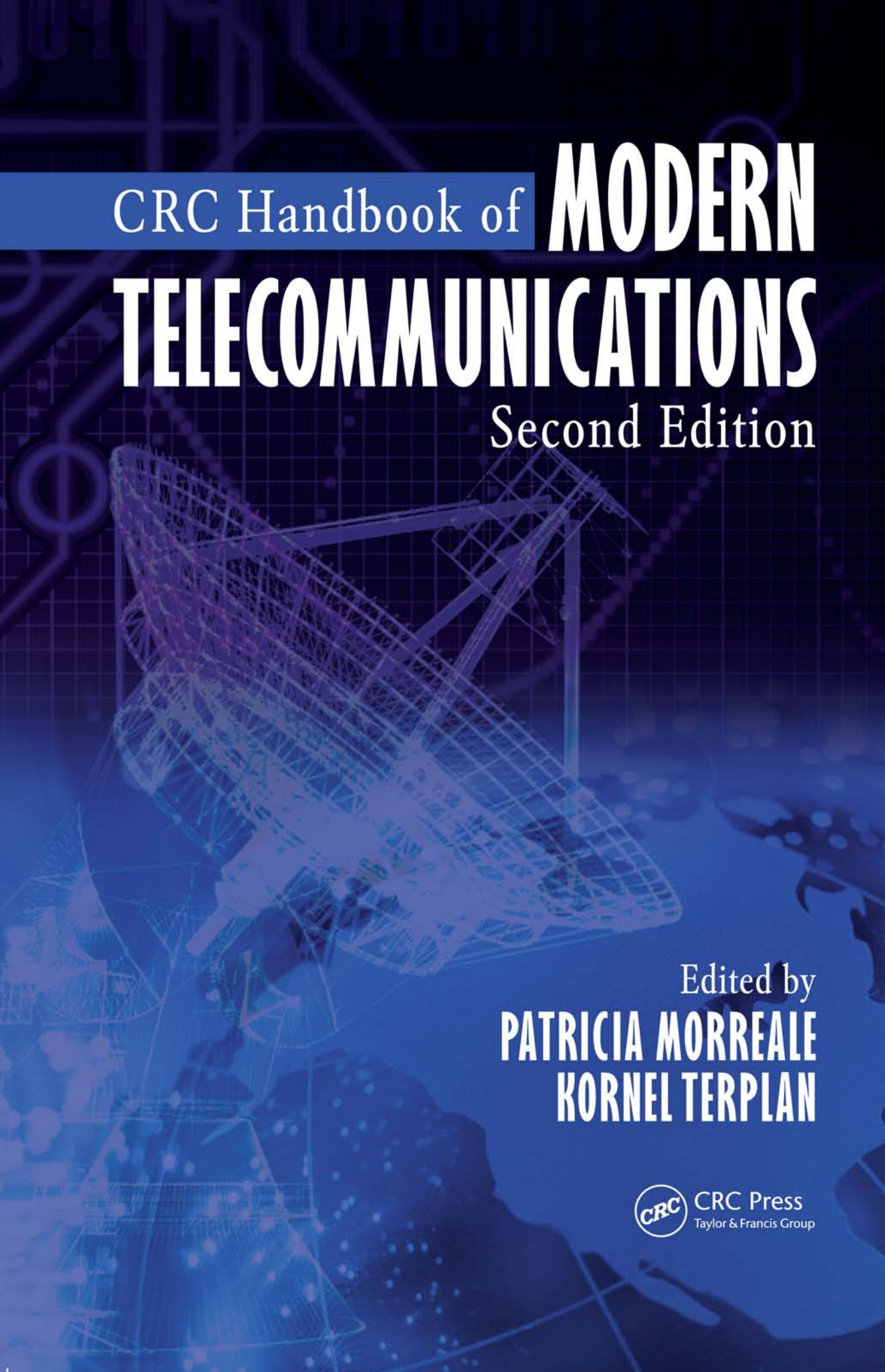
CRC Handbook of MODERN TELECOMMUNICATIONS

Second Edition



Edited by

PATRICIA MORREALE
KORNEL TERPLAN

CRC Press
Taylor & Francis Group

CRC Handbook of MODERN TELECOMMUNICATIONS

Second Edition

Edited by
PATRICIA MORREALE
KORNEL TERPLAN

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-7800-8 (Hardback)

Library of Congress Cataloging-in-Publication Data

CRC handbook of modern telecommunications / editors, Patricia A. Morreale, Kornel Terplan. -- 2nd ed.
p. cm.
"A CRC title."
Includes index.
ISBN 978-1-4200-7800-8 (hardcover : alk. paper)
1. Telecommunication--Handbooks, manuals, etc. I. Morreale, Patricia. II. Terplan, Kornel. III. Title: Handbook of modern telecommunications.

TK5101.C72 2010
621.382--dc22 2009027279

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Foreword v
Acknowledgments vii
Editors ix
Contributors xi

1 Voice and Data Communications
Introduction **1**-1
1.1 Computer Telephone Integrated (CTI) *Michel Gilbert* **1**-2
1.2 Voice over IP *Matthew Kolon and Patricia Morreale* **1**-13
1.3 Local Area Networks *John Amoss* **1**-21
1.4 RFID Architecture and Protocols *Chonggang Wang, Mahmoud Daneshmand, and Kazem Sohraby* **1**-39
1.5 Design of Wireless Sensor Network Applications, Hardware and Software *Sajid Hussain* **1**-51
1.6 Multimedia Applications for Cognitive Radio Networks *Sajid Hussain and Muhammad Farhat Kaleem* **1**-58
Summary *Patricia Morreale* **1**-70

2 Intranets
Introduction **2**-2
2.1 Internet and Intranet Management Concepts *Teresa Piliouras and John Braun* **2**-2
2.2 Virtual Private Networking Solutions *Endre Sara* **2**-30
2.3 Web-Enabled Data Warehousing *Dermot Murray* **2**-38
2.4 Web Performance Management *Kornel Terplan* **2**-46
2.5 Application Performance Management *Vadim Rosenberg* **2**-81
2.6 Electronic Technologies *Patricia Morreale and Mihir Parikh* **2**-99
2.7 Internet Protocols *John Braun* **2**-105
2.8 Role of Open Source Software *Tivadar Szemethy* **2**-115
Summary and Trends *Patricia Morreale and Kornel Terplan* **2**-125

3 Network Management and Administration
Introduction **3**-2
3.1 Management Concepts *Joe Ghetie* **3**-4
3.2 Management of Emerging Technologies *Tivadar Szemethy* **3**-19
3.3 Management-Related Standards *Tivadar Szemethy* **3**-51
3.4 Management Function *József Wiener* **3**-84

3.5 Support Systems for Service Providers *József Wiener* **3**-104
3.6 Support Processes for Service Providers *Árpád Bakay and József Wiener* **3**-132
3.7 Management Frameworks and Applications *Árpád Bakay, Tivadar Szemethy, and József Wiener* **3**-169
3.8 Intelligence Support Systems *Paul Hoffmann and Kornel Terplan* **3**-200
3.9 Management of Sensor Networks *Jim Frey* **3**-226
3.10 Solution Architectures *Norman Kincl* **3**-241
Summary and Trends *Kornel Terplan* **3**-258

4 Network Organization and Governance
Introduction **4**-2
4.1 Information Life Cycle Management *Kornel Terplan* **4**-3
4.2 Information Technology Alignment with Businesses *Kornel Terplan* **4**-17
4.3 Business Intelligence and Analytics *Patricia Morreale and Deepak Pareek* **4**-27
4.4 Service-Level Management *Christian Voigt and Kornel Terplan* **4**-51
4.5 Management Services and Outsourcing *Kornel Terplan and Christian Voigt* **4**-77
4.6 Network Management Organization *Kornel Terplan* **4**-107
4.7 Best Practices Benchmarks for Service Providers *Kornel Terplan* **4**-123
Summary and Trends *Kornel Terplan* **4**-147

5 Future Telecommunications Services
Introduction **5**-1
5.1 User Needs *James Anderson and Patricia Morreale* **5**-2
5.2 Application Trends *James Anderson and Patricia Morreale* **5**-13
5.3 Systems and Service Integration for Management *James Anderson and Kornel Terplan* **5**-22
5.4 New Produce and Service Creation *James Anderson* **5**-33
5.5 Telecommunications Tariffing *James Anderson* **5**-38
5.6 Telecommunications Strategies *James Anderson and Patricia Morreale* **5**-42
Summary *Patricia Morreale* **5**-48

Index **I**-1

Foreword

In the preparation of this book, our objective was to provide an advanced understanding of emerging telecommunications systems, their significance, and the anticipated role these systems will play in the future. In addition to our new discussions of radio frequency identification (RFID) and wireless sensor networks, this book addresses network management and administration, as well as network organization and governance, topics that were not as clearly defined during the development of the first edition. With the help of our talented contributors, we believe we have accomplished this. By addressing voice, Internet, network and traffic management, along with future trends, we feel our readers will be knowledgeable about current and future telecommunications systems.

Section 1 outlines the techniques of voice communication systems, with attention paid to both basic and advanced systems. Voice over IP and the integration of voice and IP data are closely examined. The second part of this section concentrates on state-of-the-art solutions for local area networks and RFID architectures. Wireless sensor network applications and multimedia applications for cognitive radio networks are presented in detail.

Section 2 provides an explanation of the Internet, including elements of its structure and consideration of how future services will be handled on the Internet. Internet management and security are presented. A detailed discussion of virtual private networks (VPNs) is provided, as well as Web-enabled data warehousing concepts. Web and application performance management, along with electronic commerce and Internet protocols are reviewed, permitting the reader to understand and select with insight from the available Web-based technology choices.

Section 3 focuses on network management and administration. As the services and features provided cause the network to become larger in scale and scope, network management will become even more crucial and important than it is today. The telecommunications support process is outlined, including management of emerging technologies, support systems and processes for service providers, and management frameworks and applications. A detailed consideration of intelligent support systems is presented. The management of sensor networks is detailed.

Section 4 addresses network organization and governance. Information life cycle management and the importance of information technology alignment with business is stressed, and business intelligence and analytics are reviewed. The importance of management services and outsourcing is clear and exemplified by best practice benchmarks for service providers.

Finally, in Section 5, future trends and directions are considered, with a view toward satisfying user needs in parallel with application trends, which will require system and service integration.

We hope our readers find this book an excellent guide to emerging telecommunications trends.

Patricia Morreale
Department of Computer Science
Kean University
Union, New Jersey

Acknowledgments

The editors would like to thank all their contributors for their excellent, timely work. Without their help, we would not have been able to submit this manuscript.

We are particularly grateful to Nora Konopka, who has supported our editorial work by providing significant administrative help from CRC Press. We would also like to thank Marsha Pronin, who greatly assisted the co-editors with the care and attention she provided to many details of the book.

Editors

Patricia Morreale, Ph.D. is a faculty member in the Department of Computer Science at Kean University, Union, New Jersey, where she conducts research in network management and design. Since joining Kean University, she has established the Network Laboratory, building on her prior work at Stevens Institute of Technology, and has continued her research in wireless network design and applications.

Her research has been funded by the National Science Foundation (NSF), the U.S. Navy, U.S. Air Force, Allied Signal, AT&T, Lucent, Panasonic, Bell Atlantic, and the New Jersey Commission on Science and Technology (NJCST). She is a senior member of the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronic Engineers (IEEE), and a member of Sigma Xi. She has served as a guest editor for *IEEE Communications* magazine and has served as vice chair for *INFOCOM*.

She has lectured internationally on network design and telecommunications service delivery. Prior to joining academia, she was in industry, working on network management and performance. She has been a consultant on a number of government and industrial projects.

Dr. Morreale holds a B.S. from Northwestern University, an M.S. from the University of Missouri, and a Ph.D. from Illinois Institute of Technology, all in computer science. She holds a patent in the design of real-time database systems and has numerous journal and conference publications. With Dr. Terplan, she co-edited *The Telecommunications Handbook* (2000).

Kornel Terplan, Ph.D. is a telecommunications expert with more than thirty years of highly successful multinational consulting experience. His books, *Communication Networks Management*, published by Prentice Hall (now in its second edition, 1992); *Effective Management of Local Area Networks* (now in its second edition, 1996), published by McGraw-Hill; and *Operations Support System Essentials*, published by John Wiley (2001); are viewed as the state-of-the-art compendium throughout the community of international telecom and corporate users.

Over the last twenty-five years, he has designed five network management–related courses and conducted over eighty seminar presentations in fifteen countries. He received his doctoral degree at the University of Dresden and completed advanced studies, researched, and lectured at the University of California at Berkeley, at Stanford University, at the University of California at Los Angeles, at Clemson University (North Carolina), and at Rensselaer Polytechnic Institute in Troy, New York.

Dr. Terplan's consulting work concentrates on network management products and services, operations, business and intelligence support systems, traffic management, service management, outsourcing, central administration of LANs, network management centers, strategy of network management integration, implementation of network design and planning guidelines, product comparisons, and benchmarking network management solutions.

For the last three years, he has been concentrating on intelligence support systems, supporting both telecommunications service providers and law enforcement agencies. His consulting work in this area

includes the selection of technologies for lawful intercepts, the integration of data collection and forensics procedures, and the building of monitoring centers.

His most important clients include AT&T, BMW, Boole & Babbage, Coca-Cola, Commerzbank (Germany), Creditanstalt Austria, Dresdner Bank, Fiducia, Ford Europe, France Telecom, Georgia Pacific Corporation, German Telekom, Groupe Bull, Gruener and Jahr, GTE, Hungarian Telecom, Kaiser Permanente, Salomon Brothers, Siemens, Slovak Telecom, the state of Washington, Swiss Credit, Telcel Venezuela, Union Bank of Switzerland, Unisource Switzerland, and Walt Disney World.

He is Industry Professor at Brooklyn Polytechnic University in New York and at Stevens Institute of Technology in Hoboken, New Jersey. Dr. Terplan has educated and trained over 3,500 subject matter experts in seventeen countries, and consulted with over 2,000 persons in twenty-seven countries.

Contributors

John Amoss
Lucent Technologies
Holmdel, New Jersey

James Anderson
Verizon
New York, New York

Árpád Bakay
Netvisor
Budapest, Hungary

John Braun
Industry Consultant
Weston, Connecticut

Mahmoud Daneshmand
AT&T Labs
Middletown, New Jersey

Jim Frey
NetScout Systems
Westford, Massachusetts

Joe Ghetie
Telcordia
Piscataway, New Jersey

Michel Gilbert
Hill Associates, Inc.
Colchester, Vermont

Paul Hoȧmann
Datakom
Ismaning, Germany

Sajid Hussain
Acadia University
Nova Scotia, Canada

Muhammad Farhat Kaleem
Acadia University
Nova Scotia, Canada

Norman Kincl
Hewlett-Packard
San Jose, California

Matthew Kolon
Hill Associates, Inc.
Colchester, Vermont

Patricia Morreale
Kean University
Union, New Jersey

Dermot Murray
Iona College
New Rochelle, New York

Deepak Pareek
Consultant
Bangalore, India

Mihir Parikh
Polytechnic University
Brooklyn, New York

Teresa Piliouras
TCR, Inc.
Weston, Connecticut

Vadim Rosenberg
CA-Wily
Islandia, New York

Endre Sara
Goldman Sachs & Co.
New York, New York

Kazem Sohraby
University of Arkansas
Fayetteville, Arkansas

Tivadar Szemethy
Netvisor
Budapest, Hungary

Kornel Terplan
Industry Consultant and
Professor
Hackensack, New Jersey

Christian Voigt
Siemens AG
Muenchen, Germany

Chonggang Wang
University of Arkansas
Fayetteville, Arkansas

József Wiener
T-Com
Budapest, Hungary

1

Voice and Data Communications

Michel Gilbert
Hill Associates, Inc.

Patricia Morreale
Kean University

Matthew Kolon
Hill Associates, Inc.

John Amoss
Lucent Technologies

Chonggang Wang
University of Arkansas

Mahmoud Daneshmand
AT&T Research

Kazem Sohraby
University of Arkansas

Sajid Hussain
Acadia University

Muhammad Farhat Kaleem
Acadia University

Introduction **1**-1
1.1 Computer Telephone Integrated (CTI) **1**-2
Basic Definitions • A Brief History of CTI • Components and Models • CTI Applications and Trends • Conclusion
1.2 Voice over IP **1**-13
The Integration of Voice and IP Data • Applications for Voice over IP (VoIP) • A Component-Based Overview • Keys to Successful Deployment
1.3 Local Area Networks **1**-21
Overview • IEEE 802.3 (CSMA/CD) Specifics • IEEE 802.2 Logical Link Control Layer • Building Cabling Specifications
1.4 RFID Architecture and Protocols **1**-39
Introduction • RFID Architecture • Gen-2 RFID Protocol • Gen-2 Performance Improvement • Conclusions
1.5 Design of Wireless Sensor Network Applications, Hardware and Software **1**-51
WSN Versus Conventional Networking • Design of WSNs • WSN Research • Motes • Hardware Components • TinyOS • Summary
1.6 Multimedia Applications for Cognitive Radio Networks **1**-58
Cognitive Radios and Cognitive Radio Networks • Dynamic Spectrum Access • Cognitive Radio Devices • Policies for Cognitive Radio Operation • Quality of Service (QoS) • Pricing Schemes for Multimedia Applications • Summary
Summary **1**-70

Introduction

The Internet started as a technological revolution, designed to protect national interests by ensuring redundancy and resiliency in governmental networks, particularly in time of war. It has spawned a worldwide cultural revolution, fostering universal communication exchange with limitless geographic, time, and subject matter boundaries. The extent and ease of the Internet's adoption has had profound implications on all—including personal, business, and governmental—aspects of life. There is no place on Earth that cannot be reached by the Internet. The success and complexity of the Internet is continuing to be realized. The building blocks for today's Internet are presented as well as detailed outlines of advanced services such as radio frequency identification (RFID) and wireless sensor networks, with multimedia applications. The evolution of communications networks and services continues.

1.1 Computer Telephone Integrated (CTI)

Michel Gilbert

In the universe of telecommunications, the worlds of voice and data have long been resistant to unification. The basic principles that underlie the two worlds have led to, at best, an uneasy truce. In recent times, however, integration has become the buzzword. The industry has seen the emergence of one technology after another that attempts to draw these two domains into closer proximity. Computer telephone integration (CTI) is yet another arena in which data and voice encounter one another. In the CTI arena, however, voice and data appear to be on the cusp of a working relationship. This paper introduces and reviews the concepts that underlie the world of CTI, the elements that comprise a CTI application, and the standards that have emerged.

1.1.1 Basic Definitions

There are four key elements to this definition: (1) identifying CTI as a technology, (2) a focus on the integration of voice and data, (3) specifying a functional integration, and (4) the need to derive tangible benefits in a business environment.

First, some would dispute the notion that CTI is a new technology. They would suggest that CTI is actually a new application for preexisting technologies. This is indeed the case. Not only is CTI simply a place to reuse existing technologies, it is also not (as we shall see) particularly new.

Second, the integration of voice and data is a key element in CTI, as the name itself implies. CTI builds on some remarkable convergence points in the evolution of computing and telephony. One of the earliest telephone exchanges was designed in 1889 by a frustrated funeral director! Almond B. Strowger was tired of seeing his competitor get the bulk of the funeral business by virtue of the fact that his competitor's spouse happened to operate the local telephone exchange. To deal with the problem, Strowger designed a telephone exchange that became generally known as a step-by-step (or stepper) exchange. Fifty-four years later, with funding from IBM, Howard Aiken created the Harvard Mark I. Both systems were entirely electromechanical, monstrous in size, and highly rigid in their design. Over the years, however, both computers and switches became entirely electronic and based on solid-state technologies.

Where early switches and computers tended to be hardwired, modern switches and computers are both stored-program machines and very flexible. The switch uses a stored-program model to handle call routing operations. The computer uses a variety of stored programs to support end-user applications. Both depend on a data communications infrastructure to exchange control information. Finally, the telephone network is rapidly converging to the digital communications model, which computers have used almost from the outset.

Telephone switches have become specialized computers designed to provide a switching function, and exchanging information via a complex digital data communications infrastructure.

The third major part of the definition, functional integration, requires a brief sidetrack to examine the anatomy of a phone call. A phone call can be divided into two logical activities, commonly referred to as call control and media processing. Call control is concerned with originating, maintaining, and terminating a call. It includes activities like going off-hook, dialing the phone, routing a call through a network, and terminating a call. Media processing is concerned with the purpose of the phone call. It deals with the type of information being conveyed across the call, and the format in which that information is presented.

Functional integration means the computer and switch collaborate in call control or media processing operations. They may actually interchange functions to meet the needs of an application. Data stored in the computer might be useful for routing incoming and/or outgoing calls. Perhaps the simplest example is an autocall application where the user can click on a name stored in a local

application and the computer retrieves the associated phone number and dials the call automatically. Alternatively, call-related data can be used to trigger information retrieval from the computer. For example, automatic number identification (ANI) can provide the calling number, which can be used to key a database lookup to retrieve a particular customer's account information before the phone even rings. In both examples, the data of the computer and the routing of a call are bound together to do work.

Another form of functional integration is when computer and telephone peripherals begin to be used interchangeably. For example, computer peripherals can become alternative call control elements instrumental in call monitoring, and telephone network peripherals can become an alternative method for moving data between people and computers. There is even a degree of functional integration achieved when the computer and telephone system are managed from a single point.

The fourth and final element of Levick's definition concerns the benefits CTI brings to business applications. One of the obvious goals of any business application is to provide better service to customers. CTI can increase responsiveness, reduce on-hold waiting times, provide the customer with a single point of contact, and make it easier to provide a broader range of services.

CTI can also increase effectiveness by eliminating many of the mechanical tasks associated with telephony (e.g., dialing phones, looking up phone numbers, etc.), providing a better interface to the telephone system, and integrating control of the phone system into a familiar and regularly used computer interface (e.g., the familiar Windows desktop).

Perhaps the most telling benefit CTI brings to the corporate world (and the one most likely to garner the attention of the decision makers) is the potential for reductions in operating costs. Correctly applied, CTI can mean faster call handling, which translates to reduced call charges. Automation of call-related tasks means potentially fewer personnel or greater capacity for business with existing personnel. Some CTI implementers have claimed 30% improvement in productivity.

1.1.2 A Brief History of CTI

Although CTI appears to be a recent introduction into the telecommunications arena, there were attempts to integrate voice and data into competitive business applications as early as the 1960s. In his book *Computer Telephone Integration* (ISBN 0-89006-660-4), Rob Walters describes an application put together by IBM for a German bookstore chain.

The bookstores were looking for a way to automate their ordering process. IBM produced a small, hand-held unit that each store manager could use to record the ISBN numbers of books they needed, together with the desired quantity of each. These small units were then left attached to the telephone at the end of the day. Overnight, an IBM 360 located at company headquarters would instruct the IBM 2570 PABX to dial each store in turn.

Once the connection was formed, the IBM mainframe would download the order and then instruct the PABX to release the connection and proceed to the next store. The link between the IBM 360 and the 2750 PABX was called *teleprocessing line handling* (TPLH). By the end of the night, the 360 would produce a set of shipping specifications for each store, the trucks would be loaded, and the books delivered.

In 1970, a Swedish manufacturer of ball bearings (SKF) replaced its data collection infrastructure with a CTI application that was also based on the IBM 360/2570 complex. Rather than using data collectors who would travel from shop to shop, local shop personnel provided the data directly. On a daily basis, they would dial a number that accessed the IBM 360/2750 complex at headquarters. Data was entered using push-button phones. The switch would pass an indicator of the numbers pressed to the 360 via the TPLH connection, and the computer would return an indication of acceptance or rejection of the data to the switch. The switch would, in turn, produce appropriate tones to notify the user of the status of the information exchange.

These two examples underscore the flexibility of this early system. Note that both outbound (IBM 360 initiates the calls) and inbound (users call the IBM 360) applications were supported. This system exhibited two classic hallmarks of a CTI application. First, the phone connection is used for media processing (i.e., the information being passed back and forth). Second, there is a linkage between the computer and the switch to exert call control.

Amazingly, after IBM's introduction of the 360/2570 applications, there was an attempt at a form of electromechanical CTI, albeit a short-lived one. In 1975, and largely in response to the IBM 360/2570 solution, the Plessey Company designed a computer link to their crossbar PABX. Every line and every control register of the switch was wired to the computer so its status could be monitored and controlled. The computer could intercept dialed digits, make routing decisions, and instruct the switch to route a call in a particular fashion. Called the System 2150, only two were deployed before electronic switching rendered the technology obsolete.

At about the same time, a group of Bellcore researchers formed the Delphi Corporation to build a system for telephone answering bureaus. These bureaus were essentially answering services for multiple companies. At the end of the day, the company phones were essentially forwarded to these bureaus, where a person would answer the line and take a message. However, it was important for the person answering the phone to know what company was being called, and to be able to answer the phone as a representative of that company. Delphi 1, released in 1978, was the answer to the problem.

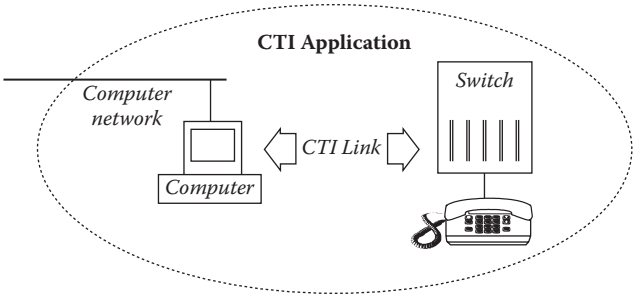
All calls were rerouted to a computer that could tell by the specific line being rung which company was being called. The computer would then retrieve the text for that company's standard greeting, as well as any special instructions for handling the call, and pass the call and instructions to an attendant. The answering bureaus saw a 30% increase in efficiency and the concept caught on quickly.

Through the 1980s, niche applications continued to appear, and new players entered the market. These included British Telecom (a telemarketing application), Aircall (paging), and the Telephone Broadcasting Systems (a predictive dialing system). Perhaps one of the best-known CTI applications to emerge in the 1980s was Storefinder™. The results of collaboration between Domino's Pizza and AT&T, Storefinder™ used ANI to route a call to the Domino's Pizza nearest that customer. Before the phone in the store could ring, Storefinder™ provided the personnel at that store with the customer's order history, significantly enhancing the level of customer service.

Many early attempts to integrate computers and telephony focused on the media processing aspect of communication. This includes early versions of voice mail and interactive voice response (IVR) systems. These simple technologies did not need much more than specialized call receiving hardware in a computer system, and a hunt group. When a caller dialed in to the service, the telephone network switched the call to one of the access lines in the hunt group. The computer then proceeded to provide voice prompts to guide the user through the service. In the case of voice mail, the user was prompted to leave or retrieve recorded messages. In the case of IVR, the user was prompted to provide, by touch-tone or voice, the information necessary to perform a database lookup (e.g., current credit card balances, history of charges, mailing address, payment due dates, etc.).

Modern voice mail and IVR systems, and more advanced CTI applications, include a strong call control component. They can transfer calls, provide outward dialing, and even paging. This requires a more complex physical and logical integration of the computer and telephony worlds. The two worlds must be physically connected, making it possible for data from the telephone network to be passed to the computer and call control information from the computer to be passed to the network. Logically, the integration of data from both the telephone network and the computer must be used to create new applications that give the corporation a competitive edge.

Today, the call center scenario dominates that CTI world. Resulting applications typically utilize the most advanced call control and media processing functions. CTI enables new call center models. A single call center can be logically partitioned to function as multiple smaller call centers, or multiple distributed call centers can be logically integrated to act as one. Modern CTI applications provide the knife, or the glue, to make these models possible.



FIGu RE 1.1.1 Basic components of a CTI application.

1.1.3 Components and Models

The basic components of a CTI application are depicted in Figure 1.1.1. At the heart of the application lie the computer and the switch. The computer houses end-user data and hosts the end-user interface to the CTI application. The switch provides the ability to make and receive calls and hosts the network interface to the CTI application. The computer provides a set of peripherals (e.g., keyboard, screen, etc.) by which the user accesses the CTI application, and the switch provides the peripheral (e.g., telephone) by which the user communicates. Between the computer and switch there must exist a connection or link, the nature of which differs depending on the type of CTI application.

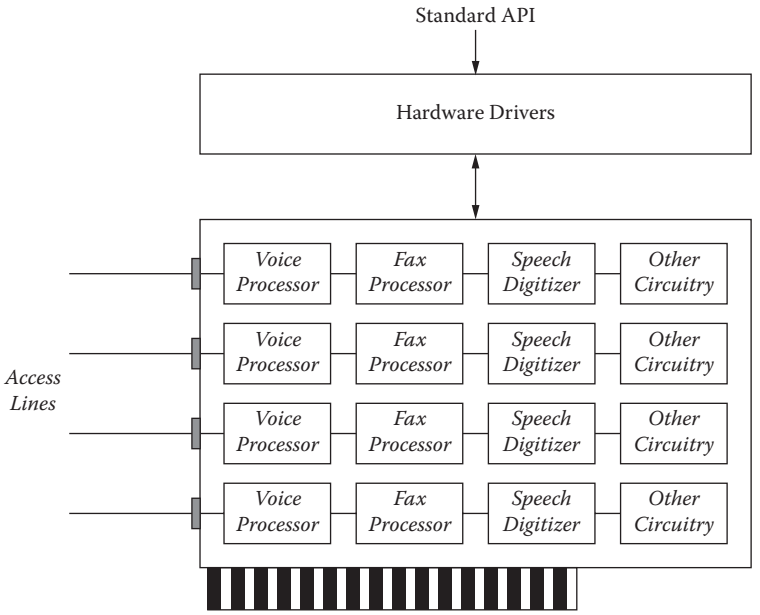
Consider the automated attendant application. A person needing to speak with someone within the company dials the company's published phone number. The switch routes the call to a computer that begins to play back a recorded message. The message prompts the caller to use the touch-tone buttons to select from an array of options. The caller can enter the extension of the person they wish to reach, in which case the computer directs the switch to reroute the call to that extension. The caller can use the keypad to enter the name of the person being reached. The computer has to translate each tone to the associated letter values, and determine if there is a match in the company personnel listing. If there is none, or if the match is ambiguous (e.g., "Sam" and "Pam" use the same key combination), the computer asks the caller to hold and transfers the call to an operator. If a single, unambiguous match is found, the computer can ask the caller to confirm the match, retrieve the extension from the database, and direct the switch to transfer the call. At any point the caller can force the computer to transfer the call to an operator by pressing 0.

1.1.3.1 Media Processing

As has been noted, any phone call can be broken down into two broad activities: media processing and call control. CTI applications typically support both, albeit in different degrees of complexity and by using different strategies. However, a complete suite of CTI services requires both media processing and call control services.

Media processing is perhaps the easiest to understand. When a fax machine calls another fax machine, the transmission of the encoded image across the connection is media processing. When an end user uses their modem to dial in to the local Internet service provider (ISP), the exchange of data across the connection is also media processing.

In the CTI arena, the hardware required for media processing is relatively simple. It often takes the form of voice processing, speech digitization and playback, and fax circuitry. Many products integrate these functions into a single printed circuit board that can be installed in a desktop computer. Many of these integrated boards support multiple lines and hardwire the circuitry to each channel. This is sometimes referred to as dedicated media processing hardware (see Figure 1.1.2). Companies that provide such integrated boards include Dialogic Corporation (www.dialogic.com), Pika Technologies, Inc. (www.pika.ca), and Rhetorex (www.rhetorex.com). Rhetorex is now a subsidiary of Lucent Technologies (www.lucent.com).



FIGu RE 1.1.2 Dedicated media processing hardware.

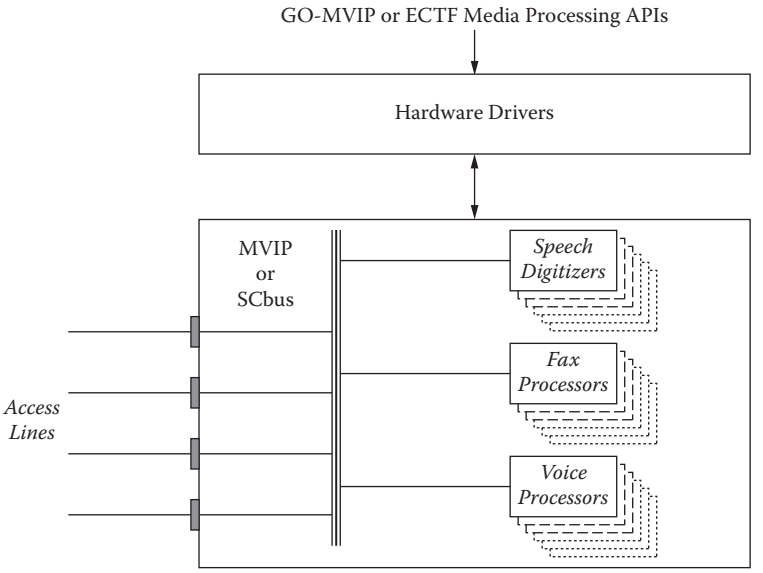
This approach is appropriate for small-scale applications. For example, a company providing voice mail services in a small town might equip a standard desktop system with a four-line integrated board. A user dialing into the service would be switched by the network to one of the four lines. Based on the tones provided by the user (e.g., "Please enter your mailbox number") or ANI information provided by the network, the user can retrieve recorded messages from the computer and play them back.

In these simple environments, standard application programming interfaces (API) are often adequate for controlling the resources. For example, the Microsoft Windows or Solaris APIs that are used to play sound files through a local speaker can also be used to send and receive multimedia content over a telephone connection.

Large-scale applications, however, are more complex. In these environments, sharing resources is more economically viable. A businessperson may be willing to purchase four complete sets of media processing circuitry, knowing that at any given time only a few components associated with any particular line are going to be used. However, equipping every line in a large application with all of the circuitry it might be called upon to use is not cost effective. For example, consider a large-scale application that implements a pool of four T1 circuit interfaces (96 voice channels). Usage patterns may show that this application needs 96 voice digitizers and playback units, but only 16 speech recognizers, 16 fax processing circuits, and 36 analog interfaces for headsets.

Assembling components at a more modular level is more cost effective and can scale more easily, but it also places new demands on the system. New APIs and standards are required for interconnecting, using, and managing these resources. There are two leading architectures for building such systems: the multivendor integration protocol (MVIP) and SCbus. In addition to describing the hardware architecture needed to interconnect telephony-related components, both Global Organization for MVIP (GO-MVIP) and Signal Computing System Architecture (SCSA™) define software APIs required to use and manage those resources (see Figure 1.1.3). The SCSA Telephony Application Objects (TAO) Framework™ is the API defined by the SCSA.

On the hardware side, both MVIP and SCbus describe a time-division bus for talk-path interconnection, and a separate communication mechanism for coordinating the subsystems. MVIP (www.mvip.org) is administered by the Global Organization for MVIP. SCbus was originally developed by the



FIGu RE 1.1.3 Architecture for sharing media processing hardware.

Signal Computing System Architecture (SCSA) working group (www.scsa.org). SCSA has since become part of the Enterprise Computer Telephony Forum (ECTF), a nonprofit organization actively prompting the development of interoperability agreements for CTI applications (www.ectf.org). SCbus, announced in 1993, is now also an American National Standards Institute (ANSI) standard.

Both GO-MVIP and the ECTF also define a set of APIs for media processing.

1.1.3.2 Call Control

The other major activity a CTI application needs to support is call control. Call control is concerned with the successful establishment, maintenance, and termination of calls. To support these activities, the switching nodes in the telephone network must communicate with one another and with the end user's terminal equipment. The process by which the switches do this is called *signaling*. Signaling can be done in-band or out-of-band. In-band signaling occurs on the same channel occupied by user information. This is common for terminal equipment (i.e., telephones), and has become less common within the network itself. Out-of-band signaling occurs on a separate channel from that occupied by user data. This approach is common within the telephone network, and less common between the user and the network (ISDN notwithstanding).

