$, a primitive element, user A generates a random $a \in [1, p-1]$ and computes α^a which is sent to user B. User B chooses $b \in [0, p-1]$ at random and sends α^b to A. Both parties are now able to compute α^{ab}, which is now used to form the common key. Typically this might be used to encrypt larger messages with a faster symmetric key cryptosystem.

An eavesdropper observing the information transmissions in this protocol sees α^a, α^b and would like to compute α^{ab} and this is referred to as the *Diffie–Hellman problem* (DHP). Certainly if one is able to compute discrete logarithms in $\mathbb{F}_p$, then this can be accomplished. The more general question as to whether the DHP and DLP are computationally equivalent problems remains open, although some progress has been reported in special cases.

It is clear that the DH protocol can take place in any group, for example, the additive group of the points on an elliptic curve. In this structure the DH protocol (ECDH) is, for a fixed curve and point P on the curve of known order, all public information, user A transmits aP to A and receives bP from which the common key abP is computed by both users. The previous comments on the security of the system, specifically, the complexity of the ECDLP problem, as compared to the DLP in $\mathbb{F}_p^*$, apply to this case as well. For security reasons, the cyclic group is always taken to have prime order since any small factors of the group order tend to weaken the system for the given bit length.

A weakness of the DH key exchange protocol is the *person-in-the-middle attack*, where a user E inserts herself in the middle; users A and B believe they are exchanging keys with each other but in fact are each exchanging keys with E in the middle. The remedy for such an attack is to use an *authenticated key exchange*, where some public and private information as exists with a public key system, discussed in the next section, is incorporated in the key exchange to authenticate the user identity.

The key management problem extends far beyond the simple key exchange protocol noted above. In a large user community of users, the use, distribution, and management of keys is a critical aspect of system operation. Questions of the privileges associated with keys, the revocation and updating of keys, the generation and storage of keys, and so on are all vital aspects of the problem.

5. ASYMMETRIC OR PUBLIC KEY CRYPTOGRAPHY

The notions of public key cryptography, where the encrypting and decrypting keys are different, were introduced in the landmark paper of Diffie and Hellman mentioned previously. The paper included applications as to what might be achieved with such an asymmetric system under certain assumptions. These included the notion of one-way authentication and digital signatures as well as the use of the discrete logarithm for DH key exchange discussed earlier.

Apart from the results in that paper, perhaps the most important public key system is the RSA system. To introduce this system let $\mathbb{Z}_n$ denote the set of integers modulo the positive integer n, $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$. Let $\mathbb{Z}_n^*$ denote the set of invertible elements in $\mathbb{Z}_n$, namely, those elements in $\mathbb{Z}_n$ that have a gcd with n of 1:

$$\mathbb{Z}_n^* = \{x \in \mathbb{Z}_n \mid \gcd(x, n) = 1\}$$

Such a set forms a multiplicative group, the group of units of $\mathbb{Z}_n$ and has order $\phi(n)$, where $\phi(\cdot)$ is the Euler phi function. For an arbitrary positive integer n with prime factorization $n = \prod_i p_i^{e_i}$, e_i a positive integer, $\phi(n) = \prod_i p_i^{e_i - 1}(p_i - 1)$. Euler's theorem then states that:

$$a^{\phi(n)} \equiv 1 \pmod{n}, \qquad a \in \mathbb{Z}_n^*$$

It is actually true that $a^r \equiv a^s \pmod{n}$ for all $r \equiv s \pmod{\phi(n)}$ and all $a \in \mathbb{Z}_n$. The case of interest for the RSA system is when $n = pq$, where p and q are large primes on the order of $\sqrt{n}$. An encryption exponent $e \in \mathbb{Z}_{\phi(n)}^*$ is chosen and a decryption exponent d such that $ed \equiv 1 \pmod{\phi(n)}$ is computed. A message m is interpreted as an element in $\mathbb{Z}_n$ via a suitable embedding from the message space to the integers. The encryption of m is then

$$c \equiv m^e \pmod{n}$$

By Euler's theorem the decryption of c is

$$c^d \equiv m^{ed} \equiv m^{k\phi(n)+1} \equiv m \pmod{n}$$

for some integer k, and the message m is recovered. To use such a system, user A places in a public directory the information n_A, e_A, their RSA modulus and encrypting modulus, respectively. The user retains as secret information the factorization of n_A, p_A, q_A, and the decryption exponent d_A. Another user wishing to send a message to user A, sends $c_A \equiv m^{e_A} \pmod{n_A}$. As only user A knows the decryption exponent d_A, only they are able to decrypt the message.

The security of this system is believed, in general, to be equivalent to the difficulty in factoring the modulus n. Certainly if the decryption exponent can be found, the system is broken. If one is able to find $\phi(n)$, one can determine the decryption exponent and, indeed, factor the modulus. The system is an example of a trapdoor one-way function in that determining m from c is computationally infeasible for a properly chosen modulus and encryption exponent, but knowing the factorization of the modulus makes it easy.

Another public key encryption system based on the difficulty of finding square-roots modulus a composite number is the Rabin encryption scheme. Again, assume that user A publishes a modulus $n_A = p_A q_A, p_A, q_A$, primes, where n_A is placed in a public directory with user A's identity and the factorization is kept secret. Another user wishing to send $m \in \mathbb{Z}_n^*$ to A sends $c \equiv m^2 \pmod{n_A}$. User A, on receiving c, computes the square roots of c modulo the factors p_A and q_A, which is a simple task. The (four) square roots of $c \pmod{n}$ are then found by computing the square roots modulo p_A and modulo q_A and combining them via the CRT noted earlier. The correct square root is then found by context (it is unlikely that more than one of the square roots makes sense), or else some prearranged type

of message padding is used prior to encryption. As noted previously, the operation of finding modular square roots is a trapdoor one-way function in that finding square roots modulo n, a product of two primes, is formally equivalent to factoring if the factorization is not known, yet simple if the factorization is known.

The final public key encryption system discussed is that of ElGamal [2], based on the DLP. The DLP is a one-way function with no trapdoor and a little more work is needed to make a cryptosystem (i.e., an encryption system) out of it. User A's public key consists of a prime p_A, a primitive element α (or else a generator of a prime order cyclic subgroup) in $\mathbb{F}_{pA}$, and an element $y = \alpha^a \pmod{p_A}$ for a secret exponent a. For another user to send to user A an encrypted message m, assuming an element of $\mathbb{F}_p^*$, they first choose a random (one-time) integer $x \in [1, p-2]$ (where $[1, p-2]$ denotes the range of integers $\{1, 2, \ldots, p-2\}$) and compute $\gamma \equiv \alpha^k \pmod{p_A}$ and $\delta \equiv my^k \pmod{p_A}$. The ciphertext is then the pair of elements (γ, δ). Since user A knows the secret exponent a, they are able to compute $\gamma^{-a} \equiv \alpha^{-ak} \pmod{p_A}$ and then $\delta\alpha^{-ak} \equiv m \pmod{p_A}$. Notice the system has message expansion since two elements are passed to convey one message element.

These public key systems have uses well beyond the encryption functions outlined here. For example, while the Diffie–Hellman key exchange of the previous section requires two passes, with the notion of a public directory containing user public keys, a user can transport the key to be used in a symmetric system by encrypting with the intended users public key in a single pass. Other, more symmetric, protocols use the notion of public keys for authenticated key exchange to defeat the person-in-the-middle attack noted earlier. The public key systems also have critical properties that can be used for more general authentication and the establishment of trust in a network.

6. DIGITAL SIGNATURES AND HASH FUNCTIONS

The notion of a written signature is central to transactions of all kinds, especially to financial ones. Such signatures have characteristics in terms of forgeability and detection of forgeries that are central to many legalities. The idea of a digital signature is, in a sense, even stronger in that it is composed by a machine and constructing a forgery requires the solution of a computationally infeasible problem. It has obtained a legal standing in several countries in terms of acceptability in court actions.

To illustrate the notion of a digital signature, suppose that a message m can be embedded in some unique way into $\mathbb{Z}_n$. Let n_A, e_A, and d_A be the public and private parameters of user A's RSA system. For user A to sign the message m, the quantity $s_A(m) \equiv m^{d_A} \pmod{n_A}$ is computed as A's signature for m. Any other user may verify the signature by computing $s_A(m)^{e_A} \equiv m \pmod{n_A}$, thereby recovering the message at the same time. It is clear that only user A could have created the signature and that the signature is bound with the message m; thus, for example, the signature is invalid when associated with any other message. This is referred to as a *signature with message recovery* scheme since the message is recovered in the process of authenticating the signature.

This system has two disadvantages in that it required the message to be embedded into $\mathbb{Z}_n^*$, and so had to be relatively short to achieve this. In addition, the system described above is susceptible to an *existential forgery attack*, meaning that it is possible for an adversary to produce another bit stream, which is likely not intelligible as a real message, that has the same signature. To overcome both of these problems the notion of hash functions is introduced. A *hash function* is a mapping from binary strings of arbitrary length to strings of fixed length (usually much shorter than the original length, hence providing compression)

$$h: \{0,1\}^* \to \{0,1\}^n$$

$$x \mapsto h(x)$$

with certain properties. The function should be computationally efficient to compute on long strings. It should also have the property of being one-way in the sense that for any given hash value $y = h$ it should be computationally infeasible to determine any x, such that $y = h(x)$, since that would allow forgeries, for example, when the hash function is used for signature generation. This property is often referred to as *preimage resistance*. Furthermore it should have the property that, given x and $y = h(x)$, it should be computationally infeasible to find a second value x' so that $y = h(x) = h(x')$, sometimes referred to as *second preimage resistance*. A slightly different notion is that h is said to be *collision-resistant* if it is computationally infeasible to find any two inputs x, x' that hash to the same value.

The design of hash functions with these properties requires careful attention. There are many such hash functions, but perhaps the most widely used is the SHA-1 (secure hash algorithm) (FIPS 180-1, April 1995), a modification of the earlier FIPS 180. This produces a hash of length 160 bits and is mandated for use in the DSA (digital signature algorithm, to be discussed below). A new family of hash functions is currently under consideration, to be called SHA-256, SHA-384, and SHA-512 with the suffix indicating hash size, to be FIPS 180-2, currently in draft form. These increased size hash functions will be used in an updated DSA in the future.

To return to digital signatures, for user A to sign a long message m using RSA, one might first hash the message to produce $h(m)$, embed this hash value into the integers modulo n_A, and then sign by using as the signature $h(m)^{d_A} \pmod{n_A}$. In such a system, the message is no longer recovered when verifying the signature and so must be transmitted, usually in the clear, along with the message; that is the pair $(m, s_A(h(m)))$ is produced as the signed message by A and verified by any other user. Such a scheme is referred to [5] as a *signature with appendix*, where the signature is an appendix to the message.

Other signature schemes are also of importance. In particular the use of the DLP to produce signatures originated with the work of El Gamal [2]. Since the DLP is simply a one-way function with no trapdoor, more work is required to achieve a signature. In this scheme user A generates a large prime p and finds an α that generates $\mathbb{F}_p^*$. A random integer a is chosen, $y = \alpha^a \pmod{p}$ computed,

and the information (p, α, y) made public with a kept secret. To sign a message m, a random integer k is chosen in $[1, p-2]$ such that $\gcd(k, p-1) = 1$ and let k^{-1} be its inverse modulo $p-1$. The element $r = \alpha^k \in \mathbb{F}_p^*$ is formed and the integer $s \equiv (h(m) - ar)k^{-1} \pmod{p-1}$ computed. The signature for the message m is then the pair (r, s). The scheme is clearly a signature with appendix since it requires the message m to be transmitted in order for the signature to be verified.

To verify the signature, a user retrieves the public information of user A (p, α, y) from the public directory and, using the signature (r, s) and the message m computes

$$u \equiv y^r r^s \equiv \alpha^{ar}\alpha^{ks} \equiv \alpha^{ar+k(h(m)-ar)k^{-1}} \equiv \alpha^{h(m)} \pmod{p}.$$

The hash value $h(m)$ is computed independently and if $v \equiv \alpha^{h(m)}$ is equal to u, the signature is accepted.

This scheme is the forerunner of the Digital Signature Algorithm (Standard) (FIPS 186-2, Feb. 2000) to be described now. For the DSA each user chooses a prime number p with a number of bits $512 + 64\ell$, where ℓ is an integer between 1 and 8. A second prime q with 160 bits is chosen so that $q \mid p-1$ and α is chosen as a generator of the unique subgroup $\mathcal{G}$ of order q in $\mathbb{F}_p^*$. An integer a is chosen and $y = \alpha^a \pmod{p}$. User A's public information is then (p,q,α,y), and secret information is the integer a. From this point the DSA is very similar to the El Gamal signature scheme. A random integer k is chosen in $[1, q-1]$ and $r \equiv (\alpha^k \pmod{p}) \pmod{q}$; thus the element is first taken modulo p and then modulo q. The signature for message m is then the pair (r, s), where $s \equiv k^{-1}(h(m) + ar) \pmod{q}$. The verification is very much as with El Gamal. The point of this system, first suggested by Schnorr, is that even though the cyclic group in which the signature takes place, $\mathcal{G}$ is of order q, the arithmetic in the subgroup is in $\mathbb{F}_p^*$; that is, it is modulo p arithmetic. Any attempt to break this system would also take place in modulo p arithmetic, which is more difficult since p is very much larger than q.

There is also a version of the DSA, in the FIPS 186-2, which uses elliptic curves, referred to as ECDSA. In this system, restricting attention to curves over $\mathbb{F}_p$ as before, let E be an elliptic curve and $P = (x, y)$ a point of prime order q with both the curve parameters, where point P and q represent public information. User A chooses a random integer $a \in [1, q-1]$ and computes $Q = aP$, which is user A's public key; a is maintained secret. To sign a message m a random integer k in $[1, q-1]$ is chosen and $kP = (x_1, y_1)$ computed. Let $x \equiv x_1 \pmod{q}$, recalling that $x_1 \in \mathbb{F}_p$. The signature for the message m is then (r, s), where $s \equiv k^{-1}(h(m) + xr) \pmod{q}$ and if either r or s are 0, the procedure is run again.

To verify the signature, the user obtains A's public key Q and from the accompanying cleartext message m computes $h(m)$. With r, s and $h(m)$ as integers modulo q, the quantities $u \equiv h(m)s^{-1} \pmod{q}$ and $v \equiv rs^{-1} \pmod{q}$ are computed and the point multiple $uP + vQ = s^{-1}(h(m) + rx)P = k \cdot P = (x_2, y_2)$ found. If $x_2 \equiv r \pmod{q}$, the signature is accepted.

The secret keys a and k in the El Gamal, DSA, and ECDSA schemes are referred to as "static keys" and "ephemeral keys," respectively. In particular it is important that the ephemeral keys k be chosen independently and randomly for each message to be signed.

7. AUTHENTICATION AND IDENTIFICATION

The problem of establishing trust in a network environment is central to the application of public key cryptography. The trust includes not only the integrity of a datastream but also the identity of the person communicating. A very large number of techniques are available to address these and similar problems. This section is a brief introduction to some of these.

A simple *manipulation detection code* (MDC) is typically a short appendix to the message formed by an unkeyed hash function. Its only purpose is to detect whether any changes have occurred in the message portion of the transmission, that is, to determine message integrity. If any changes have occurred then with very high probability the MDC computed at the receiver will differ from the appendix attached. A *message authentication code* (MAC) is similar except that the hash function used is keyed, thereby allowing for the verification of the sender, since it is assumed the sender and receiver are in possession of a common key and that it is computationally infeasible for anyone without the key to compute the hash appendix.

The constructions of either type of code begins with a hash function, such as SHA-1 mentioned earlier. A MAC, however, requires the incorporation of a key in some manner and experience has shown that this must be done very carefully — several such hash functions have been broken when subtle faults in their construction were found. A particular type of MAC is the *keyed hash MAC* or HMAC, currently in draft for a FIPS. It recommends the following construction, using any FIPS-approved hash function. For an input block size of B bytes to the hash function (e.g., for SHA-1 $B = 64$), construct an *ipad* of B bytes consisting of the byte (in hexadecimal) of $0 \times$ '36' and an *opad* consisting of the byte $0 \times$ '5c' each repeated B times. If K is the secret shared key, let K_0 be K with zeros appended to form a B byte key. The recommended formation of the HMAC is then the leftmost t bytes of the hash value

$$H((K_0 \oplus opad) \| H((K_0 \oplus ipad) \| text))$$

where $\|$ indicates catenation.

The subject of MACs is much wider than touched on here with many such constructions and concepts used. However, most of them share common features with the above.

Most approaches to the establishment of trust in a network involve public key concepts and the use of a *central authority* (CA) [these are also referred to as a *trusted third party* (TTP) or *trusted authority* (TA)]. The idea here is for the CA to have a public and a private key, with the public key known to all subscribers of the network. The CA establishes in some secure manner the identity of a user and their public key and binds the two together by encrypting the pair with the CA's private key. This is a simple example of a certificate — typically such certificates include other information such as level of authority

given to the individual and time limit for validity of the certificate. The crucial point here is that the CA has verified the information relating the identity of the individual and their public key and bound it together with the CA's private key, as for a signature. Anyone who has received the certificate can verify the information by use of the CA's public key and can assume that the information is as reliable as the CA. The construction of certificates and their maintenance and issuance to clients is the idea behind a *public key infrastructure* (PKI). A client of the PKI may request a certificate for any user in the system thereby assuring identity.

Many cryptographic protocols rely on the existence of a CA, although sometimes the requirements of the protocol are less than the need for a full CA. As an example, consider the Diffie–Hellman key exchange described earlier. It was noted it is susceptible to the person-in-the-middle attack where a third party injects themselves in between users A and B, who end up communicating with each other through the third party, C, who controls the flow of information. A remedy for this situation is the use of a CA where each user obtains sufficient information from the CA and the common key is established using both private and public information of both users. Such a scheme is termed an authenticated key exchange protocol, noted earlier.

Identification protocols [5] usually involve the notions of what a person knows [e.g., a PIN (personal identification number) for an ATM card], what they have (the ATM card with their account number and name on it) and a physical attribute (such as fingerprint, ocular iris characteristics, facial features, and even DNA). Two identification protocols are briefly described here for the interesting features they introduce. Both require the use of a CA although not explicitly the notions of certificates.

In the Schnorr identification protocol ([5,9]), the CA chooses a large prime q such that q is a large divisor of $p-1$ and α is an element in $\mathbb{Z}_p$ of order q. The CA has a public and private key, which allows the creation of signatures for information, $s_{CA}(\cdot)$, and the creation of certificates, $C_{CA}(\cdot)$. Each user A chooses a random integer $a \in [1, q-1]$, a private key, and computes $v \equiv \alpha^{-a} \pmod{p}$ and v is sent to the CA, who creates the certificate $C_{CA} = (I(A), v, s_{CA}(I(A), v))$. For user A to identify his/herself to B, A chooses a random number $k \in [1, q-1]$ and sends $x \equiv \alpha^k \pmod{p}$ to B, along with his/her certificate. User B verifies the certificate, which has bound the public key of A, v, to his/her identity. User B then chooses a random integer $r \in [1, 2^t]$ for some suitably large integer t, say 2^{40} (t is referred to as the security parameter of the scheme), and sends it to user A. User A sends back $y \equiv k + ar \pmod{q}$, which is sent to B. User B then verifies that $\gamma \equiv \alpha^y v^r \pmod{p}$. The protocol uses, in an essential way, the notion of certificates and a CA to establish identity in a reliable manner.

The Fiat–Shamir identification protocol introduces the notions of a *zero-knowledge proof* and a *statistical, interactive proof* which consists of multiple rounds of a three-pass challenge–response protocol. It operates as follows (basic idea only). For user A to prove his/her identity to user B, the CA first chooses the (secret) primes p and q and publishes the modulus $n = pq$. Each user chooses a secret s relatively prime to n, $s \in [1, n-1]$ and registers $v \equiv s^2 \pmod{n}$ with the CA as its public key. For A to identify his/herself to B, the following three steps are performed (and repeated t times):

$$A \to B: x \equiv r^2 \pmod{n}$$

$$A \leftarrow B: e \in \{0, 1\}$$

$$A \to B: y \equiv rs^e \pmod{n}$$

The purpose of user B choosing e in the second step of the protocol is to prevent cheating by A in the sense that if user B always chose $e = 1$, then A could compute $x \equiv r^2/v$ and answer the challenge $e = 1$ with the correct answer $y \equiv r \pmod{n}$; thus, any user knowing s can always answer either challenge while another user can only answer the one question — hence an impostor has a probability of $\frac{1}{2}$ of answering a given question. Repeating this basic protocol t times reduces the probability of success by the impostor to less than $(\frac{1}{2})^t$.

This basic protocol, which can be made much more efficient [e.g., 5,9], illustrates the two notions of a statistical interactive proof and a zero-knowledge proof. The term *zero knowledge* refers to the fact that the protocol reveals no knowledge about the factorization of the modulus n or the value of a users secret key s except for the fact that the public modulus v is in fact a square modulo n.

Many more sophisticated zero-knowledge protocols exist with the preceding example a small introduction to the area.

8. PRIME-NUMBER GENERATION AND PSEUDORANDOMNESS

Much of public key cryptography depends on the difficulty of certain number theoretic problems, such as factoring, discrete logarithms, and taking modular square roots. The setup procedure for these problems invariably involves the generation of large primes, often of several hundred digits. Many standards suggest the use of probabilistic methods to achieve this, and perhaps the most commonly used of these techniques and the most efficient is the *Miller–Rabin* method, which is briefly described [e.g., 5].

In this method, as with many others, an integer n is chosen at random and tested for primality. By the *prime-number theorem*, if $\pi(x)$ is the number of primes less than x, then for large values of x this quantity is well approximated by $x/\ln x$. This gives an estimate of the number of trials that must be made to locate a prime; thus, in the neighborhood of n the average spacing between primes is approximately $\ln(n)$. To find a *probable prime* choose an integer n at random. It is often convenient to trial-divide n for a stored table of small primes to avoid work on integers that are thus easily dismissed. Let $n - 1 = 2^s r$, where $2 \nmid r$. Let a be a randomly chosen integer in $[1, n-2]$ and compute $y \equiv a^r \pmod{n}$. If $y \equiv \pm 1 \pmod{n}$, declare n to be prime. Otherwise if $y \not\equiv \pm 1 \pmod{n}$, then successively square $y \pmod{n}$ up to $s-1$ times. If the result of the squaring never results in $y \equiv \pm 1 \pmod{n}$, then n is declared composite. The procedure rests on the observation that if p is a prime, then for $p - 1 = 2^s r$, r odd, either $a^r \equiv 1 \pmod{n}$ or $a^{2^j r} \equiv -1 \pmod{n}$ for some $j \in [1, s-1]$.

An integer a that fails the test proves conclusively that the integer n is composite and is called a "witness" to the compositeness of n. It can be shown that the probability the test declares a composite number to be prime on any given trial is less than $\frac{1}{4}$. If the test is run t times the probability the test always declares a composite number a prime is less than $(1/4)^t$. Thus for $t = 40$, this probability is $1/2^{80}$, which is usually the recommended level of certainty in standards that suggest the use of this technique, for example, the DSA (FIPS 186-2). Primes produced with this test are termed *probable primes*.

The Miller–Rabin test produces a probably prime with a probability of failing so low that it is regarded as reliable and finds wide use in practice. Maurer [e.g., 5] has devised a test, however, that produces a *provable* prime in almost the time it takes to run one iteration of the Miller–Rabin test. It is based on a modification of Pocklington's theorem of computational number theory. While Maurer's technique on provable primes removes the uncertainty of primality inherent in the Miller–Rabin test, this latter test is still widely used, and with the number of iterations proposed for this test, it yields very acceptable results.

Another frequently needed facility in any cryptographic application is the ability to generate random or pseudorandom numbers. These are used to initialize many functions. Indeed, the security of a system might often be equated to the degree of uncertainty in the random seed. While pure random generators are to be preferred, these are usually only found in hardware devices and chips where the outputs of so-called noisy diodes are often used. Such systems invariably include self-checking devices to ensure that the system has not failed and is still producing sequences that pass recognized statistical tests. In software, it is desirable to use as many different sources of randomness available, such as time between keystrokes, and to put the catenation of such information through a hash function to produce a random seed for a random-number generator. This is still often the weakest point of security of the system. Many systems in use are proprietary, making an assessment of their security difficult.

A commonly used technique to produce pseudorandom numbers is the linear congruential technique. Typically these produce sequences of numbers from an initial seed x_0, often kept secret, with the recursion

$$x_n \equiv ax_{n-1} + b \pmod{m}$$

where a and b are fixed parameters and m is a modulus. Sequences derived from such a system are in fact predictable in that given a part of the sequence, the parameters can often be derived and the entire sequence generated. However, variations on these sequences are sometimes used. Typically it is required that the seed value be sufficiently large that an exhaustive search over all possibilities be infeasible. In addition it is required that the sequence pass statistical properties of randomness in that failure to do so may lead to ways of breaking the sequence.

Other pseudorandom bit generators can be formulated that depend on the use of one-way functions, for example, the Blum–Blum–Shub generator, which depends on modular square roots for its security. While very secure, such generators tend to be very inefficient.

9. COMMENTS

Current cryptographic research goes beyond the standard techniques discussed here and is concerned with applications that can be achieved with cryptographic techniques. A few examples of these are given. One such area is that of zero-knowledge proofs, noted briefly previously, where one user proves to a verifier that they are in possession of knowledge and proves to another that this is so without revealing any information. Protocols exist showing, for example, that a prover knows the discrete logarithm of a given element $y \in \mathbb{F}_p^*$ to some known base α, p also known, without revealing what the logarithm is. Similarly, it is possible to prove to a verifier that a given RSA modulus is indeed a product of two primes without revealing the factorization. Another such protocol shows how to compute an RSA modulus and an encryption exponent, among a number of parties without any of the parties knowing the factorization. At the same time each party obtains a share of the decryption exponent, which must be combined with other user shares to either create a signature or decrypt an encryption. Such a protocol is an example of a secure multiparty distributed computation, and many other such examples exist.

Electronic cash is a payment technique whereby a user is able to spend electronic cash obtained from a bank and spent at a store. The protocol preserves anonymity from the bank as to how the e-cash was spent and prevents double-spending of it by either the individual or the merchant. It involves the notion of a blind signature that also arises in the context of electronic voting. A distributed voting scheme has among its requirements the anonymity, the ability for any participant to verify the tally and the ability to ensure that no one votes more than once.

An example of the key exclusion problem is where a pay-TV provider encrypts a movie and provides sufficient information for those users who paid to decrypt the movie but those who did not pay cannot. Of course, the set of users who view each movie changes frequently, and the challenge is to derive an efficient scheme to achieve this. There is also a problem of "traitor tracing," whereby users who collaborate to form new valid keys can be detected and identified. The research on protocols shows a power and elegance that can be exploited to achieve interesting goals.

Perhaps the single topic of greatest interest to public key cryptography is that of quantum computation. While a description of the idea of a quantum computer is beyond the scope of this article, a few of the implications of the existence of such a computer might be mentioned, and for this the article by Gottesman and Lo [4] is used as a guide. The state of a quantum computer, rather than being a deterministic state as in a classical computer, is actually a superposition of exponentially many basis states, and each of these corresponds to a state of a classical computer of the same size. The result of this is that such a computer would take a very small amount of time to do tasks that are not possible to contemplate on a classical computer. In a celebrated result, Peter Shor of AT&T Laboratories has shown that the two problems of central interest in

cryptography, integer factoring and discrete logarithms, have a complexity in the quantum computing model that is polynomial in the integer lengths, rather than the subexponential complexity, $L(\frac{1}{3}, c)$, mentioned earlier, for the conventional computing model. In a similarly spectacular result, Grover of Bell Laboratories (Lucent Technologies), has shown that a searching algorithm to find one of N items has a classical complexity of order N but a quantum algorithm of complexity of order $\sqrt{N}$. This fact could, for example, be used to search for keys for a block cipher. While a quantum computer has not yet been constructed, informed opinion considers them likely and indeed several corporations are in the process of attempting to construct them. The implications for cryptography are quite clear. Without the notions of one-way functions, the field would have to reconstitute itself in a very different direction.