In addition to differentiating between in-band and out-of-band signaling, it is important to note that signaling between the network and the user is bidirectional. The user signals the network by going off-hook, dialing a phone number, and hanging up a phone. This signaling is well standardized. The most common standard today is dual tone multifrequency (DTMF), the familiar tones we hear as we press buttons on a touch-tone phone. The network signals the user in-band by providing dial tone, busy signals, ringing tones, fast busy, and so forth. Each of these has a distinct meaning, but the sounds have not been well standardized internationally. This is a significant challenge for the CTI environment. Out-of-band network-to-user signaling is somewhat more standardized. Examples include the D-channel on an integrated services digital network (ISDN) interface, the proprietary interfaces defined by digital telephones, and dedicated CTI interfaces to private branch exchanges (PBX) and switches.

Perhaps the most challenging aspect of CTI applications is achieving accurate and reliable call control. In most applications, out-of-band signaling is preferred. Each option, however, has its scope, strengths, and weaknesses. In an ISDN environment, D-channel signaling can be used by the CTI application.

One possible CTI application is a network-based automatic call distributor (ACD). Naturally the scope is limited to the domain for which the ISDN signaling is meaningful. For example, the ACD application may not be completely effective when calls cross some public network boundaries.

A CTI application could also leverage the proprietary signaling between a PBX and a digital telephone. Again, such an application may be limited to the scope of the PBX or a group of PBXs from the same manufacturer.

In the public network, the switch-to-switch signaling protocol is called *Signaling System 7* (SS7). The domain for SS7 signaling can be as large as an entire public telephone network. Unfortunately, SS7 is usually not available to the CTI application. Closely associated with the internal operation of the public network, SS7 access is jealously guarded by most carriers. Where access is available to the corporate customer, a CTI application based on SS7 requires sophisticated customer premises equipment (CPE) that can handle the complexity of SS7. As a result, this signaling option is usually only appropriate for call centers handling large volumes of calls.

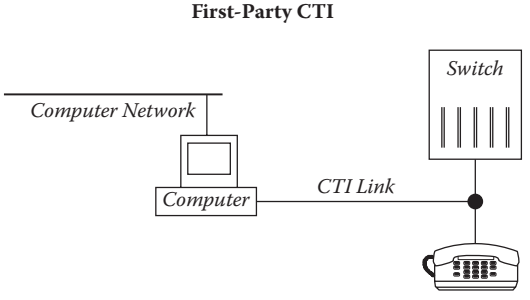
One of the most popular strategies for CTI applications is the dedicated CTI link implemented by many modern PBXs and some public exchange switches. The domain for a dedicated CTI link is a single telephone switch or a small number of tightly integrated switches or PBXs. These facilities are designed for CTI, and tend to offer the rage of signaling options best suited to this environment. These dedicated facilities can implement proprietary or standard call control strategies. Examples of proprietary strategies include Nortel's Meridian Link Protocol (MLP) and AT&T's Adjunct Switch Application Interface (ASAI) Protocol.

Naturally, the industry is leaning strongly to standards-based strategies. The predominant standard is the Computer-Supported Telephony Application (CSTA) from the ECMA (European Computer Manufacturers Association, formerly the European Computer Manufacturers Association). Adopted in 1990, the CSTA protocol (www.ecma.ch) has now been implemented by such major players as Siemens ROLM, Ericsson, and Alcatel, to name a few. It is important to note that, although CSTA is a standard, the features any particular vendor elects to implement can vary. As a result, CSTA implementations from different vendors are not necessarily interoperable.

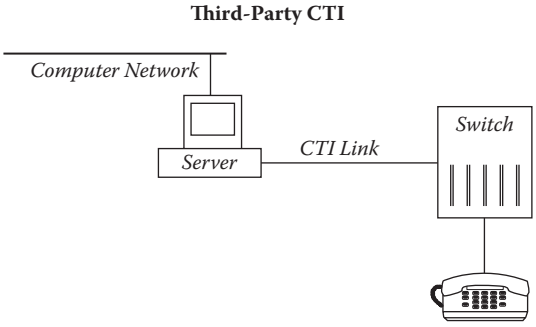
1.1.3.3 First-Party and Third-Party CTI

CTI applications can be broken into two broad classes based on the relationship between the computer and the switch. In first-party CTI, the computer is essentially on an extension to the line on which a call is being received. The computer can exert the same call control functions a human attendant could exert via a standard telephone set attached to the telephone system. This implies that call control is on a call-by-call basis. First-party CTI call control includes such activities as going off-hook, detecting dial tone, dialing a call, monitoring call status signals (e.g., ring, ring no-answer, answer, busy, and fast busy) conditions, and terminating the call.

In the first-party CTI model (Figure 1.1.4) the computer, the keyboard and screen, and the phone are all on the same line. The computer will tend to use the dedicated media processing hardware model, and tend to be a user end-system (as opposed to being a server). First-party CTI is further subdivided into



FIGu RE 1.1.4 First-party CTI model.



FIGu RE 1.1.5 Third-party CTI model.

basic and enhanced flavors. Essentially, basic systems use in-band signaling and have limited capability. Enhanced systems use out-of-band signaling, usually either ISDN or proprietary signaling to the PBX. While there are basic first-party CTI platforms on the market, the industry is more interested in enhanced first-party CTI systems.

The classic example of an inbound first-party CTI application is the voice mail system. In a voice mail application, an inbound call is received by the computer. The computer activates the local voice mail software to record and store, or retrieve and playback, voice mail. The simplest example of an outbound first-party CTI application is autocall.

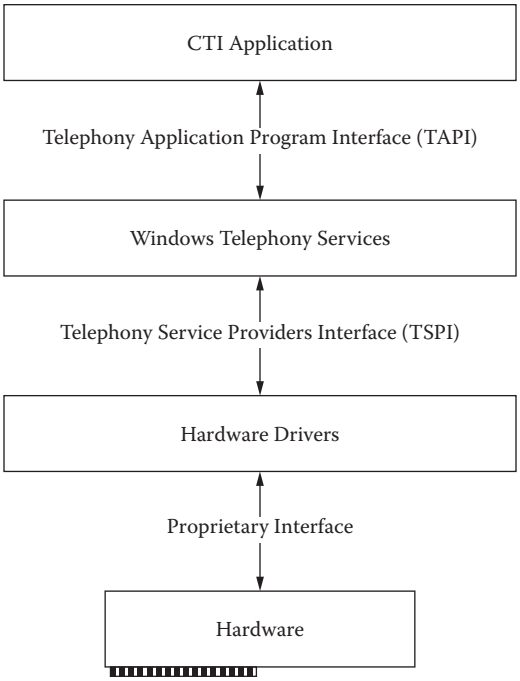
APIs for first-party call control first appeared from the manufacturers of network access equipment (e.g., modems, fax boards, etc.). The only such API that achieved de facto standards status was the Hayes modem command set. Now universally understood by modem products, the Hayes command set defines basic commands for initiating and terminating calls, and altering the configuration of the modem.

Third-party CTI is the more sophisticated model. In third-party CTI, the computer exerts call control via a dedicated connection to the switch or PBX (Figure 1.1.5). This naturally implies out-of-band signaling. It also implies that call control can be exerted over several calls, or over the switch itself. The call control functions third-party CTI application could exert are similar to those a human attendant could exert using a specialized telephone set with enhanced privileges, such as an operator's console.

In the third-party CTI application, the computer, the keyboard and screen, and the phone have no relationship to one another unless the computer establishes one. These environments tend to use the shared media processing hardware model, and tend to perform signaling via SS7 or (more commonly) dedicated CTI links implementing the CSTA protocol. The CTI link typically terminates in a server rather than a specific application end-system.

There are three basic flavors of third-party CTI, which reflect the essential relationship between the computer and the switch. In the *compeer* model, the computer and switch are on equal terms. Each operates as the master of its own realm, passing information and receiving instructions from the other across a specialized interface. In the *dependent* model, the computer rules and the switch obeys. The switch has no innate call handling capability, and is actually incapable of processing calls without receiving instructions from the computer. Finally, the *primary* model is virtually identical to the compeer model, but the computer and switch do not share a specialized link. Rather, the computer attaches via a standard trunk or line port. Over the years, the dependent and primary models have seen diminishing emphasis as the market moves toward the compeer model. Unless explicitly identified as dependent or primary, third-party CTI is usually assumed to operate on the compeer model.

Automatic call routing applications are classic examples of third-party CTI. A server-based application is alerted, by the switch, to the arrival of a call. Based on ANI information, or the specific Dialed Number Identification Service (DNIS; i.e., called number), the computer directs the switch to divert the call to a specific line.



FIGu RE 1.1.6 The TAPI architecture.

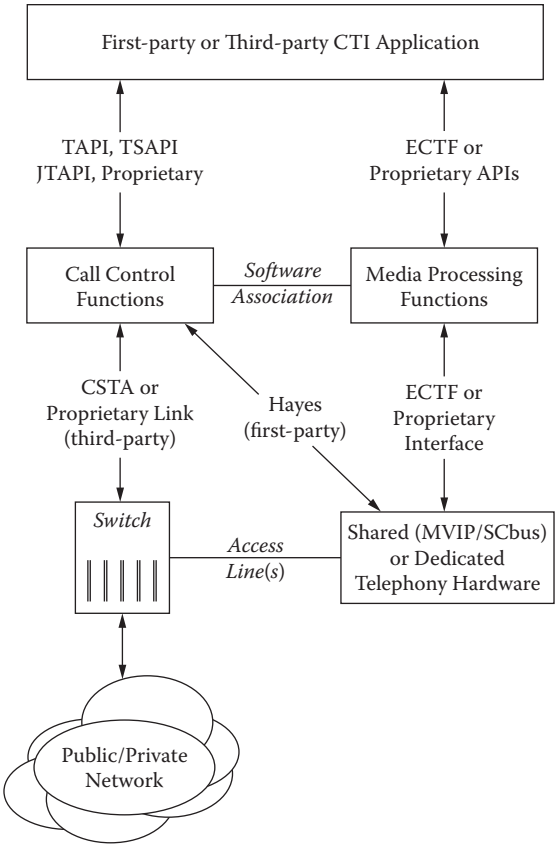
As with first-party CTI, the first third-party APIs were developed by manufacturers to support applications running on their own systems. Examples included the CallPath API from IBM, and the Computer-Integrated Telephony (CIT) API from Digital Equipment Corporation (DEC). Unlike the Hayes command set, however, none of these have achieved de facto standard status.

In the 1990s, three major APIs emerged, all strongly associated with a particular computing environment. Novell (www.novell.com) and Lucent collaborated to create the Telephony Services API (TSAPI). Novell's commercial product based on TSAPI is called NetWare Telephony Services, which links applications on remote clients with telephone system driver modules. TSAPI defines the boundary between CTI application software, and the drivers that control the links and signaling into the network.

Microsoft (www.microsoft.com) and Intel collaborated to create the Telephony API (TAPI). Like TSAPI, TAPI is concerned with call control. However, the TAPI architecture actually defines two distinct interfaces (see Figure 1.1.6). The first interface resides between CTI applications and the Windows operating system (OS). This interface, which unfortunately has the same name as the overall architecture, provides a standard means for CTI applications to access the telephony services provided by the Windows OS.

The second interface resides between the Windows OS and the CTI hardware drivers. Known as the *telephony service providers interface* (TSPI), this interface provides a standard mechanism for hardware vendors to write drivers that can support the telephony services provided by Windows. It is Microsoft's job to ensure that TAPI-compliant applications can access all of the resources provided by TSPI-compliant hardware drivers.

The third call control API is the more recent and brings CTI into the world of the Internet and the World Wide Web (WWW). Developed jointly by design teams from Sun, IBM, Intel, Lucent, Nortel, and Novell, the Java Telephony API (JTAPI) defines a call control interface for CTI applications running as Java applets. This opens the door to creating Web-based CTI applications. The Sun Microsystems product that implements this API is called JavaTel™.



FIGu RE 1.1.7 Combining the standards and components.

Figure 1.1.7 integrates the various standards and concepts introduced in this paper into a single CTI model. A CTI application can be either first-party or third-party. First-party applications tend to use local, proprietary APIs (e.g., the Windows APIs) to access local call control and media processing services, and the Hayes command set to control dedicated telephony hardware.

Third-party CTI applications tend to use sophisticated call control APIs like TAPI, TSAPI, or JTAPI, and standardized media processing APIs like those defined by the ECTF. The link between the CTI server and the switch commonly implements the CSTA protocols. The server typically uses shared telephony hardware that is interconnected using the MVIP or SCbus architecture.

It is also possible to build a CTI server that supports several APIs and standards simultaneously. Such a product would have to map requests from all APIs into a single common function set. Dialogic's CT-Connect product takes this approach. It supports both the TAPI and TSAPI interfaces and includes built-in drivers for the ECMA CSTA link protocol and several other proprietary CTI link protocols.

1.1.4 CTI Applications and Trends

A few of the more common, and simpler, CTI applications have already been noted: voice mail, autocall, and automatic attendant. Each of these is commonly implemented as first-party CTI applications using dedicated media processing hardware. Digital dictation is another CTI application that is virtually identical to voice mail, but typically supports longer record times. The recorded dictation is usually retrieved and transcribed locally.

Many companies are beginning to provide interactive or on-demand fax services. For example, the real estate company could provide automated faxes of current properties for sale. In such a service, the user dials in and, using a touch-tone-driven menu system, requests a particular fax or group of faxes and provides the number to which the fax is to be sent. The service retrieves the fax from a local file, initiates an outbound call to the specified number, and transmits the fax. As with the automated attendant application, interactive fax could be implemented as a first-party or third-party application.

Many pay-per-call applications are CTI applications. This is a common strategy for implementing fee-for-access Internet services. The user dials a 900 number and the PBX routes the call to the CTI application. The user is prompted to provide a code identifying the service they are trying to access. The CTI application provides an access code that permits the user to access the web site. The phone service bills the user for the 900 call and passes the majority of the fee to the pay-per-call service provider. The pay-per-call service provider takes an additional cut and passes the remainder of the fee to the company hosting the web service.

Perhaps the most common third-party CTI application is the inbound and outbound call center. Inbound call centers typically integrate an automatic attendant to collect initial customer information (i.e., credit card numbers, zip codes, pin numbers, etc.) and provide core services (e.g., account balances, mailing addresses, account histories, a list of service or product options, automated order taking, etc.). The caller always has the option, however, to abandon the automated system and speak to a person. In this case, the CTI application routes the call to an available attendant and provides all information the user has submitted. The application may also provide any call information provided by the phone network and any customer data retrieved from the computer's database.

The CTI market is showing clear signs of accelerated growth, fueled by a number of enabling factors in the industry. The pervasive deployment of LANs and internetworks provides the infrastructure over which many first-party and third-party CTI applications operate. The growth in digital communications and integrated networks that provide enhanced signaling capabilities (e.g., ISDN and digital telephones) create a rich set of network information on which CTI applications can be built.

The emergence of standard APIs in both the media processing and call control arenas has furthered equipment and service interoperability. Furthermore, the increasing maturity of voice processing technology makes interactive voice response (IVR) systems easier to deploy and use. Finally, the industry is seeing a broad array of CTI application development toolkits. Examples of these include OmniVox from Apex Voice Communications (www.apexvoice.com), Visual Voice from Artisoft (www.artisoft.com), MasterVox from Mastermind Technologies (www.mastermind-tech.com), and IVS Builder and IVS Server from Mediasoft Telecom (www.mediasoft.com).

1.1.5 Conclusion

The CTI market is a young one, but the technologies coming together into this application environment are relatively mature. As the CTI-related standards themselves mature, interoperability agreements emerge, and economies of scale begin to apply, CTI applications are likely to become pervasive. Furthermore, with the emergence of JTAPI and the increasing drive toward voice over IP (and hence over the Internet), CTI applications are finding a new niche in which to grow. The Internet is a significant niche indeed!

For further information, the reader is recommended to visit the various web sites identified in this chapter. There are also two periodical publications dedicated to CTI, both of which can be accessed via the Internet: *Computer Telephony* (www.computertelephony.com) and *CTI Magazine* (www.tmcnet.com).

1.2 Voice over IP

Matthew Kolon and Patricia Morreale

1.2.1 The Integration of Voice and IP Data

Although voice over IP (VoIP) is an existing technology, it has only recently gained wide acceptance as an alternative to traditional voice systems and public switched telephone networks (PSTN). Many domestic and international corporations spend billions annually on long-distance and international telephony services. Most of that money goes to the basic transit of voice and fax from one location to another. With the continued pervasiveness of intelligent peripheral (IP) networking, a new class of products and services has evolved to move some of that traffic from its traditional home on the public switched telephone network to a variety of packet-switched networks. While many of these new "voice" networks have not previously been considered telephony-class, they are nonetheless attractive because of their low cost.

Interest in VoIP has developed as the technology has been recognized as being capable of helping both service providers and corporations reduce costs by using a single IP network for both data and voice applications. Continued improvements in digital signal processor (DSP) technology, voice packetization techniques, and the networks that IP voice runs over have combined to make the start of the 21st century into the era in which IP telephony begins the transition to a mainstream solution for business.

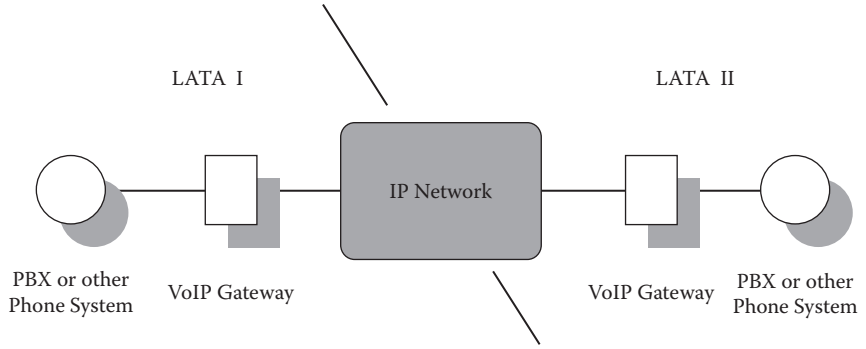
There are a number of reasons for the inevitability of this transformation, but all of them come back to the relief of high-cost long-distance telephone services. Reviewing a few comparative facts regarding the PSTN and VoIP presents some compelling realities:

- **One can fit more voice on an IP network than one can on the PSTN.** The Bell System definition of a single voice channel as a 64-kbps DS-0 has led to a long-standing institutional belief that 64 k is necessary to carry a voice conversation. Thus a T1 is commonly referred to as supporting 23 "voice" channels over its 1.544 Mbps. Yet today's VoIP products can carry hundreds of voice conversations over that same amount of unchannelized bandwidth.
- **Packet networks are much better than they used to be.** Improvements in the quality of physical-layer packet networks over the past 30 years have resulted in a large general improvement in data integrity. The same forces that make simple frame relay an effective replacement for the robust X.25 protocol mean that even connectionless IP data—and voice—may be entrusted to today's connectionless networks and still have an excellent chance of getting through in a reasonable amount of time and with few errors (or little delay) of consequence.
- **Control of IP data networks rests largely in the hands of the customer.** As long as a minimum quality of service—particularly the establishment of maximum delay guidelines—is met, virtually every service available over IP is controllable from the sending and receiving stations. For example, packets may be routed over the Internet for free if tolerant of lower quality, over a private IP network if demanding of higher quality, or even over the PSTN if necessary—all at the discretion of the originating node.

These are just a few of the reasons why many network managers are examining the current and future options for placing portions of their voice traffic into IP networks.

1.2.2 Applications for Voice over IP (VoIP)

Of course, with long-distance services being the single most expensive portion of any company's telephony budget, the application of VoIP to the interexchange carrier (IEC) realm is taking the forefront



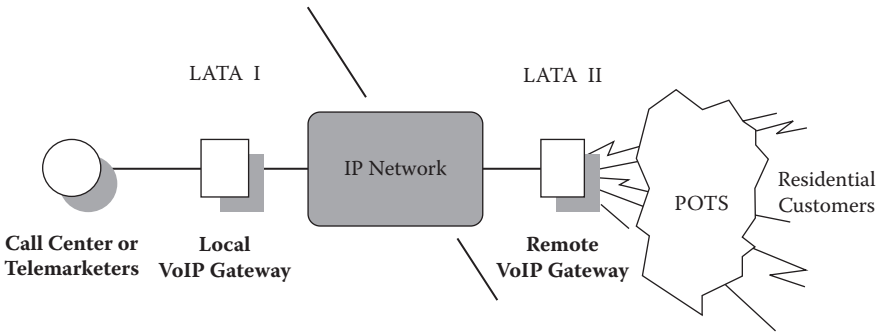
FIGu RE 1.2.1 Business IEC replacement using VoIP.

when it comes to the immediate application of the technology. The basic design of such a network is rather simple: gateways within local calling areas connected by an IP network that spans the distance previously covered by the IEC.

While a company implementing VoIP for the purpose of saving charges on interoffice communications may have a design as simple as that in Figure 1.2.1, it is more likely that the IP network will connect multiple sites, each with its own gateway, each of which may then contact another dynamically when it has a voice call destined for that site. The connectionless nature of IP ensures that new gateways may be added at will, with little need for reconfiguration at the other stations.

Many variations of this scheme are possible, depending upon the nature of the service one is trying to implement. For tie-line replacement and business-to-business calls, the simplest to exploit is that shown in Figure 1.2.1, that is, two or more gateways connected by an IP network. The reason that most pundits consider this setup to be the first area to exploit VoIP is because the difficult part—getting the voice to a few places where it can be digitized and packetized into IP—is already done. The PBX that currently connects via a leased line or IEC to another PBX can easily have that connection replaced by IP, with no changes in how users place calls.

Another application that is generating a large amount of industry interest is that of business-to-residential telephony (Figure 1.2.2), to allow telemarketers or call centers to physically centralize while obtaining low-cost long-distance service via VoIP. In this scenario, residential customers are able to dial a local number and access a VoIP gateway that connects them to the implementer's customer support or sales office, wherever it may be. The customer makes a free call, and receives the same service had an 800 number been dialed, but the company avoids the cost of maintaining 800 service. It is also able to supply customers with a "local" number to call for service, which can enhance the company's image.



FIGu RE 1.2.2 Business-to-residential VoIP network.

Reversing the above strategy—that is, using the remote gateway to *place* local calls rather than accept them—allows telemarketers access to large, yet distant, markets without the need to place large numbers of long-distance calls to get to them.

Yet another option exists for those eager to exploit the possibility of VoIP at their businesses or campus: replacing the PBX and its network with an IP network. Most businesses are already halfway there; they have local area networks (LANs), routers, and digital wide area network (WAN) facilities capable of handling IP traffic. New products, such as 100- and 1000-Mbps Ethernet, as well as the cost-effective speed of LAN switching, mean that network managers can build an enormous amount of capacity into their local and enterprise networks—capacity that might well be used to carry voice traffic. Traditional models for business traffic have always involved the creation and management of two separate networks, one for voice and one for data. The encapsulation of voice in IP packets means that the consolidation of voice into the data network is now possible, with the corresponding reduction in the need for equipment, data facilities, staffing, and expertise in several types of systems. Consolidation of voice traffic and data traffic into the same end-to-end network opens the door to true integration of messaging and telephony systems, such as integrated e-mail and voice mail, and IP-based fax messaging.

The final area of interest for VoIP proponents is that of residential-to-residential connectivity, that is, friends and relatives speaking to each other from handsets or speakerphones integrated into Internet-connected PCs. While this is the application that "proved" the possibility of VoIP, it remains the most difficult application for which to ensure acceptable quality. The difficulty of obtaining quality voice this way has nothing to do with the equipment at the ends of the link, but rather with the lack of guaranteed, or even reliable, values for delay and delay variation over the Internet. Indeed, improvements in low-cost digitization hardware and "Internet telephony" software have made it possible to have a full-featured, high-quality VoIP gateway for the cost of a new PC. But even the best-quality digital voice will be unintelligible if only half of it arrives at the intended destination.

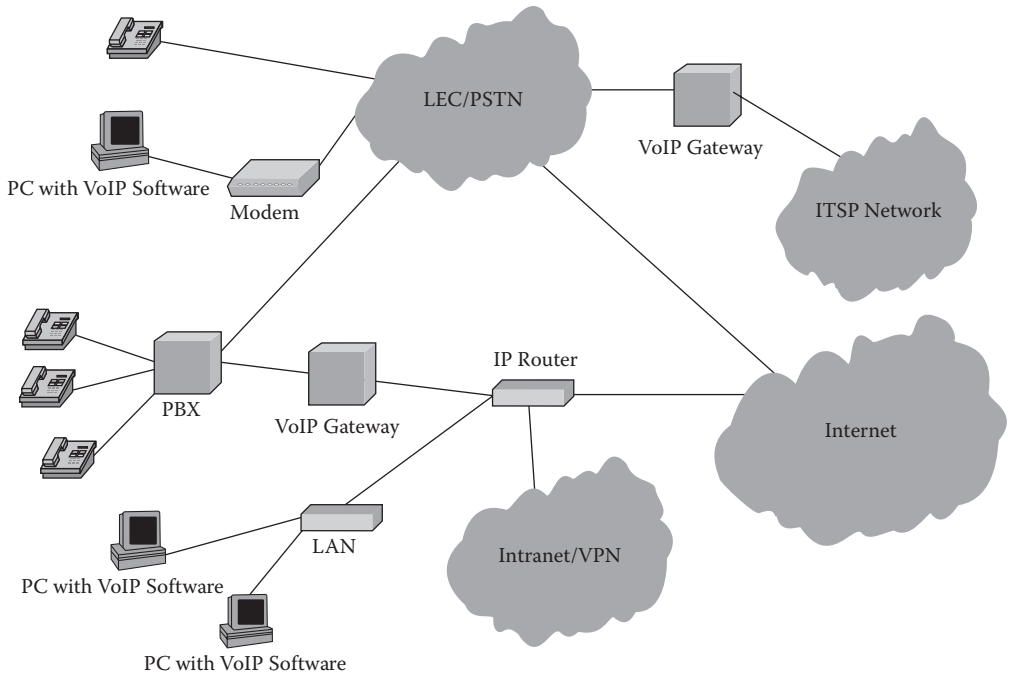
These are just the basic categories into which some of the most obvious applications for VoIP fall. But applications are as numerous as those for the telephone itself—perhaps even more so. The lower cost of VoIP means that some uses for telephony that were once deemed uneconomical may now be justified. And the integration of voice and data traffic over a single IP network may make some forms of integration possible that were unthinkable just a few years ago.

1.2.3 A Component-Based Overview

What are the components of a successful IP telephony system? While there are of course a number of different approaches, there are a few basic ingredients that all systems must implement, although the use and location of parts changes with different network designs.

The VoIP Network: In the list of VoIP components (Figure 1.2.3), the IP network(s) over which the voice will travel is of primary importance. IP is first and fundamentally a connectionless protocol, with no guarantees concerning the traffic that it carries, and a VoIP service in this environment is understood to be "best-effort" network service. It cannot ensure a minimum, maximum, or variability of delay; cannot retransmit errored or lost packets; and does not even promise that its payload will arrive at all. The quality of service one receives from the PSTN, and that provided by even the most carefully managed and overbuilt IP network, do not bear comparison. And for those thinking about using the Internet as the equivalent of their current expensive IEC service, suffice it to say that when a web page often takes 60 seconds to download, sending real-time voice traffic over that same series of links will be a challenge. Until the Internet infrastructure is managed under an agreement that includes concrete plans to provide some limited and predictable delay—in an interprovider fashion—voice traffic cannot travel the Internet and maintain the quality that business customers demand. It's worth mentioning that this agreement is nowhere in sight.

That does not mean that today's Internet has no place in the voice network, however. VoIP gateways can use the Internet to provide the non-real-time services that constitute much of today's "voice"



FIGu RE 1.2.3 VoIP network components.

traffic. The most obvious one of these is facsimile transmission. While fax machines thrive on the dedicated lines of the circuit-switched PSTN, there is no reason why their transmissions cannot be placed in IP for long-distance transit. Delay—the reason why interactive voice is so difficult over the Internet—doesn't affect fax transmissions at all, and transmission control protocol/Internet protocol (TCP/IP) can resend data until the network gets it right without bothering the receiver. The same could be said for voice mail messages.

The next step between the very public Internet and a completely private IP network is the ISP backbone itself, which is nothing more than a single provider's portion of the Internet. If this network extends close to the points where gateways will be placed, IP traffic between them may remain solely on that network. In almost all circumstances, this will result in less delay and better predictability for traffic of all types. But while the statistics for network performance may improve in a single-provider environment, the lack of user control over these fundamentally public networks may be unacceptable for the network manager who seeks to have some influence over the environment in which his traffic travels. Single Internet service provider (ISP) IP telephony, though, has the lowest cost of any of the non-Internet options, and therefore is attractive as long as acceptable quality can be achieved. This may be a matter of simply trialing a number of ISP networks and choosing the one with the best performance, or may actually involve a level of performance, with stated delay and throughput characteristics, to be specified in the user contract.

Luckily, the Internet and its constituent networks are not the only options for long-distance carriage of VoIP. Many of the larger ISPs offer, in addition to their public Internet network, access to a separate IP network designed for virtual private network (VPN), intranet, extranet, and other semiprivate usage. These networks are not any more remarkable in concept than an average ISP's network, except for their managed nature; that is, the knowledge the provider has of just how much traffic any one user is likely (or allowed) to subject the network to at any one time—something unheard of on an Internet access network. This knowledge allows the provider to predict and maintain a high level of quality, which can result in service level agreements in which end-to-end delay is specified to be well below 0.5 seconds,

the point at which telephony starts becoming reasonable. In this environment, service-level agreements (SLAs) are becoming the rule rather than the exception.

The ultimate VoIP network, however, is the one where all aspects of IP traffic and performance can be managed by the users—a completely private intranet. Formed from private (leased) lines, with perhaps some links composed of frame relay or asynchronous transfer mode (ATM), the distinguishing characteristic of these networks is that they are completely under the control of the network managers who deploy and run them. Therefore, the amount of bandwidth reserved for voice traffic can be strictly controlled, as can the throughput of routers and other connectivity equipment. How those resources are actually apportioned may vary from protocol-based reservation systems like reservation protocol (RSVP) to completely manual intervention, but whatever the method, the manager has the ability to restrict the effect of data traffic that interferes with voice. While this sounds like, and in fact is, the ideal environment for packetized voice, it comes with a price. Completely private IP networks are by far the most expensive way to ship IP from one location to another. Whether the establishment of such a network is worth the ability to carry voice effectively depends on how much money can be saved by eliminating IEC charges from the IT budget.

If the number of options and the headaches of managing another network service are a serious disincentive, another possibility is to leave the network and its management to the specialists — that is, to contract with one of the growing number of Internet (or IP) telephony service providers (ITSPs). An ITSP functions as a plug-and-play replacement for a traditional IEC, by providing the gateway, network, and management needed to make VoIP successful. The trade-off here, of course, is that since the ITSP does all the work, it also reaps some of the rewards. Typically, ITSPs function like an IEC in terms of billing, with per-minute rates that range from one half to three quarters that of comparative IECs.

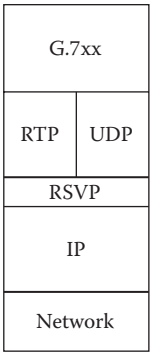
That level of discount may change before long, however. Much of the savings that ITSPs are able to pass on to their customers are possible because of a May 1997 Federal Communications Commission (FCC) ruling that classifies ISPs and ITSPs as end users of the PSTN rather than as carriers. This classification currently makes it impossible for local exchange characters (LECs) to charge ITSPs the same access charges they demand from traditional IECs. Those access charges, when passed on to the IEC customer, can account for as much as one half of the average IEC bill. It is the lack of these charges, more than the technological benefits of VoIP that allows ITSPs to sell services for so much less than their IEC counterparts.

While the level of savings on recurring charges is the least with the ITSP option, it may well be compensated for by the simplicity of setup and management, and the lack of gateway hardware or software costs. The users who benefit from the access charge loophole, however, may have some hard decisions to make if, as many believe will occur, the FCC reverses itself and decides to consider ITSPs as carriers. In that market, much of the price differential would disappear, and users would have to make their decisions based more on quality, service, and other points rather than price (Figure 1.2.4).

All of these networks can and will benefit from work currently underway to allow efficient prioritization of packets containing voice over those containing non-real-time data. Gigabit-speed routers, faster

Network	Gateway	Cost	User Control	Performance
Internet	User-provided	Least	Least	Worst
Single ISP	User-provided	↓	↓	↓
Managed IP	User-provided			
Private IP	User-provided		Most	Best
ITSP	Included in contract	Most	N/A	N/A

FIGu RE 1.2.4 VoIP network options compared.



FIGu RE 1.2.5 VoIP protocol components.

switches, better routing and path-reservation protocols, and the continued addition of cheap bandwidth are all reasons why VoIP quality will continue to increase.

In summary, there are a number of network options for VoIP. Which one best suits a particular need depends on a number of factors, primarily revolving around the level of expected quality. For those looking for a way to lower the cost of interoffice communications—an application where the "internal" aspect may allow slightly lower quality than that required for communications with customers—some of the lower-cost options like single-ISP VoIP networking may suffice. Those wishing to completely replace their IEC contract with an IP-based IEC solution are faced with replacing a complex network from the ground up, and will have to plan, and pay for, a much more robust service. And for the time being, at least, voice over the public Internet remains in the realm of a hobby for those willing to tolerate indifferent and completely unpredictable voice quality.

Gateway Software and Hardware: The hard work of actually taking analog voice and sending it over an IP network, as well as receiving IP and converting it back into voice, is the job of the gateway. It is easiest to examine the issues related to this complex task if we break it down into its components (Figure 1.2.5).

Accept analog or digital voice: A gateway must have some connection to the non-IP world where the voice traffic originates, usually consisting of either a bank of dial-in plain old telephone service (POTS) ports or a digital connection to a PBX.

Prepare the voice signal: In order to use the available bandwidth as efficiently as possible, the voice signal must go through a number of transformations before it is ready to be digitized. First, it must be cleaned up, so that it has as much noise and echo removed as possible. The techniques for doing this have been well established in the traditional telephony world for years, but the cooperation of the various systems and gateways through which voice may pass is essential. This means that calls traveling through an LEC on their way to the VoIP gateway may need to be treated differently than those coming directly from a PBX.

Second, it must be stripped of unnecessary silence, to avoid making the gateway send hundreds or thousands of packets per second carrying nothing. Most gateways have adjustable options for when silence suppression "closes off" and stops transmitting on behalf of a user, but the effectiveness of default settings may depend on usage characteristics that are themselves dependent on cultural factors. Some adjustment of this setting to achieve the best compromise between quality and throughput is usually necessary. Related to the subject of silence suppression is the modeling and regeneration (at the remote end) of background noise, without which users can become disconcerted.

Compress and digitize the voice signal: The standard compression and digitization of voice provided by traditional 64-k pulse-code modulation (PCM) produces a stream of digital data that is enormous compared to that available with many newer codecs. While some vendors have achieved good results with proprietary schemes, most of the industry is settling down to the use of one or another International

Telecommunications Union (ITU) G-series codecs, as specified in their H.323 standard. H.323 is a complex specification for point-to-point and multipoint teleconferencing, data sharing, and telephony over IP. While the full effect of this standard on VoIP-only products remains to be seen, the G.711, G.723, and G.729 codec specifications referenced by it are current favorites for coding voice.