In spite of the results noted above, it has also been shown that quantum key distribution using quantum systems is entirely practical, and indeed such systems have been successfully constructed and demonstrated by several research groups around the world. While such systems have been shown to be theoretically secure, it is not yet clear how they might withstand practical attacks.

Other aspects of quantum cryptography have been investigated, and some, like quantum bit commitment, have been shown to be insecure. It is clear that quantum computing will have an important impact on cryptography although at this point it is not quite clear precisely what this impact will be.

BIOGRAPHY

Ian F. Blake received his undergraduate education at Queen's University in Kingston, Ontario and his Ph.D. at Princeton University in New Jersey. From 1967 to 1969 he was a Research Associate with the Jet Propulsion Laboratories in Pasadena, California. From 1969 to 1996 he was with the Department of Electrical and Computer Engineering at the University of Waterloo, in Waterloo, Ontario, where he was Chairman from 1978 to 1984 and Director of the Institute of Computer Research from 1990 to 1994. He is currently with the Department of Electrical and Computer Engineering at the University of Toronto, where he is Director of the Bell University Labs program.

His research interests are in the areas of cryptography, algebraic coding theory, digital communications, and spread-spectrum systems.

BIBLIOGRAPHY

1. W. Diffie and M. Hellman, New directions in Cryptography, *IEEE Trans. Inform. Theory* **22**: 644–654 (1976).
2. T. El Gamal, A public key cryptosystem and signature scheme based on discrete logarithms, *IEEE Trans. Inform. Theory* **31**: 469–472 (1985).
3. D. Kahn, *The Codebreakers*, Macmillan New York, 1967.
4. D. Gottesman and H.-K. Lo, From quantum cheating to quantum security, *Physics Today* **53**(11): 22–27 (Nov. 2000).
5. A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, FL, 1996.
6. R. L. Rivest, A. Shamir, and L. M. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Commun. ACM* **21**: 120–126 (1978).
7. B. Schneier, *Applied Cryptography*, 2nd, ed., Wiley, New York, 1996.
8. C. E. Shannon, Communication theory of secrecy systems, *Bell Syst. Tech. J.* **28**: 656–715 (1949).
9. D. Stinson, *Cryptography: Theory and Practice*, CRC Press, Boca Raton, FL, 2002.

CYCLIC CODES

STEPHEN B. WICKER
Cornell University
Ithaca, New York

1. INTRODUCTION

Cyclic codes are a class of highly structured algebraic block error control codes. This class includes Golay, BCH (Bose–Chaudhuri–Hocquenghem), and Reed–Solomon codes — codes that have arguably seen greater application than any other error control codes save the simple parity check. A Reed–Solomon decoder, for example, can be found in every compact-disk player. Golay, BCH, and Reed–Solomon codes have also been used in paging, mobile data systems, and deep-space telecommunications. Such cyclic codes have become an important basis for practical error control for a variety of reasons. First, they can be implemented using shift-register-based encoders and decoders, implementations that are of particular importance in high-speed data communication systems. Also, nonbinary cyclic codes, such as Reed–Solomon codes, provide a form of error trapping that is highly effective on fading channels. For this reason alone, Reed–Solomon codes have been used extensively in wireless data applications. In this article I will describe the structure of cyclic codes, paying particular attention to the special cases of Golay, BCH, and Reed–Solomon codes. I will mention several theoretical results, but will provide no proofs — the mathematically inclined reader is referred to Wicker [1] or Lin and Costello [2]. The true disciple is referred to MacWilliams and Sloane [3].

I begin with an historical overview. The general class of cyclic codes was first discussed in a series of technical notes and reports published from 1957 to 1959 by E. Prange at the Air Force Cambridge Research Labs [4–6]. In his work Prange identified cyclic codes with an algebraic structure called an "ideal," a connection that would lead to the development of BCH codes and a reinterpretation of Golay and Reed–Solomon codes a few years later.

Prange himself introduced a class of cyclic codes whose construction is based on quadratic residues. Quadratic residue (QR) codes are linear cyclic codes that generally have rates close to $\frac{1}{2}$ and have large minimum distances. A great deal of recent work on algebraic decoding algorithms for quadratic residue codes has increased their utility in a number of applications [e.g., 7,8].

Prange's work on QR codes showed that they include a small but very important group of codes — the Golay codes — that had been discovered almost 10 years earlier. In his 1948 paper that, among other things, gave birth to the field of information theory, Shannon gave a brief description of Hamming's perfect [7,4] binary code [9]. A code is said to be perfect if it has the maximum possible number of code words for a given error-correcting capability. After reading Shannon's paper, a number of people began searching for other perfect codes, with varying degrees of success. Golay, an engineer at the Signal Corps Engineering Laboratories in Fort Monmouth, New Jersey, published one of the first follow-up papers in June, 1949 [10]. In what is almost certainly the best short paper ever written (it fits on one side of an $8\frac{1}{2} \times 11$-in. sheet of paper, with room to spare), Golay extended the (7,4) Hamming code to a general class of p-ary codes of length $(p^n - 1)/(p - 1)$, where p is a prime.[1] In this same paper, Golay went on to describe a binary triple-error-correcting code and a ternary double-error-correcting code, both of which are perfect. Golay deduced the existence of the Golay codes through an examination of Pascal's triangle and the recognition of the relationship between the triangle's entries and the parameters of perfect codes. He then proceeded to find what we now call the Golay codes through a "limited search" of the triangle. The resulting code has served as fodder for specialists in abstract algebra and combinatorics ever since. On the practical side, Golay codes have seen frequent application in the United States space program, most notably with the *Voyager* I and II spacecraft. The extended binary Golay code served as the primary *Voyager* error control system, providing clear color pictures of Jupiter and Saturn between 1979 and 1981.[2] Golay codes were also used as the basis for a paging standard (called, not surprisingly, Golay).

Several efficient decoding algorithms for the Golay codes have been discovered. The most important was discovered in 1964, when Kasami described a shift register-based "error trapping" decoder [11]. The error trapping decoder is an extension of the shift register decoding techniques that will be discussed in this article.

The next subclass of cyclic codes that I will consider in detail is the BCH codes. Two independent research teams conducted the fundamental work on BCH codes and published their results at roughly the same time. A. Hocquenghem discussed binary BCH codes as "a generalization of Hamming's work" in a 1959 paper entitled "Codes correcteur d'erreurs" [12]. This was followed in March and September 1960 by Bose and Ray-Chaudhuri's publications on error-correcting binary group codes [13,14]. Given their simultaneous discovery of these codes, all three have given their name to what are now called BCH codes. Shortly after these initial publications, Peterson proved that BCH codes were cyclic and presented a moderately efficient decoding algorithm [15]. Gorenstein and Zierler then extended BCH codes from the binary field to arbitrary fields of size p^m, where p is a prime number [16].

Reed-Solomon codes were first described in a June 1960 paper entitled "Polynomial codes over certain finite fields," published in the *SIAM Journal on Applied Mathematics* by Irving Reed and Gus Solomon [17]. Through the work of Gorenstein and Zierler it was later discovered that Reed–Solomon codes and BCH codes are closely related, and that Reed–Solomon codes can be described as nonbinary BCH codes.

In 1960 Peterson provided the first explicit description of a decoding algorithm for binary BCH codes [15]. His "direct solution" algorithm is quite useful for correcting small numbers of errors, but becomes computationally intractable as the number of errors increases. Reed and Solomon discussed a decoding algorithm in their original paper on Reed–Solomon codes [17], but that algorithm was also inefficient for large codes and large numbers of errors corrected. Peterson's algorithm was improved and extended to nonbinary codes by Gorenstein and Zierler [16], Chien [21], and Forney [22], but it was not until 1967 that Berlekamp introduced the first truly efficient decoding algorithm for both binary and nonbinary BCH codes [18]. In 1969 Massey showed that Berlekamp's algorithm was a general solution to the problem of synthesizing the shortest linear feedback shift register capable of generating a given sequence [23]. Massey then demonstrated a fast shift-register-based decoding algorithm for BCH and Reed–Solomon codes that is equivalent to Berlekamp's algorithm.

In 1975 Sugiyama et al. showed that Euclid's algorithm can be used to decode BCH and Reed–Solomon codes [24]. Reed et al. then showed in 1978 that a related technique based on continued fractions and Fermat-theoretic transforms resulted in a fast decoding algorithm for Reed–Solomon codes [25]. Completing the decoding picture, a frequency domain approach to decoding BCH codes was introduced by Gore in 1973 [26]. Blahut provided a more general discussion of spectral decoding techniques in 1979 [27].

In the remainder of this article, I will provide a brief overview of the general theory of cyclic codes, and then turn to the three important classes (the Golay codes, BCH, and Reed–Solomon codes). I will close with a discussion of several important applications.

2. GENERAL THEORY

A code is said to be cyclic if it satisfies a very simple property–every codeword must be the right cyclic shift of another codeword. More formally, an (n, k)[3] code $\boldsymbol{C}$ is said to be cyclic if for every codeword $\mathbf{c} = (c_0, c_1, \ldots, c_{n-2}, c_{n-1}) \in \boldsymbol{C}$, there is also a codeword $\mathbf{c}' = (c_{n-1}, c_0, \ldots, c_{n-3}, c_{n-2}) \in \boldsymbol{C}$. Since the codeword $\mathbf{c}$ in this definition has been

[1] It is a virtual certainty that this result was already known to Hamming [18]. A more detailed discussion of the general class of codes now known as *Hamming codes* is included in Hamming's 1950 paper [19], which was probably delayed due to patent considerations.

[2] A secondary error control system based on Reed–Solomon codes was substituted for the Golay system for the *Voyager* 2 flybys of Uranus and Neptune in the mid-1980s [20].

[3] In this article the usual conventions are adopted — an (n, k) code over GF(q) is a collection of vectors of length n that form a vector space of dimension k over the Galois field GF(q).

arbitrarily selected from among all the codewords in $\boldsymbol{C}$, it follows that all n of the distinct cyclic shifts of $\mathbf{c}$ must also be codewords in $\boldsymbol{C}$. To see this, replace $\mathbf{c}$ with $\mathbf{c}'$ and apply the definition again.

The key to the underlying structure of cyclic codes lies in the association of a *code polynomial* $c(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{n-1}x^{n-1}$ with every codeword $\mathbf{c} = (c_0, c_1, \ldots, c_{n-2}, c_{n-1}) \in \boldsymbol{C}$. If C is an (n, k) code over Galios Field GF(q), it follows that the set of code polynomials associated with $\boldsymbol{C}$ form a vector space as well. The terms *codeword* and *code polynomial* are henceforth used interchangeably. When we look at the definition of cyclic in terms of the code polynomials, some interesting structure comes to light. If the code word $\mathbf{c}'$ is the right cyclic shift of the code word $\mathbf{c} \in \boldsymbol{C}$, then $c'(x) = x \cdot c(x)$ modulo $(x^n - 1) \in \boldsymbol{C}$. This can be seen as follows.

Continuing along this line, we can show that two right cyclic shifts of a code word are equivalent to the multiplication modulo $(x^n - 1)$ of the associated code polynomial by x^2

$$\begin{aligned} x \cdot c(x) \bmod (x^n - 1) &= (c_0x + c_1x^2 + \cdots + c_{n-1}x^n) \bmod (x^n - 1) \\ &\equiv c_{n-1} + c_0x + \cdots + c_{n-2}x^{n-1} \\ &= c'(x) \end{aligned}$$

where three right cyclic shifts are equivalent to multiplication modulo $(x^n - 1)$ by x^3, and so on. Let the cyclic shifts of $\mathbf{c}$ and the associated polynomials be represented as follows.

$$\begin{aligned} \mathbf{c} &= (c_0, c_1, \ldots, c_{n-1}) \leftrightarrow c(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1} \\ \mathbf{c}' &= (c_{n-1}, c_0, \ldots, c_{n-2}) \\ &\leftrightarrow c'(x) = c_{n-1} + c_0x + \cdots + c_{n-2}x^{n-1} \\ \mathbf{c}'' &= (c_{n-1}, c_0, \ldots, c_{n-2}) \\ &\leftrightarrow c''(x) = c_{n-2} + c_{n-1}x + \cdots + c_{n-3}x^{n-1} \\ &\vdots \\ \mathbf{c}^{(n-1)} &= (c_{n-1}, c_0, \ldots, c_{n-2}) \\ &\leftrightarrow c^{(n-1)}(x) = c_1 + c_2x + \cdots + c_0x^{n-1} \end{aligned}$$

Let $a(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}$ be an arbitrary polynomial in GF(q)[x]/$(x^n - 1)$ [the ring of polynomials with coefficients in GF(q) and maximum degree $n - 1$]. The product $a(x)c(x)$ is a linear combination of cyclic shifts of $\mathbf{c}$. Since $\boldsymbol{C}$ forms a vector space, $a(x)c(x)$ must be a valid code polynomial. This result shows that the space formed by the code polynomials of $\boldsymbol{C}$ has an interesting structure — it is an *ideal* in the ring of polynomials of degree n or less with coefficients in GF(q) (normally written as GF(q)[x]/$(x^n - 1)$). Several interesting, practical properties follow immediately from this result.

2.1. The Basic Properties of Cyclic Codes

Let $\boldsymbol{C}$ be a (n, k) linear cyclic code over GF(q):

- Within the set of code polynomials in $\boldsymbol{C}$ there is a unique monic polynomial $g(x)$ with minimal degree $r < n$. $g(x)$ is called the *generator polynomial* of $\boldsymbol{C}$.
- Every code polynomial $c(x)$ in $\boldsymbol{C}$ can be expressed uniquely as $c(x) = m(x)g(x)$, where $g(x)$ is the generator polynomial of $\boldsymbol{C}$ and $m(x)$ is a polynomial of degree less than $(n - r)$ in GF(q)[x].
- The generator polynomial $g(x)$ of $\boldsymbol{C}$ is a factor of $(x^n - 1)$ in GF(q)[x].

The proof for these statements follows from the definition of ideal; the interested reader is referred to Wicker [1]. or MacWilliams [3]. The last of the three properties leads to some very interesting structural issues for cyclic codes, but to see this, we have to introduce some additional abstract algebra.

Let β be an element in the Galois field GF(q^m). The *conjugates of β with respect to the subfield GF(q)* are the elements $\beta, \beta^q, \beta^{q2}, \beta^{q3}$, and so on. Note that, since the field is finite, this series of elements has to start repeating at some point. These conjugates form a set called a *conjugacy class*. The elements in a Galois field can be partitioned into conjugacy classes with respect to a subfield of the Galois field. For example, the elements in the field GF(8) can be partitioned into conjugacy classes with respect to GF(2) in the following way. Let α be a primitive element in GF(8) (a "primitive" element is an element whose powers generate all the nonzero elements in the field). The various powers of α fall into three conjugacy classes with respect to GF(2): $\{\alpha^0 = 1\}$, $\{\alpha, \alpha^2, \alpha^4\}$, and $\{\alpha^3, \alpha^5, \alpha^6\}$. The fourth and final conjugacy class is $\{0\}$.

Rather than deal with a given conjugacy class itself, it is often easier to focus on the exponents of the powers of α that form the conjugacy class. The result is a *cyclotomic coset*. The cyclotomic cosets associated with the nonzero conjugacy classes in the above example are $\{0\}$, $\{1, 2, 4\}$, and $\{3, 5, 6\}$. If we want to include the zero element, we can adopt a symbol, say, an asterisk $\{^*\}$, to formally represent $\log_\alpha 0$, but this will not be necessary for what follows. If we ignore the zero element, the cyclotomic cosets modulo n with respect to GF(q) are a partitioning of the integers $\{0, 1, \ldots, n - 1\}$ into sets of the form $\{a, aq, aq^2, aq^3, \ldots, aq^{d-1}\}$.

With a bit of effort, the preceding development and some Galois field mathematics yield the following key result: the roots of a polynomial with coefficients in GF(q) must be the union of one or more conjugacy classes with respect to GF(q) (see, e.g., Wicker [1]). To see what this means, consider a generator polynomial $g(x)$ for a binary cyclic code of length 7. According to our properties of cyclic codes, $g(x)$ must be a divisor of $x^7 - 1$. The key result says that any binary polynomial that is a divisor of $x^7 - 1$ must have roots that are the union of one of the conjugacy classes with respect to GF(2) that we listed above. Since zero is not a root of $x^7 - 1$, we have to focus on the three nonzero conjugacy classes. Each of these classes is associated with a *minimal polynomial*, a polynomial whose roots are the elements of a single conjugacy class.

Conjugacy Class	Associated Minimal Polynomial
$\{\alpha^0 = 1\}$	$M_0(x) = (x - 1) = x + 1$
$\{\alpha, \alpha^2, \alpha\}$	$M_1(x) = (x - \alpha)(x - \alpha^2)(x - \alpha^4)$ $= x^3 + x + 1$
$\{\alpha^3, \alpha^6, \alpha^5\}$	$M_3(x) = (x - \alpha^3)(x - \alpha^6)(x - \alpha^5)$ $= x^3 + x^2 + 1$

We conclude that the only generator polynomials for binary, cyclic codes of length 7 are $M_0(x)$, $M_1(x)$, $M_3(x)$, $M_0(x)M_1(x)$, $M_0(x)M_1(x)$, and $M_1(x)M_3(x)$. In general, the binary polynomials that are available for use as generator polynomials for cyclic codes are the products of one or more minimal polynomials.

2.2. Systematic Encoding

An encoding for a given error control code is said to be *systematic* if the codeword consists of a set of parity symbols followed by the message itself. This greatly facilitates decoding, for if a parity check indicates that there are no errors in the received word, the parity symbols can be deleted and the remaining message forwarded to the application. By exploiting the algebraic structure of cyclic codes, it is possible to develop a simple systematic encoder.

Consider an (n, k) cyclic code $\boldsymbol{C}$ with generator polynomial $g(x)$. A k-symbol message $\mathbf{m} = (m_0, m_1, \ldots, m_{k-1})$ is encoded as follows. Multiply the corresponding message polynomial $m(x)$ by x^{n-k}, obtaining $x^{n-k}m(x) = m_0x^{n-k} + m_1x^{n-k+1} + \cdots + m_{k-1}x^{n-1}$. This product is associated with an n-symbol block $(0, 0, \ldots, 0, m_0, m_1, \ldots, m_{k-1})$ whose first $(n-k)$ coordinates are zero. Now divide $x^{n-k}m(x)$ by $g(x)$ to obtain $x^{n-k}m(x) = q(x)g(x) + d(x)$, where $d(x)$ is the remainder. Since $c(x) = [x^{n-k}m(x) - d(x)] = q(x)g(x)$ is a multiple of $g(x)$, it must be a valid code polynomial. Now note that the remainder $d(x)$ has a degree less than $(n-k)$, the degree of the generator polynomial $g(x)$. The term $-d(x)$ can be associated with an n-symbol block whose last k coordinates are zero: $-d(x) \leftrightarrow (-d_0, -d_1, \ldots, -d_{n-k-1}, 0, 0, \ldots, 0)$. The codeword associated with the code polynomial $c(x) = [x^{n-k}m(x) - d(x)]$ thus has the form

$$c(x) = [x^{n-k}m(x) - d(x)] \leftrightarrow (-d_0, -d_1, \ldots, -d_{n-k-1}, m_0, m_1, \ldots, m_{k-1})$$

where $m(x)$ has been systematically mapped to a codeword. The encoding algorithm is summarized below:

1. Multiply the message polynomial $m(x)$ by x^{n-k}.
2. Divide the result of step 1 by the generator polynomial $g(x)$. Let $d(x)$ be the remainder.
3. Set $c(x) = x^{n-k}m(x) - d(x)$.

As an example, consider the systematic encoding of a (7,3) binary code with generator polynomial $g(x) = x^4 + x^3 + x^2 + 1$; we will encode the message block (101):

1. $x^{n-k}m(x) = x^4(x^2 + 1) = x^6 + x^4$
2. $$\begin{array}{r|l} x^4+x^3+x^2+1 & x^2+x+1 = q(x) \\ & \underline{x^6+x^4} \\ & \underline{x^6+x^5+x^4+x^2} \\ & \underline{x^5+x^2} \\ & \underline{x^5+x^4+x^3+x} \\ & \underline{x^4+x^3+x^2+1} \\ & x+1 = d(x) \end{array}$$
3. $c_m(x) = x^{n-k}m(x) - d(x) = 1 + x + x^4 + x^6$ $\leftrightarrow \boldsymbol{c}_m = (1100101)$

2.3. Shift Register Encoders and Decoders for Cyclic Codes

Data rates in the hundreds or even thousands of megabits per second are common in many applications. Unfortunately, such data rates severely limit the device technologies that can be used to implement error control systems, and within a given technology, the complexity of the circuits. It is extremely important to note that encoders and decoders for cyclic codes can be implemented using simple exclusive-OR gates, switches, shift registers, and, in the case of nonbinary encoders and decoders, finite-field adder and multiplier circuits. Shift registers are among the simplest of digital circuits, consisting of a collection of flipflops connected in series. They are operable at speeds quite close to the maximum speed possible for a single gate using a given device technology.

The systematic encoding procedure described above has a simple shift register implementation, as shown in Fig. 1. The first step — multiplication of the message polynomial — is implemented by inserting the message into the shift register circuit on the right side [this is equivalent to padding the message block with an initial $(n-k)$ zeros]. Polynomial division is then performed through the use of a linear feedback shift register (LFSR).

Encoding proceeds as follows. During the first step of the encoding operation the three switches are placed in position X and the k message symbols are fed into the

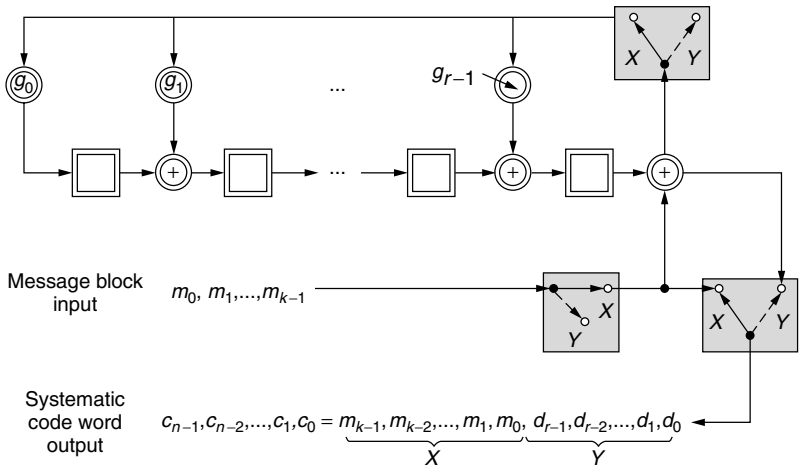


Figure 1. Systematic shift register encoding circuit for cyclic codes [1].

encoder in order of decreasing index. The k symbol bits are simultaneously sent to the transmitter, for they represent the last k coordinates of the systematic codeword. This, by the way, is the primary rationale for placing the message symbols at the end of a systematic code word. After the kth message symbol has been fed into the shift register, the switches are moved to position Y. At this point the shift register cells contain the remainder generated by the division operation. These symbols are then shifted out of the shift register and to the transmitter, where they constitute the remaining systematic codeword coordinates.

Cyclic codes allow for a number of highly convenient techniques for detecting errors using shift register circuits. For example, consider the following:

1. Since the transmitted codeword was systematically encoded, we can construct an estimated message block $\mathbf{m}'$ and estimated remainder block $\mathbf{d}'$ using the values in the message and parity positions of the received word $\mathbf{r}$.
2. Encode $\mathbf{m}'$ using an encoder identical to that used by the transmitter and obtain an estimated remainder block $\mathbf{d}''$.
3. Compare $\mathbf{d}'$ to $\mathbf{d}''$. If they are not the same, then $\mathbf{r}$ is not a valid codeword, indicating the presence of errors in the received word.

This approach to error detection has a significant advantage. The encoder and error detection circuits are essentially identical, and the design process is correspondingly simplified.

The error detection technique described above can be used as the first steps in an error-correcting algorithm. The difference $\mathbf{s} = \mathbf{d}' - \mathbf{d}''$ is called the *syndrome* for the received word $\mathbf{r}$. Maximum-likelihood error correction is performed by computing the syndrome for a received vector, determining the most likely error pattern among those associated with the syndrome, and subtracting this error pattern from the received vector. The primary drawback to this approach is usually the size of the syndrome lookup table. For an arbitrary (n, k) q-ary code, the syndrome table must contain $q^{n-k}n$-tuples — one for each possible syndrome. Cyclic codes have an interesting property that allows us to cut the size of the syndrome table to $1/n$th its original size.

> Let $s(x)$ be the syndrome polynomial corresponding to a received polynomial $r(x)$. Let $r^{(1)}(x)$ be the polynomial obtained by cyclically shifting the coefficients of $r(x)$ once to the right. Then the remainder obtained when dividing $xs(x)$ by $g(x)$ is the syndrome $s^{(1)}(x)$ corresponding to $r^{(1)}(x)$.

Let's consider this result in terms of the shift register syndrome circuit. Given a received vector $\mathbf{r}$, the corresponding syndrome $\mathbf{s}$ is obtained by entering $\mathbf{r}$ into a shift register division circuit. When the last symbol of $\mathbf{r}$ has been shifted into the circuit, the shift register cells contain the syndrome. The input to the circuit of an additional zero at this point is equivalent to multiplying s(x) by x and dividing by $g(x)$. The remainder $s^{(1)}(x)$ is now in the shift register cells. According to the result given above, $s^{(1)}(x)$ is the syndrome for $r^{(1)}(x)$. This process can be repeated n times, bringing us back to the starting point with the original syndrome in the shift register cells. We need only store one syndrome $\mathbf{s}$ for an error pattern $\mathbf{e}$ and all cyclic shifts of $\mathbf{e}$. A syndrome decoder with a reduced-size lookup table is used as follows.

2.4. Syndrome Decoder for Cyclic Codes

1. Set a counting variable j to the value 0. Compute the syndrome $\mathbf{s}$ for a received vector $\mathbf{r}$.
2. Look for the error pattern $\mathbf{e}$ corresponding to $\mathbf{s}$ in the syndrome lookup table. If there is such a value, go to step 7.
3. Increment the counter j by 1 and enter a zero into the shift register circuit at the input, computing $\mathbf{s}^{(j)}$.
4. Look for the error pattern $\mathbf{e}^{(j)}$ corresponding to $\mathbf{s}^{(j)}$ in the syndrome lookup table. If there is such a value, go to step 6.
5. Go to step 3.
6. Determine the error pattern $\mathbf{e}$ corresponding to $\mathbf{s}$ by cyclically shifting $\mathbf{e}^{(j)}j$ times to the left.
7. Subtract $\mathbf{e}$ from $\mathbf{r}$, obtaining the codeword $\mathbf{c}$.

So in general, there are fast, simple encoding and decoding circuits for cyclic codes of modest size. To get truly powerful error control with limited complexity, however, it is necessary to turn to one of the special cases.

3. QUADRATIC RESIDUE CODES AND GOLAY CODES

The nonzero squares modulo p, p a prime, are called the quadratic residues modulo p. Quadratic residues can be found by simply squaring every integer modulo p. For example, note that, using modulo 7 arithmetic, $1^2 = 6^2 = 1$, $2^2 = 5^2 = 4$ and $3^2 = 4^2 = 2$. So 1, 2, and 4 are the three quadratic residues modulo 7. In showing how quadratic residues can lead to some interesting codes, we have to indulge in some abstract algebra. The reader interested solely in applications and other practical matters can safely skip the next few paragraphs.