These three standards differ primarily in the amount of work that the DSP must do in order to process the analog signal, and the number of bits that it takes to represent a given amount of voice. While recent advances in DSP design and manufacture have allowed vast improvement in these areas, there remains an inverse relationship between them, and also therefore a higher cost for greater efficiency. Nevertheless, the most aggressive of the standards (G.729) can represent 10 msecs of voice with only 10 octets of IP data. The less intensive G.711 and G.723 trade higher traffic volume for higher quality. Many gateways can be configured to use whichever one of these standards provides the most acceptable tradeoff between quality and traffic level.

Route the call: Once a gateway has a potential stream of packets ready to send, it must have some way to identify the address of the gateway it will send them to, and to inform that gateway of which local user it is destined for (or what local number to dial.) For simple point-to-point applications, IP address can be a manually configured variable, since there is only one destination possible. But in cases where a multipoint network means that packets may be simultaneously distributed among a number of destinations, there must be a process in which the called number is translated into an IP address.

Informing the destination gateway of the called phone number has its complications, too, because many of the codecs used in current gateways compress the analog signal so much that the dual-tone multifrequency (DTMF) tones produced by phones become unreliable. Therefore, the calling gateway must be able to transform those DTMF tones into a code representing the called number and transmit them to the destination gateway for correct routing at the called end.

Packetize and send digital voice in IP datagrams: At first glance, this is the simple part. After all, IP stacks on end stations and routers have been performing this function since the late 1960s. Yet some of the characteristics of packet-switched networks with regard to real-time traffic are different than those regarded as common knowledge by those used to thinking of IP as data-only transport.

For example, the flexible size of an IP datagram, while an advantage in the transmission of data, complicates the problem of achieving low variability of delay, since IP routers handle packets of various sizes differently, and may tend to process smaller packets more quickly than larger ones. The destination gateway would then need to account for the tendency of larger packets to take longer, and thus delay reassembly. In practice, VoIP gateways by default transmit packets of a single size or small range of sizes in order to obviate this problem, but this is one area where the capabilities of the gateways and the network(s) over which they will transmit must be closely matched. Setting the maximum packet size of the gateway to any amount higher than the maximum transmission unit (MTU) of the underlying network will introduce latency as routers fragment datagrams that are too big to travel through the networks attached to them.

Enabling routers to prioritize packets containing voice can enable voice and data to coexist on the same network more easily. Methods for doing this include enabling priority queuing based on transport layer port number, packet size, and source and destination addresses. RSVP can be used to reserve router bandwidth and processing capability, as well as network segment bandwidth, for packets that meet certain criteria, but implementing RSVP demands a network path in which all routers are RSVP-compliant, something that is not likely in a multiprovider (or even some single-provider) scenarios.

Receive, buffer, and decode the incoming stream of VoIP data: Again, this is a well-understood process for data that generally depends upon the IP suite's TCP protocol to retransmit lost data and reassemble segments in the proper sequence before it is passed to the application. VoIP software seldom makes use of TCP, largely because the services it provides introduce far too much latency into the transmission process for them to be useful (an exception to this rule is fax transmission, for which TCP makes sense given the lack of need for real-time treatment of data. Instead, most gateways can use real-time protocol (RTP) as the protocol in which voice data rides. While having no control over delay imposed

by the network, RTP makes it possible to trade a small amount of additional delay for a reduction in the amount of delay variation. This is accomplished by transmitting each packet with a time stamp that can be read by the receiver and used to pass data to the upper layers of the VoIP software with something like the transmitted amount of interpacket delay.

Alternatively, some gateways have the option to send digitized voice in user datagram protocol (UDP) packets, which travel in an unstructured stream, free of sequence numbers, time stamps, and acknowledgments, but also free of the delay imposed by processing these variables. Since the audio stream at the remote end must go on regardless of the actual receipt of data, large numbers of packets that are lost en route simply result in *holes* or *dropouts* in the audio signal. While this sounds as though it would spell the end for reproduction of any reasonable quality, in fact it takes the loss of a relatively large number of packets to create noticeable holes in outbound audio at anything but the highest compression levels. Whether the control and complexity of RTP or the simplicity and speed of UDP will prove to be the most effective way to carry datagram voice remains to be seen.

1.2.4 Keys to Successful Deployment

The large number of configurable variables and the many options within each make configuring VoIP networks a considerable challenge, especially since these networks' main role is to replace some of the most bulletproof networks in the world: those of the PSTN. Aside from performance issues, questions of interoperability abound, particularly for those users who wish to deploy distributed VoIP networks consisting of hardware and software from more than one vendor, and networks from more than one provider.

One thing is certain, though: IP telephony is here to stay. Despite the challenges that network managers face in order to reduce their IEC bills, in at least some applications the payoff is great enough to make the decision to at least trial the technology obvious. The astute manager, however, remembers a few things:

- Few, if any, of the products currently available for VoIP networking work well "out of the box." Nearly everyone who has implemented gateways on either a point-to-point or multipoint basis has a story to tell about the setup and configuration of their system, and the shakedown and subsequent adjustments that had to occur before the network settled down. Almost as invariably, though, they can recount the time that things began to work well, and now can point to users who are happy with the price and performance of the VoIP network.
- Not all VoIP products are the same. Vendors are scrambling to improve quality and add features, and that translates into large variations in product lines—at least until the next revision is introduced.

The good news is that there are many positive signs for those considering putting their trust into VoIP. The current standards situation for components of VoIP products seems to be stabilizing. While any emerging technology—especially ones with such high visibility—generates a large number of proprietary solutions, which are narrowed down by the market, VoIP is one example of how vendors can cooperate. Most of the standards for encoding (the ITU G-series) seem to be settling down for a long period of maturity.

With regard to the network technologies in use, a new generation of network designers and engineers feels more comfortable with IP than with any other technology, including voice traffic. The ubiquity of the Internet and of IP itself have created a large pool of experience from which managers can draw when deploying VoIP. As for the future, knowledge of the workings of Internet protocols is commonplace among graduates of almost any technical program.

While the public telephone network has existed for years, fast public data networks have not existed until recently, and new data networks are being constructed at a staggering rate. Many of these networks will be suitable for voice traffic, and thus can extend the reach of VoIP networking. And the rapid pace of network improvement means that end-to-end latency will continue to drop, which can only mean good things for the quality and success of VoIP.

Acronyms

ATM	Asynchronous Transfer Mode
DSP	Digital Signal Processor
DTMF	Dual-Tone Multifrequency
FCC	Federal Communications Commission
IEC	Interexchange Carrier
IETF	Internet Engineering Task Force
IP	Internet Protocol or Intelligent Peripheral
ITSP	Internet (IP) Telephony Service Provider
LAN	Local Area Network
LEC	Local Exchange Carrier
PBX	Private Branch Exchange
PCM	Pulse Code Modulation
PSTN	Public Switched Telephone Network
RSVP	Reservation Protocol
RTP	Real-Time Protocol
SLA	Service Level Agreement
UDP	User Datagram Protocol
VoIP	Voice over IP
WAN	Wide Area Network

1.3 Local Area Networks

John Amoss

1.3.1 Overview

1.3.1.1 Standards

The Institute of Electrical and Electronics Engineers (IEEE) 802 Local and Metropolitan Area Network Standards Committee has the basic charter to create, maintain, and encourage the use of standards for local and metropolitan networks. In the IEEE 802 Committee context, the term *local* implies a campus-wide network and the term *metropolitan* implies intracity networks. The IEEE 802 Committee defines interface and protocol specifications for access methods for various local area network (LAN) and metropolitan area network (MAN) technologies and topologies. The project has had a significant impact on the size and structure of the LAN market.

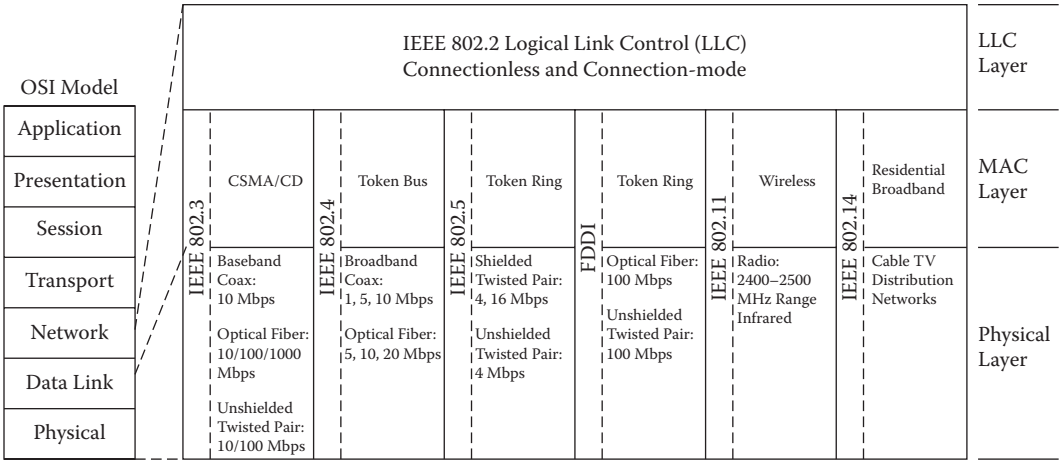
The standards are jointly published by the IEEE, the International Organization for Standardization (ISO), and the International Electrotechnical Commission (IEC). An overview of the standards is published by these bodies [1,2].

1.3.1.2 Reference Model

Figure 1.3.1 relates the specific protocol layers defined by the IEEE 802 Committee, which include Physical, Media Access Control (MAC) and Logical Link Control (LLC) layers, to the layers of the Open Systems Interconnection (OSI) Reference Model [3]. The protocol architecture shown in Figure 1.3.1, including the Physical, MAC, and LLC layers, is generally referred to as the IEEE 802 Reference Model.

Working from the bottom up, the Physical layer of the IEEE 802 Reference Model corresponds to the Physical layer of the OSI Reference Model and includes the following functions.

- Encoding/decoding the signals to be transmitted in a manner appropriate for the particular medium, e.g., the use of Manchester or non-return to zero encoding schemes



FIGu RE 1.3.1 IEEE 802 reference model.

- Achievement of synchronization, e.g., by the addition of a preamble field at the beginning of a data frame
- Bit transmission and reception
- Specification of the physical and electro/optical characteristics of the transmission media (e.g., fiber, twisted pair wire)
- Network topology (e.g., bus, ring)

Above the Physical layer are functions concerned with providing the frame transmission service to LAN users. Such functions include the following.

- Governing access to the LAN transmission medium
- Performing error detection (e.g., via addition of a Frame Check Sequence field)
- Assembling the frame for transmission
- Upon reception, performing address recognition

These functions are collectively associated with a MAC sublayer, shown in Figure 1.3.1. As indicated in the figure, a number of MAC layers are defined within the IEEE 802 Reference Model including access control techniques such as Carrier Sense Multiple Access/Collision Detection (CSMA/CD)—also generally referred to as Ethernet—Token Bus and Token Ring.

Finally, the Logical Link Control (LLC) layer is responsible for providing services to the higher layers regardless of media type or access control method (such as those specified for CSMA/CD, Token Bus, Token Ring, and so on). The LLC layer provides a High-Level Data Link Control (HDLC)–like interface to the higher layers and essentially hides the details of the many MAC schemes shown in Figure 1.3.1 from the higher layers. The LLC layer provides a multiplexing function, supporting multiple connections, each specified by an associated destination service access point (DSAP) and source service access point (SSAP), discussed later. As shown in Figure 1.3.1, the LLC layer provides both connectionless and connection-oriented services, depending on the needs of the higher layers.

1.3.1.3 Overview of the Major MAC Standards

Since its inception at Xerox Corporation in the early 1970s, the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) method, also commonly termed *Ethernet*, has been the dominant LAN access control technique. The CSMA/CD method was the first to be specified by the IEEE, under the IEEE 802.3 Working Group, and was closely modeled after the earlier joint Digital/Intel/Xerox (DIX) Ethernet specification [4]. Ethernet has, by far, the highest number of installed ports and provides

the greatest cost performance relative to other access methods such as Token Ring, Fiber Distributed Data Interface (FDDI) and the newer Asynchronous Transfer Mode (ATM) technology. Recent and in-progress extensions to Ethernet include Fast Ethernet, which, under the auspices of the IEEE 802.3u Working Group, increased Ethernet speed from 10 Mbps to 100 Mbps, thereby providing a simple, cost-effective option for higher speed backbone and server connectivity, and Gigabit Ethernet, which under the auspices of the IEEE 802.3z Working Group, increased the speed to 1000 Mbps.

The IEEE 802.4 Token Bus specifications were developed primarily in response to requirements for the deterministic performance of a token passing scheme, coupled with a bus-oriented topology. The use of a broadband technology option provided the additional benefits of increased bandwidth, geographic coverage, and number of terminations.

IEEE 802.5 Token Ring specification was developed with major support from IBM and reflected IBM's perspective on local area networking. Improvements over the IEEE 802.3 scheme include deterministic performance and the specification of a priority mechanism.

As shown in Figure 1.3.1, work has been completed in several new technology areas including wireless LANs (IEEE 802.11) [5] and Cable Modems (IEEE 802.14) [6].

Due to their wide market acceptance, this section focuses on the details of the IEEE 802.3 (CSMA/CD) and 802.5 (Token Ring) specifications. The section also addresses the Logical Link Control layer and presents an overview of building wiring considerations that would ensure that the building cabling meets the requirements of the various LAN types.

1.3.2 IEEE 802.3 (CSMA/CD) Specifics

1.3.2.1 Frame Structure

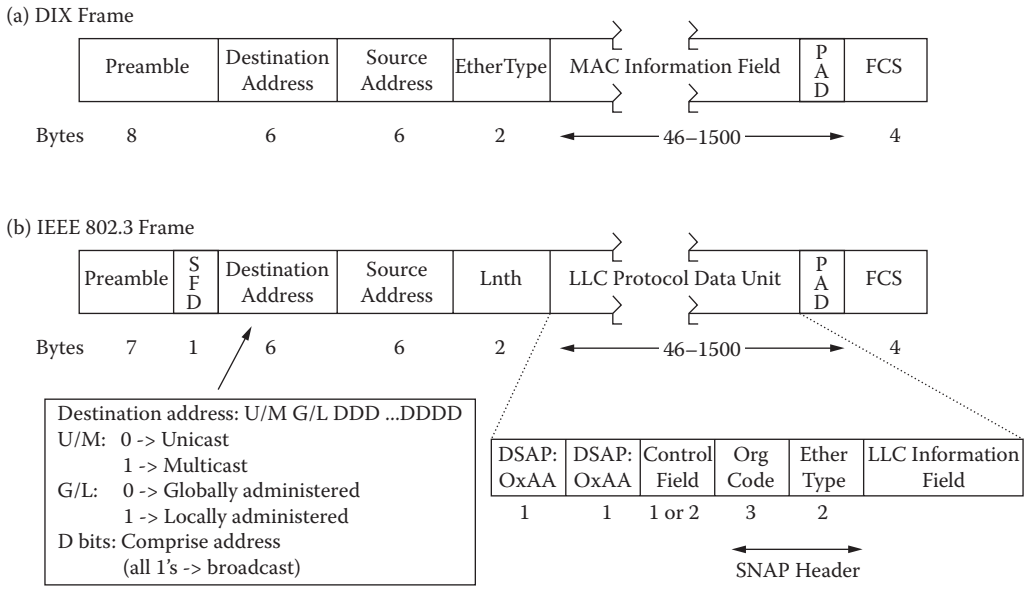
As mentioned, the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) method was the first to be specified by the IEEE and was closely modeled after the Digital/Intel/Xerox (DIX) Ethernet specification. Although there are differences between the Ethernet and the 802.3 specifications, manufacturers now typically produce hardware that can support both, so that effectively the two are compatible. Differences in the packet format are resolved in firmware for a particular implementation. We use the terms Ethernet and IEEE 802.3 CSMA/CD interchangeably.

The frame format in the original DIX specification is shown in Figure 1.3.2(a). Frame fields are as follows.

- Preamble: To allow synchronization by the receiving station and to indicate the start of frame, the frame starts with an eight-byte sequence, the first seven of which have the format (10101010), and the eighth the format (10101011).
- Source and destination addresses are 48 bits each (a little-used option allows for 16 bits) and have the structure shown in Figure 1.3.2(b) except for a minor variation in the second bit of the address.
- EtherType: The EtherType field (16 bits) allows for the multiplexing of data streams from different higher-level protocols and identifies the particular higher-level protocol data steam carried by this frame, e.g., an EtherType of Ox08-00* indicates a frame carrying an IP datagram. Values for the EtherType field can be found in [7].
- Data: The Data field carries the service data unit from the higher-layer protocol entity and ranges in length from 46 (including an added PAD field to meet the minimum field size of 64 bytes if the service data unit is less than 46 bytes) to a maximum of 1500 bytes.
- Frame Check Sequence (FCS): Finally, a four-byte FCS field is added for error detection purposes.

The IEEE 802.3 frame format is shown in Figure 1.3.2(b). The major difference in format arises from the need to accommodate other MAC specifications under the IEEE umbrella, which may have

* This notation indicates a string of bytes (groups of eight bits) with the values of the bytes given in hexadecimal form; thus Ox08-00 represents the two bytes 00001000–00000000.



FIGu RE 1.3.2 DIX and IEEE 802.3 frame formats.

no equivalent of the EtherType field. As a result, this multiplexing capability is included in the next higher layer of the IEEE 802 Reference Model, the LLC layer (see Figure 1.3.1). The method used to provide this additional protocol information is the Subnetwork Access Protocol (SNAP). A SNAP encapsulation is indicated by the LLC layer SSAP and DSAP fields both being set to OxAA. The SNAP header is five bytes long: the first three bytes consist of an organization code, which is assigned by the IEEE; the second two bytes use the EtherType value set from the Ethernet specifications. Using this scheme, the multiplexing service afforded by the EtherType field is available at the LLC layer, independent of the individual MAC layer capabilities. Note that several layers of multiplexing are available at the LLC layer; one provided by the LLC Destination Address/Source Address fields in Figure 1.3.2(b), and the other by the LLC/SNAP fields shown in the figure (which include the EtherType field). Again, when the length of MAC layer data field is less than 46 bytes, a PAD field is added to ensure a minimum data plus PAD field length of 46 bytes. The PAD field consists of an arbitrary array of bits.

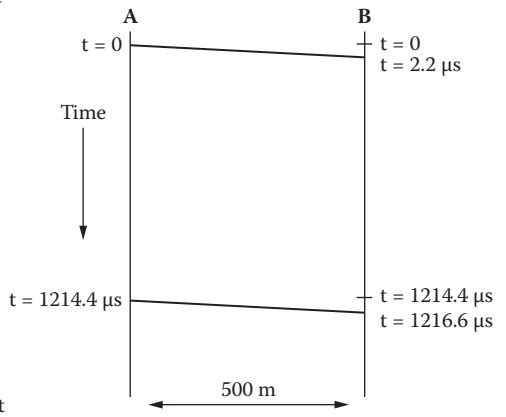
1.3.2.2 Sample Frame Transmission

For a transmission media operating at a data rate of 10 Mbps, typical of many 802.3 specifications, Figure 1.3.3 shows the successful transmission of a frame between two stations at the ends of the cable, from station A (shown on the left) to station B (shown on the right). Cable length is assumed to be 500 meters, the approximate maximum length for a number of IEEE 802.3 configurations (per Section 13 of [8]). A frame size of 1518 bytes is assumed, also the maximum as per the IEEE 802.3 specification. From the figure, station A begins transmitting at time t = 0 and some time later the leading edge of the signal appears at station B. This time is determined by the propagation speed of the signal on the particular media, with the speeds for a number of media shown in Table 1.3.1. Assuming a propagation speed of .77c, where c is the speed of light (3×10^8 m/s), yields a propagation delay of about 2.2 μs for the example in Figure 1.3.3.

The total signal transmission time, neglecting a short initial synchronization period when the preamble and start of frame delimiter are transmitted is

$$\left(1518 \text{ bytes}\right) \times \left(8 \text{ bits/bytes}\right) / 10 \text{ Mbps} = 1214.4 \text{ μs}$$

- Consider two stations (A and B) at the ends of an Ethernet network.
- Assume the maximum allowed frame size of 1518 bytes (12144 bits).
- At 10 Mbps, the resulting frame transmission time is 1214.4 us.
- Assume a 500 m cable; propagation time is thus about 2.2 μs (using propagation speed of .77c, where c is the speed of light).
- This figure shows successful transmission of frame from A to B. Station A starts to send at t = 0 and completes transmission at t = 1214.4 μs; station B starts to receive at t = 2.2 μs and has received the entire frame at t = 1216.6 μs.



FIGu RE 1.3.3 Example of successful frame transmission.

TABl E 1.3.1 Minimum Propagation Speeds for Sample Media

Media Type	Minimum Propagation Speed
Coax (10BASE5)	0.77 c
Coax (10BASE2)	0.65 c
Twisted-Pair (10BASE-T)	0.585 c

Thus station A completes transmitting the signal at t = 1214.4 μs and station B begins receiving the signal at t = 2.2 μs and receives the entire signal at time t = 1216.6 μs.

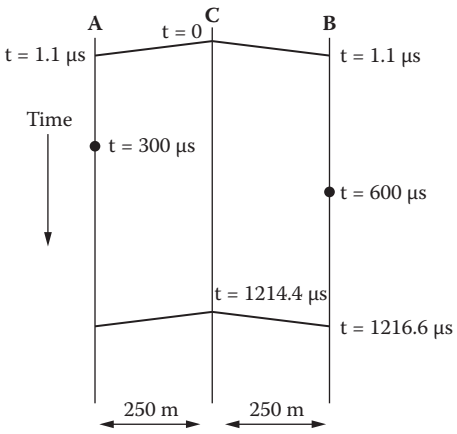
After a brief delay period to allow recovery time for other stations and the physical medium, termed the *interframe gap*, another frame can be transmitted if available. An interframe gap of 9.6 μs or 96 bit times for a 10-Mbps implementation is specified by the standard. This value is chosen to account for variability in the gap as frames travel over the media and connecting repeaters (discussed below). This variability occurs because two successive frames may experience different bit loss in their preambles. If the first packet experiences greater bit loss than the second, the gap will shrink as the repeater reconstructs the preamble and therefore introduces delay. If the second frame experiences greater bit loss, the gap will expand.

1.3.2.3 Carrier Sense Multiple Access

A simple addition to the above scheme is to require each station to "listen before talking," that is, require a station to sense the medium to determine if another station's signal is present and defer transmission if this is the case. This situation is shown in Figure 1.3.4 where a third station at the middle of the cable begins sending at time t = 0. Due to signal propagation delays, signal reception begins at both A and B at time t = 1.1 μs. In the figure, while sensing the presence of the carrier from C, A and B both receive frames from higher layers to transmit, but adhering to the CSMA scheme, defer transmitting until some time after station C's transmission is completed.

For typical CSMA schemes, a number of strategies can be employed to determine when to begin transmitting after deferring to a signal already on the medium. These strategies typically involve invoking one of the persistency schemes shown in Table 1.3.2. The persistency parameter *p* relates to the probability that a station sends its frame immediately after the medium is sensed idle. To obtain maximum channel utilization, the choice of the persistency value, 0.1, 0.2, 0.3, … , etc., is dependent on the traffic offered by the stations. A low level of traffic would operate best with a persistency value, *p*, near 1.0 (here,

- With Carrier Sense Multiple Access (CSMA), Station A would check that the media is idle before sending. If idle, it would generally send (as in last example) and if busy, it would perform a backoff algorithm (persistent or non-persistent).
- For example, suppose station C (in middle of network) began transmitting a 1518 byte packet at t = 0. If Station A received a frame to send at t = 300 μs and B at t = 600 μs, both would sense the media busy and perform a backoff algorithm.



FIGu RE 1.3.4 Use of carrier sense multiple access (CSMA).

TABl E 1.3.2 Typical Persistency Algorithms

Persistency Scheme	Description
Non-persistent	• idle—transmit • busy—wait random time and repeat
1-persistent	• idle—transmit • busy—wait until idle then transmit immediately (Note that if 2 or more stations are waiting to transmit, a collision is guaranteed)
p-persistent*	• idle—transmit with probability p and delay one time unit with probability $1-p$; time unit is typically the maximum propagation delay • busy—continue to listen until channel is idle and repeat above for idle • delayed one time unit—repeat above for idle

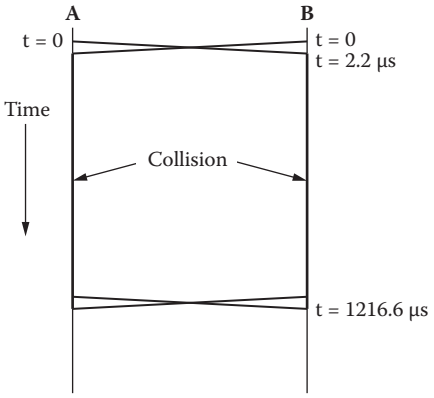
* Issue is choice of p
- Need to avoid instability under heavy load.
- If n stations are waiting to send, the expected number transmitting is np. $np > 1$ fi collision is likely.
- New transmissions will also begin to compete with retries and network will collapse: all stations waiting to transmit, constant collisions, no throughput.
- Thus np must be <1; but heavy load means p must be small and time will be wasted even on a lightly loaded line, e.g., $p = 0.1$ fi on average, will transmit in tenth interval on an idle line.

typically only a single station will be ready to send and thus should send immediately with high probability), and a high level of traffic would operate best with a lower value of p (here multiple stations will likely be ready to send and the lower value of p will make it more likely that only one station attempts to transmit). It should be noted that the above retransmission algorithm is not related to the binary exponential backoff algorithm discussed below, associated with resolving collisions. Also of note is that the IEEE 802.3 standard specifies the 1-persistent scheme, ensuring a collision if two or more stations are deferring to an ongoing transmission.

1.3.2.4 Adding Collision Detection

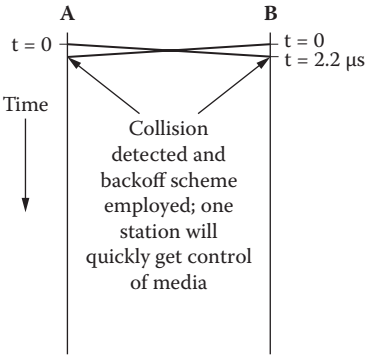
A problem with the CSMA scheme is depicted in Figure 1.3.5, where stations A and B both have something to send at t = 0 and, sensing the medium idle (no carrier), both begin transmission. For example, this case would occur if both have been deferring to another station transmitting on the medium and used the 1-persistent backoff scheme. At some short time later, the signals will collide at stations A and B (and at all other stations on the medium). In this case, no useful information is transferred for the entire transmission time of the frame, approximately 1200 μs for a frame of maximum length.

- Problem: Suppose both stations decide to transmit in the time frame t = 0 to t = 2.2 μs. With no collision detection, both will continue to transmit for the entire frame time.
- Considerable time will be wasted (over 1200 μs in this example).
- Also, a scheme is needed to determine who transmits when the media becomes idle or else, with probability 1, they will collide again, e.g., an n-persistent scheme (e.g., .1-persistent) could be used, i.e., a particular station transmits with probability n (e.g., probability of .1).



FIGu RE 1.3.5 Wasted resources with CSMA.

- Collision Detection (CD) solves this problem; the stations stop transmitting when they sense a collision. After they stop transmitting, they implement a *binary exponential backoff* scheme (random retransmission interval doubled each time a collision is detected).
- CSMA/CD also employs the 1-persistent scheme for stations accessing a media that just became idle. This can raise the chance of collision since multiple stations waiting for the media to become idle will collide but the collision detection scheme will minimize wasted resources.

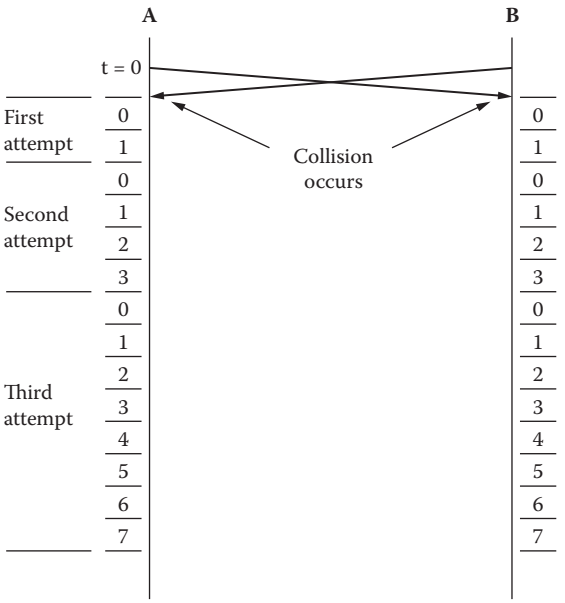


FIGu RE 1.3.6 CSMA/CD.

A solution to this problem is the addition of the collision detection mechanism depicted in Figure 1.3.6. The addition of such a mechanism reduces the wasted transmission time as both stations will stop transmitting upon detection of the collision. Here the stations have the added capability of detecting the occurrence of a collision of the two signals on the medium. With this added functionality, the stations can stop transmitting upon detecting collisions and immediately undertake a backoff scheme to allow one station to capture the medium.

1.3.2.4.1 Collision Backoff Scheme

Two stations, A and B, implementing the CSMA/CD media access control technique are shown in Figure 1.3.7. If, as shown in the figure, they both sense the media idle and begin to transmit at t = 0, a collision will occur and IEEE 802.3 specifies a *truncated binary exponential backoff* randomization scheme so that one of the stations can obtain control of the media. As shown in the figure, with this scheme, first one of two "slots" is chosen randomly by each station to attempt to capture the medium. The slot time is chosen based on factors that include the round-trip transmission time between two stations at the ends of the medium, and the time required to detect a collision. It is specified in bit times;



FIGu RE 1.3.7 Retransmission attempts with the binary exponential backoff scheme.

the IEEE 802.3 standard specifies a slot time of 512 bit times (51.2 μs for a 10 Mbps system). If a collision occurs again (i.e., they both choose the same slot), one of four slots is chosen in the next attempt; then eight; then sixteen; etc. The number of slots grows in this manner to 2^{10} and truncates at this value. After a total of 16 retransmission attempts fail, this event is reported as an error.

With the CSMA/CD scheme, for reasonable traffic levels, a station should capture the medium in a rather short time, especially when compared to the CSMA scheme. For instance, consider the two stations in Figure 1.3.6 implementing a 1-persistent scheme (the recommended IEEE 802.3 scheme used after deferring to a transmitting station). Each will begin to transmit when station C stops transmitting and thus will suffer a collision on this first transmission. The binary exponential backoff scheme of Figure 1.3.7 will yield the following results.

For the two stations competing for the medium, the following outcomes are equally likely during the first retransmission.

(0,0)—station A picks slot 0 and station B also picks slot 0,
(0,1)—station A picks slot 0 and station B picks slot 1,
(1,0)—station A picks slot 1 and station B picks slot 0,
(1,1)—station A picks slot 1 and station B also picks slot 1.

Two of these outcomes, (0,1) and (1,0), will result in a station capturing the medium, station A in the first of these and station B in the latter. Two of the outcomes, (0,0) and (1,1), result in collisions in slots 0 and 1, respectively. Thus with a probability of $p = 0.5$, one of the stations will capture the medium in this first retransmission period.

If there is a collision for the first retransmission, the second retransmission uses 4 slots chosen at random by the stations (numbered 0, 1, 2, and 3 in Figure 1.3.7), resulting in 16 possible outcomes. Using similar notation as above, these outcomes are:

(0,0), (0,1), (0,2), (0,3)
(1,0), (1,1), (1,2), (1,3)
(2,0), (2,1), (2,2), (2,3)
(3,0), (3,1), (3,2), (3,3)

Only four of these, (0,0), (1,1), (2,2), and (3,3), result in collisions yielding a probability of 12/16 or p for a successful outcome. Thus the probability of exactly two retransmissions is (probability of collision on first retransmission) × (probability of success on second) = p = 3/8 = .375.

Similarly, the probability of exactly three retransmissions is

$$1/2 \times 1/4 \times 7/8 = 0.109$$

and the probability for four is

$$1/2 \times 1/4 \times 1/8 \times 240/256 = 0.0146$$

The likelihood of more than four transmissions is rather small.

Finally, the average number of retransmissions is

$$\sum_{i=1}^{\infty} i \times (\text{prob of i retransmission}) = (1 \times 0.5) + (2 \times 0.375) + (3 \times 0.109) + (4 \times 0.0146) \ldots$$

$$= 0.5 + 0.75 + 0.327 + 0.058 \ldots = 1.635$$

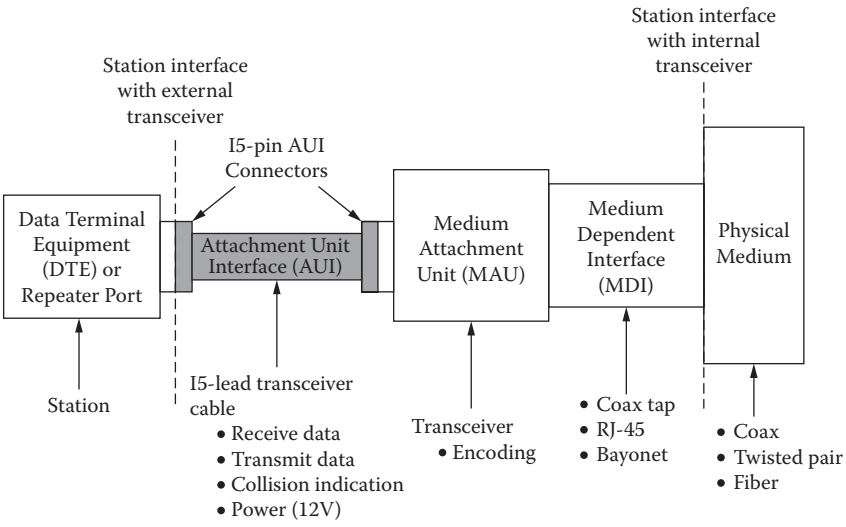
On average then, with two stations competing for the medium, one will capture the medium during the second retransmission attempt. Note that this saves medium resources when compared to Figure 1.3.5, where over 1200 μs are wasted due to the collision and additional time will be spent in some sort of backoff scheme.

It is interesting to note that for three stations competing, media capture will occur more quickly. Here, the three stations will collide on the first transmission attempt and the eight possible outcomes for the first retransmission are (0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0) and (1,1,1), with only (0,0,0) and (1,1,1) resulting in unsuccessful outcomes. For example, the (0,0,1) outcome will result in media capture: stations A and B will collide in slot 0 and station C will capture the medium in slot 1. Thus the probability of one station capturing the medium in the first retransmission is p = .75, greater than the case with two stations competing. The average number of retransmissions for this case can be shown in the above manner to be on the order of 1.27. The reduced average number of retransmissions in the case of three stations competing is somewhat of an anomaly; for more stations competing, the average number of retransmissions steadily increases. Of course, the average waiting time for a particular station to capture the medium will increase with the number of stations competing for the medium.

1.3.2.5 CSMA/CD System Components

As mentioned, the IEEE 802.3 specifications support multiple media types, including coaxial cable, twisted pair, and fiber. Thus one of the component interfaces will of necessity vary with the media type. For example, the physical media dependent interface for twisted pair differs in a number of respects from that for fiber, including the physical connector (or plug), the electrical vs. optical nature of the interface and the encoding scheme (translating a logical sequence of bits to the electrical/optical signal on the medium), e.g., Manchester or non-return to zero (NRZ) schemes. In a number of IEEE implementations, this interface is remote from the station itself and the IEEE specification provides a media-independent manner of extending the station interface to the media-dependent interface.