The set of integers modulo p, p a prime, form the field GF(p) under modulo p addition and multiplication. Let Q be the set of quadratic residues modulo p and N the set of corresponding nonresidues. Since GF(p) is a Galois field, there must exist at least one primitive element $\gamma \in \mathrm{GF}(p)$ that generates all the elements in Q and N. It follows that γ must be a quadratic nonresidue; otherwise there exists some element $\sqrt{\gamma} \in \mathrm{GF}(p)$ that generates $2(p-1)$ distinct elements in GF(p), contradicting the order of GF(p). One can then see that $\gamma^e \in Q$ if and only if e is even; otherwise, $\gamma^e \in N$. We conclude that all the elements in Q correspond to the first $(p-1)/2$ consecutive powers of γ^2, and that Q is a cyclic group under modulo p multiplication.

Now consider a field GF(s^m) that contains a primitive pth root of unity. Such a field exists for a given s, m, and p whenever $p|s^m$-1. We add the further restriction that s must be a quadratic residue modulo p. This can be somewhat restrictive; for example, if $s = 2$, then p must be of the form $p = (8k \pm 1)$. Since Q is a cyclic group,

multiplication of any element in Q by any other element in Q must result in an element in Q. It follows that the conjugates with respect to GF(s) of any element in Q must also be in Q. We conclude that Q is the union of one or more cyclotomic cosets modulo p with respect to GF(s).

Let α be primitive in GF(s^m). The results above show that the following polynomials have coefficients in the subfield GF(s):

$$q(x) = \prod_{i \in Q} (x - \alpha^i)$$

$$n(x) = \prod_{i \in N} (x - \alpha^i)$$

The *quadratic residue codes* of length p are defined by the generator polynomials $q(x)$, $(x-1)q(x)$, $n(x)$, and $(x-1)n(x)$, respectively.

The *binary Golay code* $\mathbf{G}_{23}$ is the (23,12,7) quadratic residue code with $p = 23$ and $s = 2$. The construction of this code proceeds as follows. The quadratic residues modulo 23 are $Q = 1, 2, 3, 4, 6, 8, 9, 12, 13, 16, 18$. Let β be a primitive 23rd root of unity [22 different β values can be found in GF(2^{11}) and its extensions]. The distinct powers of β form two cyclotomic cosets modulo 23 with respect to GF(2):

$$C_1 = \{1, 2, 3, 4, 6, 8, 9, 12, 13, 16, 18\}$$

$$C_5 = \{5, 7, 10, 11, 14, 15, 17, 19, 20, 21, 22\}$$

The term $x^{23} + 1$ factors into three binary irreducible polynomials:

$$x^{23} + 1 = (x + 1)(x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1) \times (x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1)$$

Depending on the selection of β, there are two possible generator polynomials for $\mathbf{G}_{23}$:

$$g_1(x) = x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1$$

$$g_2(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$$

Using either of these generator polynomials, the resulting code can be shown to be triple-error correcting. Given that $\mathbf{G}_{23}$ has dimension 12, it can be shown that $\mathbf{G}_{23}$is perfect. Each codeword is associated with a decoding sphere containing all vectors that are Hamming distance ≤ 3 from the codeword. Since the vectors have length 23, the decoding spheres have cardinality:

$$\begin{aligned} V_2(23, 3) &= \binom{23}{0} + \binom{23}{1} + \binom{23}{2} + \binom{23}{3} \\ &= 1 + 23 + 253 + 1771 \\ &= 2^{11} \end{aligned}$$

Through the addition of a parity-check bit, $\mathbf{G}_{23}$ can be extended to form $\mathbf{G}_{24}$, the triple-error-correcting, quadruple-error-detecting code that was used on the *Voyager* spacecraft.

The ternary [i.e., defined over GF(3)] Golay code ternary Golay code $\mathbf{G}_{11}$ is the quadratic residue code with $p = 11$ and $s = 3$. An analysis similar to that for $\mathbf{G}_{23}$ yields the factorization of $x^{11} - 1$ into three irreducible polynomials in GF(3)[x]:

$$x^{11} - 1 = (x - 1)(x^5 + x^4 - x^3 + x^2 - 1) \times (x^5 - x^3 + x^2 - x - 1)$$

Again there are two possible generator polynomials:

$$g_1(x) = x^5 + x^4 - x^3 + x^2 - 1$$

$$g_2(x) = x^5 - x^3 + x^2 - x - 1$$

$\mathbf{G}_{11}$ has length 11, dimension 6, and minimum distance 5, and is perfect.

4. BCH AND REED–SOLOMON CODES

When constructing an arbitrary cyclic code, there is no guarantee as to the resulting minimum distance. Given a generator polynomial $g(x)$, we must conduct a computer search of all corresponding nonzero code words $c(x)$ to determine the minimum-weight codeword and thus the minimum distance of the code. BCH codes, on the other hand, take advantage of a useful result that ensures a minimum "design distance" given a particular constraint on the generator polynomial. This result is known as the *BCH bound*.

4.1. The BCH Bound

Let $\boldsymbol{C}$ be a q-ary (n, k) cyclic code with generator polynomial $g(x)$. A *primitive* nth *root of unity* is defined to be an element β such that $\beta^n = 1$, but there is not smaller nonzero integer j such that $\beta^j = 1$. If GF(q^m) is the smallest extension field of GF(q) that contains a primitive nth root of unity, then we say that m is the multiplicative order of q modulo n. The BCH bound developed as follows. Let α be a primitive nth root of unity. Select $g(x)$ to be the minimal degree polynomial in GF(q)[x] such that $g(\alpha^b) = g(\alpha^{b+1}) = g(\alpha^{b+2}) = \cdots = g(\alpha^{b+d-2}) = 0$ for some integers $b \geq 0$ and $d \geq 1$. This $g(x)$ has $(d-1)$ consecutive powers of α as zeros. The BCH bound states that the code $\boldsymbol{C}$ defined by such a $g(x)$ has minimum distance $d_{\min} \geq d$.

We can use the BCH bound to construct a t-error-correcting q-ary BCH code of length n in the following way:

1. Find a primitive nth root of unity α in a field GF(q^m), where m is minimal.
2. Select $(d-1) = 2t$ consecutive powers of α, starting with α^b for some nonnegative integer b.
3. Let $g(x)$ be the least common multiple of the minimal polynomials for the selected powers of α with respect to GF(q). (Each minimal polynomial should appear only once in the product.)

Step 1 follows from our design procedure for general cyclic codes. Steps 2 and 3 ensure, through the BCH bound,

that the minimum distance of the resulting code equals or exceeds d and that the generator polynomial has the minimal possible degree. Since $g(x)$ is a product of minimal polynomials with respect to GF(q), $g(x)$ must be in GF(q)[x] and the corresponding code is q-ary with $d_{\min} \geq d$.

If $b = 1$, then the BCH code is said to be *narrow-sense*. If $n = q^m - 1$ for some positive integer m, then the BCH code is said to be *primitive*, for the nth root of unity α is a primitive element in GF(q^m). Let's consider a few binary BCH codes of length 31 to see how the design rule derived from the BCH bound fits in with the earlier discussion of conjugacy classes. Let α be a root of the primitive polynomial $x^5 + x^2 + 1$. It is thus a primitive element in the field GF(32). Since 31 is of the form $2^m - 1$, our BCH codes in this example are primitive. We begin by determining the cyclotomic cosets modulo 31 with respect to GF(2) and the associated minimal polynomials.

Cyclotomic Cosets	Minimal Polynomials
$C_0 = \{0\}$	$\leftrightarrow M_{(0)}(x) = x + 1$
$C_1 = \{1, 2, 4, 8, 16\}$	$\leftrightarrow M_{(1)}(x) = x^5 + x^2 + 1$
$C_3 = \{3, 6, 12, 24, 17\}$	$\leftrightarrow M_{(3)}(x) = x^5 + x^4 + x^3 + x^2 + 1$
$C_5 = \{5, 10, 20, 9, 18\}$	$\leftrightarrow M_{(5)}(x) = x^5 + x^4 + x^2 + x + 1$
$C_7 = \{7, 14, 28, 25, 19\}$	$\leftrightarrow M_{(7)}(x) = x^5 + x^3 + x^2 + x + 1$
$C_{11} = \{11, 22, 13, 26, 21\}$	$\leftrightarrow M_{(11)}(x) = x^5 + x^4 + x^3 + x + 1$
$C_{15} = \{15, 30, 29, 27, 23\}$	$\leftrightarrow M_{(15)}(x) = x^5 + x^3 + 1$

Recall that if a code $\mathbf{C}$ is to be a binary cyclic code, then it must have a generator polynomial $g(x)$ that is the product one or more of the minimal polynomials listed above. According to the BCH bound, if $\mathbf{C}$ is to be t-error correcting BCH code, then $g(x)$ must have as zeros $2t$ consecutive powers of α.

- *One-Error-Correcting Narrow-Sense Primitive BCH Code*. Since the code is to be narrow-sense and single-error-correcting, $b = 1$ and $\delta = 3$. The generator polynomial must thus have α and α^2 as zeros. $M_{(1)}(x)$ is the minimal polynomial of both α and α^2. The generator polynomial is thus

$$g(x) = \text{LCM}(M_{(1)}(x)M_{(2)}(x)) = M_{(1)}(x) = x^5 + x^2 + 1$$

  Since the degree of the generator polynomial $g(x)$ is 5, the dimension of the resulting code is $31 - 5 = 26$. Thus $g(x)$ defines a (31,26) binary single-error-correcting BCH code.
- *Two-Error-Correcting Narrow-Sense Primitive BCH Code*. Again $b = 1$, but δ has been increased to 5. $g(x)$ must thus have as roots α, α^2, α^3, and α^4.

$$\begin{aligned} g(x) &= LCM(M_{(1)}(x), M_{(2)}(x), M_{(3)}(x), M_{(4)}(x)) \\ &= M_{(1)}(x)M_{(3)}(x) \\ &= (x^5 + x^2 + 1)(x^5 + x^4 + x^3 + x^2 + 1) \\ &= x^{10} + x^9 + x^8 + x^6 + x^5 + x^3 + 1 \end{aligned}$$

  Since the degree of $g(x)$ is 10, it defines a (31,21) binary double-error-correcting code.

There are a number of ways to define Reed–Solomon codes. Reed and Solomon's initial definition focused on the evaluation of polynomials over the elements in a finite field [17]. Reed–Solomon codes can also be viewed as a natural extension of BCH codes. Simply put, a Reed–Solomon code is a q^m-ary BCH code of length $q^m - 1$.

Consider the construction of a t-error-correcting Reed–Solomon code of length $(q^m - 1)$. The first step is to note that the required primitive $(q^m - 1)$st root of unity α can be found in GF(q^m) (every finite field of size q contains an element with order $q - 1$). Since the code symbols are to be from GF(q^m), the next step is to construct the cyclotomic cosets modulo $(q^m - 1)$ with respect to GF(q^m). This is a trivial task, for $(s \cdot q^m) \equiv s$ modulo $(q^m - 1)$. The cyclotomic cosets are singleton sets of the form $\{s\}$ and the associated minimal polynomials are of the form $(x - \alpha^s)$.

The BCH bound indicates that $2t$ consecutive powers of a are required as zeros of the generator polynomial $g(x)$ for a t-error-correcting Reed-Solomon code. The generator polynomial is the product of the associated minimal polynomials:

$$g(x) = (x - \alpha b)(x - \alpha^{b+1})(x - \alpha^{b+2}) \cdots (x - \alpha^{b+2t-1}).$$

Consider a two-error-correcting 8-ary Reed–Solomon code of length 7. Let α be a root of the primitive binary polynomial $x^3 + x + 1$ and thus a primitive 7th of unity. The Galois field GF(8) can be represented as consecutive powers of α:

$$\begin{aligned} \alpha &= \alpha & \alpha^5 &= \alpha^2 + \alpha + 1 \\ \alpha^2 &= \alpha^2 & \alpha^6 &= \alpha^2 + 1 \\ \alpha^3 &= \alpha + 1 & \alpha^7 &= 1 \\ \alpha^4 &= \alpha^2 + \alpha & 0 &= 0 \end{aligned}$$

If the resulting code is to be double-error-correcting, it must have $2t = 4$ consecutive powers of a as zeros. A narrow-sense generator polynomial is constructed as follows:

$$\begin{aligned} g(x) &= (x - \alpha)(x - \alpha 2)(x - \alpha 3)(x - \alpha 4) \\ &= x^4 + \alpha^3 x^3 + x^2 + \alpha x + \alpha^3 \end{aligned}$$

Since the generator polynomial has degree 4, the (7,3) Reed–Solomon code it defines has dimension 3 over GF(8) and thus $8^3 = 512$ codewords.

Reed–Solomon codes have a number of interesting properties that are not shared by the other BCH codes. Recall that the minimum distance for BCH codes is in general lower-bounded by the design distance, but in many cases the actual minimum distance exceeds the design distance. One of the most significant properties of Reed–Solomon codes is the fact that an (n, k) Reed–Solomon code always has minimum distance exactly equal to $(n - k + 1)$.

BCH and Reed–Solomon codes can be decoded in a number of different ways. Perhaps the most efficient technique is Berlekamp's algorithm. In this article we provide a

brief description of Berlekamp's algorithm. Readers interested in a detailed exposition are referred to the source in Ref. 18.

The definition of a Reed–Solomon generating polynomial requires that, for some b and t, $g(\alpha^b) = g(\alpha^{b+1}) = \cdots = g(\alpha^{b+2t-1}) = 0$. A binary vector $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})$ is a codeword if and only if its associated polynomial $c(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}$ has as zeros these same $2t$ consecutive powers of α. Now consider a received polynomial $r(x)$, which can be expressed as the sum of the transmitted code polynomial $c(x)$ and an error polynomial $e(x) = e_0 + e_1 x + \cdots + e_{n-1}x^{n-1}$. A series of syndromes is obtained by evaluating the received polynomial at the $2t$ zeros. To minimize the complexity of our notation, it is assumed henceforth that all codes under discussion are narrow-sense ($b = 1$). The syndromes are computed as follows:

$$\begin{aligned} S_j &= r(\alpha^j) = c(\alpha^j) + e(\alpha^j) = e(\alpha^j) \\ &= \sum_{k=0}^{n-1} e_k (\alpha^j)^k, \qquad j = 1, 2, \ldots, 2t \end{aligned}$$

The computations in this expression are performed in GF(2^m), the field containing the primitive nth root of unity. Now assume that the received word $\mathbf{r}$ has ν errors in positions $i_1, i_2, \ldots, i_\nu$. We will assume that the code is binary, so the errors in these positions have value $e_{ij} = 1$. The syndrome sequence can be reexpressed in terms of these error locations:

$$S_j = \sum_{l=1}^{v} e_{i_l} (\alpha^j)^{i_l} = \sum_{l=1}^{v} (\alpha^{i_l})^j = \sum_{l=1}^{v} X_l^j, \qquad j = 1, \ldots, 2t$$

The $\{X_l\}$ are *error locators*, for their values indicate the positions of the errors in the received word. Expanding this equation we obtain a sequence of $2t$ algebraic *syndrome equations syndrome equations* in the ν unknown error locations:

$$\begin{aligned} S_1 &= X_1 + X_2 + \cdots + X_v \\ S_2 &= X_1^2 + X_2^2 + \cdots + X_v^2 \\ S_3 &= X_1^3 + X_2^3 + \cdots + X_v^3 \\ &\vdots \\ S_{2t} &= X_1^{2t} + X_2^{2t} + \cdots + X_v^{2t} \end{aligned}$$

Equations of this form are called *power-sum symmetric functions*. Since they form a system of nonlinear algebraic equations in multiple variables, they are somewhat difficult to solve in a direct manner.[4] Peterson showed, however, that the BCH syndrome equations can be translated into a series of linear equations that are much easier to work with [15]. Let $\Lambda(x)$ be the *error locator polynomial* that has as its roots the inverses of the ν error locators $\{X_l\}$:

$$\Lambda(x) = \prod_{l=1}^{v} (1 - X_l x) = \Lambda_v x^v + \Lambda_{v-1} x^{v-1} + \cdots + \Lambda_1 x + \Lambda_0$$

This equation can be used to express the coefficients of $\Lambda(x)$ directly in terms of the $\{X_l\}$. The resulting expressions are the *elementary symmetric functions* of the error locators. Power-sum symmetric functions and elementary symmetric functions are related by *Newton's identities*, which are generally expressed as follows for polynomials over arbitrary fields:

$$\begin{aligned} S_1 + \Lambda_1 &= 0 \\ S_2 + \Lambda_1 S_1 + 2\Lambda_0 &= 0 \\ S_3 + \Lambda_1 S_2 + \Lambda_2 S_1 + 3\Lambda_3 &= 0 \\ &\vdots \\ S_v + \Lambda_1 S_{v-1} + \Lambda_2 S_{v-2} + \cdots + \Lambda_{v-1} S_1 + v\Lambda_v &= 0 \\ S_{v+1} + \Lambda_1 S_v + \Lambda_2 S_{v-1} + \cdots + \Lambda_v S_1 &= 0 \\ &\vdots \\ S_{2t} + \Lambda_1 S_{2t-1} + \Lambda_2 S_{2t-2} + \cdots + \Lambda_v S_{2t-v} &= 0 \end{aligned}$$

If we assume that the codes in question are binary, we can reduce these expressions to the following:

$$\begin{aligned} S_1 + \Lambda_1 &= 0 \\ S_3 + \Lambda_1 S_2 + \Lambda_2 S_1 + \Lambda_3 &= 0 \\ S_5 + \Lambda_1 S_4 + \Lambda_2 S_3 + \Lambda_3 S_2 + \Lambda_4 S_1 + \Lambda_5 &= 0 \\ &\vdots \\ S_{2t-1} + \Lambda_1 S_{2t-2} + \Lambda_2 S_{2t-3} + \cdots + \Lambda_t S_{t-1} &= 0 \end{aligned}$$

Now suppose, for a moment, that we had an infinite number of syndromes available. We could then define an infinite-degree syndrome polynomial as follows:

$$S(x) = S_1 x + S_2 x^2 + \cdots + S_{2t} x^{2t} + S_{2t+1} x^{2t+1} + \cdots$$

Clearly we do not know all of the coefficients of $S(x)$, but fortunately the first $2t$ coefficients are entirely sufficient. $S(x)$ is made into an infinite degree polynomial so that it can be treated as a *generating function*. Define a third polynomial as follows:

$$\begin{aligned} \Omega(x) &\triangleq [1 + S(x)]\Lambda(x) \\ &= (1 + S_1 x + S_2 x^2 + \cdots)(1 + \Lambda_1 x + \Lambda_2 x^2 + \cdots) \\ &= 1 + (S_1 + \Lambda_1)x + (S_2 + \Lambda_1 S_1 + \Lambda_2)x^2 \\ &\quad + (S_3 + \Lambda_1 S_2 + \Lambda_2 S_1 + \Lambda_3)x^3 + \cdots \\ &= 1 + \Omega_1 x + \Omega_2 x^2 + \cdots \end{aligned}$$

where $\Omega(x)$ is called the *error magnitude polynomial*, and is useful in nonbinary decoding. For now we will simply

[4] The general problem of finding solutions to systems of algebraic equations in multiple variables is NP-hard, and is the basis for a number of nice cryptosystems, including the Data Encryption Standard.

note that if the syndrome and error locator polynomials are to satisfy this expression, then the odd-indexed coefficients of $\Omega(x)$ must be zero (see the Newton's identities for the binary case presented above). Given that we know only the first $2t$ coefficients of $S(x)$, the decoding problem then becomes one of finding a polynomial $\Lambda(x)$ of degree less than or equal to t that satisfies

$$[1 + S(x)]\Lambda(x) \cdots (1 + \Omega_2 x^2 + \Omega_4 x^4 + \cdots + \Omega_{2t} x^{2t}) \bmod x^{2t+1}$$

Berlekamp's algorithm proceeds iteratively by breaking this equation) down into a series of smaller problems of the form

$$[1 + S(x)]\Lambda^{(2k)}(x) \cdots (1 + \Omega_2 x^2 + \Omega_4 x^4 + \cdots + \Omega_{2k} x^{2k}) \bmod x^{2k+1}$$

where k runs from 1 to t. A solution $\Lambda^{(0)}(x) = 1$ is first assumed and tested to see if it works for the case $k = 1$. If it does work, we proceed to $k = 2$; otherwise, a correction factor correction factor is computed and added to $\Lambda^{(0)}$, creating a new solution $\Lambda^{(2)}(x)$. The genius of the algorithm lies in the computation of the correction factor. It is designed so that the new solution will work not only for the current case but for all previous values of k as well. We first consider the binary case.

4.2. Berlekamp's Algorithm for Decoding Binary BCH Codes

1. Set the initial conditions: $k = 0$, $\Lambda^{(0)}(x) = 1$, $T^{(0)} = 1$.
2. Let $\Delta^{(2k)}$ be the coefficient of x^{2k+1} in the product $\Lambda^{(2k)}(x)[1 + S(x)]$.
3. Compute

$$\Lambda^{(2k+2)}(x) = \Lambda^{(2k)}(x) + \Delta^{(2k)}[x \cdot T^{(2k)}(x)]$$

4. Compute

$$T^{(2k+2)}(x) = \begin{cases} x^2 T^{(2k)}(x) & \text{if } \Delta^{(2k)} = 0 \text{ or if } \deg[\Lambda^{(2k)}(x)] > k \\ \dfrac{x\Lambda^{(2k)}(x)}{\Delta^{(2k)}} & \text{if } \Delta^{(2k)} \neq 0 \text{ and } \deg[\Lambda^{(2k)}(x)] \leq k \end{cases}$$

5. Set $k = k + 1$. If $k < t$, then go to 2.
6. Determine the roots of $\Lambda(x) = \Lambda^{(2t)}(x)$. If the roots are distinct and lie in the right field, then correct the corresponding locations in the received word and *stop*.
7. Declare a decoding failure and *stop*.

For an example, consider a narrow-sense double-error-correcting code of length 31 with generator polynomial $= 1 + x^3 + x^5 + x^6 + x^8 + x^9 + x^{10}$. Let the received vector and associated polynomial be as follows:

$$\mathbf{r} = (0010000110011000000000000000000)$$

$$\updownarrow$$

$$r(x) = x^2 + x^7 + x^8 + x^{11} + x^{12}$$

A bit of number crunching in GF(32) yields the following syndrome polynomial:

$$S(x) = \alpha^7 x + \alpha^{14} x^2 + \alpha^8 x^3 + \alpha^{28} x^4$$

Applying Berlekamp's algorithm, we obtain the following sequence of solutions.

k	$\Lambda^{(2k)}(x)$	$T^{(2k)}(x)$	$\Delta^{(2k)}$
0	1	1	α^7
1	$1 + \alpha^7 x$	$\alpha^{24} x$	α^{22}
2	$1 + \alpha^7 x + \alpha^{15} x^2$	—	—

$\Lambda^{(4)}(x) = 1 + x\alpha^7 x + \alpha^{15} x^2$ is the error locator polynomial. The error locators are $X_1 = \alpha^5$ and $X_2 = \alpha^{10}$, indicating errors at the fifth and tenth coordinates of $\mathbf{r}$. The corrected word, with the corrected positions underlined, is

$$\mathbf{c} = (0010010110\underline{1}11\underline{0}00000000000000000)$$

$$\updownarrow$$

$$c(x) = x^2 + x^5 + x^7 + x^8 + x^{10} + x^{11} + x^{12} = x^2 g(x)$$

For the nonbinary case, we first note that the syndromes are now a function of the magnitude of the errors as well as their locations. Assuming that some ν errors have corrupted the received word, the syndromes are as follows:

$$S_j = e(\alpha^j) = \sum_{k=0}^{n-1} e_k (\alpha^j)^k = \sum_{l=1}^{v} e_{i_l} X_l^j$$

This expression defines a series of $2t$ algebraic equations in 2ν unknowns:

$$\begin{aligned} S_1 &= e_{i_1} X_1 + e_{i_2} X_2 + \cdots + e_{i_v} X_v \\ S_2 &= e_{i_1} X_1^2 + e_{i_2} X_2^2 + \cdots + e_{i_v} X_v^2 \\ S_3 &= e_{i_1} X_1^3 + e_{i_2} X_2^3 + \cdots + e_{i_v} X_v^3 \\ &\vdots \\ S_{2t} &= e_{i_1} X_1^{2t} + e_{i_2} X_2^{2t} + \cdots + e_{i_v} X_v^{2t} \end{aligned}$$

We can reduce this system of equation to a set of linear functions in the unknown quantities. We first assume an error locator polynomial $\Lambda(x)$ whose zeros are the inverses of the error locators $\{X_i\}$:

$$\Lambda(x) = \prod_{l=1}^{v} (1 - X_l x) = \Lambda_v x^v + \Lambda_{v-1} x^{v-1} + \cdots + \Lambda_1 x + \Lambda_0$$

It follows that for some error locator X_l

$$\Lambda(X_l^{-1}) = \Lambda_v X_l^{-v} + \Lambda_{v-1} X_l^{-v+1} + \cdots + \Lambda_1 X_l^{-1} + \Lambda_0 = 0$$

Since the expression sums to zero, we can multiply through by a constant:

$$e_{i_1} X_l^j (\Lambda_v X_l^{-v} + \Lambda_{v-1} X_l^{-v+1} + \cdots + \Lambda_1 X_l^{-1} + \Lambda_0) =$$

$$e_{i_1} (\Lambda_v X_l^{-v+j} + \Lambda_{v-1} X_l^{-v+j+1} + \cdots + \Lambda_1 X_l^{j-1} + \Lambda_0 X_l^j) = 0$$

Now sum both sides of over all indices l, obtaining an expression from which Newton's identities can be constructed:

$$\sum_{l=1}^{v} e_{i_1}(\Lambda_v X_l^{j-v} + \Lambda_{v-1} X_l^{j-v+1} + \cdots + \Lambda_1 X_l^{j-1} + \Lambda_0 X_l^j)$$

$$= \Lambda_v \sum_{l=1}^{v} e_{i_1} X_l^{j-v} + \Lambda_{v-1} \sum_{l=1}^{v} e_{i_1} X_l^{j-v+1} + \cdots + \Lambda_1 \sum_{l=1}^{v} e_{i_1} X_l^{j-1}$$

$$+ \Lambda_0 \sum_{l=1}^{v} e_{i_1} X_l^j$$

$$= \Lambda_v S_{j-v} + \Lambda_{v-1} S_{j-v+1} + \cdots + \Lambda_1 S_{j-1} + \Lambda_0 S_j = 0$$

From the earlier expressions it is clear that Λ_0 is always one. The preceding equation can thus be reexpressed as

$$\Lambda_v S_{j-v} + \Lambda_{v-1} S_{j-v+1} + \cdots + \Lambda_1 S_{j-1} = -S_j$$

This expression shows that the syndrome S_j can be expressed in recursive form as a function of the coefficients of the error locator polynomial $\Lambda(x)$ and the earlier syndromes $S_{j-1}, \ldots, S_{j-v}$. In 1969 Massey showed that this expression can be given a physical interpretation through the use of a linear feedback shift register (LFSR) [23]. The double-lined elements in Fig. 2 denote storage of and operations on nonbinary field elements.

Part of the problem of decoding BCH and Reed–Solomon codes can be reexpressed as follows. Find an LFSR of minimal length such that the first $2t$ elements in the LFSR output sequence are the syndromes $S_1, S_2, \ldots, S_{2t}$. The taps of this shift register provide the desired error locator polynomial $\Lambda(x)$.

Let $\Lambda^{(k)}(x) = \Lambda_k x^j + \Lambda_{k-1} x^{k-1} + \cdots + \Lambda_1 x + 1$ be the *connection polynomial* of length k whose coefficients specify the taps of a length-k LFSR. Massey's construction of Berlekamp's algorithm starts by finding $\Lambda^{(1)}$ such that the first element output by the corresponding LFSR is the first syndrome S_1. The second output of this LFSR is then compared to the second syndrome. If the two do not have the same value, then the *discrepancy* between the two is used to construct a modified connection polynomial. If there is no discrepancy, then the same connection polynomial is used to generate a third sequence element, which is compared to the third syndrome. The process continues until a connection polynomial is obtained that specifies an LFSR capable of generating all $2t$ elements of the syndrome sequence.