This general situation is shown in Figure 1.3.8, where the various functional components that make up the CSMA/CD system are also shown. The medium dependent interface (MDI) is shown on the right of the figure and provides the direct electrical/optical connection. Examples include an eight-



FIGu RE 1.3.8 CSMA/CD system components.

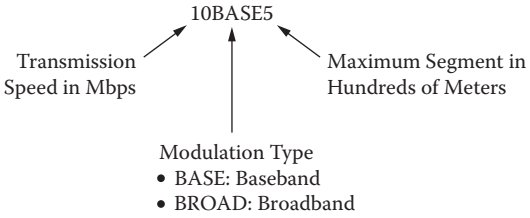
pin connector (RJ-45 telephone style jack) for twisted pair, a coaxial cable clamp for coax medium, and a spring-loaded bayonet connector (termed an *ST connector*) for fiber. The Medium Attachment Unit (MAU) shown in the figure, also commonly known as a *transceiver*, is also specific to the type of medium and performs signal encoding (e.g., Manchester or NRZ). Transceivers also contain a jabber protection circuit that protects the network from a station or transceiver that is transmitting frames whose length exceeds the maximum allowed.

The interconnection of the station to the remote transceiver is accomplished by the attachment unit interface (AUI), also termed the transceiver cable. This 15-lead cable is media independent and carries data receive, data transmit, and collision detection signals along with power from the station to the transceiver. The maximum length of the transceiver cable is 50 m. The associated 15-pin AUI connector is commonly found on Ethernet station cards.

It is important to note that many implementations include the transceiver as part of the Ethernet station card. For example, twisted pair implementations use this scheme. In this case, only the media-dependent interface is visible on the station Ethernet card; for example, the RJ-45 type interface is commonly found on interface cards connecting to twisted pair medium.

1.3.2.6 Example Implementations

IEEE 802.3 standards are characterized by a shorthand notation to facilitate their description. The notation (e.g., 10BASE5) is composed of three elements as shown in Figure 1.3.9, indicating the transmission rate of the system in Mbps, the modulation scheme, baseband or broadband, and the maximum length of the segment in hundreds of meters. With standards adopted more recently, such as 10BASE-T, the



FIGu RE 1.3.9 IEEE 802.3 nomenclature.

IEEE has been more descriptive with its notation. For example, the "T" in the 10BASE-T notation is short for "twisted-pair wiring."

This section describes the most commonly implemented versions of the IEEE 802.3 specification.

1.3.2.6.1 10BASE5

10BASE5 was the first version of the IEEE specification to be developed and it most closely resembles the earlier DIX versions 1 and 2 [4]. The 10BASE5 specification employs a "thick Ethernet" 50-ohm coaxial cable. While this cable is difficult and relatively expensive to install, it provides advantages over other implementations in terms of distance and the number of terminations permitted for each segment.

This specification uses the standard outboard transceiver option discussed above: the station adapter board connects to the standard AUI cable; this in turn is connected to the transceiver, which is connected to the Ethernet trunk cable via a "vampire" tap. Up to 100 devices can be placed on a 500-meter segment, with a maximum of 1024 devices on a multisegment network, discussed below.

1.3.2.6.2 10BASE2

10BASE2 (also known as *thin Ethernet* or *Cheapernet*) employs a thin flexible coaxial cable (RG-58). In earlier implementations, the transceiver functions were onboard the station and the connection to the station was by means of a media-dependent BNC "T" connector. To provide the flexibility to use the station board for either the 10BASE5 or 10BASE2 systems, station boards have been developed that provide options for both external 10BASE5 vampire taps and 10BASE2 BNC connectors. Board manufacturers commonly provide boards with built-in transceivers that can be switched on or off for a particular application.

The standard 10BASE2 LAN can support only 30 terminations on each coaxial cable segment of 185 meters. While this may seem like a constraint, it is often adequate for most work area environments. Where a requirement exists for interconnecting multiple work areas, or work areas with multiple 10BASE2 segments, a backbone 10BASE5 segment can be employed to provide intersegment connectivity. Table 1.3.3 highlights the differences between the 10BASE5 and 10BASE2 systems.

1.3.2.6.3 1BASE5

This standard approach was contributed by AT&T to accommodate its earlier Starlan products. It operates at 1 Mbps, and as such is often most useful for small work areas or low-traffic environments. 1BASE5 also employs inexpensive twisted-pair wire interconnected through a hierarchical system of concentrator hubs. The hubs emulate a bus configuration by broadcasting data and collision information on all ports.

1.3.2.6.4 10BASE-T

One of the most important developments in the IEEE 802.3 area was the specification of the 10-Mbps unshielded twisted-pair (UTP) Ethernet system, 10BASE-T. Virtually every vendor active in the Ethernet market now offers 10BASE-T products.

TABl E 1.3.3 802.3 10BASE5/10BASE2 Comparison

	10BASE5	10BASE2
Common name	802.3 "Ethernet"	Cheapernet, THIN Ethernet, THINWIRE Ethernet, etc.
Type of cable	50 Ω Thick dual shield	50 Ω RG-58
Maximum segment length	500 m	185 m
Spacing of devices on cable	2.5 m minimum	0.5 m. minimum
Maximum number of taps for a segment	100	30
Maximum number of full repeaters in a path between two stations	2	2
Type of taps	Vampire or coax	BNC "T" connector for "daisy chaining"

Like the 1BASE5 specification, this system uses a hub concentrator to interconnect multiple stations and emulate bus operation. These implementations are limited to 100-meter segments due to the greater attenuation and signaling difficulties of twisted pair. This does not present any unusual problems since these connections only reach to the communications closet. From there, fiber and coax segments can be used to concatenate and extend the LAN system. 10BASE-T systems use one twisted pair for transmitting data and a separate pair for receiving. Collisions are detected by sensing the simultaneous occurrence of a signal on both the transmit and receive pairs.

It is imperative, however, that organizations planning these networks have their existing twisted-pair wire certified for both attenuation and capacitance before making any assumptions on its reuseability.

1.3.2.6.5 10BASE-FL

The 10BASE-FL specification allows for the use of two fiber-optic cables as the medium, one for signal transmission and one for reception. Such a medium allows advantages of greater distances, for example, the standard allows segment lengths of up to 2000 meters, evolution to higher transmission speed, and isolation from electromagnetic radiation. The system components are identical to those shown in Figure 1.3.8 with the use of a fiber-optic MAU.

1.3.2.6.6 10BROAD36

The 10BROAD36 implementation uses much of the same hardware as the baseband implementations. The specification enables an organization to use its existing workstation boards for connection to either a baseband or broadband system. The essential difference is the substitution of a broadband electronics unit and a passive broadband tap for the baseband MAU. The primary function of the broadband electronics unit is to create the frequency-derived data channels and to monitor for collisions. It also converts the signals from the baseband-coded signal on the AUI to the analog signal necessary on the broadband channel. Workstations can be placed up to 1800 meters from the "head-end" of the broadband cable plant. By placing the head-end in the center of the configuration, workstations can be installed up to 3600 meters from each other. In recent years, this standard has been less frequently used.

1.3.2.7 Topology Extensions

1.3.2.7.1 Repeaters

Repeaters regenerate the signals from one LAN segment for retransmission to all the others. The earliest repeaters were simple two-port devices that linked a couple of coaxial cable segments. Later, repeaters evolved to multiport devices deployed as the hub of a star topology. Since with repeaters, all segments are part of a unified LAN, the nature of the shared channel must be preserved by broadcasting all information to all attached devices. An aspect of these repeaters is that they must be capable of retransmitting collisions as well as data frames. In the case of IEEE 802.3 CSMA/CD LANs, physical LAN segment connection standards for repeaters are well developed and mature. The latest specifications for implementation of 10BASE-T repeaters are contained in the IEEE 802.3 specifications (see Section 9 of [8]).

In addition to the functions described above, repeaters can provide an optional "partitioning" feature between segments. This function is designed to address an abnormal situation such as a cable break or network card failure. Thus, if conditions on a given segment are causing an extensive proliferation of collisions, the rest of the LAN can be protected from this anomaly. The repeater will count the number of collisions from the source segment and when excessive, isolate these from transmission to the next segment.

1.3.2.7.2 Switched Hubs (Ethernet Switches)

In Ethernet switching, the interconnecting device (termed a *switching hub*) has intelligence to use the MAC-layer address of a received frame to determine the specific port on which the destination station is attached and transmit the frame on only that port. No other stations are aware of the frame. If frames arrive destined for a busy port, that port can momentarily hold them in its buffer; the size of port buffers

differs by vendor from a few hundred to more than a thousand packets. When the busy port becomes free, frames are released from the buffer and sent to the port. This mechanism works well unless the buffer overflows, in which case packets are lost. To avoid this, some vendors offer a throttling capability; when a port's buffer begins to fill up, the hub begins to transmit packets back to the workstations. This effectively stops the stations from transmitting and relieves the congested state.

Some LAN switching products offer a choice of packet-switching modes, for example, fast forward and "store and forward." These modes affect the amount of packet latency, the time from which the first byte of a packet is received until that byte is forwarded. Each mode reads a certain number of bytes of a packet before forwarding. This creates a trade-off among latency, collision, and error detection. The greater number of bytes read, the greater the latency, but the fewer errored or collision-terminated frames propagated through the network. The fast-forward mode passes packets shortly after receiving the destination address portion of the Ethernet frame (see Figure 1.3.2) and, based on this field, determines the appropriate destination port. In this mode, typical latencies are on the order of tens of μs for a 10-Mbps Ethernet system. Store-and-forward mode receives entire frames and performs error detection via the FCS field (see Figure 1.3.2) before forwarding, resulting in increased latency but maximizing frame error detection. Here, for a maximum-length frame, latency will be greater than 1200 μs. Some LAN switching devices support only one address per port, while others support 1500 or more. Some devices are capable of dynamically learning port addresses and allowing or disallowing new port addresses. Disallowing new port addresses enhances hub security; in ignoring new port addresses, the corresponding port is disabled, preventing unauthorized access. These and other techniques used in conjunction with port switching enhance overall network performance by eliminating the contention problem that occurs in shared Ethernet networks.

1.3.2.7.3 Multisegment Guidelines

The IEEE 802.3 specifications provide guidelines for the number and types of segments that can be interconnected via repeaters and switch hubs. A number of example configurations are presented that would be typical of many implementations. For example, one possible configuration that taxes the CSMA/CD configuration guidelines has a distance of about 1800 meters between stations, with three 500-meter segments and two shorter segments interconnected by four repeaters. For more complex cases, the specifications provide guidelines for calculating system parameters such as the round-trip delay time and interframe gap shrinkage to ensure that these parameters are within allowed limits.

1.3.2.8 Higher Speed Extensions (100 Mbps and 1000 Mbps)

1.3.2.8.1 Fast Ethernet (100 Mbps) Overview

The 100-Mbps CSMA/CD operation, also termed Fast Ethernet, is specified in the IEEE 802.3u supplement to the standard. While attaining a tenfold increase in transmission speed, other important aspects, including the frame format and the CSMA/CD access control scheme, remain unchanged from a 10-Mbps system, making the transition to higher speeds straightforward for network managers. The 100-Mbps specification uses Ethernet's traditional CSMA/CD protocol and is designed to work with existing medium types, including Category 3 and Category 5 twisted-pair and fiber media. In addition, Fast Ethernet will look identical to lower-speed Ethernet from the LLC layer upward. Since Fast Ethernet significantly leverages existing Ethernet technology, network managers will be able to use their existing knowledge base to manage and maintain Fast Ethernet networks.

The Fast Ethernet specifications include mechanisms for autonegotiation of the media speed. This makes it possible to provide dual-speed Ethernet interfaces that can be installed and run automatically at either 10 Mbps or 100 Mbps.

There are three media varieties that have been specified for transmitting 100-Mbps Ethernet signals: 100BASE-T4, 100BASE-TX, and 100BASE-FX. The third part of the identifier provides an indication of the segment type.

T4 segment type is a twisted-pair segment that uses four pairs of telephone-grade twisted-pair wire.
TX segment type is a twisted-pair segment that uses two pairs of wires and is based on data-grade twisted-pair physical medium standards, which are covered later in this section.
FX segment type is a fiber-optic link segment that uses two strands of fiber cable and is based on the fiber-optic physical medium standard developed by ANSI.

The TX and FX medium standards are collectively known as 100BASE-X. The 100BASE-TX and 100BASE-FX media standards used in Fast Ethernet are both adopted from physical media standards first developed by ANSI for the Fiber Distributed Data Interface (FDDI) LAN standard (ANSI standard X3T9.5), and are widely used in FDDI LANs. Rather than reinventing the wheel when it came to signaling at 100 Mbps, the Fast Ethernet standard adapted these two ANSI media standards for use in the new Fast Ethernet medium specifications. The T4 standard was also provided to make it possible to use lower-quality twisted-pair wire for 100-Mbps Ethernet signals.

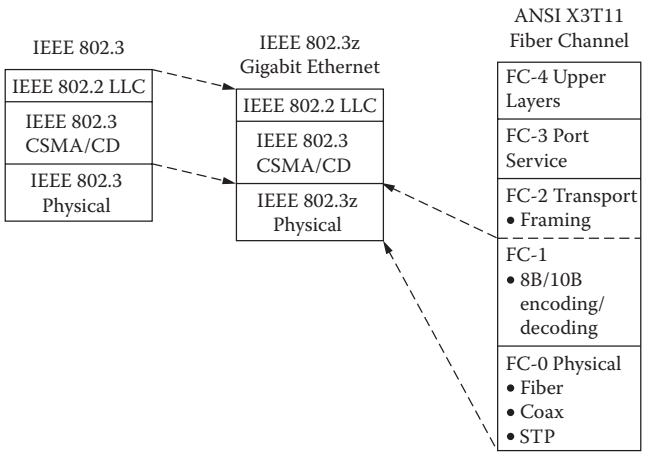
1.3.2.8.2 Gigabit Ethernet (1000 Mbps) Overview

Gigabit Ethernet, under the auspices of the IEEE 802.z Working Group, builds on the CSMA/CD MAC scheme and increases the transmission speed to 1000 Mbps. A key feature of Fast Ethernet implementations is the autoconfiguration capability, and Gigabit Ethernet solutions providing 10-/100-/1000-Mbps operation allow comparable features.

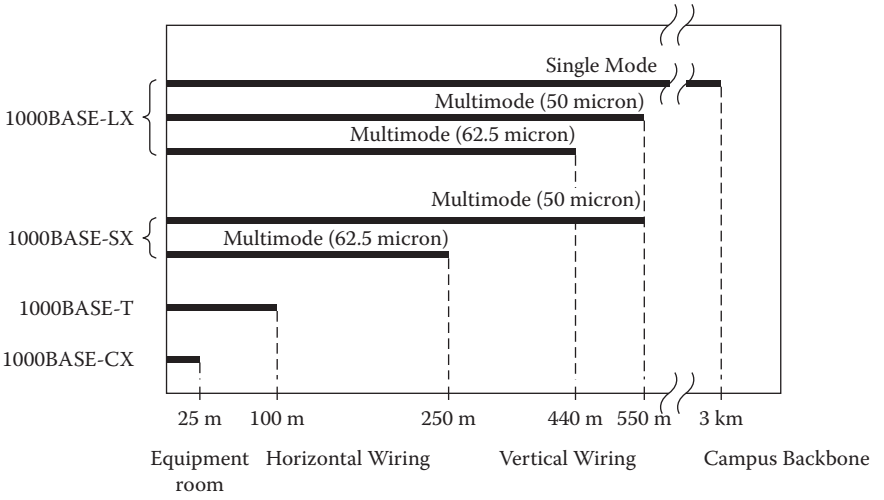
It should be noted that, as in the case for Fast Ethernet, a number of challenges involved in achieving rapid time to market for Gigabit Ethernet were resolved by merging existing technologies:

1. IEEE 802.3 CSMA/CD and
2. ANSI X3T11 Fiber Channel; Fiber Channel encoding/decoding integrated circuits (ICs) and optical components were readily available and optimized for high performance at relatively low cost.

Leveraging these two technologies meant that the Gigabit Ethernet standard could take advantage of the existing, proven high-speed physical interface technology of Fiber Channel while maintaining the IEEE 802.3 Ethernet frame format, backward compatibility for installed media, and use of CSMA/CD. This strategy helped minimize complexity and resulted in a technology that could be quickly standardized. Figure 1.3.10 shows how key components from each technology have been leveraged to form Gigabit Ethernet. As a result, the Gigabit Ethernet standard based on fiber optics for the MAC and Physical layers has progressed rapidly. Unshielded twisted-pair media did not have the advantage of a proven, existing technology base and the standards remained in further development; this work is under the auspices of the IEEE 802.3 Working Group and referred to as 1000BASE-T.



FIGu RE 1.3.10 Gigabit Ethernet and the ANSI Fiber Channel Standard.



FIGu RE 1.3.11 Distance specifications for gigabit ethernet media.

1.3.2.8.2.1 Physical Layer Characteristics of Gigabit Ethernet The initial Gigabit Ethernet specification from the IEEE 802.3z Working Group calls for three transmission media: single-mode and multimode fiber and balanced shielded 150-ohm copper cable. There are two supported types of multimode fiber: 62.5-micron and 50-micron diameter fibers. The IEEE 802.3ab committee is examining the use of unshielded twisted-pair (UTP) cable for Gigabit Ethernet transmission (1000BASE-T). The distances for the media supported under the IEEE 802.3z standard and those projected for the IEEE 802.3ab are summarized in Figure 1.3.11.

1.3.2.8.2.2 Fiber Optic Media (1000Base-SX and 1000Base-LX) As mentioned, the Fiber Channel physical medium dependent specification was employed for Gigabit Ethernet to speed standardization. This standard provides 1.062 gigabaud in full-duplex mode and Gigabit Ethernet will increase this rate to 1.25 gigabaud with an 8B/10B encoding scheme allowing a data transmission rate of 1000 Mbps. In addition, the connector type for Fiber Channel was also specified for both single-mode and multimode fiber.

The standard supports two laser types, a short-wave laser type (termed 1000Base-SX) and a long-wave laser type (termed 1000Base-LX). Both short-wave and long-wave lasers are supported over multimode fiber. There is no support for short-wave laser over single-mode fiber. The key issues between the use of long-wave and short-wave laser technologies are cost and distance. Short-wave lasers are readily available since variations of these lasers are used in compact-disc technology. Long-wave lasers take advantage of attenuation dips at longer wavelengths in the cable and suffer lower attenuation. The net result is that short-wave lasers will cost less, but traverse a shorter distance. In contrast, long-wave lasers will be more expensive but will traverse longer distances.

The 62.5-micron fiber is typically seen in vertical campus and building cable plants and has been used for Ethernet, Fast Ethernet, and FDDI backbone traffic. However, this type of fiber has a lower modal bandwidth (the ability of the cable to transmit light), especially with short-wave lasers. This means that short-wave lasers over 62.5 micron will generally traverse shorter distances. The 50-micron fiber has significantly better modal bandwidth characteristics and will be able to traverse longer distances with short-wave lasers relative to 62.5-micron fiber.

1.3.2.8.2.3 150-Ohm Balanced Shielded Copper Cable (1000Base-CX) For shorter cable runs (of 25 meters or less), Gigabit Ethernet will allow transmission over a new type of special balanced shielded 150-ohm cable (termed 1000Base-CX). Because of a distance limitation of 25 meters, this cable will likely have limited use.

1.3.2.9 Full-Duplex Operation

The previous discussion focused on half-duplex operation, where only a single communications channel was available and thus data could be transmitted in only one direction at a time. The CSMA/CD Media Access Control mechanism was required to determine which station could use the single channel. Full-duplex is an optional point-to-point mode of operation between a pair of devices allowing simultaneous communication between the devices and thus a doubling of aggregate capacity. Two separate communications channels are needed in this case to allow both stations to simultaneously transmit and receive. Thus the allowed physical media for this operation are only those with the capability of supporting two simultaneous channels; for example, 10BASE-T provides independent transmit and receive data paths that can be simultaneously active. Full-duplex operation cannot be supported on coaxial cable systems since they do not provide independent transmit and receive data paths. The optional full-duplex mode of operation is specified by the 802.3x supplement to the standards.

1.3.3 IEEE 802.2 Logical Link Control Layer

The IEEE 802.2 Logical Link Control (LLC) layer specifications [9] include those Data Link Layer functions that are common to all 802 LAN MAC sublayer alternatives. The LLC frame format is shown in Figure 1.3.2.

Three basic types of service are defined in the standard.

1.3.3.1 Type 1 (Connectionless)

This service provides a best-effort delivery mechanism between origin and destination nodes. No call or logical circuit establishment procedures are invoked. Each frame is treated as an independent entity by the network. The type of frame used to provide this service is the unnumbered type; no flow control or acknowledgments are provided with this service. If the packet does not arrive at the destination, it is the responsibility of higher layers to resolve the problem through time-outs and retransmission. This type of service would be provided to the IP network layer protocol.

1.3.3.2 Type 2 (Connection Oriented)

Many wide area network protocols require that a logical circuit or call be established for the duration of the exchange between the origin and destination nodes. Packets usually travel in sequence over this logical circuit and are not routed as independent entities. LLC Type 2 provides this type of service. The service involves a number of control frames to manage the logical circuit (establishment, disconnection) and numbered frames for information transfer. Positive acknowledgments and flow control mechanisms based on this frame numbering are an integral part of this service. LLC Type 2 is commonly found in implementations of IBM's Systems Network Architecture (SNA).

1.3.3.3 Type 3 (Acknowledged Connectionless)

No circuit is established in this service variation, but acknowledgments are required from the destination node. This type of service adds additional reliability to Type 1, but without the overhead of Type 2. These LLC alternatives are summarized in Table 1.3.4.

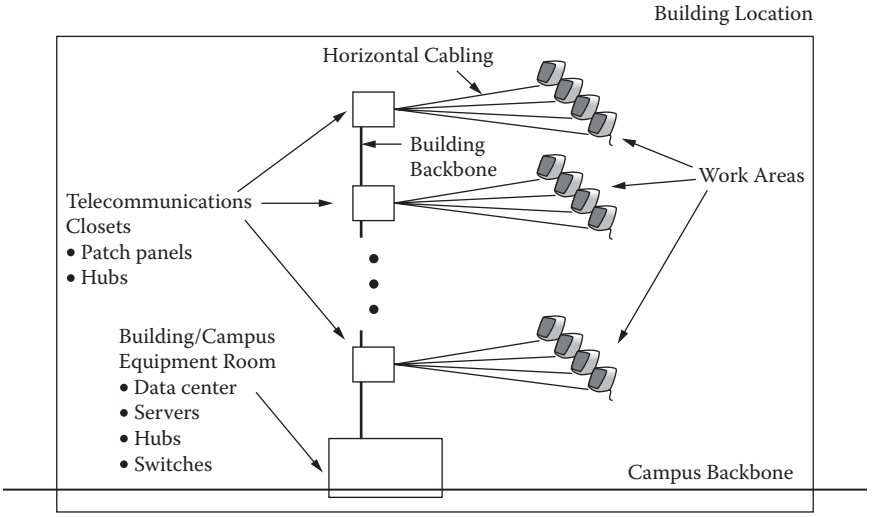
1.3.4 Building Cabling Specifications

The major components of a building cabling architecture that apply to the implementation of LANs are shown in Figure 1.3.12 and include the following:

- Equipment room: This location houses major data center processing and communications equipment for the building including servers, routers, and LAN switches. For a campus environment involving a number of buildings, one such location would serve as a data processing and

TABl E 1.3.4 Summary of Logical Link Control Alternatives

Service Type	Type 1	Type 2	Type 3
Description	Connectionless	Connection	Acknowledged connectionless
Acknowledgments	No	Yes	Yes
Error recovery	No	Yes	Yes
Flow control	No	Yes	No



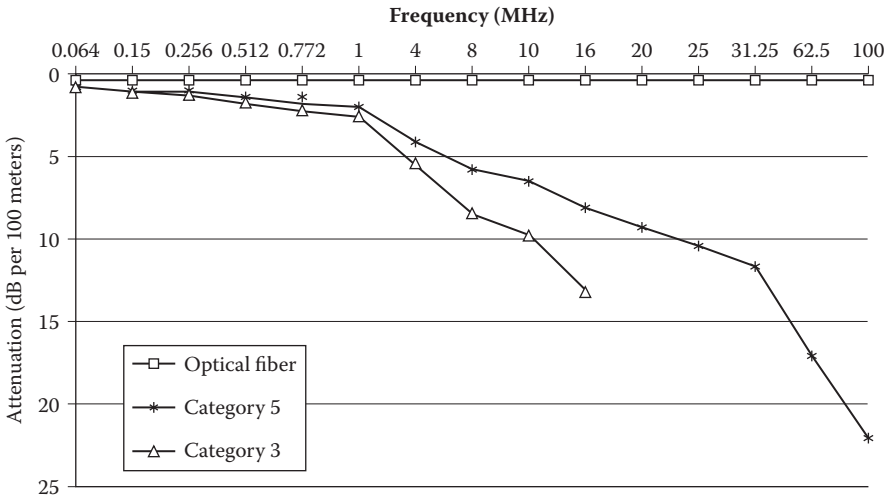
FIGu RE 1.3.12 Architecture for building/campus telecommunications cabling.

communications center for the campus; other equipment rooms on the campus would serve specific buildings.

- Telecommunications closet: This is an area, typically located on each floor of a building, that houses data and telecommunications equipment providing wiring concentration, cross-connect and hubbing functions.
- Backbone cabling: This cabling provides connectivity between equipment in the equipment room and the telecommunications closets. It includes vertical connections between the floors and connections between buildings.
- Horizontal cabling: This cabling extends from the telecommunications closet to the individual work areas on the building floors.

The American National Standards Institute (ANSI), the Electronics Industry Association (EIA), and the Telecommunications Industry Association (TIA) develop specifications for commercial building cabling standards. This set of standards, referred to as ANSI/EIA/TIA 568A, defines the installation practices, certification, and physical, electrical, and optical characteristics for various physical media such as unshielded twisted-pair and fiber-optic cable [10]. The intent of the standard is to provide a guideline by which a cabling system can be designed and implemented as part of the overall design of a new building, even if the systems that the cabling must support are not yet defined. It guides the user toward the selection of cabling that will support current and future communications needs. The 568A standard and other EIA standards are recognized by architectural and engineering firms as definitive guidelines to use during the design phase of a building.

The 568A standard defines the requirements of a cabling system on a generic level that is appropriate to a commercial environment. The standard allows certain options, such as the use of various cabling media, including the following:



FIGu RE 1.3.13 Attenuation of various media—EIA 568A.

- 100-ohm unshielded twisted-pair cable in a four-pair configuration;
- 150-ohm shielded twisted-pair cable in a two-pair configuration;
- 50-ohm coaxial cable (not recommended for new installations); and
- 62.5 micron optical fiber cable in a two-pair configuration.

These cables exhibit performance that varies greatly depending on the frequency of the signal that is carried. For example, at Mbps speeds, a signal on a twisted-pair cable deteriorates in quality over a fairly short distance. The 568A standard provides performance criteria for the above cabling which must be met to be classified as 568A compliant. A summary of the attenuation limits specified by 568A for certain cable types is shown in Figure 1.3.13.

References

[1] International Standard ISO/IEC TR 8802-1, *Part 1: Overview of LAN/MAN Standards*. New York, 2001.
[2] IEEE Std. 802, *IEEE Standards for Local and Metropolitan Networks: Overview and Architecture*. New York, 1990.
[3] International Standard ISO/IEC 7498-1, *Open Systems Interconnection Basic Reference Model*. New York, 1994.
[4] *The Ethernet, A Local Area Network, Data Link Layer and Physical Layer Specifications,* Digital Equipment Corp., Maynard, MA; Intel Corp., Santa Clara, CA; Xerox Corp., Stamford, CT; Version 1.0, Sept. 30, 1980, and Version 2.0, Nov. 1982.
[5] IEEE 802.11, *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*. New York, 2007.
[6] IEEE 802.14, Working Group, *Standard Protocol for Cable-TV Based Broadband Communication Network*. New York, 1995.
[7] RFC 1700, *ASSIGNED NUMBERS*, J. Reynolds, J. Postel. Oct. 1994.
[8] International Standard ISO/IEC 8802-3: 2000-12-15 IEEE Std 802.3, Sixth Edition, *Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications*.
[9] ANSI/IEEE Std 802.2 [ISO/IEC 8802-3], *Logical Link Control*.
[10] ANSI/EIA/TIA 568A, *Commercial Building Telecommunications Cabling*.

1.4 RFID Architecture and Protocols

Chonggang Wang, Mahmoud Daneshmand, and Kazem Sohraby

Radio frequency identification (RFID) refers to a wireless protocol standard for the physical encoding of the signal transmitted, as the data link layer framing of the information transmitted. An RFID tag is an object that can be applied to or incorporated into a product or being (animal or human) for the purpose of identification using radio waves. Most RFID tags contain at least two parts. An integrated circuit for storing and processing information is joined with an antenna for receiving and transmitting the signal. Chipless RFID allows for discrete identification of tags without an integrated circuit, which allows tags to be printed directly onto assets at a lower cost than traditional tags. RFID technology is a significant element in enterprise supply chain management, as the use of RFID for inventory tracking and management supports global information technology and efficiency.

In this section, the general architecture of RFID and its basic elements are explained. Then, the latest ultra high-frequency (UHF) passive Generation-2 (Gen-2) RFID air-interface protocol, specified by EPC (Electronic Product Code) Global is presented.

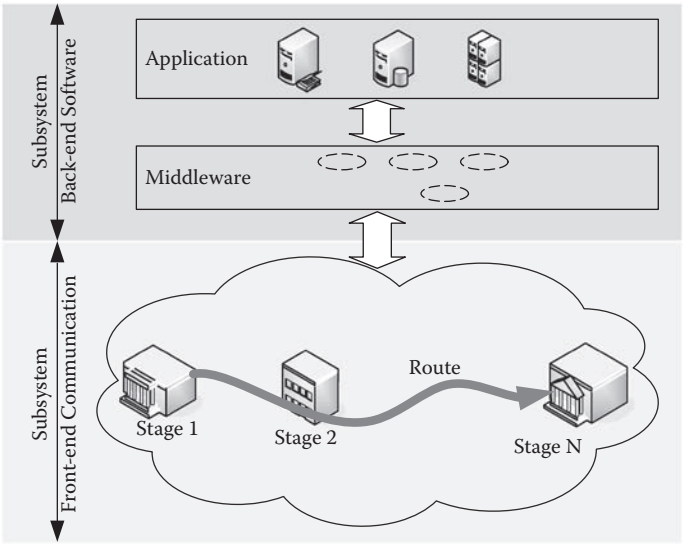
1.4.1 Introduction

The term RFID contains two key concepts: *radio frequency* (RF) and *identification* (ID). The first concept indicates that RFID uses radio waves for information transmission, while the second concept shows that the purpose of RFID is to identify entities. The radio waves approach used by RFID to identify entities implies two advantages of RFID over traditional barcode techniques: it is contactless and non-line-of-sight (NLOS). However, entity identification depends on RFID architecture and corresponding protocols, which will be explained below.

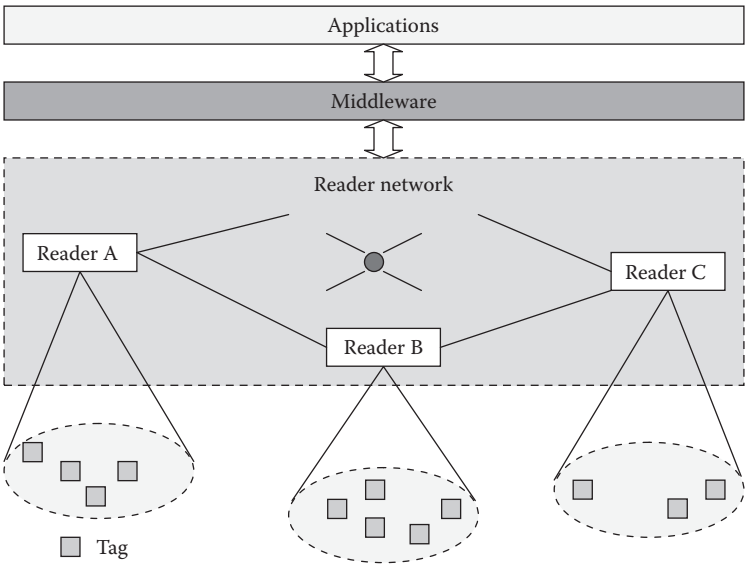
Although the first model of RFID deployed could be as early as World War II, where the "friend or foe" system used RFID techniques for plane identification [1], it is in recent years that RFID applications have been created and keep emerging in broad fields such as asset tracking and management, supply chain management, access control, object identification and positioning, and so on. Depending on the number of *tags* and *readers*, RFID applications can be classified as: single-tag-single-reader (STSR), such as an RFID-based car immobilizing system; single-tag-multiple-readers (STMR), such as active tag-based positioning [2], multiple-tags-single-reader (MTSR), such as in RFID-based smart refrigerators; and multiple-tags-multiple-readers (MTMR), such as in RFID-based smart shelves. The multiplicity of tags and readers has some implications. Multiple tags inevitably cause tag collisions and tag identification speed abates; multiple readers can provide benefits such as improved coverage and reliability as well as faster tag reading. In addition, in RFID-based positioning systems, the more readers that are deployed, the more precise the location inference. According to the mobility of tags and readers, four kinds of applications exist: immobile-tag-immobile-readers (ITIR), such as in RFID-based file management; immobile-tag-mobile reader (ITMR), such as in RFID-based vehicle lane-level positioning system [3], RFID mobile phone, mobile-tag-immobile-readers (MTIR), such as in RFID-based ePass systems for automatic toll payment; and mobile-tag-mobile-readers (MTMR), such as the friend and foe system. RFID system performance and reliability can however, be impacted under mobile environments due to the fact that it might be hard to guarantee right antenna orientation when for mobile tags and/or readers.

1.4.2 RFID Architecture

A RFID network generally has two subsystems (See Figure 1.4.1): a front-end communication subsystem and a back-end software subsystem.



(a) RFID System Logical Architecture including N Stages/Places



(b) Communications Subsystem at Stage *i*

FIGu RE 1.4.1 RFID system architecture.

1.4.2.1 Front End Communication Subsystem

The front-end communication subsystem, responsible for data collection, consists of readers (or interrogators) and tags (or transponders).

- **RFID Tags:** Each tag contains its own unique identity (ID). Tags cab be classified as: semipassive tags, which have a power source, such as an embedded battery or environmental power supply; active tags, which in addition to being semipassive, also have active communications capabilities, such as tag-to-tag ad hoc networking; and passive tags, which do not have their own power source and must be energized and activated by radio waves from readers. Although semipassive tags have

TABl E 1.4.1 RFID Air-Interface Protocols for Item Management

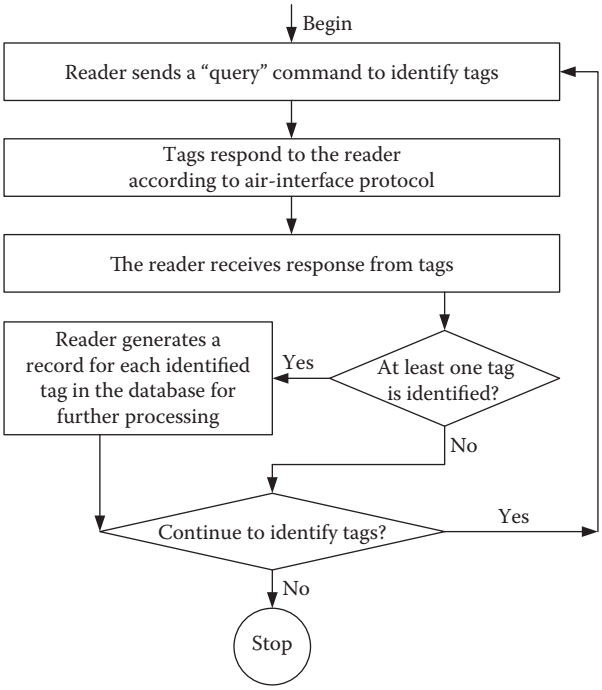
Air-Interface Protocols	Features
ISO/IEC 18000-1	Generic parameters for the air interface for globally accepted frequencies
ISO/IEC 18000-2	<135 KHz, short range
ISO/IEC 18000-3	13.56 MHz, middle range
ISO/IEC 18000-4	2.45 GHz, long range
ISO/IEC 18000-6	860–960 MHz, long range, Three types (A, B, and C)
	A: Aloha-based collision resolution
	B: Adaptive binary tree-based collision resolution
	C (EPC Gen-2): Adaptive aloha-like collision resolution
ISO/IEC 18000-7	433 MHz, long range, active tags

their own energy, it is only used for computation, not communication. Active tags have better radio performance but finite lifetimes constrained by the power source, while passive tags can be used forever and have no lifetime limit.

- **RFID Readers:** RFID readers detect RFID tags and read/write their ID and other information, which is referred to as RFID raw data. This is accomplished using transmission of a modulated RF signal to the tags first and acquiring the tags' responses later. Since radio signals attenuate along with the distance it travels, each reader can only interrogate/identify/read tags in its vicinity, which is called an *interrogation region*.
- **Antenna:** An antenna is usually embedded in each tag, as well as separately connected in pairs to a reader. After receiving the RF signal from the reader, the tag modulates its antenna reflection coefficient with the information to be transmitted back to the reader.
- **Air-Interface Protocols:** The RFID air-interface protocol defines the physical layer, line coding, protocol structure, modes and related parameters, and the collision arbitration algorithm for reader-to-tag and tag-to-reader communication. On the basis of the applications used, the reader and tags implement different air-interface protocols and different radio frequencies for communications. As a result, different standards for RFID systems have appeared in recent years. Among them, Electronic Product Code (EPC) UHF Generation-2, which is commonly known as the Gen-2 protocol [4], and the ISO/IEC 18000 series [5] are important standards for item management, which describe and define air-interface protocols working at different frequencies (See Table 1.4.1). As reported in a 2006 *RFID Journal* article [6], the Gen-2 protocol was incorporated into the ISO/IEC 18000 series in the middle of 2006, and called ISO/IEC 18000-6 Type C. However, in this paper we use Gen-2 instead of ISO/IEC 18000-6 Type C to refer to this protocol. As indicated in Table 1.4.1, this protocol operates in the range of 860 to 960 MHz.

In reality, multiple readers can be deployed and networked together. The process of identifying tags can be seen as a handshake process between readers and tags. For passive tags, readers initiate tag identification by issuing query commands. Tags reply to query commands and therefore readers successfully identify tags. On the other hand, active tags can either initiate the handshake process or wait for readers to initiate this handshake (whether active tags or readers initiate the handshake, depends on applications). In a case where multiple tags are scattered around a reader, there is a chance that more than one tag will simultaneously respond to the reader and therefore signals collide. This is referred to as *tag collision* [7]. In case of multiple readers placed close to each other, it is possible that the signal from one reader will interfere with that of other readers. Such an interference, which potentially exists among multiple readers, is called *reader collision* [8].

Communications between a reader and its tags are specified in an air-interface protocol, which usually provides an algorithm to resolve tag collision by means of an anticollision protocol. This defines the format of the commands, the timing between the reader and tags, and determines how the frequency



FIGu RE 1.4.2 The common process of tag identification.

and time resources are best utilized by the reader and tags. Although air-interface protocols could be different for various applications, the general approach to identifying tags is the same. In general, the tag identification process shown in Figure 1.4.2 consists of the following steps:

1. The reader begins the identification process by sending a *query command* to the tags.
2. The tags passively receive the query command and respond to the reader according to the specifications of the air-interface protocol.
3. The reader receives responses from tags and determines whether or not a tag is successfully identified.
4. If a tag is successfully identified, the reader generates a record of the tag information in the database for further processing.
5. If no tag is identified, the reader will arbitrate whether it is necessary to continue to identify tags or not.
6. If the answer is yes, it keeps sending query commands; otherwise, it stops. The arbitration rules depend on the application.

A reader identifies its tags, generates reading records, and forwards the readings to the software subsystem after preprocessing. A reader can execute commands from the software subsystem on a specified tag, such as writing, locking, unlocking, and even decommissioning it.

1.4.2.2 Back-End Software Subsystem

The back-end software subsystem consists of middleware and applications to fulfill tasks including data storage, data processing, data utilization, data mining, and so forth. The middleware can provide device management, data management, and other necessary management functions. For device management, traditional Simple Network Management Protocol (SNMP) can be utilized to realize online and real-time management of readers and tags; for data management, the middleware receives raw RFID data

from the reader network and inserts the data into a database for application-level querying. In addition, the middleware can also perform data cleaning, compression, and transformation. Applications residing above the middleware utilize the processed data to perform functions and execute tasks defined by the application logic. Database servers, directory servers, and application servers are necessary in the back-end subsystem. Database servers store the processed data; directory servers provide mapping between tags' identities and the identified product details; application servers receive commands from users, execute them, and return results to users.

1.4.3 Gen-2 RFID Protocol

Gen-2 protocol [4] offers a new air-interface protocol including physical and Media Access Control (MAC) specification for UHF RFID passive tags, which operates in the range of 860 to 960 MHz. This protocol provides advanced new features designed for fast tag identification, flexibility, and security. First, the specification describes an adaptive slot-count (Q) selection algorithm, referred to in this chapter as the *adaptive Q algorithm*, to resolve tag collisions adaptively. Second, the capacity of the radio link between the tags and the reader can be adjusted. For example, Gen-2 offers various tag-to-reader data rates (referred to as TRrate) through flexible and configurable modulation in the tag-to-reader direction. Third, Gen-2 provides 32-bit password-protected access control and designs *Kill* and *Lock* commands for security consideration. In the following, we briefly introduce the tag identification process in Gen-2, including basic operation and the adaptive Q algorithm.

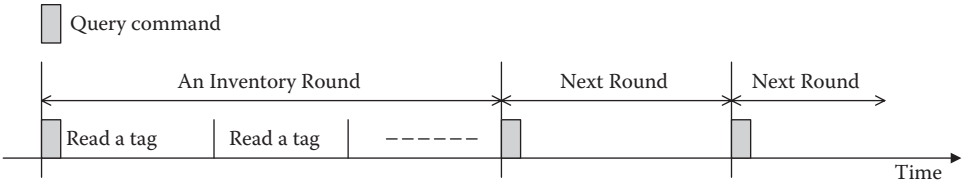
1.4.3.1 Basic Operations of Identifying Tags

In practice, an RFID system consists of *one* or *multiple readers* and a *population of tags*. Readers must identify tags and read/write their ID as well as other information called RFID raw data. Efficient tag identification requires clever algorithms and operational procedures to avoid reader *collisions* as well as *tag collisions*. The Gen-2 protocol uses an efficient operational procedure to increase tag identification *success rate* and *speed*. Here, we provide a high-level operational procedure of the Gen-2, necessary for a performance analysis (a detailed procedure is given in [4]):

Effectively, the Gen-2 protocol partitions a tag population into distinct subpopulations so that tags can associate separately and independently with each of the several readers. As described in [4, p. 37], "readers (interrogators) support, and tags provide four sessions denoted as S0, S1, S2, and S3. Tags participate in one and only one session during an *inventory round*. Two or more readers can use sessions to independently inventory a common tag population. Tags shall maintain an independent *inventoried flag* for each session. Each of the four inventoried flags has two values, denoted *A* and *B*. At the beginning of each and every inventory round a reader chooses to inventory either A or B tags in one of the sessions. Tags participating in an inventory round in one session shall neither use nor modify the inventoried flag for different session. The inventoried flags are the only resource a tag provides separately and independently to a given session; all other resources are shared among sessions."

Readers manage tag populations using three basic operations: *Select*, *Inventory*, and *Access*. Each of these operations consists of one or more commands. *Select* commands select a particular tag population prior to *Inventory*; *Inventory* commands facilitate inventorying tags prior to *Access*; and *Access* commands are used to read from or write to individual tags. In the rest of this section we focus on describing certain aspects of *Inventory* operation (the process by which a reader identifies tags) necessary for performance analysis of Gen-2 protocol tag identification *success rate* and *speed*.

The inventory command set consists of *Query*, *QueryAdjust*, *QueryRep*, *ACK*, and *NAK* which will be explained later in this section. The Query command initiates an *inventory round* and decides which tags participate in the round. An *inventory round* is the period between successive Query commands. Figure 1.4.3 depicts a basic tag identification/inventory process according to the Gen-2 protocol specification [4]. The process consists of a number of inventory rounds. In [4], the term *inventory round* is defined as the period between successive Query commands, which is issued by the reader and therefore



FIGu RE 1.4.3 Tag identification process in Gen-2 protocol.

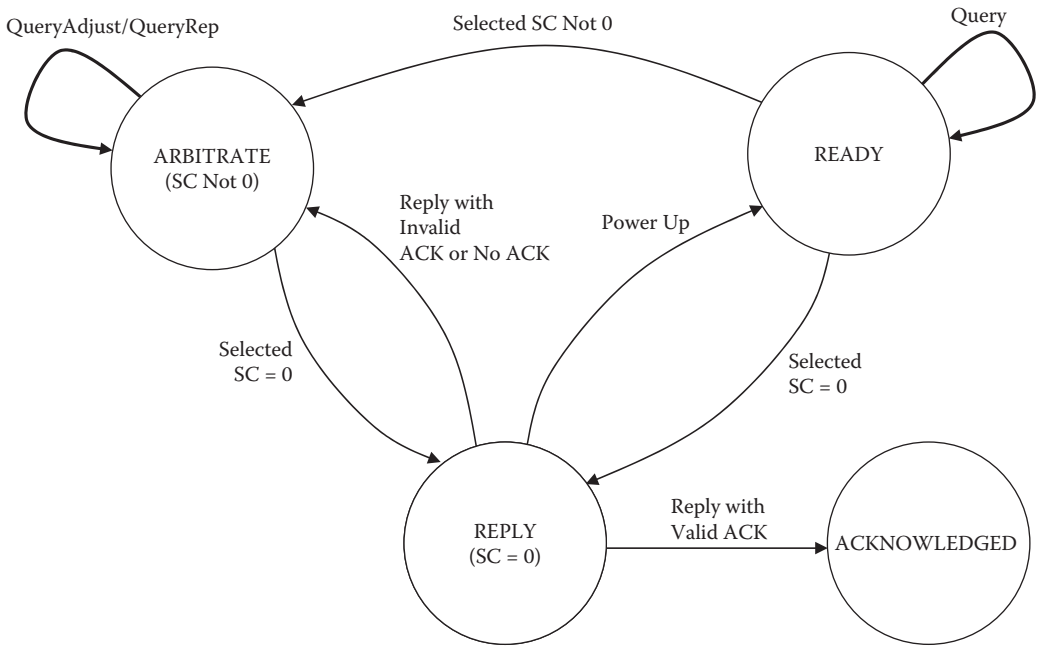
the issuing of a new Query command implies the ending of the current inventory round and the beginning of new inventory round. According to [4], the reader can issue a Query command when the system is powered up or when the channel is idle; therefore, if there are no tags identified in a particular round, the reader can issue a Query command to start a new inventory round. During each inventory round, the reader issues a set of Query, QueryAdjust, or QueryRep commands to identify tags. When a tag is identified in a particular inventory round, it will cease to respond to commands from the reader in the same round.

The basic operations of successfully identifying a tag can be summarized as follows:

1. **Reader → Tags:** The reader initiates the process of an inventory round by sending a Query command to the population of tags it has chosen to participate in the round. The Query command sends an integer-valued parameter Q (0 <= Q <= 15) and instructs tags to independently select a random integer from the uniform distribution [0, 2^Q–1] and report the result back to the reader. The reader waits for replies from the tags.
2. **Tags → Reader:** Tags' responses to the reader lead to either a *Success (S)* or a *Failure (F).* It is a Success if exactly one tag has selected number 0, and thus that tag has been identified by the reader. Otherwise, it is a Failure and no tag is identified.
3. **Reader → Tag:** The reader waits for a reply from tags. There are three possibilities: no reply, a single and successful reply from a single tag, and collided replies because more than one tag responds to the tag at the same time. If there is a single and successful reply, the reader sends an ACK for acknowledgment; otherwise, it keeps sending query commands. An ACK issued by the reader could be heard by all tags due to the broadcast nature of the wireless channel; however, only the tag that responded previously is eligible to receive it and process it.
4. **Tag → Reader:** The tag that responded previously processes the received ACK and reports its ID to the reader.
5. **Reader → Tags:** In the event of either a Success or a Failure, the reader continues with new query commands and the tag identification process.

How tags respond to the reader is dependent on and controlled by the reader's commands. Since all tags respond independently, collision could occur among tags' responses and a *slotted random anticollision* algorithm is described in [4] for its resolution. The following is a summary of the algorithm: Upon receiving a Query or QueryAdjust command, each tag deposits an integer-valued number in its *Slot Counter* (SC), which is selected at random from a uniform distribution [0, 2^Q–1], where Q is an integer-valued parameter. Q varies in the 0 to 15 range, and is designated and adjusted by the reader. The value of Q is embedded in the Query command, and is updated using the QueryAdj command. After selecting the SC, tags that have SC = 0 respond to the reader command. The reader continues to issue new Query commands that instruct those unidentified tags to either reselect or reduce their SC, or to restart and choose a new SC with a new Q value. If multiple tags happen to choose SC = 0, tag collision occurs again, which can be resolved under the control of the reader. In the following, we explain the resolution process and give an example.

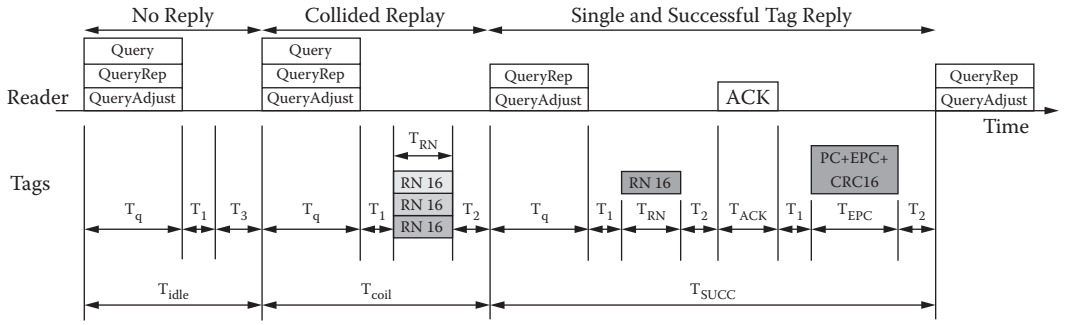
Gen-2 protocol defines three types of query commands: Query, QueryAdjust, and QueryRep. Query carries the value of parameter Q and initiates an inventory round; Query triggers all tags to select



FIGu RE 1.4.4 Major states diagram in Gen-2.

their SC uniformly from [0, 2^Q–1]. QueryAdjust is used to ask all tags to adjust the value of Q and reselect their SC. With this command, Q is incremented by 1, decremented by 1, or remains unchanged according to the adaptive Q algorithm, which will be explained later. QueryRep is used by the reader to notify all tags to decrement their SC by 1. Those tags that contain SC = 0 will decrement to 7FFF. In summary, Query and QueryAdjust inform all tags of the latest Q value and trigger them to reselect SC, while QueryRep instructs tags to decrement their SC by 1. The sending of Query by the reader implies that a new inventory round has begun. Within an inventory round, several QueryAdjust and/or QueryRep commands can be transmitted by the reader in order to identify the remaining tags. According to [4], the Query command can be issued when the system is powered on or if there are no replies from the tags; QueryAdjust (or QueryRep) can be issued either after a tag is successfully identified or when a channel is in collision. Such operations are described in [4] as options for implementation consideration.

The detailed procedure of tag identification is as follows [4] (See Figure 1.4.4). As shown in this figure, a tag can be found in one of four major states. In [4], additional states are described, but for the purpose of understanding system operation and performance analysis, which is included in future sections, these four states will suffice. Tags are normally in a holding state called the *ready* state. In this state, tags receive a Query command (start of a new inventory round), which contains an integer-valued parameter Q in the range (0,15). Tags in this state select an SC as an integer-valued random number from [0, 2^Q–1]. If the selected random number is not zero, the tag moves to the *arbitrate* state and waits for a QueryRep or QueryAdjust command, which causes the SC to decrement by 1, or change the value of Q and choose a new SC, respectively. While in the ready or arbitrate states, if SC = 0, the tag moves to the reply state. In this state, the tag selects a 16-bit random number referred to as RN16, which will be used as a means of identification by the reader and tag for the duration of command exchange between the two. In this state, the tag replies to the reader with an acknowledgment message containing this RN16, its PC, EPC, and CRC-16. If the reader receives this message correctly, the successful tag is assumed to have been correctly identified and is moved to the acknowledged state. Otherwise, the tag state changes to the arbitrate state. In this state, the tag waits until a

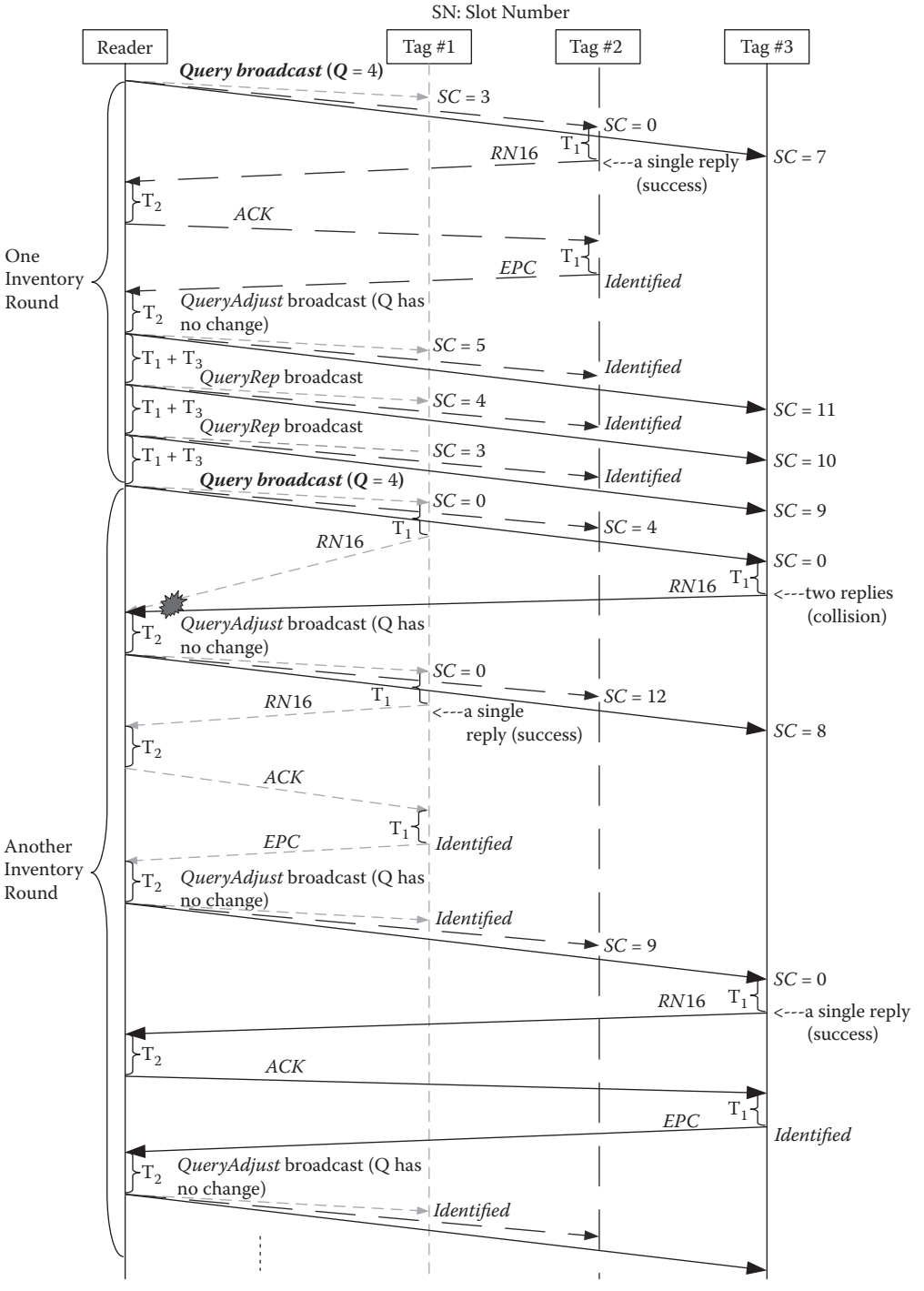


FIGu RE 1.4.5 Tag identification process and timing relationship in Gen-2 protocol.

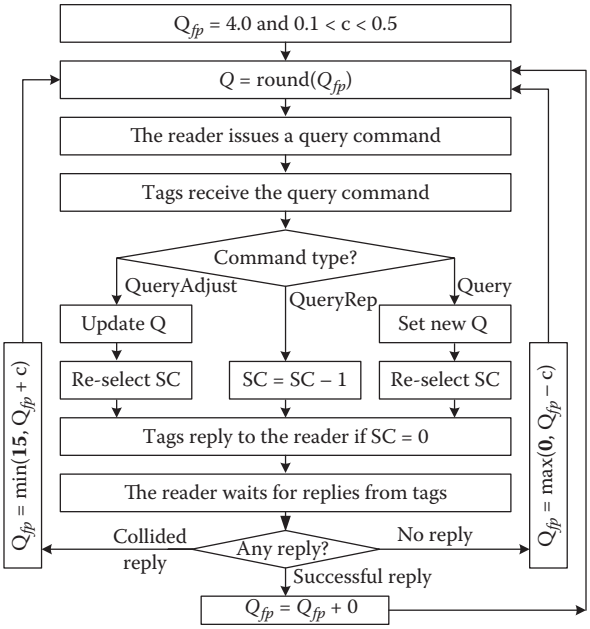
QueryRep or QueryAdjust command from the reader is received, and when its SC = 0, moves back to the reply state. While in the ready state, a new inventory round begins, a Query command is issued by the reader resulting in selection of a new SC value by the tag, and its subsequent movement to the reply or arbitrate state. Figure 1.4.5 shows the process involved in an inventory round including timing parameters.

Following the procedures just described, we present a simple example of Gen-2 tag identification in Figure 1.4.6, where it is assumed that there is one reader and three tags. As shown in Figure 1.4.6, the reader initiates an inventory round by broadcasting a Query command that contains $Q = 4$ (note that the Gen-2 protocol requires $Q = 4$ as the initial value). All three tags receive the Query command and choose their SC as a random number uniformly distributed in the 0 to 15 range since $Q = 4$. Suppose Tag 1 selects SC = 3, Tag 2 selects SC = 0, and Tag 3 selects SC = 7. Tag 2, which has SC = 0, responds to the reader by generating an RN16 and sending it to the reader. This value serves as the identification mechanism during the command–response exchange between the reader and the tag. Once the RN16 has been received from Tag 2, the reader acknowledges it by sending an ACK to the tag. Tag 2 receives the ACK and sends its RN16 (along with several other pieces of information) back to the reader. Therefore, Tag 2 is successfully identified and will not respond to the reader in this inventory round again. The reader continues to identify the remaining tags by sending QueryAdjust commands, which causes Tag 1 and Tag 3 to reselect their SC; Tags 1 and 3 are in arbitrate state. Suppose Tag 1 selects SC = 5 and Tag 3 selects SC = 11. Therefore this time, none of the tags respond and the reader will continue to send QueryRep commands, which causes Tag 1 and Tag 3 to decrement their SCs by 1. After sending two consecutive QueryRep commands, Tag 1 has SC = 3 and Tag 3 has SC = 9. At this time, the reader has detected an idle channel for the third time. We suppose the reader starts a new inventory round at this time by sending another Query command. Then all three tags are eligible to be identified again. All of them reselect their SC. Suppose Tag 1 selects SC = 0, Tag 2 selects SC = 4, and Tag 3 selects SC = 0. The SCs of both Tags 1 and 3 are 0, and therefore they both respond to the reader by sending their RN16s, which inevitably collide. The reader detects this collision and sends QueryAdjust commands to the tags. At this time, suppose Tag 1 chooses 0, Tag 2 chooses 12, and Tag 3 chooses 8; therefore Tag 1 is eligible to respond to the reader and will be successfully identified by the reader after following the same procedures described above to identify Tag 2. Tag 1 will not respond in the current inventory round. The reader continues to identify tags by sending QueryAdjust commands to all tags. Following the same procedures above, Tag 3 will be successfully identified after exchanging messages.

From the above example, it can be noted that parameter Q plays a critical role in the process of tag identification. If Q is too small while the number of tags is too large, many collisions are bound to occur. On the other hand, if Q is too large but the number of tags is small, the channel will frequently be in idle state and many query commands will have to be issued. In either case, the tag identification speed is negatively affected. Therefore, Gen-2 protocol introduces an adaptive Q algorithm to adjust the value of Q dynamically.



FIGu RE 1.4.6 Illustration of tag identification according to Gen-2 protocol. (In this example, we assume the reader begins a new inventory round after detecting an idle channel for three consecutive times.)



FIGu RE 1.4.7 Adaptive algorithm in Gen-2 protocol.

1.4.3.2 Adaptive Q Algorithm

As described previously, communications between the reader and the tags follow a four-way handshake. The reader initiates the handshake by sending a Query command and controls communications with the tags. Independently, tags receive query commands from the reader and respond to them. If more than one tag responds simultaneously, collision occurs. In this case the reader, through query commands, instructs the tags to use an algorithm to resolve collisions. Each tag chooses its SC in the range $[0, 2^Q–1]$ whenever it receives a Query or QueryAdjust command from the reader. Only a tag with an SC = 0 responds to the reader. As shown in Table 1.4.2, if the number of tags to be identified is large while Q is small, many collisions may occur. On the other hand, if Q is large while the number of tags is small, the channel could be idle most of the time. Both cases influence the speed by which tags are identified. In order to overcome these problems, an adaptive Q algorithm is proposed in Gen-2, which adjusts Q dynamically according to the flow chart in Figure 1.4.7. When the reader receives a reply from the tags after issuing a query command or the time $T_1 + T_3$ has expired before receiving a reply (see Figure 1.4.5), the algorithm in Figure 1.4.7 is triggered by the reader to update Q based on the following rule. In this flow chart, suppose Q_{fp} is the float-point representation of Q. The value of Q is determined based on the integer nearest to Q_{fp}. The detailed operation is as follows:

- **Collided Reply:** This is due to the fact that more than one tag has selected SC = 0, which in turn could imply that Q is too small and that the number of remaining tags is too large. In this case Q_{fp} is incremented by the value of parameter c, a real number.* After this operation, if Q_{fp} exceeds 15, it is set to 15. The value of Q is the integer that is nearest to Q_{fp}; that is, $Q = round(Q_{fp})$.
- **No Reply:** This can be due to the fact that none of the tags has selected SC = 0, which in turn could imply that Q is too large and that the number of remaining tags is too small. In this case Q_{fp} is decremented by c, a real number.[2] After this operation, if Q_{fp} is negative, we let $Q_{fp} = 0$. Then the value of Q is $Q = round(Q_{fp})$.

* The typical values for suggested by the Gen-2 standard [1] are in the range of (0.1, 0.5). It is suggested in [1] to use a small when Q is large, and a large when it is small.

TABl E 1.4.2 An Example of the Adaptive *Q* Algorithm

Sequence of Query Commands	Type of Replies from Tags	Q_{fp}	Q
1	Collided	4.2	4
2	Collided	4.4	4
3	Collided	4.6	5
4	Collided	4.8	5
5	Collided	5.0	5
6	Idle	4.8	5
7	Idle	4.6	5
8	Idle	4.4	4
9	Idle	4.2	4
10	Idle	4.0	4
11	Idle	3.8	4
12	Successful	3.8	4

Note: In this example, *c* is supposed to **be** 0.2 and $Q_0 = 4$.

- **Successful Reply:** This means that only one tag has selected SC = 0 and that the current value of *Q* is proper. In this case, Q_{fp} and *Q* remain unchanged.

According to the specification in [4] (also shown in Figure 1.4.7), the initial value of *Q* is $Q_0 = 4$. Since $c < 1$, there are three possibilities after each update: *Q* increments by 1, *Q* decrements by 1, or *Q* remains unchanged. The reader uses the QueryAdjust command to notify the tags of which possibility may happen. An example of values of *c* and Q are given in Table 1.4.2, where replies from tags are assumed to have collided, been successful, or there is no reply and $c = 0.2$. As shown in Table 1.4.2, the first five Query commands receive a collided reply, and the third command leads *Q* to be incremented from 4 to 5.

1.4.4 Gen-2 Performance Improvement

In the Gen-2 adaptive Q algorithm, Q is adjusted with the same increment/decrement *c* no matter whether the reader receives a collided reply or no reply. We argued that this is not the best way and designed a new algorithm (in [9]) that used difference values of *c* for scenarios "collided reply" and "idle reply," respectively. The new algorithm uses operations similar to those in Figure 1.4.7 to adjust Q with the following changes: (1) when there is no reply, $Q_{fp} = Q_{fp} - c1$; (2) when there is a collided reply, $Q_{fp} = Q_{fp} + c2$; and (3) when there is a successful reply, $Q_{fp} = Q_{fp}$. Here, c1 and c2 are determined as follows:

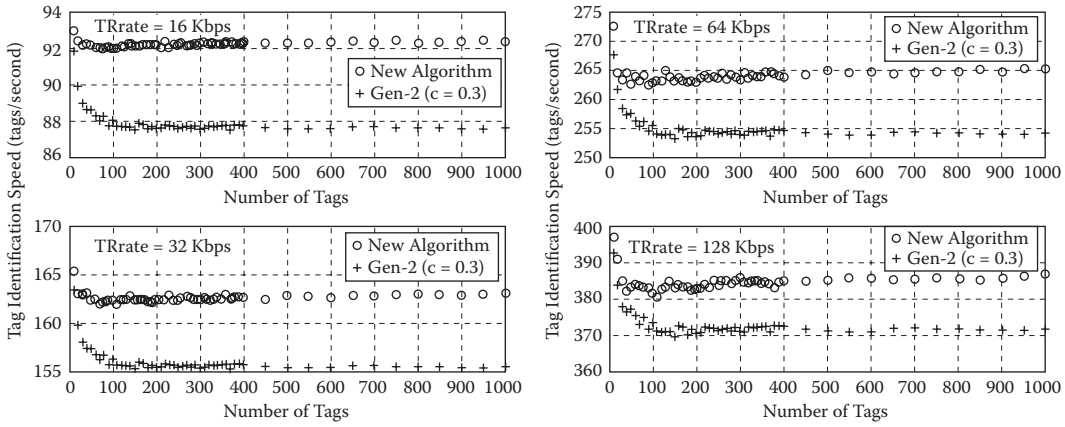
$$c1 = 0.1 \tag{1.4.1}$$

$$c2 = \min (1.0, c1 * \text{Ave\_T}_{coll}/\text{Ave\_T}_{idle}) \tag{1.4.2}$$

where Ave_T_{coll} and Ave_T_{idle} are mean values of T_{coll} and T_{idle}, respectively. Computations for determining Ave_T_{coll} and Ave_T_{idle} are given in [9]. The philosophy of the new algorithm is that we need different values for *c* because the duration of a collided reply is different from and usually longer than the duration of an idle reply, so as to improve tag identification speed.

We conducted simulations as follows:

1. Reader-to-Tag rate (RTrate) is fixed at 64 Kbps and Tag-to-Reader rate (TRrate) has four values: 16, 32, 64, and 128 Kbps.
2. The number of tags N_t varies between 10 and 1000; below $N_t = 400$, increments are 10 while between $N_t = 400$ and $N_t = 1000$ the increments are 50.



FIGu RE 1.4.8 Tag indemnification speed of Gen-2 new proposed algorithm; (a) TRrate = 16 and 32 Kbps; (b) TRrate = 64 and 128 Kbps.

3. A channel with zero bit error rate is assumed.
4. When there is a collided reply, the reader sends a QueryAdjust command; when there is a successful reply, the reader issues a QueryRep command; when there is no reply, the reader uses a QueryRep command if Q has no change or a QueryAdjust command if Q is changed.

Tag identification speed is calculated as the ratio of the total number of identified tags over the total time consumed and is shown in Figure 1.4.8, which demonstrated that the new slot-counter algorithm achieves higher tag identification speed than the Gen-2 adaptive Q algorithm, especially when the number of tags is large.

1.4.5 Conclusions

This section discusses RFID general architecture, which consists of a front-end communication subsystem, a back-end software subsystem, and a specific air-interface protocol, the EPC Global generation-2 UHF RFID protocol. A new algorithm was also introduced to improve the performance of Gen-2's adaptive Q algorithm.

References

[1] J. Landt, "The history of RFID," *IEEE Potential* 24, no. 4 (October–November 2005): 8–11.

[2] L. M. Ni, Y. Liu, Y. C. Lau, and A. Patil, "LANDMARC: Indoor location sensing using active RFID," *ACM Wireless Networks* 10, no. 6 (November 2004): 701–710.

[3] M. Donath, "Future directions in RFID application and research in transportation," *Conference on Research Opportunities in RFID Transportation Applications*, October 17–18, 2006, National Academies Keck Center, Washington, DC, http://www.trb.org/conferences/rfid/pdf/Donath.pdf.

[4] *EPCglobal Specification for RFID Air Interface*, "EPC™ radio-frequency identity protocols class-1 generation-2 UHF RFID protocol for communications at 860 MHz–960 MHz," version 1.0.9, January 2005.

[5] ISO/IEC 18000: *Information technology—Radio frequency identification for item management*, December 2008.

[6] M. C. O'Connor, "Gen 2 EPC Protocol approved as ISO 18000-6C," http://www.rfidjournal.com/article/articleview/2481, July 2006.

[7] J. Myung, W. Lee, and T. K. Shih, "An adaptive memoryless protocol for RFID tag collision arbitration," *IEEE Transactions on Multimedia* 8, no. 5 (October 2006): 1096–1101.
[8] D. W. Engels and S. E. Sarma, "The reader collision problem," *Proc. of IEEE International Conference on Systems, Man and Cybernetics (SMC'02)*, October 6–9, 2002.
[9] C. Wang, M. Daneshmand, and K. Sohraby, "A new slot-count selection algorithm for RFID Protocol," *Proc. of Chinacom 2007*, August 22–24, 2007, Shanghai, China.

1.5 Design of Wireless Sensor Network Applications, Hardware and Software

Sajid Hussain

A wireless sensor network (WSN) generally consists of a large number of sensor nodes, called *motes*, which could be either densely or sparsely deployed. The deployment of the sensors and their locations are not necessarily predetermined, but could be random to allow deployment in any kind of terrain. Sensor nodes can be used for continuous sensing, event detection, location sensing, and local control of actuators. The concept of microsensing and wireless connection of these nodes promises many new application areas [1, 2].