Massey showed that, given an error pattern of weight $\leq t$, the connection polynomial resulting from the Berlekamp algorithm uniquely specifies the correct error locator polynomial.

The algorithm has four basic parameters: the connection polynomial $\Lambda^{(k)}(x)$, the correction polynomial $\mathrm{T}(x)$, the discrepancy $\Delta^{(k)}$, the length L of the shift register, and the indexing variable k. The algorithm proceeds as follows.

4.3. The Berlekamp–Massey Shift Register Synthesis Decoding Algorithm

1. Compute the syndrome sequence $S_1, \ldots, S_{2t}$ for the received word.
2. Initialize the algorithm variables as follows:

$$k = 0, \Lambda^{(0)}(x) = 1, \quad L = 0, \quad \text{and} \quad \mathrm{T}(x) = x$$

3. Set $k = k + 1$. Compute the discrepancy $\Delta^{(k)}$ by subtracting the kth output of the LFSR defined by $\sigma^{(k-1)}(x)$ from the kth syndrome:

$$\Delta^{(k)} = S_k - \sum_{i=1}^{L} \Lambda_i^{(k-1)} S_{k-i}$$

4. If $\Delta^{(k)} = 0$, then go to step 8.
5. Modify the connection polynomial: $\Lambda^{(k)}(x) = \Lambda^{(k-1)}(x) - \Delta^{(k)}\mathrm{T}(x)$
6. If $2L^3 k$, then go to step 8.
7. Set $L = k - L$ and $T(x) = \Lambda^{(k-1)}(x)/\Delta^{(k)}$.
8. Set $T(x) = x \cdot T(x)$
9. If $k < 2t$, then go to step 3.
10. Determine the roots of $\Lambda(x) = \Lambda^{(2t)}(x)$. If the roots are distinct and lie in the right field, then determine the error magnitudes, correct the corresponding locations in the received word, and *stop*.
11. Declare a decoding failure and *stop*.

The Berlekamp–Massey algorithm allows us to find the error locator polynomial, but there remains the problem of finding the error magnitudes. Forney [22]. showed that the following expression will do the trick:

$$e_{i_k} = \frac{-X_k \Omega(X_k^{-1})}{\Lambda'(X_k^{-1})}$$

Consider the following example of double-error correction [1]. using the Berlekamp–Massey algorithm and a (7,3) Reed–Solomon code (from Wicker [1]). Let the received polynomial be $r(x) = \alpha^2 x^6 + \alpha^2 x^4 + x^3 + \alpha^5 x^2$, giving the syndrome sequence $S_1 = \alpha^6$, $S_2 = \alpha^3$, $S_3 = \alpha^4$, $S_4 = \alpha^3$. The algorithm generates the following set of connection polynomials, discrepancies, and correction polynomials [the last column, $T(x)$, Follows at conclusion of step 8:

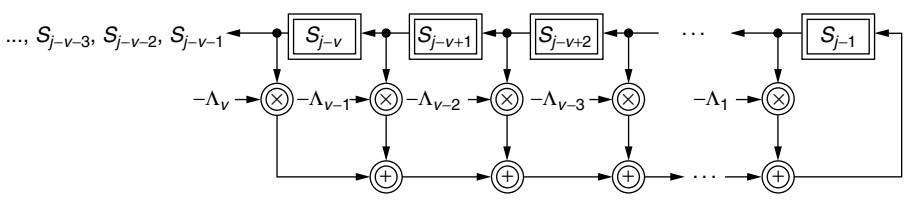


Figure 2. LFSR interpretation.

k	S_k	$\Lambda^{(k)}(x)$	$\Delta^{(k)}$	L	$T(x)$
0	—	1	—	0	x
1	α^6	$1+\alpha^6 x$	$S_1 - 0 = \alpha^6$	1	αx
2	α^3	$1+\alpha^4 x$	$S_2 - \alpha^5 = \alpha^2$	1	αx^2
3	α^4	$1+\alpha^4 x+\alpha^6 x^2$	$S_3 - 1 = \alpha^5$	2	$\alpha^2 x+\alpha^6 x^2$
4	α^3	$1+\alpha^2 x+\alpha x^2$	$S_4 - \alpha^4 = \alpha^6$	—	—

We obtain the error locator polynomial $\Lambda(x) = 1 + \alpha^2 x + \alpha x^2$.

The LFSRs corresponding to the connection polynomials $\Lambda^{(1)}(x)$ through $\Lambda^{(4)}(x)$ are drawn in Fig. 3, along with the initial conditions and the generated output sequence. Consider the LFSR with connection polynomial $\Lambda^{(3)}(x)$. This LFSR correctly generates the first three syndromes, but its fourth output, α^4, is not equal to $S_4 = \alpha^3$. The discrepancy $\Delta^{(4)} = \alpha^6$ is the difference between the two values, as shown in Fig. 3. This discrepancy is used to determine the connection polynomial $\Lambda^{(4)}(x)$, which, as shown in Fig. 3, correctly generates all four syndromes.

The syndrome sequence provides the syndrome polynomial $S(x) = \alpha^6 x + \alpha^3 x^2 + \alpha^4 x^3 + \alpha^3 x^4$. We used the Berlekamp–Massey algorithm to obtain the error locator polynomial $\Lambda(x) = 1 + \alpha^2 x + \alpha x^2$. We can now compute the error magnitude polynomial:

$$\begin{aligned}\Omega(x) &\equiv \Lambda(x)[1 + S(x)] \bmod x^{2t+1} \\ &\equiv (1 + \alpha^{2x} + \alpha^{x^2})(1 + \alpha^{6x} + \alpha^{3x^2} + \alpha^{4x^3} + \alpha^{3x^4}) \bmod x^5 \\ &\equiv (1 + x + \alpha^{3x^2}) \bmod x^5\end{aligned}$$

The errors locators were found to be $X_1 = \alpha^3$ and $X_2 = \alpha^5$ in the first part of this example. Using the Forney algorithm, the error magnitudes are found to be

$$\begin{aligned}e_{i_x k} &= \frac{-X_k \Omega(X_k^{-1})}{\Lambda'(X_k^{-1})} = \frac{-X_k[1 + X_k^{-1} + \alpha^3 X_k^{-2}]}{\alpha^2} \\ &= \alpha^5 X_k + \alpha^5 + \alpha X_k^{-1}\end{aligned}$$

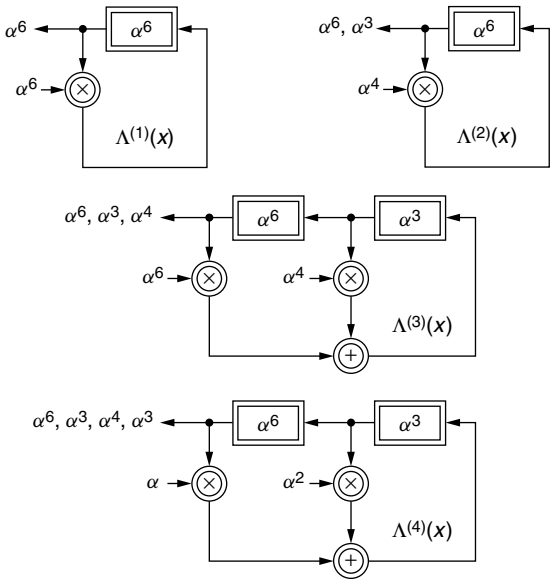


Figure 3. Several different LFSR syndromes.

$$e_3 = \alpha^5 \alpha^3 + \alpha^5 + \alpha\alpha^4 = \alpha$$

$$e_5 = \alpha^5 \alpha^5 + \alpha^5 + \alpha\alpha^2 = \alpha^5$$

The error polynomial is thus $e(x) = \alpha x^3 + \alpha^5 x^5$.

5. APPLICATIONS

The applications of cyclic codes are legion. In this final section we will focus on two of the more interesting applications: digital audio and deep-space telecommunications.

5.1. The Compact-Disk (CD) Player

The most ubiquitous application of cyclic codes (or possibly any error control codes) lies in the CD player. The channel in a CD playback system consists of a transmitting laser, a recorded disk, and a photodetector. Assuming that the player is working properly, the primary contributor to errors on this channel is the contamination of the surface of the disk (e.g., fingerprints and scratches). As the surface contamination surface contamination affects an area that is usually quite large compared to the surface used to record a single bit, channel errors occur in bursts when the disk is played. As we shall see, the CD error control system handles bursts through cross-interleaving and through the burst error-correcting capability of Reed–Solomon codes.

Figure 4 shows the various stages through which music is processed on its way to being recorded on a disk. Each channel is sampled 44,100 times per second, allowing accurate reproduction of all frequencies up to 22 kHz. Each sample is then converted into digital form by a 16-bit analog-to-digital converter.

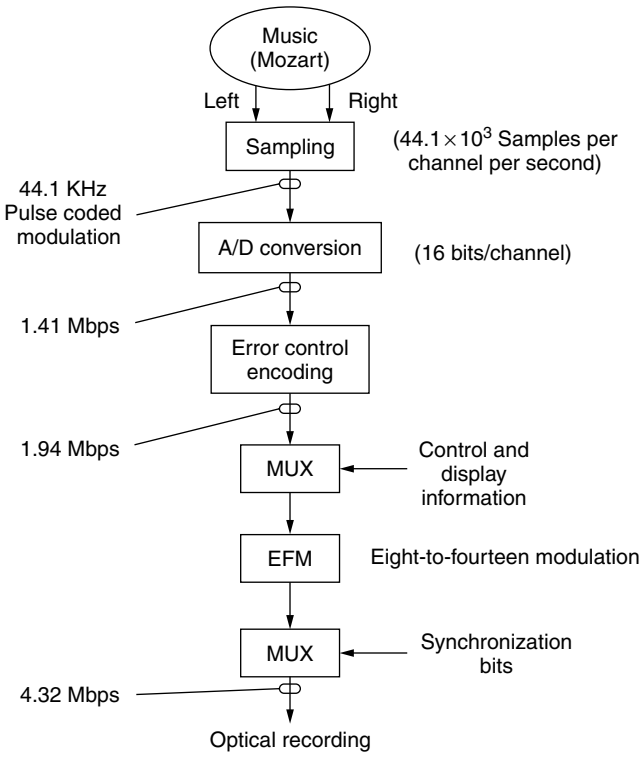


Figure 4. Data processing in the creation of a compact disk [1].

The output of the A/D converter forms a 1.41-Mbps (megabits per second) datastream, which is passed directly to the CIRC encoder. The CIRC encoder, as shown in Fig. 4, uses two shortened Reed–Solomon codes, $\mathbf{C}_1$ and $\mathbf{C}_2$. Both codes uses 8-bit symbols from the code alphabet GF(256). This provides for a nice match with the 16-bit samples emerging from the A/D converter. The "natural" length of the RS code over GF(256) is 255, which would lead to 2040-bit codewords and a relatively complicated decoder. It should be remembered that the decoder will reside in the retail player, and it is extremely important that its cost be minimized. The codes are thus shortened significantly: $\mathbf{C}_1$ is a (32,28) code and $\mathbf{C}_2$ is a (28,24) code. Both have redundancy 4 and minimum distance 5.

Each 16-bit sample is treated as a pair of symbols from GF(256). The samples are encoded 12 at a time by the $\mathbf{C}_2$ encoder to create a 28-symbol codeword. The 28 symbols in each $\mathbf{C}_2$ codeword are then passed through a cross-interleaver ($m = 28$, $D = 4$ symbols) before being encoded once by the $\mathbf{C}_1$ encoder. The resulting 32-symbol $\mathbf{C}_1$ codewords are then processed as shown in the figure above.

The CIRC encoding process for the CD system is standard; no matter where you buy your CDs (standard size), they will play on any CD player. However, the CIRC decoding process is not standardized and can vary from player to player [28]. This was done intentionally to allow the manufacturers to experiment with various designs and to speed the player to market. The basic building blocks of the decoder are shown in the figure. The $\mathbf{C}_1$ decoder is followed by a cross-deinterleaver and a $\mathbf{C}_2$ decoder. Since both codes have minimum distance 5, they can be used to correct all error patterns of weight

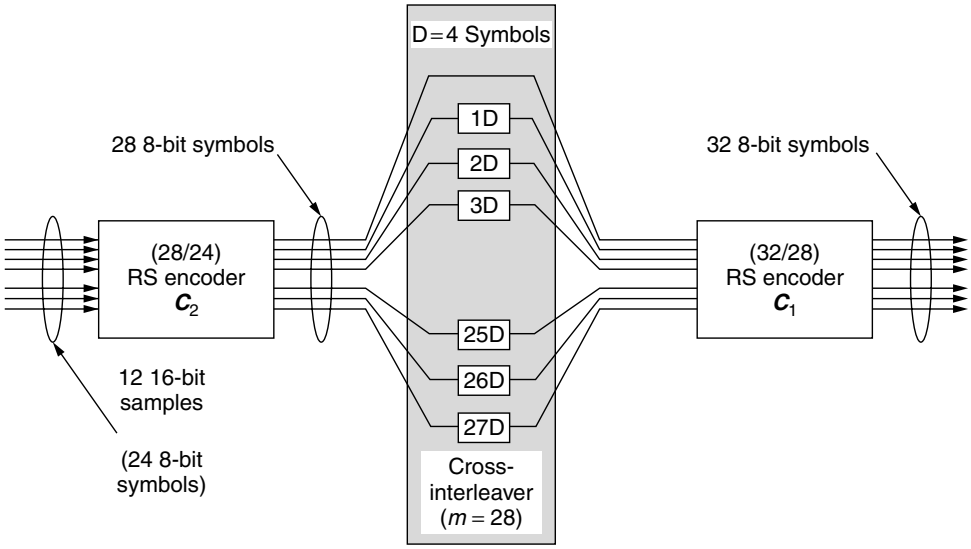


Figure 5. Cross-interleaved encoding [1].