These tiny sensors have the capability of sensing and processing data, as well as transmitting data over short distances. Advances in sensor technology, low-power analog and digital electronics, have allowed the development of small, relatively inexpensive and low-power sensors, called motes. These motes are equipped with: (i) sensor modules (acoustic, seismic, image sensor, motion, temperature, humidity, and light) to monitor these variables in the environment in quantitative terms; (ii) a digital processor to process the signals from the sensors and to perform network protocol functions; (iii) radio modules for communication; and, (iv) a battery to provide operating energy. Despite the limited range of the sensor nodes, accurate and reliable information can be acquired by aggregating the readings of a group of nodes. To enable remote sensing of an environment, the nodes must send the acquired data to a distant base station (BS), where a user can access the information. Moreover, using the strength of the received signal, these motes can estimate the distances of their neighboring nodes.

Sensor networks may consist of many different types of sensors such as seismic, low sampling rate magnetic, thermal, visual, infrared, acoustic and radar, which are able to monitor a wide variety of ambient conditions that include the following: temperature, humidity, vehicular movement, light intensity, pressure, noise levels, the presence or absence of certain kinds of objects, mechanical stress levels on attached objects, and the current characteristics such as speed, direction, and size of an object.

Due to large network size, limited power supply, and inaccessible remote deployment environments, the WSN-based protocols have different requirements than traditional wireless protocols. The routing in WSN, for instance, uses multihop routing, which is not based on the principles of uniformity and fairness. The motes can filter the redundant, inaccurate, or conflicting sensor data. Instead of address-based and point-to-point communications as in traditional wireless protocols, the routing decisions are data-centric, where the goal is to efficiently disseminate queries and query results into the network.

The core operation of a WSN is to collect and process data at the network nodes, and transmit the necessary data to the base station for further analysis and processing. Currently, there are several energy-efficient communication models and protocols that are designed for specific applications, queries, and topologies. The cluster-based routing protocol proposed in this thesis is suitable for continuous monitoring of numerous widespread sensors, which are at a large distance from the base station.

1.5.1 WSN Versus Conventional Networking

WSNs have specific advantages over conventional wireless networks, which are usually large, expensive, directly connected to the end user, and need to be placed accurately and in predetermined positions.

On the other hand, WSN motes are scattered over a large area, including environments that are harsh, hostile, inaccessible, and remote. The network, however, requires easy deployment and maintenance-free strategies. Due to a large number of redundant nodes, these sensor networks could be designed as fault tolerant. Furthermore, wireless communication eliminates the need of expensive installation and maintenance of wired infrastructure. In summary, WSN has four basic advantages:

1. *Extended range of sensing.* The dense deployment of sensor nodes enables precise monitoring of an extended area.
2. *Fault tolerance.* Since several nodes can be located close to one another, the sensor data could be correlated data, such that a failure of a few of nodes would not result in significant damage, and the system as a whole could still produce acceptable quality information.
3. *Improved accuracy.* Since nodes in close proximity monitor information about the same event, the collected data can reduce uncertainty.
4. *Lower cost.* WSNs are less expensive than the conventional networks.

1.5.2 Design of WSNs

Due to the above-mentioned characteristics of WSN, a lot of interest about its utility and further improvement has been generated [3]. For efficient, effective, and durable WSN systems, the following parameters need to be considered.

1.5.2.1 Ease of Deployment

Since a WSN could contain hundreds or thousands of nodes, they have to be small and cheap enough and the method of deployment should be viable so as to allow placing them in remote, and at times harsh and even dangerous environments; and the nodes should be able to communicate with each other even in the absence of an established network infrastructure. Further, since there is no uniformity of placement, to function in such ad-hoc settings, sensor networks should be self-configuring, requiring no global control to set up or maintain the network. Extraction of information from the whole of the defined territory would not be possible otherwise.

1.5.2.2 System Lifetime

Since redeployment would be expensive, these networks should function for a long time. The nodes should provide acceptable quality results for a long time. Moreover, while operating in remote or dangerous territory, it may be impossible to retrieve the nodes in order to recharge or replace batteries. Therefore, there is a need for energy-efficient mechanisms and techniques for cluster-based routing and transferring data from one node to another, which would obviously increase the network lifetime [4,5]; however, due to limited computing and storage resources, the sensors are not equipped with an enriched operating system that can provide energy-efficient resource management. Hence, application developers are responsible to design efficient communication, clustering, and routing strategies with low-level energy consumption.

1.5.2.3 Latency

Data from sensor networks are typically time sensitive; thus, it is important to receive the data in a timely manner because long delays due to processing or communication may be unacceptable.

1.5.2.4 Quality

The quality would depend on how much data is sent by the sensor network; the more data the base station (BS) receives, the more accurate the estimation of the remote environment. Unlike in cellular networks or WLANs (wireless local area networks), in a sensor network, data sensed by each node is required at a remote base station, rather than at other nodes. Therefore, the notion of *quality* in sensor

network is very different from a WLAN or cellular network. For sensor networks, the end user does not require all the data in the network because: (i) the data from neighboring nodes are highly correlated, making the data redundant, and (ii) the end user typically cares about a higher-level description of events occurring in the environment the nodes are monitoring. The quality of the network is therefore based on the quality of the aggregated data set rather than the quality of the individual data signals.

1.5.3 WSN Research

The WSN design and implementation can be optimized using various reliable, high-performance, and cost-effective technologies [6]. In recent years, there has been a great recognition and interest in WSNs among researchers in academia and industry, as well as by practitioners in many organizations and businesses for developing an optimized network in terms of energy consumption, query processing, data dissemination, and sensor topology [3]. As each mote has sufficient computing, storage, and communication resources, motes can be treated as autonomous units that can create self-organized clusters by coordinating their activities and working collectively to achieve their goals [7].

The WSN research is most often performed at a high-level assumption that all hardware devices are identical. In reality, there are a decent amount of hardware devices (known as motes) that one might use for a wireless sensor network. Each of these motes has different underlying hardware such as its processor, memory storage, and radio devices. This work will discuss the hardware of the many motes available and discuss the benefits and drawbacks of each. Further, the need for a specialized operating system is given by the low system resources available on wireless sensor devices. TinyOS is an operating system designed specifically for wireless sensor network devices. The system architecture and design of this operating system is described and demonstrates how it is suited for resource-limited sensor devices.

1.5.4 Motes

In this section we briefly describe commonly used motes, though the list is not exhaustive.

1.5.4.1 Tmote Sky

The Tmote Sky is a wireless sensor manufactured by Moteiv (now known as Sentilla). Its size is roughly that of the two AA batteries that are used to power the device. It is equipped with the Texas Instruments MSP430 processor, which runs at 8 kHz. The radio module on this device is the ChipCon 2420, which is a 802.15.4-compliant radio module. The main attraction of the Tmote Sky is the integrated USB connection, which allows for ease of programming and computer connections. The Tmote Sky is limited to light, temperature, and humidity sensor readings. It is, however, possible to equip the board with custom sensors [8].

1.5.4.2 Mica2 Dot

The Mica2Dot is a wireless sensor device approximately the size of a quarter and weighs 3 grams. It is produced by Crossbow and is a member of the Mica wireless sensor product line. It runs off a single 3V coin cell battery. This device is based on the Mica2 mote and is equipped with the Atmel ATmega128L microcontroller and ChipCon 1000 radio module. This enables it to communicate with Mica2 devices and take part in a large heterogeneous network. The size of this device makes it easy to deploy and it can be easily hidden [9]. It can be equipped with a number of sensors such as light, sound, and movement [10].

1.5.4.3 Mica2

The Mica2 mote is another device produced by Crossbow and is equipped with the same processor, Atmel ATmega128L, and radio module, ChipCon 1000, as the Mica2Dot. It is larger in size than the Mica2Dot as it is powered by two AA batteries. Unlike the Mica2Dot, the Mica2 can be used as a base

station gateway when connected to a computer. Mica2 devices were a main wireless sensor device used for research for a number of years [11]. Mica2 motes can be equipped with sensors that measure movement, light, temperature, sound, and magnetic fields [12] or movement, barometric pressure, light, temperature, humidity, and GPS location [13].

1.5.4.4 Micaz

The Micaz mote is an improved version of the Mica2 mote by Crossbow. The sole improvement over the Mica2 is a new radio transceiver. It supports the ChipCon 2420 rather than the ChipCon 1000. It is a ZigBee-compliant 802.15.4 device that allows for transmitting at data rates up to 250 kbps. It is the same size as the Mica2 device and also uses two AA batteries. It supports the same sensoring devices as the Mica2 [14].

1.5.4.5 BTnode

The BTnode is a product similar to the Mica2 created by the Swiss Federal Institute of Technology Zurich. It is equipped with the same processor and radio hardware as the Mica2. The main feature of this device is that it is also equipped with a Bluetooth communication module. It is able to communicate with other nearby Bluetooth devices including a computer equipped with Bluetooth support. Both the radio and Bluetooth system can be used for communication at the same time. These components can be disabled to conserve energy if required. This device is powered by two AA batteries like most other motes [15].

1.5.4.6 Imote2

The Imote2 is a wireless sensor device intended for applications with increased processing and storage requirements. It runs a Marvell PxA271 XScale processor, which can run at 13 to 416 MHz, which is far greater than any of the other motes discussed above. It is equipped with the ChipCon 2420 radio module like many other motes. It has the capability to capture and process image data. It also has 32 MB flash memory and 32 MB SDRAM, which is a significantly greater than the memories of the other motes. Due to the increased capabilities of this mote, it requires 3 AAA batteries in comparison to the 2 AA batteries used by other motes. This mote is capable of running operating systems other than TinyOS, such as Linux [16].

1.5.5 Hardware Components

The many motes available today use a combination of independent hardware components. This includes the microprocessor and radio devices.

1.5.5.1 Processors

MSP430: The MSP430 is a 16-bit RISC microprocessor manufactured by Texas Instruments. The MSP430 runs at frequencies up to 8 MHz; however it can lower its frequency in sleep mode in order to reduce power consumption. The MSP430 has a mere 27 instructions available for use. The MSP430 has 48 KB flash memory used for program storage and 10 KB of RAM for program execution. It has 16 registers of which 4 are reserved for special functions. The wake-up time from sleep mode is very fast at less than 6 μs [17].

ATmega103: The Atmel ATmega103 is an 8-bit RISC microcontroller designed to be energy efficient. The ATmega103 has far more instructions than the MSP430 (121 instructions). This processor runs at a frequency of 6 MHz and has 32 registers available for use. It has 128 KB programmable flash memory used for storing programs and 4 KB RAM [18].

1.5.5.2 Radio Transceivers

CC1000: The ChipCon 1000 is a radio transceiver designed for low-power devices. It can be used at frequencies between 30 and 1000 MHz. It allows for variable transmission power levels from –20 to 10 dBm. It supports data rates up to 76.8 kbps [19].

CC2420: The ChipCon 2420 is a ZigBee-compliant radio transceiver for low-power devices. Boasting data rates of 250 kbps, it is quite fast for a low-power radio device. It follows the IEEE 802.15.4 specifications and broadcasts at frequencies between 240 and 2483.5 MHz. Signals are output at programmable power levels between –24dBm to 0 dBm in 8 steps [20].

1.5.6 TinyOS

TinyOS is an operating system designed specifically for use with sensor mote devices. It was developed by the University of California, Berkeley, EECS Department in order to meet requirements needed for wireless sensor networks.

1.5.6.1 Requirements

An operating system designed for wireless sensor networks must support low power consumption as it is intended for small devices with a limited power supply. Further, these devices are often deployed in large scale in areas that may not be easily accessible by humans. As it is not practical to consistently replace the batteries in the devices, conserving energy consumption is a major concern. As with energy consumption, processing and storage is also very limited. The system must be able to minimize the resources needed in order to allow the application on the device to run without issues.

Concurrency is another major concern for the operating system on these devices. The application must be able to manage multiple streams of data incoming and outgoing efficiently. As previously mentioned, the storage on the device is limited so it is not possible for the device to buffer large amounts of data from these streams for an extended period of time. They must be dealt with in a timely manner. Everything on these devices is limited including the space on the board itself. There is a limit to the number of controllers that can be placed onboard. Due to this, it is unlikely that multiple levels of controllers are possible when a single microcontroller is used. This microcontroller would need to connect directly to all the hardware chips and sensors. This microcontroller has a large responsibility to maintain as it must support the all these chips and be able to handle them all directly.

The system must also be very modular and an efficient modular operating system. These devices can be equipped with a variety of sensing devices and may be run on different types of devices. Due to this, it is required that it support a plug-and-play approach for new hardware chips. This allows the system to support many of today's devices and allows the operating system to be easily extended for any future components that may need to be included. The system should also be easy enough to allow the programmer to specify which components will be used. The final requirement is that the system should be robust.

Robustness is very important as these devices could potentially be used in mission-critical projects such as health monitoring or security. A failure of one device could lead to a complete system failure if it was an important mote in the network. These applications are expected to run for extended amounts of time so preventing failures and minimizing maintenance costs is a very important part of the system [21].

1.5.6.2 Other Operating Systems

While there are many operating systems available these days, there are none that completely satisfy the requirements for WSNs. The limited resources, modularity, and concurrency issues are too much for any single one of these operating systems to cover. Quite often, if they support the modularity required, they require excess storage and processing for the limited devices. Alternatively, those systems that support the low computing resources are much too constrained in terms of concurrency and modularity. For instance, there are embedded operating systems for appliances that have very low resource requirements. These systems can easily match the resource requirements needed for a WSN device. The problem with these systems is that they are very static in that all chip integration is hard coded and does not allow programmers to easily include integration with new chips. Also, they cannot support multiple streams of data incoming and outgoing, which is not suitable for WSNs. On the other hand, there are many operating systems designed for mobile devices such as PDAs and cell phones. These devices

have far greater storage and processing ability and the OS is similar to that of a desktop computing OS. However, these operating systems simply have a far greater resource drain and are not suitable for WSN devices [21].

1.5.6.3 Execution Model

The TinyOS execution model is structured as an event-invoked architecture. There are three tiers to its architecture: the scheduler, components, and frame/handlers. The scheduler included with TinyOS is a First In, First Out (FIFO)–based scheduler, which processes tasks in the order that they are submitted for processing. The scheduler can be interchanged with another type of scheduler if one so wishes. It must be an efficient implementation that takes energy consumption into consideration. When no tasks are to be performed, it can sleep the processor in order to conserve energy.

A component can be thought of as a physical device or encapsulated set of commands and events. Components are able to send commands to lower-level components, such as a hardware device, and receive an event notification when the command has completed. For instance, an application component can make a request to a sensor device asking for its sensor data through the Read command. When the command completes, it notifies the application component by invoking the event *readDone*, which passes the status and result as parameters. By using this architecture, commands do not block program execution, therefore saving wasted CPU cycles on stalls. Unlike commands, tasks are similar to normal functions of languages like C. These tasks stall the calling program until the task is completed. These should be used for computations that must be completed in full before the program continues. Tasks may be interrupted if a hardware event is signaled as those must be dealt with quickly.

The frame is a representation of the component's handlers, which describes commands, events, tasks, and memory requirements needed by the component. Components that employ the command triggers notification event technique are referred to as *split-phase components*. The term split-phase refers to the fact that the command is called but does not return its result directly. The result is later returned by signaling an event back to the caller. This allows for a command to be delayed without stalling the calling application [22].

The event model suits wireless sensor network devices quite well as task switching can be efficiently implemented. It also prevents CPU cycles from being wasted and the CPU switched into sleep mode. This allows for a longer sleep period and less mode switching. These two reasons increase battery conservation significantly and make the design attractive for wireless sensor networks. Since each component contains a static frame that describes its commands, events, and memory requirements, it allows for memory allocation at compile time. Avoiding dynamic memory allocation for these components saves CPU cycles and memory requirements, further indicating that this approach is ideal for WSN devices. Interestingly, static memory allocation can in some cases cause wasted memory if the allotted memory is more than that used by the device at any given time. This is often dependent on the amount of data that is flowing through the device, as in periods with little data flow; the memory allocated may be excessive. With dynamic allocation, it is possible that this will cause problems for motes with high usage. If multiple components were using dynamic allocation, they could run up a large allocation of memory causing starved components if all memory had been exhausted.

1.5.6.4 Event-Based Radio Model

Active message communication is the radio model used by most TinyOS sensor applications. Interestingly, the radio model typically used with TinyOS is another event-based model similar to that of TinyOS itself. Each packet sent across the network specifies an active message (AM) type. These AM types allow for applications to register listeners to specific types of messages. The beauty of this model is that applications can easily ignore messages of other AM types to prevent any invalid data from being accepted into the application. Also, packets of different AM types can be discarded earlier than the application layer of a wireless sensor device, which saves valuable CPU cycles. Not only does this save CPU cycles, but also allows the processor to sleep more often if those cycles are not needed by other tasks running on

the device. With the limited amount of energy that these devices can store, this is a useful bonus. Active messages (AM) originated from distributed systems where computations are shared among numerous networked computers. The AM allowed them to quickly and efficiently share data with little overhead [23]. Applying this technique in TinyOS works extremely well as reducing overhead is a significant factor. Since TinyOS is an event-based operating system, using active messages meshes very well. The ability for different portions of an application to subscribe to certain message types also prevents each application subscriber from determining whether they are interested in the packet itself [24].

1.5.6.5 NesC Programming Language

NesC is the programming language used to implement TinyOS. It is an extension of the well-known C programming language [25]. C was chosen as a base language due to its large programmer base, efficient implementation, and support for a large number of hardware devices. NesC allows the modular design of TinyOS components to be statically wired together at compile time. The compiler also supports checking that all modules are wired together properly and all required commands and events defined by interfaces are implemented. As components require commands to be implemented in the component itself and events to be implemented within other components that use it, having these checks available at compile time is very useful for the programmer and avoids problems during execution. It is difficult to debug applications on the sensor devices since there is no screen to output debugging statements to, so having these compile time checking is very valuable. Referring back to the requirements of a wireless sensor network architecture, robustness was considered a requirement. This compile time checking assists with meeting this robustness requirement [26].

1.5.7 Summary

This section provides an introduction to WSN applications and a few design issues. It summarizes several of the leading wireless sensor mote devices used for research and industry.

The sensor nodes are commonly designed for specific applications. For instance, the Mica2Dot would work well for real-world deployments as they are small, easily hidden, and easy to place. Compared to all the other devices, which are larger, difficult to hide, and simply a bit nonpervasive for deployments. The Mica2 and Micaz motes can be equipped with GPS modules, which would allow for sensor tracking and localization. The Micaz is a Zig-Bee-compliant device, which makes it an attractive device. The Tmote Sky is excellent for testing applications as it is equipped with a USB connector, which facilitates programming of the sensor. The Imote2 is ideal for applications with large processing requirements; however, it has a higher drain on the batteries when operating at full capacity. The BTnode is equipped with Bluetooth support, which could be a great alternative to radio communication for a small amount of devices.

The main architecture of the TinyOS operating system was discussed. It shows that this event model architecture works well for resource-constrained devices. It meets the low storage and processing requirements, along with being robust and supporting multiple data sources at once.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Computer Networks* 38, 4 (March 2002): 393–422.

[2] D. Estrin, D. Culler, K. Pister, and G. Sukhatme, "Connecting the physical world with pervasive networks," *IEEE Pervasive Computing* (January–March 2002): 59–69.

[3] D. Culler, D. Estrin, and M. Srivastava, "Querying the physical world," *Overview of Sensor networks* 37, 8 (August 2004): 41–49.

[4] V. Mhatre, C. Rosenber, D. Kofman, R. Mazumdar, and N. Shroff, "A minimum cost heterogeneous sensor network with a lifetime constraint," *IEEE Transactions on Mobile Computing (TMC)* 4, 1 (2005): 4 –15.

[5] N. Trigoni, Y. Yao, A. Demers, J. Gehrke, and R. Rajaramany, "Wavescheduling: Energy-efficient data dissemination for sensor networks," *ACM Proceedings of the First International Workshop on Data Management for Sensor Networks (DMSN), in Conjunction with the International Conference on Very Large Data Bases (VLDB)*. Toronto, Canada: August 2004, pp. 48–57.
[6] K. P. Ferentinos, T. A. Tsiligiridis, and K. G. Arvanitis, "Energy optimization of wireless sensor networks for environmental measurements," in *IEEE Proceedings of the International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA)*. July 2005, pp. 20–22.
[7] M. Younis, P. Munshi, and E. Al-Shaer, "Architecture for efficient monitoring and management of sensor networks," in *Proceedings of Workshop on End-to-End Monitoring Techniques and Services, with IEEE/IFIP Management of Multimedia Networks and Services (MMNS)*, September 2003.
[8] *Tmote Sky*, Ultra low power IEEE 802.15.4 compliant wireless sensor module, Moteiv, 2006.
[9] *Mica2Dot,Wireless Microsensor Mote*, Crossbow, accessed March 22, 2008. www.xbox.com.
[10] *MTS510, Mica2Dot Sensor Board*, Crossbow, accessed March 22, 2008. www.xbox.com.
[11] *Mica2, Wireless Measurement System*, Crossbow, accessed March 22, 2008. www.xbox.com.
[12] *MTS/MDA, Sensor Data Acquisition Boards*, Crossbow, accessed March 22, 2008. www.xbox.com.
[13] *MTS420/400, Environmental Sensor Board*, Crossbow, accessed March 22, 2008. www.xbox.com.
[14] *Micaz, Wireless Measurement System*, Crossbow, accessed March 22, 2008. www.xbox.com.
[15] *BTnode rev3 - Product Brief*, Swiss Federal Institute of Technology Zurich, 2006.
[16] *Imote2, High-Performance Wireless Sensor Network Node*, Crossbow, accessed March 22, 2008.
[17] *MSP430x1xx Family, User's Guide*, Texas Instruments, 2006.
[18] *8-bit AVR Microcontroller with 128K Bytes In-System Programmable Flash, ATmega103 / ATmega103L*, ATMEL, 2007.
[19] *CC1000, Single Chip Very Low Power RF Transceiver*, ChipCon, 2002.
[20] *CC2420, 2.4 GHz IEEE 802.15.4 / ZigBee-ready RF Transceiver*, ChipCon, 2004.
[21] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. E. Culler, and K. S. J. Pister, "System architecture directions for networked sensors," in *Architectural Support for Programming Languages and Operating Systems*, 93–104, 2000.
[22] P. Levis, S. Madden, D. Gay, J. Polastre, R. Szewczyk, A. Woo, E. Brewer, and D. Culler, "The emergence of networking abstractions and techniques in tinyos," in NSDI'04: *Proceedings of the 1st Conference on Symposium on Networked Systems Design and Implementation*. Berkeley, CA, USA: USENIX Association, 1–1, 2004.
[23] A. M. Mainwaring and D. E. Culler, "Design challenges of virtual networks: fast, general-purpose communication," *SIGPLAN Not.* 34, 8 (1999): 119–130.
[24] B. Phillip, H. Jason, and C. David, "Active message communication for tiny networked sensors," 2001. www.tinyos.net/media.html/.
[25] B. W. Kernighan and D. M. Ritchie, *The C Programming Language*. Upper Saddle River, NJ: Prentice Hall Press, 1988.
[26] D. Gay, P. Levis, R. von Behren, M. Welsh, E. Brewer, and D. Culler, "The nesc language: A holistic approach to networked embedded systems," in PLDI '03: *Proceedings of the ACM SIGPLAN 2003 Conference on Programming language design and implementation*. New York, NY, USA: ACM, 2003, pp. 1–11.

1.6 Multimedia Applications for Cognitive Radio Networks

Sajid Hussain and Muhammad Farhat Kaleem

The prevalent model of licensed radio spectrum is much established in business and government tradition [1]. At the same time, there is successful application of nonlicensed bands for short-, medium-, and long-range radio networks [2]; for example, Bluetooth, WLAN, and WiMAX. While the unlicensed bands are experiencing heavy spectrum utilization, the licensed bands have low to medium utilization

[3]. However, research in cognitive radio [4] and cognitive radio networks (Thomas et al. [5]), has led to many possibilities for opportunistic usage of unused spectrum in the licensed band. Not only can the widespread use of cognitive radio technology lead to more efficient use of the radio spectrum [6], it can also open the way for a plethora of applications; for example, multimedia and broadband wireless networking applications [7].

In this section, we discuss the relevance of cognitive radio networks for multimedia applications, and how making efficient use of unused spectrum can enable widespread use of multimedia applications. Although military usage is a main driving force behind cognitive radios and networks [8], the ordinary consumer market will also emerge as very significant for cognitive radio technologies [9]. Indeed, it is this market where multimedia applications will assume much importance. The success of this segment of the consumer market will depend on technical issues, such as quality of service, but also on nontechnical issues, such as pricing of applications for consumers. Similarly, while efficient design and development of cognitive radio networks is important for realizing dynamic spectrum access, the role of wireless devices that can take advantage of these networks is equally important. In the next sections, we describe different aspects relevant for multimedia applications for cognitive networks. First, we provide the characteristics of dynamic spectrum access, and elaborate different aspects of cognitive radios and cognitive radio networks. Then we discuss issues related to the design of cognitive radio networks, and consider which aspects of the design are related to the suitability of cognitive radio networks for multimedia applications. Technical issues like the role of cognitive radio devices, policies and languages for cognitive radios and networks, and various facets of quality of service (QoS) are also covered in our discussion. We also consider the importance of nontechnical issues like the economics of multimedia applications for cognitive radio networks, and discuss how these issues affect, and are affected by, the technical aspects. Finally, we conclude by summarizing the main points.

1.6.1 Cognitive Radios and Cognitive Radio Networks

There are numerous definitions of cognitive radio in literature [10,11,12,13]. All of these definitions allow us to consider a cognitive radio as a stand-alone device, or a node in a network, which has an observe→think→act cycle [9], by way of which the cognitive radio autonomously observes the radio environment and improves its sensing, awareness, and adaptation capabilities by learning intelligently. Therefore the main properties of a cognitive radio may be classified [14] as *sensing, awareness, adaptation*, and *learning*.

A cognitive radio would have to be aware of the existing radio conditions, of any cognitive radio network of which it may form a part, and any other cognitive node that may interfere with its operation. Similarly, adaptation to the existing radio conditions, for example by on-the-fly modification of the transmission and reception parameters, is necessary to achieve an acceptable QoS for users' needs and applications [12]; for example, multimedia applications.

Cognitive radios may also be categorized into two types, where the categorization is with the goal of maximizing the QoS expectations of the applications. These types of cognitive radio are *static* and *dynamic* [11] cognitive radio, where static cognitive radio represents a cognitive device that uses a fixed spectral bandwidth to transmit data, and a dynamic cognitive radio device can expand or contract its spectral bandwidth. Dynamic cognitive radios demonstrate higher statistical multiplexing and lower delay [11], which makes them more suitable for multimedia applications.

Cognitive radio networks can be considered as consisting of two major parts, which may be defined as the *cognitive engine*, and the *nodes* [15], which are cognitive radios. It is also possible to classify the cognitive radio network as *centralized, distributed,* or *partially distributed* [5]. Centralized cognitive radio networks have all the cognition capabilities in the cognitive engine with the nodes having limited cognition capabilities, whereas distributed cognitive radio networks have full cognition capability built into the nodes. Partially distributed networks are between the two extremes.

Collaboration within the cognitive radio network also allows us to classify cognitive radio networks as *cooperative, noncooperative*, and *partially cooperative*. Cooperative cognitive radio networks represent those networks where the nodes collaborate to accomplish tasks. This collaboration can be ad hoc or in the form of organized teamwork [10]. There is no collaboration between nodes in a noncooperative cognitive radio network, whereas in a partially cooperative cognitive radio network a group of nodes may collaborate, while others may not.

Finally, if a cognitive radio network spans a single type of network, it may be classified as *homogeneous*, whereas a cognitive radio network spanning multiple wireless networks of different types can be classified as *heterogeneous*. A heterogeneous network will be more challenging for cognitive radios, but more beneficial for multimedia applications, due to the availability of more resources (e.g., radio spectrum) and more mechanisms (e.g., transmission technologies) spread over various networks.

It is important here, however, to view a cognitive radio network as having an end-to-end scope [5], where *end-to-end* refers to all network nodes involved in the transmission of a dataflow. It is therefore possible to attribute the properties of cognitive radio mentioned previously to cognitive radio networks as well, though in the context of cognitive radio networks the sensing, awareness, adaptation, and learning are with respect to end-to-end goals, thereby implying a networkwide scope.

1.6.2 Dynamic Spectrum Access

Radio spectrum has a high social and economic value, and represents a scarce and highly valuable resource that needs to be managed efficiently [16]. Given that a large part of the unlicensed spectrum may be lying unused at a given time, approaches that allow sharing of the licensed spectrum provide a good basis for flexible and efficient use of the radio spectrum. Known by different names, such as dynamic spectrum access (DSA) [17] or spectral hole-filling [18], opportunistic usage of the unused spectrum forms the core idea of cognitive radio and cognitive radio networks.

Sharing of the spectrum can be classified into two categories—*horizontal* and *vertical* [18]. The former represents the scenario where all networks and users have equal rights to the spectrum, as is the case for unlicensed spectrum bands (e.g., Bluetooth and WLAN). The latter is valid for the licensed radio spectrum case where the networks and users may be classified into primary, i.e., licensed, and secondary, i.e., opportunistic, users. Secondary users may *only* access the licensed spectrum opportunistically provided they do not affect the operation of the primary users [14].

Dynamic spectrum access can be in the context of two frameworks, namely *cooperative* and *noncooperative* dynamic spectrum access [19]. In cooperative dynamic spectrum access, a cognitive radio node may only access the spectrum with permission from the licensed user of the spectrum, which may also involve monetary agreements. Noncooperative dynamic spectrum access, on the other hand, refers to cognitive radio nodes accessing the licensed spectrum opportunistically.

The main issues with respect to dynamic spectrum access that a cognitive radio has to deal with, either independently or within a cognitive radio network, can be listed as [20]:

1. Spectrum sensing
2. Spectrum management
3. Spectrum sharing
4. Spectrum mobility

Next we discuss each of these issues briefly.

1.6.2.1 Spectrum Sensing

In order for opportunistic use of unused spectrum, a cognitive radio would need to continuously scan the spectrum and detect any holes, or white spaces (the unused spectrum), which may be used by cognitive radios and cognitive radio networks. However, scanning for holes in the spectrum is a process that consumes both time and power [21]. This makes it unsuitable for certain types of multimedia applications, especially

time-sensitive multimedia applications. The time spent in spectrum sensing can compromise the QoS [17] and could instead be used for data communication. If a cognitive radio node is operating in the cooperative mode, as described in the last section, the cost associated with dynamic spectrum access also increases.

Numerous methods are available for making dynamic spectrum access efficient with respect to time, power requirements, and cost. These include mechanisms [22] and efficient algorithms [23,24] for optimal detection of available spectrum, performing data transmission in parallel with spectrum sensing [25], having separate channels known to all cognitive radios that can provide spectrum sensing information [14, 21], or having an information registry containing authoritative data about primary users [19]. Employing these methods can enable more effective dynamic spectrum access, better utilization of the spectrum, and hence enable better QoS for multimedia applications.

1.6.2.2 Spectrum Management

The unused spectrum that has been sensed can possibly be spread over a wide frequency range, and will likely have different characteristics, such as operating frequency and available bandwidth. Spectrum management is related to selecting the frequency band(s) most suited to the QoS requirements of the end-user application. Additionally, spectrum management requires flexibility from the cognitive radio to support user-demanded services under different radio conditions [26], a task that may need to be performed in real time. However, this also implies that the node performing the spectrum management function be aware of the user QoS requirements. As we will discuss later, different mechanisms are available; for example, as policies for cognitive radios, which aid the task of spectrum management.

1.6.2.3 Spectrum Sharing

Spectrum sharing represents a significant challenge in the opportunistic usage of unused spectrum. This is due to the presence of licensed users, and also due to the presence of other opportunistic users vying for the same unused spectrum. The wide range of the available spectrum also contributes to the challenge.

The part of the spectrum on which a cognitive radio node would transmit after having sensed a hole would depend on the requirements and policies of the node. For example, a cognitive radio dedicated to bandwidth-intensive multimedia applications would want a large share of the unused spectrum. Due to the presence of other cognitive radio nodes, and to prevent collisions between nodes, access to the unused spectrum has to be coordinated to ensure fair sharing of the spectrum, a process that may be called *dynamic spectrum sharing* [17].

Different architectural principles and design techniques are used for this purpose, as summarized in [20]. One design approach is called *centralized spectrum sharing*, which takes the form of a centralized entity controlling access to the spectrum; for example, in the form of a Cognitive Resource Manager [27]. The centralized entity can also be in form of a universal control channel [4], such as a resource awareness channel [21] where information about resource usage is available to all cognitive radio nodes using the particular spectrum band. Centralized spectrum sharing allows more efficient spectrum utilization, and also takes the task of coordinating spectrum sharing away from the cognitive radio devices, thereby making the devices simpler and less expensive.

Design approaches for spectrum sharing also take the form of *cooperative* and *noncooperative* [15, 20] *spectrum sharing*. In the former approach, the cognitive radio nodes share information about resource/spectrum usage with each other, whereas in the latter approach, cognitive radio nodes operate selfishly. While the noncooperative mode can lead to less effective spectrum sharing [20], the communications overhead between the cognitive radio nodes in the cooperative mode can render it less effective for certain multimedia applications, since bandwidth and transmission time will be spent in coordination messaging. Table 1.6.1 represents the spectrum access mechanisms and the pros and cons for different parameters, as indicated by a+ or a–, respectively. For example, it can be seen from Table 1.6.1 that cooperative spectrum access is beneficial for sharing the unused spectrum, but requires a more capable, and hence expensive cognitive radio device. Furthermore, it may prove beneficial for QoS, for the case where through effective spectrum sharing the expectations of the user are fulfilled, or otherwise, where

TABl E 1.6.1 Dynamic Spectrum Access and Sharing

Access Mechanism	Spectrum Sharing	Device Cost	QoS	User Cost
Centralized	+	+	+	
Cooperative	+		+/	
Noncooperative		+	+/	+

excessive messaging overhead involved in sharing the spectrum might degrade the QoS. Given the more expensive device, and a more expensive method of spectrum sharing, such as one involving licensing the unused spectrum from a primary user, the effect on the cost to the user will be negative.

Spectrum sharing is also achieved by different transmission techniques, which can be classified as the *underlay*, *overlay* and *interweave* approaches [28]. The underlay approach allows concurrent primary and secondary transmissions, but the primary users are protected by enforcing a spectral mask on the secondary signals. This limits the underlay approach only for short-range communication. The overlay approach, also known as the *known-interference approach*, also allows concurrent primary and secondary transmissions; however in this approach, the primary and secondary transmissions cooperate with each other in a designated way. The interleave approach, or the *interference-avoidance approach*, corresponds to opportunistic sharing of the spectrum.

1.6.2.4 Spectrum Mobility

While spectrum management is associated with providing a cognitive radio with a spectrum best suited to its target applications' requirements, spectrum mobility adds to the challenges of spectrum management. Spectrum mobility comes into play when a cognitive radio node has to change the spectrum it is using during operation. This mobility can be vertical, which relates to using available spectrum from different wireless networks, or horizontal, in which case the cognitive radio changes its operating frequency within the same network type [29].

Spectrum mobility can be necessitated by various conditions. For example, the conditions of the channel that a cognitive radio might be using can deteriorate such that it does not remain possible to fulfill the QoS requirements of the applications being used on the cognitive radio. This would require the cognitive radio to adapt to these conditions, and shift to a part of spectrum more suited to the QoS requirements of the applications. In the same way, appearance of the primary user of the spectrum might necessitate the cognitive radio to vacate the channel, and shift to another available part of the spectrum. This shifting is referred to as *spectrum handoff* [29] with which numerous challenges are associated with respect to maintaining the QoS of the applications. These take the form of latency, delay, and loss of information during spectrum handoff, and can lead to performance degradation of the cognitive radio.

Furthermore, spectrum handoff might have to happen while applications are running on the cognitive radio node, such as live video being streamed to a user. Therefore robust mechanisms of spectrum handoff are required to minimize any performance degradation of the cognitive radio. Additionally, cognitive radio nodes and cognitive radio networks have to be capable of efficient spectrum handoff, which can be necessitated in time and space, the former due to change of availability of different channels over a period of time, and the latter based on the location [15] of the cognitive radio node, which may change while communication is in progress.

1.6.3 Cognitive Radio Devices

Today's wireless devices have progressed to a state where they can offer not only voice communications, but also handle high-rate multimedia traffic, enabling consumers to use services like high-quality audio and video [30]. However, these wireless devices are still not any different in function from plain old radios [6]. Therefore, the likely bottleneck in the provision of very high data rate multimedia services to consumers will be the lack of available spectrum.

This can be overcome by cognitive radios and cognitive radio networks. The path to cognitive radio for a wireless device goes by way of *software-defined radio*, which is a multiband radio capable of supporting multiple air-interfaces and protocols [4, 9, 13] by way of software reconfiguration supported by hardware design. Cognitive radios are built on software-defined radios, and with their cognitive capabilities, represent the intelligent, adaptive, and frequency agile wireless devices described in the previous section.

Just like today's wireless devices, cognitive radios can allow a variety of voice and data services [9]. The ability of cognitive radio devices to observe their environment and adapt themselves to the needs of the user is, however, a feature beyond the capability of current wireless devices. One example of this behavior could be a user's requirement for a certain QoS based on a particular application, such as streaming video. Issues that could affect the QoS, like insufficient bandwidth or user mobility in heterogeneous networks, could be overcome by the cognitive radio device as a stand-alone node or as part of a cognitive radio network.

The two main characteristics of cognitive radio devices can be described as *cognitive capability* and *reconfigurability* [20]. The cognitive capability enables the device to sense, analyze, and decide on the spectrum, as previously described. Reconfigurability is the ability of the cognitive radio device to intelligently adapt to the dynamic radio environment by changing its operating parameters, like operating frequency, modulation technique, or transmission power.

Cognitive radio devices are expected to operate over a wide range of frequencies, and within different geographical regions; therefore the devices have to be spectrum agile and policy agile [31]. Spectrum agility of a cognitive radio device depends on the cognitive capability of the device. Policy agility relates to the ability of the device to understand the constraints on its operation, and the rules driving its cognitive capability and reconfigurability. These rules may depend on the location of the device; for example, due to constraints on cognitive radio usage in a certain geographical area, or due to the type of radio network in an area, and also on the user of the device, who may want only a particular type of service with specific QoS and pricing constraints. Therefore a cognitive radio device needs the capability to understand and apply different policies, which may be defined by the cognitive radio network or by the end user, and should also have mechanisms to load the policies, either when it encounters them during the cognitive process, or when prompted to do so by the network or the user.

Cognitive radio devices are also required to be very sensitive in their operation in order to prevent the hidden node problem [32, 30] whereby a part of the spectrum is incorrectly sensed and opportunistically used as being unoccupied, when in fact a licensed user is using that spectrum, leading to unwanted and also potentially illegal interference.

The operations required for a cognitive radio's cognitive capability and reconfigurability are sufficiently complex to cause a drain on the device's battery power [9]. Coupled with the multimedia applications that might be running on cognitive radio devices, battery life assumes an important dimension for these devices. Therefore cognitive radio devices based on software-defined radios need to have energy-efficient designs [33].

1.6.4 Policies for Cognitive Radio Operation

An important aspect of cognitive radio operation is the exchange of information between the cognitive engine and the software-defined radio. Efficient exchange of information is necessary for flawless operation of the cognitive radio device. The information exchanged contains the desired behavior of the software-defined radio as determined by the cognition process. It is, however, important to note that this information remains independent of the implementation details of a particular software-defined radio. One method of achieving this is to represent the desired behavior of the software-defined radio in an intermediate format that could be translated to different software-defined radio implementations. Such an intermediate format can be defined using Radio XML, which can then be manipulated using open-source and robust parsing software.

XML files can be used to exchange information between the cognitive engine and the software-defined radio in a way that is independent of the software-defined radio platform [12]. Once the results of the cognitive process by the cognitive engine have been described in an XML file representing the desired behavior of the software-defined radio, an application programming interface (API) could be used to translate the desired behavior into commands specific to a particular software-defined radio platform being used by the cognitive radio device [34]. This allows different desired behaviors to be represented in different XML files, so that desired behavior is not bound to any particular software-defined radio platform, and by virtue of this, the cognitive engine and the software-defined radio platform remain independent of each other.

This convenience of describing the behavior of the software-defined radio in XML files comes at a cost, however. The ability to parse XML files and converting the behavior described therein into software-defined radio platform–specific commands requires time and resources. This can result in the whole process becoming a bottleneck [12], which can have negative repercussions for the QoS requirements of an application when real-time adjustments in the cognitive radio device behavior are needed. For time-sensitive multimedia applications, the negative effect just described is likely to be noticeable.

Cognitive radio devices will operate under varying radio conditions and in different geographical locations. This can affect the results of the cognition process of the cognitive radio devices, and hence the behavior of the software-defined radios within the devices. While it is beneficial to represent the results of the cognition process in XML files to keep the cognition engine and the software-defined radio loosely coupled, it is also desirable to encode the policies determining the cognition process and overall behavior of the cognitive radio device in a way that a cognitive radio device can operate based on different policies, as described earlier. This is referred to as *policy agility* [31] and the devices conforming to such a behavior can be termed *policy-agile* cognitive radio devices. An obvious advantage of policy agility is that cognitive radio devices can operate over a wide range of geographical locations, different operating conditions, and various QoS requirements without the policy determining device behavior being hard-coded into the device. This way, different desired behaviors of the cognitive radio device could be obtained by applying different policies to the device. Furthermore, more intelligent cognitive radio devices could load appropriate policies available to them, for example, from a known repository in the cognitive radio network, to constrain their operation.

Policy agility in cognitive radio devices can also be very useful for multimedia applications for these devices. Different policies targeting different types of multimedia applications can be developed, which may then be loaded by the cognitive radio device according to the requirements of the user. This could then constrain the operation of the cognitive radio device so that the best possible QoS is obtained for the particular multimedia application on the device. Furthermore, user expectations and requirements could also be made part of a policy in order to optimize the user experience. The cognitive radio device and the cognitive radio network can collaborate to deliver an optimized experience to the user based on the defined policy.

Another advantage of policy agility is the decoupling of policy definition, loading, and enforcement from device-specific implementations and optimizations [35]. This allows cognitive radio devices and policies to evolve independently, and adds to the flexibility of the devices by enabling adaptation to their behavior by changing the declarative policies. A change in policy does not require a change in cognitive radio device hardware or software. This adds versatility to cognitive radio devices.

A major issue with regard to policy languages for cognitive radio devices remains the development of a common ontology so that the policies, including especially user-defined policies, are portable over a wide range of cognitive radio devices. Unitl that happens, policy-based solutions will likely remain isolated efforts. This is also necessary to make policies extensible and applicable to different domains. One of these areas could be policies geared toward multimedia applications for cognitive radio devices and networks, for example, to express user expectations or for expressing parameters that the cognitive radio device should adapt to in order to provide an acceptable QoS for multimedia applications.

1.6.5 Quality of Service (QoS)

Quality of service is a concept relative to the context in which it is defined. With respect to multimedia applications for cognitive radio networks, quality of service can represent an experience that is acceptable to the user. However, the user's experience is limited to the interaction with the cognitive radio device. Therefore the interaction between the device and the cognitive radio network with respect to fulfilling an expected user experience is of paramount importance. This interaction involves various aspects of dynamic spectrum access as discussed earlier.

Not only are technical parameters of cognitive radio operation important in order to achieve user satisfaction, but the cost to the user is important as well. Some users, depending on the application they want to use, might be willing to compromise on the user experience in return for reduced cost, whereas others may pay more if it guarantees them the expected user experience. Therefore, QoS cannot be treated as a stand-alone aspect, but instead is an interplay among the requirements of the user, the capabilities of the device, and the flexibility offered by the network.

It is also possible to represent the QoS as the performance level of a service offered by the network to the user [36]. A network can offer different types of service to users, where a service may be characterized by measurable, prespecified service requirements. These requirements will depend on the type of application, and may be called, for multimedia applications, *bandwidth, transmission delay, jitter,* and *error rate* [36,26]. For multimedia applications, bandwidth, jitter, and delay are key QoS requirements. However, all of these are network-centric requirements.

At the same time, QoS for cognitive radio devices is influenced by resource constraints of the nodes, some of which are battery charge and processing power. If the cognition process consumes too much power, the device battery or power source may be drained while a multimedia application is running, thereby negatively impacting the user experience.

The requirements of the user, reflected in the applications that the user wants on the cognitive radio device, determine the QoS requirements as well. For example, bandwidth, delay, and jitter are the main requirements for streaming video, which do not matter much for simple web browsing or e-mail downloading, or newer applications like podcasting targeted for wireless networks [37]. Fulfilling QoS requirements ensures that the requirements of the user are fulfilled. However, a guarantee of this fulfillment can come at a price. If a user is not willing to pay that price, but instead can accept a lower QoS at a lower price, that option should also be made available to the user. Therefore economic factors also influence the QoS related to particular applications requested by the user. Also, any application that is bandwidth intense will also directly affect the network infrastructure, since bandwidth, and hence data rate, represents a QoS parameter that affects the infrastructure foremost [9]. Opportunistically claiming a large portion of the bandwidth will be expensive, both with respect to the price, and with respect to the resources consumed in the cognitive process and the resource usage in the cognitive radio device.

1.6.6 Pricing Schemes for Multimedia Applications

As we have already discussed, radio spectrum is a scarce and valuable resource. The economics of dynamic spectrum access, however, take on different characteristics depending on the type of cognitive radio network and dynamic spectrum access, as discussed in Sections 1.6.1 and 1.6.2. This will mean, for example, that users having a cognitive radio device operating in a noncooperative spectrum sharing mode in a decentralized cognitive radio network will experience a different pricing regime for the same multimedia applications as compared to users whose cognitive radio devices access unused spectrum under the coordination of a central controlling entity.

At the same time, the link between pricing and QoS is also extremely important, as discussed in this section. Users wanting a higher QoS will be expected to pay more than those users who opt for lower-priced services, where the QoS will also be lower. It has also to be kept in view, however, that a user who is willing to pay more for a higher QoS will do so depending on the application, for example, live video

streaming as opposed to downloading nonemergency e-mail. Therefore the application a user chooses will also influence the pricing regime.

It is important to consider user satisfaction in terms of QoS and pricing from a user-centric point of view, but also from a service provider or network operator–centric point of view [38]. But the paradigm shift from operator-centric services to user-centric services is undeniable [39] for cognitive radio devices and networks, whereby these devices and networks intelligently adapt to provide applications to the users at a price and quality determined by the users themselves.

Services for users of cognitive radio networks can be made cheaper in the case of unlicensed, dynamic access to unused spectrum [40]. Since there is no spectrum licensing cost involved, more users for the services can be expected, which allows for more diverse services and applications. Furthermore, this means that if any service provider becomes unavailable for any reason, there will be some other provider for the same service, since more users encourage more providers with the same services competing for users. However, in the absence of coordination among cognitive radio nodes, as will be the case in noncooperative spectrum sharing, the cognitive radio device will have to manage spectrum sensing, management, sharing, and mobility all by itself. Such cognitive radio devices with enhanced capabilities will be expensive, and the cost to the consumer will increase. Additionally, the cost to the service provider can also be expected to increase, since it will have to contend with uncoordinated spectrum users. This cost can, in turn, be transferred to the users.

While more capable cognitive radio devices are expected to cost more for consumers, their enhanced capabilities are likely to be beneficial to the user in terms of other aspects of QoS. For example, such a device can be expected to adapt its operation for a given QoS in such a way that the drain on the battery life of the device is reduced. This will result in an enhanced experience for the user. A more capable cognitive radio device can also be expected to dynamically adjust its spectral bandwidth, which can be beneficial for delay-sensitive, bandwidth-intense multimedia applications.

Access to unused spectrum can be made more efficient if the unused spectrum is licensed for short-term usage [40,13,16]. The short-term licenses may be requested by a cognitive radio network provider, or by the end user possessing a capable cognitive radio device. Such a scheme would involve a central coordinator in the form of a spectrum manager [16] or a spectrum broker [13]. Even though schemes like these would lead to more efficient use of the spectrum, and potentially more stable QoS characteristics for end users, the cost of the services by the provider will increase due to the licensing cost incurred by the service provider. This increase in cost to the end user is likely for all such scenarios in which a central entity coordinates the cognitive radio network or cognitive radio devices, as mentioned in Section 1.6.3. On the other hand, if there is a central entity that controls access to the resources in the cognitive radio network, a capable cognitive radio device can purchase resources from the entity [9], thereby enabling itself to present a higher QoS to the user, albeit at a higher price.

The role of the consumer, or the end user, is also very important in any pricing scenario. Effective communication of the requirements of the user to the cognitive radio device is essential for the device to meet the requirements of the user, either by itself or through coordination within a cognitive radio network. Once the requirements of the user, in terms of the required application and the QoS expectations associated with that application are known to the cognitive radio device, it can, as part of the cognitive process, find the best service for the best price. In this respect, the cognitive radio device may have to find a compromise between the user's requirements and the resources available within the cognitive radio network, whether it operates in the cooperative or noncooperative mode. Another possible scheme is the cognitive radio device presenting different options with respect to price and QoS to the user, and the user making a choice. This scheme can suit less capable, and hence less expensive, cognitive radio devices.

Based on the discussion in the previous paragraphs, it is possible to categorize pricing schemes in three broad categories, as reflected from a user-centric view. These categories can be called *fixed price* schemes, *contingent price* schemes, and *dynamic pricing* schemes [9]. In fixed price schemes, the price for the services and the associated QoS are nonflexible. In contingent price schemes, the user can indicate

TABl E 1.6.2 Pricing Schemes and Their Relevance for the QoS, Cost, and Flexibility for the User and Service Rrovider

			Flexibility	
Pricing Scheme	QoS	Cost	User	Service Provider
Fixed	Fixed	Fixed		–
Contingent	User selected	Variable, depends on user requirements	+	–
Dynamic	User selected depends on network conditions	Variable, depends on network conditions	+	+

the type of service it requires (e.g., video streaming or web browsing), and can be charged according to the requirement. However, while there is flexibility for the user, there is no flexibility for the service provider because the requested service and the associated QoS are nonflexible. On the other hand, in a dynamic pricing scheme the services, and hence the associated price, can be categorized with respect to the QoS and priority for the user. Pricing can be tied to network conditions, and based on these, different options can be provided to the user. In a dynamic pricing scheme, there is flexibility for the user as well as for the service provider, as depicted in Table 1.6.2.

1.6.7 Summary

Ordinary consumers will provide a big segment of cognitive radio users, and within this segment, multimedia applications will be the most widely requested over the cognitive radio network and used over the cognitive radio device. In order to meet users' expectations with respect to the applications requested by them, the QoS requirements relevant for a particular application have to be fulfilled. This requires efficient dynamic spectrum access, such that sufficient bandwidth is opportunistically claimed for the use of multimedia applications, as well as efficient coordination between cognitive radio nodes in a cognitive radio network, so that there is no collision between different nodes. Further development of cognitive radio devices will also need to be geared toward ordinary consumers, allowing them access to a wide range of multimedia applications. For this, the cognitive radio devices would need enhancement of their cognitive capability and flexibility of operation, so that these devices can not only tap into unused spectrum resources available at different locations and in different networks, but also allow intelligent adaptation of the devices' operation based on network or user policies. Success of multimedia applications for cognitive radio networks will also depend on the cost of the applications to the users. Different pricing schemes, based on conditions in the cognitive radio network, or the demands of the user, or a combination of both, will have to be put in place. Even though opportunistic usage of unused spectrum is technically more challenging, benefits to the users in terms of more choice of multimedia applications at a reduced cost will make multimedia applications for cognitive radio networks more acceptable and prevalent.

References

[1] P. Mahonen, Cognitive trends in making: Future of networks. In *Proc. Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'04)*, pp. 1449–1454, September 2004.

[2] S. Dhawan, Analogy of promising wireless technologies on different frequencies: Bluetooth, wifi, and wimax. In *Proc. 2nd International Conference on Wireless Broadband and Ultra Wideband Communications (AusWireless 2007),* August 2007.

[3] C. Cordeiro, K. Challapali, D. Birru, and S. Shankar, IEEE 802.22: The first worldwide wireless standard based on cognitive radios. In *Proc. IEEE First International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN'05),* pp. 328–337, November 2005.

[4] J. Mitola III, Cognitive radio for flexible mobile multimedia communications. In *Proc. IEEE International Workshop on Mobile Multimedia Communications (MoMuC'99),* pp. 3–10, November 1999.

[5] R. W. Thomas, D. H. Friend, L. A. DaSilva, and A. B. MacKenzie, Cognitive networks: Adaptation and learning to achieve end-to-end performance objectives. *IEEE Communications Magazine* 44, no. 12 (2006): 51–57.

[6] R. Rubenstein, Radios get smart. *IEEE Spectrum Online* (February 2007). http://www.spectrum.ieee.org/feb07/4892.

[7] *Cognitive radio technology: A study for Ofcom—volume 1*. Technical Report QINETIQ/06/00420 Issue 1.1, QinetiQ Ltd., February 12, 2007.

[8] J. L. Burbank and W. T. Kasch, Cross-layer design for military networks. In Proc. *IEEE Military Communications Conference (MILCOM 2005)*, pp. 1912–1918, October 2005.

[9] S. Ball, A. Ferguson, and T. W. Rondeau, Consumer applications of cognitive radio defined networks. In *Proc. IEEE First International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN'05),* pp. 518–525, November 2005.

[10] K. E. Nolan and L. E. Doyle, Teamwork and collaboration in cognitive wireless networks. *IEEE Wireless Communications* 14, no. 4 (2007): 22–27.

[11] S. N. Shankar, Squeezing the most out of cognitive radio: A joint mac/phy perspective. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, pp. IV–1361–IV–1364, April 2007. ISBN 1-4244-0728-1.

[12] T. W. Rondeau, B. Le, D. Maldonado, D. Scaperoth, and C. W. Bostian, Cognitive radio formulation and implementation. In *Proc. First International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM 2006),* June 2006. ISBN 1-4244-0381-2. Digital Object Identifier: http://dx.doi.org/10.1109/CROWNCOM.2006.363476.

[13] M. Nekovee, Dynamic spectrum access with cognitive radios: Future architectures and research challenges. In *Proc. First International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWN-COM),* June 2006. ISBN 1-4244-0381-2.

[14] H. Arslan and M. E. Sahin, System design for cognitive radio communications. In *Proc. First International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM 2006),* June 2006. ISBN 1-4244-0381-2. Digital Object Identifier: http://dx.doi.org/10.1109/CROWNCOM.2006.363451.

[15] H. Celebi and H. Arslan, Utilization of location information in cognitive wireless networks. *IEEE Wireless Communications* 14, no. 4 (2007): 6–13.

[16] V. Rodriguez, K. Moessner, and R. Tafazolli, Market driven dynamic spectrum allocation over space and time among radio-access networks: Dvb-t and b3g cdma with heterogeneous terminals. *Mob. Netw. Appl.* 11, no. 6 (2006): 847–860, ISSN 1383-469X. Digital Object Identifier: http://dx.doi.org/10.1007/s11036-006-0053-2.

[17] C. Cordeiro, K. Challapali, and M. Ghosh, Cognitive PHY and MAC layers for dynamic spectrum access and sharing of TV bands. In *TAPAS '06: Proceedings of the First International Workshop on Technology and Policy for Accessing Spectrum*, p. 3, New York: ACM Press. ISBN 1-59593-510-X. Digital Object Identifier: http://doi.acm.org/10.1145/1234388.1234391.

[18] N. Devroye, P. Mitran, and V. Tarokh, Cognitive decomposition of wireless networks. In *Proc. First International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWN-COM),* June 2006. ISBN 1-4244-0381-2.

[19] J. M. Chapin and W. H. Lehr, The path to market success for dynamic spectrum access technology. *IEEE Communications Magazine* 45, no. 5 (2007): 96–103.

[20] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer Networks* 50, no. 13 (2006): 2127–2159. ISSN 1389-1286. Doi: http://dx.doi.org/10.1016/j.comnet.2006.05.001.

[21] O. Holland, A. Attar, N. Olaziregi, N. Sattari, and A. H. Aghvami, A universal resource awareness channel for cognitive radio. In *Proc. 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '06),* September 2006. ISBN 1-4244-0330-8.

[22] M. Ghozzi, M. Dohler, F. Marx, and J. Palicot, Cognitive radio: Methods for the detection of free bands. *C.R. Physique* 7, no. 7 (2006): 794–804. http://dx.doi.org/10.1016/j.crhy.2006.07.009.

[23] H. Sun, J. Jiang, and M. Lin, Adaptive cooperation algorithm for cognitive radio networks. In *Proc. International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2006),* September 2006. ISBN 1-4244-0517-3.

[24] T. Yucek and H. Arslan, Spectrum characterization for opportunistic cognitive radio systems. In *Proc. Military Communications Conference (MILCOM 2006),* October 2006. ISBN 1-4244-0618-8.

[25] W. Hu, D. Willkomm, G. Vlantis, M. Gerla, and A. Wolisz, Dynamic frequency hopping communities for efficient IEEE 802.2 operation. *IEEE Communications Magazine* 45, no. 5 (2007): 80–87. ISSN 0163-6804. Digital Object Identifier: http://dx.doi.org/10.1109/MCOM.2007.358853.

[26] S. Nagel, V. Blaschke, and F. K. Jondral, Mechanisms for the adaptation of the physical layer in a cognitive radio. In *9th European Conference on Wireless Technology*, pages 43–46. IEEE, 2006. ISBN 2-9600551-5-2. Digital Object Identifier: http://dx.doi.org/10.1109/ECWT.2006.280430.

[27] P. Mahonen, M. Petrova, J. Riihijarvi, and M. Wellens, Cognitive wireless networks: Your network just became a teenager (poster). In *Proc. 25th Conference on Computer Communications (INFOCOM 2006)*, April 2006.

[28] S. Srinivasa and S. A. Jafar, The throughput potential of cognitive radio: A theoretical perspective. *IEEE Communications Magazine* 45, no. 5 (2007): 73–79. ISSN 0163-6804. Digital Object Identifier: http://dx.doi.org/10.1109/MCOM.2007.358852.

[29] X. Fu, W. Zhou, J. Xu, and J. Song, Extended mobility management challenges over cellular networks combined with cognitive radio by using multi-hop network. In *Eighth ACIS International Conference on Software Engine Ring, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, pp. 683–688. IEEE, 2007. ISBN 978-0-7695-2909-7. Doi: http://dx.doi.org/10.1109/SNPD.2007.280.

[30] W. Krenik and A. Batra, Cognitive radio techniques for wide area networks. In DAC '05: *Proceedings of the 42nd annual conference on Design automation*, pp. 409–412, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-058-2. Digital Object Identifier: http://doi.acm.org/10.1145/1065579.1065688.

[31] C. Santivanez, R. Ramanathan, C. Partridge, R. Krishnan, M. Condell, and S. Polit, Opportunistic spectrum access: Challenges, architecture, protocols. In *WICON '06: Proceedings of the 2nd Annual International Workshop on Wireless Internet*, p. 13, New York: ACM Press. ISBN 1-59593-510-X. Digital Object Identifier: http://doi.acm.org/10.1145/1234161.1234174.

[32] D. Raychaudhuri, N. B. Mandayam, J. B. Evans, B. J. Ewy, S. Seshan, and P. Steenkiste, Cognet: An architectural foundation for experimental cognitive radio networks within the future internet. In *MobiArch '06: Proceedings of First ACM/IEEE International Workshop on Mobility in the Evolving Internet Architecture*, pp. 11–16, New York, ACM Press, 2006. ISBN 1-59593-566-5. Digital Object Identifier: http://doi.acm.org/10.1145/1186699.1186707.

[33] B. Bougard, D. Novo, F. Naessens, L. Hollevoet, T. Schuster, M. Glassee, A. Dejonghe, and L. Van der Perre, A scalable baseband platform for energy-efficient- active software-defined-radio. In *Proc. First International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM 2006),* June 2006. ISBN 1-4244-0381-2. Digital Object Identifier: http://dx.doi.org/10.1109/CROWNCOM.2006.363450.

[34] D. Scaperoth, B. Le, T. Rondeau, D. Maldonado, C. W. Bostian, and S. Harrison, Cognitive radio platform development for interoperability. In *Proc. Military Communications Conference (MILCOM 2006)*, October 2006. ISBN 1-4244-0618-8.

[35] D. Wilkins, G. Denker, M.-O. Stehr, D. Elenius, R. Senanayake, and C. Talcott, Policy-based cognitive radios. *IEE Wireless Communications* 14, no. 4 (2007): 41–46. XML. Extensible markup language (xml), 2007. URL: http://www.w3.org/XML/.

[36] T. B. Reddy, I. Karthigeyan, B. S. Manoj, and C. S. R. Murthy, Quality of service provisioning in ad hoc wireless networks: A survey of issues and solutions. *Ad Hoc Networks* 4, no. 1 (2006): 83–124. ISSN 1570-8705. Digital Object Identifier: http://dx.doi.org/10.1016/j.adhoc.2004.04.008.

[37] M. May, V. Lenders, G. Karlsson, and C. Wacha, Wireless opportunistic podcasting: implementation and design tradeoffs. In *CHANTS '07: Proceedings of the Second Workshop on Challenged Networks*, pp. 75–82, New York, ACM Press. ISBN 978-1-59593-737-7. Digital Object Identifier: http://doi.acm.org/10.1145/1287791.1287806.

[38] L. Badia, M. Lindström, J. Zanderb, and M. Zorzi, An economic model for the radio resource management in multimedia wireless systems. *Computer Communications* 27, no. 11 (2004): 1056–1064. ISSN 0140-3664. Digital Object Identifier: http://dx.doi.org/10.1016/j.comcom.2004.01.011.

[39] M. Kuroda, K. Ishizu, H. Harada, and R. Komiva, A study of radio-information services for networks of cognitive radios. In *2nd IEEE Workshop on Networking Technologies for Software Define Radio Networks*, pp. 32–39, IEEE, 2007. ISBN 1-4244-1315-X. Digital Object Identifier: http://dx.doi.org/10.1109/SDRN.2007.4348971.

[40] T. X. Brown and D. C. Sicker, Can cognitive radio support broadband wireless access? In *2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks. DySPAN 2007*, pp. 123–132, April 2007. ISBN 1-4244-0663-3.

Summary

Patricia Morreale

This chapter presented the current trends in voice and data networks regarding voice services; the evolution from the connection-oriented networks to the connectionless networks, and unification between voice and data networks.

CTI is a technology platform that merges voice and data services at the functional level to add tangible benefits to business applications. In fact, CTI is a new application for preexisting technologies. The most important benefit CTI brings to the corporate world is the potential to reduce operating expenses.

The low cost of the public Internet resulted in a great increase of data traffic in the last couple of years. Besides this, the quality of service (QoS) has been considerably improved and offers acceptable delays for traffic less tolerant to delays. This raised a considerable interest in transporting packetized voice (not tolerant to delays) over data, in particular, IP networks. The support for voice communications over the Internet Protocol is usually called Voice over IP (VoIP).

Local area networks don't lose their importance for users independently of how much bandwidth is available in the wide area. They serve as local connectors between end-user devices, server farms, storage farms, and the ingress/engress nodes of wide area networks. Their limitation to data is changing by supporting multimedia applications up to the desktop. Step by step, throughput capabilities could be improved; in this respect, Giga Ethernet seems to be the winner for high-speed LAN technology.

Advancements in communications continue to occur. Some of the recent changes are best exemplified by RFID, wireless sensor networks, and multimedia applications. These emerging technologies build on the earlier foundation and offer a glimpse of the future.

2

Intranets

Introduction **2**-2

2.1 Internet and Intranet Management Concepts **2**-2
Management Overview of the Internet and Intranets • Intranet Planning and Management • Technical Overview • Intranet Components • Intranet Implementation • Intranet Deployment • Intranet Security Issues • Internet Security • Summary

2.2 Virtual Private Networking Solutions **2**-30
Layer 2 Protocols • Layer 3 Tunneling Protocols • Frame Relay • Layer 2 or Layer 3 Comparison

2.3 Web-Enabled Data Warehousing **2**-38
Introduction • Data Warehousing Overview • Web-Enabled Data Warehousing • Vendors • Future Trends

2.4 Web Performance Management **2**-46
Internet, Intranets, and Extranets • Generic Intranet Management Challenges • Specific Challenges to Intranet Performance Management • Wire Monitors and Network Analyzers • Application Performance Management • Web Performance Management Trends

2.5 Application Performance Management **2**-81
Introduction • The Need for APM in Communication Services • Application Performance Management Overview • Performance Management for Next Generation Service Delivery Platforms • Performance Management for Next Generation OSS/BSS and SDP: The Solution from CA Wily Technology • Summary and Trends

2.6 Electronic Technologies **2**-99
E-Commerce Technologies • Web Service Delivery Challenges • Management of Web Services

2.7 Internet Protocols **2**-105
Addressing for Internet • Communication Protocols in Internet • Information Transfer in Internet • Types of Internet Access • Internet E-Mail • Telnet in Internet—STD8 • File Transfer in Internet (FTP, RFC 959, RFC 990—Port number) • News and Usenet (NNTP, RFC 977) • Mailing Lists in Internet • Information Search in Internet • Netscape and Microsoft

2.8 Role of Open Source Software **2**-115
Introduction: The Nature of Open Source • Commercial and Noncommercial Applications and Tools • Opportunities and Vulnerabilities • Licensing • Cost Considerations (CapEx, OpEx) • Summary and Trends

Summary and Trends **2**-125

Teresa Piliouras
TCR, Inc.

John Braun
Industry Consultant

Patricia Morreale
Kean University

Endre Sara
Goldman, Sachs & Co.

Kornel Terplan
Industry Consultant and Professor

Dermot Murray
Iona College

Mihir Parikh
Polytechnic University

Vadim Rosenberg
CA-Wily

Tivadar Szemethy
Netvisor

Introduction

The Internet started as a technological revolution, designed to protect national interests by ensuring redundancy and resiliency in governmental networks, particularly in time of war. It has spawned a worldwide cultural revolution, fostering universal communication exchange with limitless geographic, time, and subject matter boundaries. The extent and ease of the Internet's adoption has had profound implications on all aspects of life—personal, business, and governmental. There is no place on Earth that cannot be reached by the Internet.

In this chapter, we review the basic technological underpinnings of the Internet and discuss why it is so flexible. As we explore the evolution of the Internet, which continues at an ever-increasing pace, we also examine corresponding effects on communication paradigms, particularly in a business context. Now, businesses small and large alike have joined the Internet bandwagon and have a Web presence.

The fear, uncertainty, and doubt that have surrounded open source since its inception seem to be dissipating. In order to promote technological innovation while controlling capital expenditures and operating expenses, service providers are increasingly considering the deployment of open source–based solutions. After open source products are positioned into commercial and noncommercial applications, the focus turns to opportunities available with open source. Enterprises are becoming increasingly interested in accessing the code, creating derivative works, and redistributing those works. Various areas, such as Web server management, database management, middleware, e-mail management, content filtering and management, virtualization, network management, and business intelligence, are addressed and evaluated. Particular attention is paid to the technical and legal risks associated with open source.

Web-based solutions are penetrating both telecommunications providers and enterprise infrastructures. The management of these kinds of networks is examined. Sensitive components that may cause congestion or bottlenecks are addressed in detail. Special emphasis is placed on log file analysis, wire monitoring, application performance measurements, load balancers to conserve bandwidth, and administration of Web server farms. Product examples are included for each area. In most cases, Web content is expected to drive decisions about resource facilities and equipment reservation/allocation.

Service fulfillment, service assurance, and billing support system performance levels are critical success factors for service providers. Application performance management technologies and tools are addressed in a special segment in which Introscope from Wily Technologies is used as an example. If these technologies and tools are properly implemented, service quality and user satisfaction can be improved significantly.

2.1 Internet and Intranet Management Concepts

Teresa Piliouras and John Braun

2.1.1 Management Overview of the Internet and Intranets

An Intranet is a company-specific, private network based on Internet technology, and as such, it is a form of local area network (LAN). However, one of the major distinctions between traditional LANs and Intranets is the reliance of the latter on TCP/IP, packet switching, and Internet technologies. In the case of the Internet, the technology is deployed over a public network, while in the case of Intranets, the technology is deployed within a private network.

One of the important benefits of Intranets is that they provide a cost-effective vehicle for communication, since the expense of reaching one person or one million people is essentially the same. Intranets are becoming the corporate world's equivalent of a town hall where people can meet, chat, and exchange information.

The emergence of Intranets promises to change the way companies communicate with their employees and how they conduct their business. For example, after years of using satellite feeds to disseminate information to its 208 network affiliates, CBS News now uses an Intranet to provide affiliates with point-and-click access to information on upcoming news stories. Access to this information is provided through the CBS Newspath World Wide Web home page.

2.1.1.1 Benefits of Intranets

Intranets offer many potential benefits, including:

- Reduced operating costs
- Improved employee productivity
- Streamlined processing flows
- Improved internal and external communication
- New and improved customer service
- Cross-platform capability

We will discuss some of the ways these benefits can be achieved.

2.1.1.1.1 The Paper-less Office

Many companies find that Intranets simplify corporatewide communications and reduce printed material costs by eliminating the need for many paper-based processes. For example, some organizations offer complete manuals on their corporate Web site in electronic form, instead of distributing the information in printed form. Companies can benefit immediately from an Intranet by replacing their printed materials, little by little, with electronic versions. Electronic media is cheaper to produce, update, and distribute than printed material. Oftentimes, printed material is out of date by the time it is distributed. Electronic documents, however, can be easily modified and updated as the need arises.

2.1.1.1.2 Improved Customer Service

For many organizations, having the right information at the right time can make a significant difference in their ability to close a sale or meet a deadline. In today's competitive business environment, companies are also under constant pressure to improve productivity while reducing costs. To achieve these productivity gains, companies must constantly improve their relationships with employees, customers, vendors, and suppliers. Intranets provide an important avenue for making these advancements.

Using an Intranet, vendors, employees, and customers can access information as it is needed, thus alleviating delays associated with mailing or distributing printed materials. For example, Intranets have been used to:

- Distribute software updates to customers
- Process and respond to customer inquiries and questions about products and services
- Collect customer and survey data

Using an Intranet, all these activities can be completed electronically in a matter of minutes.

2.1.1.1.3 Improved Help Desks

Intranets have been used to augment help desk services. For example, when someone in the organization learns about a new technology or how to perform a new task (for example, running virus software), he/she can put information and instructions for others on a personal Web page. Others within the organization, including help desk staff, can then access this information as needed. In an organization empowered by an Intranet, all employees can leave the imprints of their expertise.

2.1.1.1.4 Improved Corporate Culture

Intranets help to cultivate a corporate culture that encourages the free flow of information. Intranets place information directly into the hands of employees, promoting a more democratic company structure. The danger of "information democracy" is that once it is in place and taken for granted, management cannot easily revert to older, more controlled forms of communication without seriously damaging employee morale and cooperation. Every individual in an Intranet environment is empowered to access and distribute information, both good and bad, on a scale heretofore unknown in the corporate realm.

Intranets dissolve barriers to communication created by departmental walls, geographical location, and decentralized organizations. Placing information directly in the hands of those who need it allows organizations to decentralize and flatten decision making and organizational processes, while maintaining control over the information exchange. Individuals and groups can distribute ideas freely, without having to observe traditional channels of information (i.e., an individual, a printed document, etc.) that are far less effective in reaching geographically dispersed individuals.