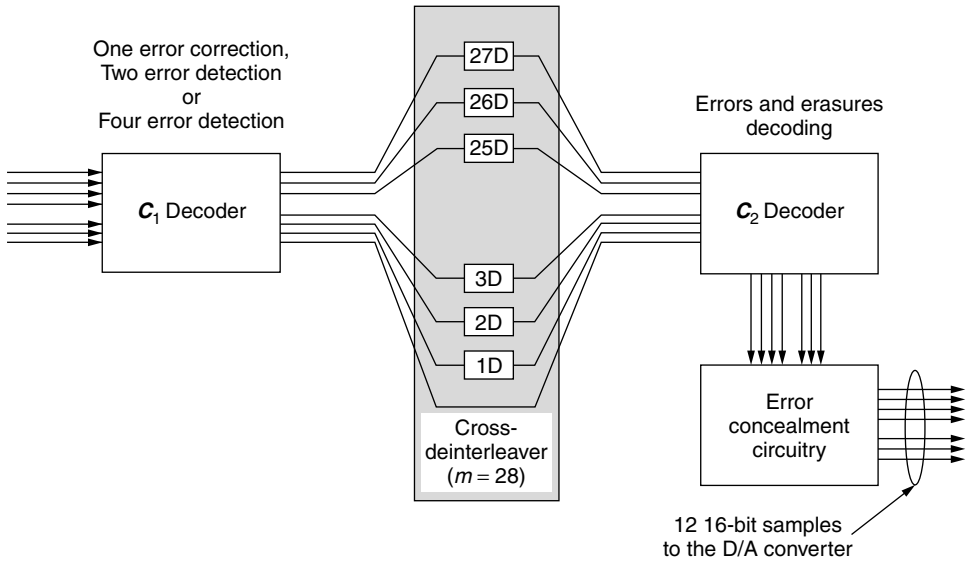


Figure 6. Basic structure of a CIRC decoder [1].

2 or less. Even when the full error correcting capacity of an RS code is used, there remains a significant amount of error detecting capacity. It is also the case that some of the error correction capacity of the RS code can be exchanged for an increase in error detection capacity and a substantial improvement in reliability performance. Most CIRC decoders take advantage of both of these principles. The $\mathbf{C}_1$ decoder is set to correct all single-error patterns. When the $\mathbf{C}_1$ decoder sees a higher-weight error pattern (double-error patterns and any pattern causing a decoder failure), the decoder outputs 28 erased symbols. The cross-deinterleaver spreads these erasures over 28 $\mathbf{C}_2$ codewords. $\mathbf{C}_1$ may also be set to correct double error patterns, or, at the other extreme, may simply be used to declare erasures (the least expensive implementation).

The $\mathbf{C}_2$ decoder can correct any combination of e errors and s erasures, where $2e + s < 5$. It is generally designed to decode erasures only (again, an inexpensive solution) due to the small probability of a $\mathbf{C}_1$ decoder error. Whenever the number of erasures exceeds 4, the $\mathbf{C}_2$ decoder outputs 24 erasures, which corresponds to 12 erased music samples. The error concealment circuitry error concealment circuitry responds by muting these samples or by interpolating values through the use of correct samples adjacent to the correct samples. A number of additional interleaving and delay operations are included in the encoding and decoding operations in order to enhance the operation of the error concealment circuitry. For example, samples adjacent in time are further separated by additional interleavers to improve the impact of interpolation [28].

5.2. Deep-Space Telecommunications

The earliest use of a cyclic code in deep-space telecommunications was in conjunction with the *Mariner* mission in 1971. *Mariner* included an infrared interferometer spectrometer (IRIS) [20,29]. The data from this instrument required a bit error rate on the order of 10^{-5}, and thus needed protection beyond that provided by the then standard Reed–Muller-based system. It was decided to precode the IRIS data using a [6,4] Reed-Solomon code with symbols from GF(2^6), thus creating the first concatenated system designed for use in deep-space telecommunications. The RS outer code was applied only to the IRIS data, thus keeping the overall code rate of the telemetry channel quite close to that of the Reed–Muller system by itself. This use of concatenated schemes to provide selective error protection was revisited in the development of the *Voyager* mission.

The *Voyager* 1 and 2 spacecraft were launched toward Saturn and Jupiter, respectively, in the summer of 1977. They carried a number of scientific instruments and imaging systems for a mission that was to be the most successful in the history of the exploration of deep space. The *Voyager* spacecraft were originally intended to explore Jupiter, Saturn, and their moons. The distances involved were substantial: Jupiter and Saturn are 483 million miles and 870 million miles from the sun, respectively. During the Jupiter flyby, *Voyager* 2 transmitted a 21.3-W signal through an antenna with $G_T = 6.5 \times 10^4$ (48.1 dB). By the time the signal was recovered by a 64-m dish on the earth, the signal measured 3.04×10^{-16} W. The telemetry signal consisted of two basic types of data: imaging and general science and engineering (GSE). The color images were reconstructed from three (800×800) arrays of 8-bit pixels, giving a total of 15,360,000 bits per image [29]. The uncompressed images required a nominal bit error rate of 5×10^{-3}. It was found that the rate-$\frac{1}{2}$ Planetary Standard was sufficient to provide the desired level of reliability. The GSE data, however, required a much lower bit error rate than the imaging data. The configuration in Fig. 7 was adopted. The GSE data was first encoded using the extended (24,12) Golay code discussed earlier in this article. The GSE and imaging data were then both encoded using the rate-$\frac{1}{2}$ Planetary Standard. At a 2.3 dB

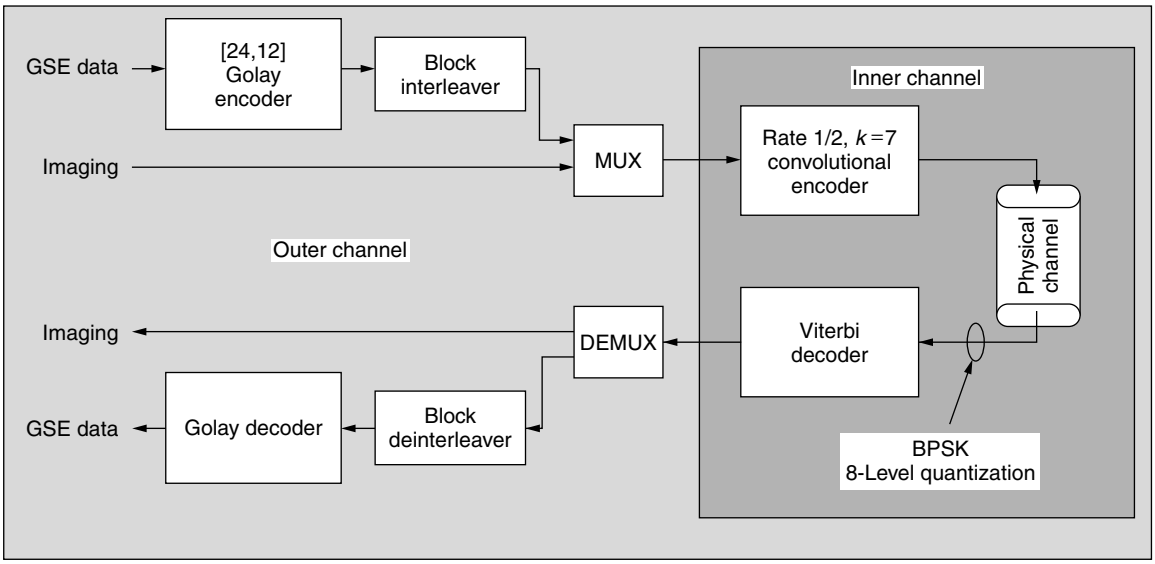


Figure 7. Convolutional/Viterbi and Golay concatenated system for the *Voyager* spacecraft (Jupiter and Saturn flybys) [20].

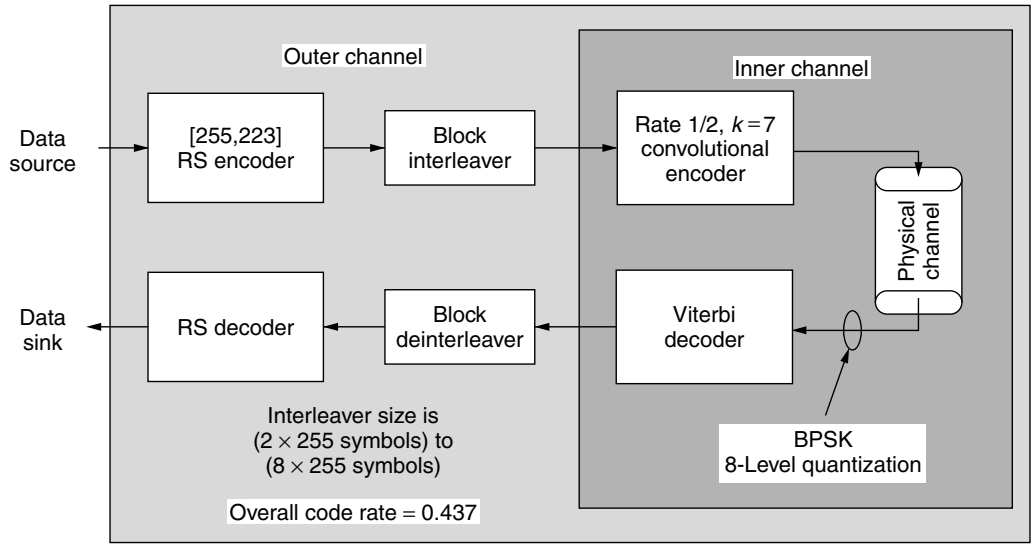


Figure 8. CCSDS Standard for deep-space telemetry links [20].

E_b/N_0, the convolutional code provided an inner channel bit error rate of 5×10^{-3}. The Golay code provides an outer channel bit error rate of less than 1×10^{-5}.

When the *Voyager* mission was extended to cover several of the outer planets, it was necessary to increase the level of error protection. As with the GSE data in the Jupiter and Saturn flybys, the additional reliability requirement was handled through the use of a concatenated system. In this case, however, the extremely low bit error rate requirement and the sensitivity of the link budget to a reduction in code rate pointed towards the use of Reed–Solomon technology for the outer code. The resulting concatenated system later became the basis for the CCSDS standard for deep-space telemetry links [30]. This standard has been used extensively by both NASA and the European Space Agency. In the CCSDS standard, the rate-$\frac{1}{2}$ Planetary Standard is joined by a (255,223) RS outer code with symbols from GF(2^8), as shown in Fig. 8.

BIBLIOGRAPHY

1. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
2. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
3. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1977.
4. E. Prange, *Cyclic Error-Correcting Codes in Two Symbols*, Air Force Cambridge Research Center Report TN-57-103, Cambridge, MA, Sept. 1957.
5. E. Prange, *Some Cyclic Error-Correcting Codes with Simple Decoding Algorithms*, Air Force Cambridge Research Center Report TN-58-156, Cambridge, MA, April 1958.
6. E. Prange, *The Use of Coset Equivalence in the Analysis and Decoding of Group Codes*, Air Force Cambridge Research Center Report TR-59-164, Cambridge, MA, 1959.
7. I. S. Reed (1990).
8. I. S. Reed (1992).
9. Shannon (1948).
10. Golay (1949).
11. Kasami (1964).
12. A. Hocquenghem, Codes correcteurs d'erreurs, *Chiffres* **2**: 147–156 (1959).
13. R. C. Bose and D. K. Ray-Chaudhuri, On a class of error correcting binary group codes, *Inform. Control* **3**: 68–79 (March 1960).
14. R. C. Bose and D. K. Ray-Chaudhuri, Further results on error correcting binary group codes, *Inform. Control* **3**: 279–290 (Sept. 1960).
15. W. W. Peterson, Encoding and error-correction procedures for the Bose–Chaudhuri codes, *IRE Trans. Inform. Theory* **IT-6**: 459–470 (Sept. 1960).
16. D. Gorenstein and N. Zierler, A class of error correcting codes in p^m symbols, *J. Soc. Indust. Appl. Math.* **9**: 207–214 (June 1961).
17. I. S. Reed and G. Solomon, Polynomial codes over certain finite fields, *SIAM J. Appl. Math.* **8**: 300–304 (1960).
18. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968 (rev. ed. Aegean Park Press, Laguna Hills, CA, 1984).
19. Hamming (1950).
20. S. B. Wicker, Deep space applications, in V. Pless and W. C. Huffman, eds., *Handbook of Coding Theory*, Elsevier, Amsterdam, 1998.
21. R. T. Chien, Cyclic decoding procedure for the Bose–Chaudhuri–Hocquenghem codes, *IEEE Trans. Inform. Theory* **IT-10**: 357–363 (Oct. 1964).
22. G. D. Forney, On decoding BCH codes, *IEEE Trans. Inform. Theory* **IT-11**: 549–557 (Oct. 1965).
23. J. L. Massey, Shift register synthesis and BCH decoding, *IEEE Trans. Inform. Theory* **IT-15**(1): 122–127 (Jan. 1969).

24. Y. Sugiyama, Y. Kasahara, S. Hirasawa, and T. Namekawa, A method for solving key equation for Goppa codes, *Inform. Control* **27**: 87–99 (1975).

25. Reed (1978).

26. W. C. Gore, Transmitting binary symbols with Reed–Solomon Codes, *Proc. Princeton Conf. Information Science and Systems*, Princeton, NJ, 1973, pp. 495–497.

27. R. E. Blahut, Transform techniques for error control codes, *IBM J. Res. Devel.* **23**: 299–315 (1979).

28. K. A. S. Immink, RS codes and the compact disc, in S. B. Wicker and V. K. Bhargava, eds., *Reed Solomon Codes and Their Applications*, IEEE Press, Piscataway, NJ, 1994.

29. R. J. McEliece and L. Swanson, Reed–Solomon codes and the exploration of the solar system, in S. B. Wicker and V. K. Bhargava, eds., *Reed–Solomon Codes and Their Applications*, IEEE Press, Piscataway, NJ, 1994, pp. 25–40.

30. Consultative Committee for Space Data Systems, *Recommendations for Space Data Systems Standard: Telemetry Channel Coding*, Blue Book Issue 2, CCSDS 101.0-B2, Jan. 1987.