2.1.1.1.5 Cross-Platform Compatibility

Since the early 1980s, organizations with private networks have struggled with connecting and disseminating information between different types of computers—such as PCs, Macintoshes, and Unix-based machines. To help manage potential barriers to electronic communication posed by hardware and software incompatibilities, many companies have instituted strict standards limiting corporate users to specific hardware and software platforms. Even today, if a company uses PCs, Macs, and Unix-based machines, sharing a simple text document can be a challenge.

Intranets provide a means to overcome many of these software and hardware incompatibilities, since Internet technologies (such as TCP/IP) are platform independent. Thus, companies using Intranets no longer need to settle on one operating system, since users working with Macintosh, PC, or Unix-based computers can freely share and distribute information. In the sections that follow, we will explain why this is so.

2.1.2 Intranet Planning and Management

To implement an Intranet, a company needs a dedicated Web server, communications links to the Intranet, and browser software. *Unfortunately, Intranets do not come prepackaged and fully assembled.* They require careful planning and construction if they are to be effective in meeting the needs of the organization. In the sections that follow, we discuss recommendations for planning and implementing an Intranet.

2.1.2.1 Gaining Support

The first step toward a successful Intranet implementation is to obtain companywide support for the project, including endorsement from upper management. A quality presentation should be made to both management and staff to explain the benefits of the Intranet project. Some of these are tangible and easy to measure, while others are intangible and difficult to measure. To gain widespread support for the Intranet project, decision makers must be shown what an Intranet is and how it will benefit the organization. There are many resources (including complete presentations) available on the World Wide Web to help promote the Intranet in a corporate environment.

2.1.2.2 Planning the Intranet Strategy

After selling upper management on the idea of an Intranet, the next step is to define the goals, purpose, and objectives for the Intranet. This is an essential part of the Intranet project planning.

The Intranet project plan should include an overview of the organizational structure and its technical capabilities. The current communication model used to control information flows within the organization

should be examined with respect to its strengths and weaknesses in supporting workflow processes, document management, training needs, and other key business requirements. It is important to understand and document existing systems within the organization before implementing the Intranet.

The Intranet plan should clearly define the business objectives to be achieved. The objectives should reflect the needs of the Intranet's potential users. Conducting interviews with employees and managers can help identify these needs. For example, the human resource department may wish to use the Intranet to display job opportunities available within the organization. If this need is to be satisfied, the Intranet should be designed to display job information and job application forms on a Web server, so applicants can apply for positions electronically. The human resource department might also wish to offer employees the ability to change their 401K information by using the Intranet. Each identified goal shapes and defines the functionality that the Intranet must support. An employee survey is also an excellent way to collect ideas on how to employ the Intranet within the organization.

In summary, the following questions are helpful in defining the requirements of the Intranet project:

- Will Intranet users need to access existing (legacy) databases?
- What type of training and support will Intranet users require?
- Who will manage, create, and update the content made available through the Intranet?
- Will individual departments create their own Web pages autonomously?
- Will there be a central authority that manages changes to the content offered on the Intranet?
- Do users need remote access to the Intranet?
- Will the Intranet need to restrict access to certain users and content?
- Will a Webmaster or a team of technicians/managers be assigned to coordinate and manage the maintenance of the Intranet?
- Will the Intranet be managed internally or will it be outsourced?

2.1.2.3 Selecting the Implementation Team

After the Intranet project plan has been developed and approved, the implementation team should be assembled. If the organization does not have an infrastructure in place that is capable of implementing the Intranet, additional staff and resources will need to be hired or the project will need to be outsourced to a qualified vendor.

It is important that the Intranet team has the requisite skills to successfully execute the project plan. A number of skills assessment checklists are provided below to help evaluate the resources available within an organization and their abilities to successfully support the Intranet implementation.

2.1.2.3.1 Technical Support Skills Checklist

The Intranet project will require staff with the technical skills needed to solve network problems, understand network design, troubleshoot hardware and software compatibility problems, and implement client–server solutions (such as integrating network databases). Thus, the following skills are required to support an Intranet:

- Knowledge of network hardware and software
- Understanding of TCP/IP and related protocols
- Experience implementing network security
- Awareness of client–server operations
- Practice with custom programming
- Abilities of database management

2.1.2.3.2 Content Development and Design Checklist

A typical organization has many sources of information: human resource manuals, corporate statements, telephone directories, departmental information, work instructions, procedures, employee records, and much more. To simplify the collection of information that will be made available through

the Intranet, it is advisable to involve people familiar with the original documentation and also those who can author content for Intranet Web pages. If possible, the original authors of the printed material should work closely with the Intranet content developers to ensure that nothing is lost in translation.

The following technical skills are needed to organize and present information (content) in browser-readable format:

- Experience in graphic design and content presentation
- Basic understanding of copyright law
- Knowledge of document conversion techniques (to convert spreadsheet data, for example, into a text document for HTML editing)
- Experience in page layout and design
- Experience with Web browsers and HTML document creation
- Knowledge of image-conversion techniques and related software
- Knowledge of programming languages and programming skills
- CGI programming and server interaction

2.1.2.3.3 Management Support Skills Checklist

As previously discussed, the company's management should be involved in the planning and implementation of the Intranet. Ideally, management should have a good understanding of the Intranet benefits, and the expected costs and time frames needed for the project completion. Managers with skills relating to quality-control techniques, process-management approaches, and effective communication are highly desirable. Thus, the following management skills are recommended:

- Understanding of the organization's document flow
- Experience with the reengineering process
- Knowledge of quality-control techniques
- Knowledge of the company's informal flow of information
- Experience with training and project coordination

2.1.2.4 Funding Growth

The initial cost of setting up a simple Intranet is often quite low and may not require top management's approval. However, when complex document management systems are needed to integrate database access, automate workflow systems, implement interactive training, and other advanced features, the Intranet should be funded with the approval of top management. To gain approval for the project, upper management must be convinced that the Intranet is an integral part of the company's total information–technology deployment strategy. This involves quantifying the tangible benefits of the Intranet to the organization. Management also needs to understand how the Intranet will change the way people work and communicate.

2.1.2.5 Total Quality Management (TQM)

Effective deployment of an Intranet often involves reengineering current process flows within the organization. Employees are usually most receptive to changes that make their jobs easier. To avoid perceptions that the Intranet is an intimidating intrusion of yet another technology, it is advisable to involve staff as early on as possible in the deployment planning. This will facilitate the transition to the Intranet, and encourage employee participation in the Intranet's success.

After migrating the company's work processes to the Intranet, it is up to managers and employees to adhere to the procedures that have been put in place to improve productivity and teamwork. Management should not assume that because employees have a new tool—the Intranet—that this alone is sufficient to ensure that the desired attitudes and service levels will be attained. Instead, managers should view the Intranet as one aspect of their quest for Total Quality Management (TQM).

TQM involves creating systems and workflows that promote superior products and services. TQM also involves instilling a respect for quality throughout the organization. TQM and the successful

deployment of Intranets represent a large-scale organizational commitment, which upper management must support.

2.1.2.6 Training Employees

If employees are expected to contribute content to the Intranet, they will need to be given tools and training so they can author HTML and XML documents. In general, it is a good idea to encourage employees to contribute to the content on display through the Intranet. To do otherwise means that the organization may have to depend on only a few people to create HTML and XML documents.

After initial training, users should be surveyed to determine if the tools they have been provided satisfy their needs. Many users find that creating HTML/XML documents is difficult. If so, then they may also need special training. In corporations, this training is often provided by one person in each department who has been given responsibility for training the rest of the department.

In summary, the following actions are recommended to help develop an effective program for training employees to author high-quality HTML/XML documents:

- Conduct a survey to assess user training needs and wants.
- Train users how to develop HTML/XML content.
- Provide users with HTML/XML authoring tools that complement what they already know (for example, the Internet Assistant for Microsoft Word is a good choice for users already familiar with Microsoft Word).
- Review the design and flow of material that will be "published" on the Intranet.
- Give feedback to HTML/XML authors on ways to improve the site appearance and ease of use.

2.1.2.7 Organizational Challenges

In addition to technological challenges, companies may also face the following organizational challenges after the initial release of an Intranet:

- Marketing the Intranet within the organization so that all employees will support its growth and continued use
- Obtaining additional funding on an ongoing basis to implement new capabilities
- Encouraging an information-sharing culture within the company so that all employees will contribute toward building a learning organization
- Merging a paper-based culture with the new culture of electronic documentation
- Ensuring that the content on the Intranet is updated on a regular basis
- Preventing one person or group from controlling (monopolizing) the content on the Intranet
- Instructing employees to author HTML/XML content so they can contribute material to the Intranet
- Informing employees on Intranet etiquette, thereby facilitating courteous online discussion forums and other forms of user interaction on the Intranet
- Using the Intranet as an integral part of working with customers and vendors
- Measuring the Intranet's overall effectiveness and contribution to the organization

As is the case when introducing any new information technology to an enterprise, Intranet deployment requires careful planning, effective implementation, and employee training. In the short term, most of the organizational focus is usually on the technical aspects of the Intranet deployment. But as time goes on, organizational issues relating to how the Intranet is used within the organization must be managed. When an organization actively examines and works toward resolving these issues, they are better able to achieve a culture of teamwork and collaboration.

2.1.2.8 Management Summary

The following list summarizes key points surrounding the use of Intranets:

- An Intranet is a company-based version of the Internet. Intranets provide an inexpensive solution for information sharing and user communication.
- An Intranet provides an easy way for users to communicate and share common documents, even if they are using different machines, such as IBM compatible and Macintosh personal computers.
- Some organizations have expanded their Intranet to allow customers to access internal databases and documents.
- Many companies can establish a functional Intranet using in-house personnel with a minimal amount of new equipment.

Internet technology adheres to open standards that are well documented. This, in turn, encourages the development of cost-effective and easy-to-implement Intranet solutions. As the popularity of Intranets has increased, so has the demand for new tools and Web-based solutions. This demand has fueled competition among software manufacturers which, in turn, has resulted in better and less-expensive Intranet products.

In summary, Intranets can be used to improve productivity, simplify workflows, and gain a competitive advantage over those who have yet to learn how to capitalize on the benefits of Intranets.

2.1.3 Technical Overview

2.1.3.1 Internet Basics

2.1.3.1.1 Packet Switching

Packet switching was introduced in the late 1960s. In a packet-switched network, programs break data into pieces, called *packets*, which are transmitted between computers. Each packet contains the sender's address, the destination address, and a portion of the data to be transmitted. For example, when an e-mail message is sent over a packet-switched network, the e-mail is first split into packets. Each packet intermingles with other packets sent by other computers on the network. Network switches examine the destination address contained in each packet, and route the packets to the appropriate recipient. Upon reaching their destination, the packets are collected and reassembled to reconstitute the e-mail message.

2.1.3.1.2 TCP/IP

The U.S. Advanced Research Projects Agency (ARPA) was a major driving force in the development and adoption of packet-switched networking. The earliest packet-switched network was called the ARPAnet. The ARPAnet was the progenitor to today's Internet. By the early 1980s, ARPA needed a better protocol for handling the packets produced and sent by various network types. The original ARPAnet was based on the Network Control Protocol (NCP). In January 1983, NCP was replaced by the Transport Control Protocol/Internet Protocol (TCP/IP). TCP/IP specifies the rules for the exchange of information within the Internet or an Intranet, allowing packets from many different types of networks to be sent over the same network.

2.1.3.1.3 Connecting to the Internet

One way to connect to the Internet is to install a link from the company network to the closest computer already connected to the Internet. When this method is chosen, the company must pay to install and maintain the communications link (which might consist of a copper wire, a satellite connection, or a fiber-optic cable) to the Internet. This method was very popular with early adopters of the Internet, which included universities, large companies, and government agencies. However, the costs to install and maintain the communications link to the Internet can be prohibitive for smaller companies.

Fortunately, specialized companies—called Internet service providers (ISPs)—are available to provide a low-cost solution for accessing the Internet. ISPs pay for an (expensive) connection to the

Internet, which they make accessible to others through the installation of high-performance servers, data lines, and modems. Acting as middlemen, the ISPs rent time to other users who want to access the Internet.

Two important decisions must be made when deciding what type of Internet connection is the most appropriate. The first decision is the company budget allocated for Internet connectivity, and the second is the Internet connection speed needed to support the business requirements. These decisions are interrelated. ISPs offer a variety of options for connecting to the Internet, ranging from a simple dialup account over phone wires to high-speed leased lines from the company to the ISP. Dialup accounts are typically available for a low, flat monthly fee, and are generally much cheaper than a leased line connection. However, the leased line connection is usually much faster than the dialup connection.

When a dialup account is used, a modem and a phone line are used to call and log into the ISP server (or computer), which, in turn, acts as the doorway to the Internet. The transmission speed of the connection is limited by the speed of the modems employed by the user and the ISP. A modem is unnecessary when a leased line connection is available to the ISP. Leased lines are offered in many different configurations with a variety of options. The most common link types are ISDN (integrated services digital network, which support transmission speeds from 56 Kbps to 128 Kbps), T1 (transmitting at speeds up to 1.54 Mbps), and T3 (transmitting at speeds up to 45 Mbps).

If a company only needs to make an occasional connection to the Internet—for example, less than 20 to 50 hours per month for all users—a dialup account should be sufficient. However, if a company needs faster data transfer speeds or has several users who must access the Internet for substantial periods of time over the course of a month, a leased line connection should be considered.

The fastest growing segment of Internet users are those who connect to the Internet through an ISP via an ordinary telephone connection. There are two major protocols for connecting to the Internet in this way: Serial Line Internet Protocol (SLIP) and Point-to-Point Protocol (PPP). SLIP is the older protocol and is available in many communications packages. The faster PPP is newer and therefore it is not as widely supported.

Principles of queuing analysis can be applied to the problem of sizing the links needed to support the Internet access, whether or not that access is to an ISP or to a direct Internet connection. The reader is referred to Chapter 2 of *Network Design: Management and Technical Perspectives* (2005) by Mann-Rubinson and Terplan, for specific techniques on how to estimate the throughput and performance characteristics associated with using different size link capacities. This analysis can be used to determine whether or not a dialup or leased line connection is sufficient to support the bandwidth requirements with tolerable transmission delays.

2.1.3.1.4 Basic Terminology

In this section, we define commonly used Internet and Intranet terminology.

The World Wide Web: The World Wide Web—or Web—is a collection of seamlessly interlinked documents that reside on Internet servers. The Web is so named because it links documents to form a web of information across computers worldwide. The documents available through the Web can support text, pictures, sounds, and animation. The Web makes it very easy for users to locate and access information contained within multiple documents and computers. *Surfing* is the term used to describe accessing (through a Web browser) a chain of documents through a series of links on the Web.

Web Browsers: To access and fully utilize all the features of the Web, special software—called a Web browser—is necessary. Its main function is to allow the user to traverse and view documents on the Web. Browser software is widely available for free, either through a download from the Internet or from ISPs. Commonly used browsers include Microsoft Internet Explorer and Mozilla's Firefox. Some of the common tasks that both support include:

- Viewing documents created on a variety of platforms
- Creating and revising content

- Participating in threaded discussions and news groups
- Watching and interacting with multimedia presentations
- Interfacing with existing legacy data (non-HTML based data) and applications
- Gaining seamless access to the Internet

It should be noted that the same Web browser software used for accessing the Internet is also used for accessing documents within an Intranet.

u niform Resourcel ocator (u Rl): The Web consists of millions of documents that are distinguished by a unique name called a URL (Uniform Resource Locator), or more simply, a Web address. The URL is used by Web browsers to access Internet information. Examples of URLs include:

http://www.netscape.com
ftp://ftp.microsoft.com

A URL consists of three main parts:

1. A service identifier (such as http)
2. A domain name (such as www.ups.com)
3. A path name (such as www.ups.com/tracking)

The first part of the URL, the service identifier, tells the browser software which protocol to use to access the file requested. The service identifier can take one of the following forms:

- http://—This service identifier indicates that the connection will use the hypertext transport protocol (HTTP). HTTP defines the rules that software programs must follow to exchange information across the Web. This is the most common type of connection. Thus, when Web addresses start with the letters *http* it indicates that the documents are retrieved according to the conventions of the HTTP protocol.
- ftp://—This service identifier indicates that the connection will use the file transfer protocol (FTP). This service identifier is typically used to download and copy files from one computer to another.
- telnet://—This service identifier indicates that a telnet session will be used to run programs from a remote computer.

The second part of the URL, the domain name, specifies which computer is to be accessed when running server software. An example of a domain name is: www.tcrinc.com.

The final part of the URL, the path name, specifies the directory path to the specific file to be accessed. If the path name is missing from the URL, the server assumes that the default page (typically the homepage) should be accessed. Large, multipage Web sites can have fairly long path names. For example, these URLs request specific pages within a given Web site:

- http://www.apple.com/documents/productsupport.html
- http://www.bmwusa.com/ultimate/5series/5series.html
- ftp://ftp.ncsa.uiuc.edu/Mac/Mosaic
- http://www.microsoft.com/Misc/WhatsNew.htm

Home Pages: Companies, individuals, and governments that publish information on the Internet usually organize that information into *pages*, much like the pages of a book or a sales brochure. The first page that people see in a sales brochure is the cover page, which may contain an index and summary of the brochure contents. Similarly, a home page is the first page that users see when they access a particular Web site. The home page is to the Web site what the cover page is to a sales brochure. Both must be appealing, concise, informative, and well organized to succeed in maintaining the reader's interest. The home page is usually used to convey basic information about the company and what it is offering in the way of products and/or services.

Many companies publish the Internet address (or URL) of their home page on business cards, television, magazines, and radio. To access a Web site, a user has merely to type the URL into the appropriate area on the Web browser screen.

Client Programs and Browsers: Across the Internet, information (i.e., programs and data) is stored on the hard disks of thousands of computers called *servers*. These are so named because, upon request, they serve (or provide) users with information. A server is a remote computer that may be configured to run several different types of server programs (such as Web server, mail server, and ftp server programs).

A client program is used to initiate a session with a server. Client programs are so named because they ask the server for service. In the case of the Web, the client program is the Web browser. All client–server interactions take the same form. To start, the client connects to the server and asks the server for information. The server, in turn, examines the request and then provides (serves) the client with the requested information. The client and server may perform many request–response interactions in a typical session.

Software programs—such as a browser—use HTTP commands to request services from an HTTP server. An HTTP transaction consists of four parts: a connection, a request, a response, and a close.

Where Web Documents Reside: When users publish Web pages, they actually store the pages as files that are accessible through a file server. Typically, Web pages reside on the same computer on which the server program is running, but this is not necessarily true. For security reasons, it may be necessary to limit accessibility to various files on the Web server. Obviously, it might be disastrous if internal documents and data were made available to competitors. To prevent this type of security risk, a Webmaster (or systems administrator) can configure the Web server so it only allows specific clients to access confidential information, based on a need-to-know basis. The Webmaster can control access to the server by requiring users to log in with a username and password that has predetermined access privileges.

HTMl —The l anguage of the World Wide Web: The European Particle Physics Laboratory at CERN, in Geneva, Switzerland, developed Hypertext Markup Language (HTML) in the late 1980s and early 1990s. HTML is the language of the World Wide Web. Every site on the Web uses HTML to display information.

Each Web document contains a set of HTML instructions that tell the browser program how to display the Web page. When you connect to a Web page using a browser, the Web server sends the HTML document to your browser across the Internet. Any computer running a browser program can read and display HTML, regardless of whether that computer is a personal computer running Windows, a Unix-based system, or a Mac.

If word processor formatted files—such as Microsoft Word—were used to create Web pages, only users with access to Microsoft Word would be able to view the Web page. HTML was designed to overcome this potential source of incompatibility. All users can access Web pages from their browser since all Web pages conform to HTML standards. An HTML Web page is a plain text file (i.e., an ASCII text file) that can be created and read by any text editor. There are many software programs available to convert document files to HTML equivalents. In addition, many standard presentation and word processing packages offer built-in routines to convert a standard document into a Web-ready HTML file. This type of conversion might be helpful, for example, if you wanted to convert a Microsoft PowerPoint presentation into a set of HTML files for display on the Web.

After HTML files are transferred to a Web site, anyone with a browser can view them. HTML provides the browser with two types of information:

1. *Mark-up information* that controls the text display characteristics and specifies Web links to other documents.
2. *Content information* consisting of the text, graphics, and sounds that the browser displays.

Hypertext and Hyperlinks: Documents on the Web can be interconnected by specifying links (called hyperlinks) that allow the user to jump from one document to another. The HTML code, which drives all Web pages, supports hypertext. Hypertext, in turn, supports the creation of multimedia documents (containing pictures, text, animation, sound, and links) on the Web.

Hyperlinks (or simply, links) are visually displayed on the Web pages as pictures or underlined text. When a user clicks on a hyperlink displayed on their browser screen, the browser responds by searching for and then loading the document specified by the hyperlink. The document specified in the hyperlink may reside on the same computer as the Web page on display or it may reside on a different computer on the other side of the world. Much of the Web's success has been attributed to the simplicity of the hyperlink point-and-click user interface.

There are four basic layouts for linking Web pages with hyperlinks: linear, hierarchical, Web, and combination. Which layout is the most appropriate depends on the type of information that is being presented and the intended audience.

FTP—The File Transfer Protocol: The FTP (file transfer protocol) is a standard protocol for transferring and copying files from one computer to another. Depending on the configuration of the FTP server program, you may or may not need an account on the remote machine to access system files. In many cases, you can access a remote computer with FTP by logging on with a username of "anonymous," and by entering your e-mail address as the password. This type of connection is referred to as an *anonymous FTP session*.

After logging in to the remote FTP server, it is possible to list a directory of the files that are available for viewing and/or copying. The systems administrator determines which files can be accessed on the remote server, and who has access privileges. When system security is a major concern, the system administrator may require a specific username and password (as opposed to allowing an anonymous log-on procedure) to gain access to system files.

FTP is very useful in accessing the millions of files available on the World Wide Web. Most browsers have built-in FTP capabilities to facilitate downloading files stored at FTP sites. To access an FTP site using your browser, you type in the FTP site address, much like entering a Web address. For example, to access the Microsoft FTP site, the address "ftp://ftp.microsoft.com" would be entered into the browser address window.

Java: Java is a programming language released by Sun Microsystems. Java is designed for creating animated Web sites. Java can be used to create small application programs, called applets, which browsers download and execute. For example, a company might develop a Java applet for their Web site to spin the company's logo, to play music or audio clips, or to provide other forms of animation to improve the appeal and effectiveness of the Web page.

2.1.4 Intranet Components

This section will provide an overview of the components necessary to create an Intranet. The final selection of the Intranet components depends on the company's size, level of expertise, user needs, and future Intranet expansion plans. In addition, we also examine some of the costs associated with the various Intranet components.

An Intranet requires the same basic components found on the Internet, including:

1. A computer network for resource sharing
2. A network operating system that supports the TCP/IP protocol
3. A server computer that can run Internet server software
4. Server software that supports hypertext transport protocol (HTTP) requests from browsers (clients)
5. Desktop client computers equipped with network software capable of sending and receiving TCP/IP packet data
6. Browser software installed on each client computer

It should be noted that if a company does *not* want to use an internal server, an ISP can be used to support the Intranet. It is very common for organizations to use an ISP, especially when there is little information content or interest in maintaining a corporate-operated Intranet server. ISPs are also used when the organizational facilities cannot support the housing of an Intranet server.

In addition to the software and hardware components listed above, HTML/XML documents must be prepared to provide information displays on the Intranet. The creation and conversion of documents to HTML/XML format is very easy using commercial software packages, such as Microsoft's FrontPage. Third-party sources are also available to provide this service at a reasonable cost.

2.1.4.1 Network Requirements

The first requirement for an Intranet is a computer network. For the purpose of this discussion, we assume that a basic computer network is in place. We now focus on the hardware and software modifications needed to support an Intranet.

Most computer networks are local-area networks (LANs). LANs are based on a client–server computing model that uses a central, dedicated computer—called the server—to fulfill client requests. The client–server computing model divides the network communication into two sides: a client side and a server side. By definition, the client requests information or services from the server. The server, in turn, responds to the client's requests. In many cases, each side of a client–server connection can perform both client and server functions.

Network servers are commonly used to send and receive e-mail, and to allow printers and files to be shared by multiple users. In addition, network servers normally have a storage area for server programs and to backup file copies. Server applications provide specific services. For example, a corporatewide e-mail system typically uses a server process that is accessible from any computer within the company's network.

A server application (or server process) usually initializes itself and then goes to sleep, spending much of its time simply waiting for a request from a client application. Typically, a client process will transmit a request (across the network) for a connection to the server, and then it will request some type of service through the connection. The server can be located at either a local or remote site.

Every computer network has a physical topology by which it is connected. The most common topologies used to connect computers are the star, Token Ring, and bus topologies.

A network interface card (NIC) is needed to physically connect a computer to the network. The network interface card resides in the computer and provides a connector to plug into the network. Depending on the network, twisted-pair wiring, fiber-optic, or coaxial cable may be used to physically connect the network components. The network interface card must be compatible with the underlying network technology employed in the network (e.g., Ethernet or Token Ring).

2.1.4.2 Network Operating Systems

The Internet supports connectivity between various hardware platforms running various operating systems. In theory, there is no reason why an organization must stay with one type of machine or operating system when implementing an Intranet. However, in practice, many organizations use only one network operating system to simplify the task of managing the network.

The primary choices for operating systems consist of UNIX and Windows NT. We now discuss each of these operating systems and important considerations surrounding their use.

2.1.4.2.1 UNIX

Many larger companies use UNIX-based machines as their primary business application server platform. UNIX is a proven operating system that is well suited for the Internet's open system model. Unfortunately, learning how to use UNIX is not easy. Also, using a UNIX-based machine limits the choices available for developing interactive Intranets and other software applications. Many program-

mers, for example, prefer to develop applications using Windows-based machines and programming languages (such as Microsoft's Visual Basic or Borland's Delphi).

2.1.4.2.2 Windows NT

Many companies choose Windows NT over UNIX because NT is easy to install, maintain, and administer. Windows NT, like UNIX and OS/2, provides a high-performance, multitasking workstation operating system. It also supports advanced server functions (including HTTP, FTP, and Gopher) and communications with clients running under MS-DOS, Windows 3.1, Windows 95, Windows for Workgroups, Windows NT Workstation, UNIX, or Macintosh operating systems. The latest version of Windows NT Server includes a free Internet Information Server (IIS) and a free Web browser (Internet Explorer). Microsoft designed the IIS so that it can be installed and up and running on a Windows NT workstation in less than ten minutes. The Windows NT Server comes with a built-in remote access services feature that supports remote access to the Intranet through a dialup phone connection.

2.1.4.3 Server Hardware

Server machines run the network operating system and control how network computers share server resources. Large businesses with thousands of users typically use high-speed Unix-based machines for their servers. Small and medium-sized companies normally use less expensive Intel-based machines. The load (i.e., the number of users and the amount of network traffic) on the Intranet server machine will influence the selection of a specific processor type.

There is considerable debate in the industry as to which machine makes a better Intranet server: a Unix workstation, an Intel-based machine, or a PowerPC-based system. In general, the server choice depends on the plans for the Intranet and the level of familiarity the company has with each of these platforms. The server hardware selection also depends on the network operating system in use.

If a Unix-based server is chosen, a company will pay more for an equivalent amount of computing power provided by an Intel-based machine. Unix machines still carry a price premium over PCs, because they are made from custom parts, while Intel-based machines are made from commodity components available from many hardware vendors and suppliers. For example, a high-end Pentium machine with the capacity to serve over 1000 client machines can be bought for about one tenth of the cost of a comparable Unix server. Macintosh-based systems are more expensive than comparable Intel-based machines, but they are still much less expensive than Unix-based machines.

The decision to use a Unix-based machine versus an Intel-based machine as the Intranet server is also influenced by maintenance costs. Maintaining a Unix-based machine requires more resources than maintaining an Intel-based machine. Hardware upgrades for Intel-based machines are also cheaper than hardware upgrades for Unix workstations. A Macintosh server will cost more to upgrade than an Intel-based machine. However, these costs are still lower than a comparable upgrade on a Unix-based machine.

The debate over Unix- and Intel-based machines focuses primarily on their performance in supporting business application servers. For example, companies that use large accounting and financial software packages often use Unix servers. On the other hand, companies that do not want to pay the price premium for Unix machines and/or are not familiar with Unix machines often select Intel-based machines as their business application servers.

Pentium-class machine supports a vast array of software applications and server software. Many industry experts believe that Pentium-class machines will take a significant amount of the market share away from Unix workstations. This, in turn, means that more and more Intranet-based applications will use Pentium-class machines in the future.

2.1.4.4 Web Server Software

A working Intranet requires server software that can handle requests from browsers. In addition, server software is needed to retrieve files and to run application programs (which might, for instance, be used to search a database or to process a form containing user-supplied information).

For the most part, selecting a Web server for an Intranet is similar to selecting a Web server for an Internet site. However, Internet servers must generally handle larger numbers of requests and must deal with more difficult security issues. The performance of the Web server has a major impact on the overall performance of the Intranet. Fortunately, it is fairly easy to migrate from a small Web server to a larger, high-performance Web server as the system usage increases over time.

2.1.4.4.1 UNIX Web Servers

One of the first Web servers for Unix-based machines was the National Center for Supercomputing Application's (NCSA) HTTP Web Server, parts of which were later initially included in the Apache Server. Much of the Internet's growth is primarily due to the popularity of this server, which was free. NCSA was always committed to the continued development of its Web server, which provided both common gateway interface (CGI) capabilities and the server side includes Software Solutions Incorporated (SSI) software. SSI software is used by Web servers to display or capture dynamic (changing) information on an HTML/XML page. For example, SSI can be used to display a counter showing the number of visitors to a Web site. The NCSA Web server also allowed the creation of virtual servers on the same machine. The virtual servers can have their own unique universal resource locator (URL). This is useful, for example, for assigning a different IP address to different departments using the same machine.

2.1.4.5 Desktop Clients Running TCP/IP

TCP/IP must be installed on each client machine running on the Internet.* To use an Internet-based application (such as a Web browser) on a Windows-based machine, a TCP/IP stack must be present. Windows 95, Windows NT, and IBM's OS/2 Warp operating systems include the TCP/IP protocol suite. Most Unix-based systems use TCP/IP as their main network communication protocol, as do the other major operating systems.

2.1.4.6 Web Browsers

A Web browser is a software application, such as Microsoft's Internet Explorer or Mozilla's FireFox, that enables users to display and interact with text, images, videos, music, and other information found on Web pages. Hyperlinks on the Web pages permit users to move through and access the content offered by the Web page. Browsers are used both on the Internet and on intranets.

2.1.4.7 Intranet Component Summary

In this section, the basic components of an Intranet were examined. We recapitulate below some of the key concepts covered in this section:

- Intranets are based on a client–server network computing model. By definition, the client side of a network requests information or services and the server side responds to a client's requests.
- The physical components of an Intranet include network interface cards, cables, and computers.
- Suites of protocols, such as TCP/IP, manage data communication for various network technologies, network operating systems, and client operating systems.

2.1.5 Intranet Implementation

2.1.5.1 Information Organization

After the physical components of the Intranet are in place, the next step is to design the information content of the Intranet and/or Internet Web pages. This task involves identifying the major categories and topics of information that will be made available on the Intranet. Information can be organized by

* If the network does not support TCP/IP, a gateway application that translates TCP/IP for the network operating system protocol must be used.

department, function, project, content, or any other useful categorization scheme. It is advisable to use cross-functional design teams to help define the appropriate informational categories that should be included on the corporate Web site. The following types of information are commonly found on corporate Intranet homepages:

- What's new
- Corporate information (history and contacts)
- Help desk and technical support
- Software and tools library
- Business resources
- Sales and marketing information
- Product information
- Human resources–related information (benefits information, etc.)
- Internal job postings
- Customer feedback
- Telephone and e-mail directory
- Quality and system maintenance records
- Plant and equipment records
- Finance and accounting information
- Keyword search/index capability

2.1.5.2 Content Structure

After the main topics of information to be displayed on the corporate Web page(s) have been identified, the flow and manner of presentation on the Intranet must be developed. Four primary flow models are used to structure the flow of presentation at an Intranet Web site: linear, hierarchical, nonlinear (or Web), and combination information structures.

A linear information structure is similar in layout to a book in that information is linked sequentially, page by page. When a linear layout is used, the Web pages are organized in a "slide show" format. This layout is good for presenting pages that should be read in a specific sequence or order. Since linear layouts are very structured, they limit the reader's ability to explore and browse the Web page contents in a nonsequential (or nonlinear) manner.

When a hierarchical layout is used to structure the information, all the Web pages branch off from the home page or main index. This layout is used when the material in the Web pages does not need to be read in any particular order. A hierarchical information structure creates linear paths that only allow up and down movements within the document structure.

A no-linear, or Web, structure links information based on related content. It has no apparent structure. Nonlinear structures allow the reader to wander through information spontaneously by providing links that allow forward, backward, up and down, diagonal, and side-to-side movement within a document. A nonlinear structure can be confusing, and readers may get lost within the content, so this structure should be chosen with care. The World Wide Web uses a nonlinear structure. The advantage of a nonlinear structure is that it encourages the reader to browse freely.

The combination Web page layout, as the name implies, combines elements of the linear, Web, and hierarchical layouts. Regardless of the type of flow sequence employed, each Web page typically has links that allow the user to move back and forth between pages and back to the home page.

Over the lifetime of the Intranet, it is likely that the layout and organization of information on the corporate Web pages will change many times. It is often helpful to use flow charting tools to help manage and document the updated information flows. Visio for Windows by Visio Corp. and ABC Flowcharter by Micrografx are two excellent tools for developing flowcharts. In addition, some of the Web authoring tools offer flowcharting and organizational tools to help design and update the information structure on the Web pages.

2.1.5.3 Interface Design

After defining the Intranet's structure, the next step is to define the functionality and user interface. The Intranet design should be consistent with the organization's corporate image. For example, items such as corporate images, logos, trademarks, icons, and related design themes add a familiar look and feel to the content. Where possible, they should be included in the Web page design. It is also advisable to work with the marketing department when designing the Web page layouts to ensure that a consistent theme is maintained in all the company communications that will be viewed by the outside world.

A technique called *storyboarding* is frequently used to design the Web page layout. Storyboards are used by film producers, story writers, and comic strip artists to organize the content and sequence of their work. A storyboard depicts the content, images, and links between pages of the Intranet in the form of a rough outline.

Software, such as Microsoft PowerPoint or a similar presentation program, can be used to develop a storyboard and sample Web pages. It is a good idea to test the interface design to ensure that the icons, buttons, and navigational tools are logical and intuitive. An Intranet without intuitive navigational tools is like a road without signs. Just as it would be difficult for drivers to find their way from one city to another without the aid of signs, street names, and directional information, Intranet users will find it difficult to retrieve information without easy-to-follow categories, buttons, and links. It is often helpful to employ graphic designers and marketing–communications staff to create effective graphics and images for the Web site.

2.1.5.4 Free Clip Art and Images

Many icons and navigational signs are available as clip art that comes with word processing, page layout, and presentation software programs. In addition, many Web sites offer images and clip art, which can be downloaded for free. However, the licensing agreements for downloading free images may have restrictions and requirements that should be observed.

2.1.5.5 Intranet Functionality

The required functionality of an Intranet dictates many of the design and user interface features. One of the goals in designing the Intranet should be to improve existing systems and infrastructures. After examining the current information structure, it may become clear which aspects of the structure work well and which ones need improvement.

Workflow processes, document management, and work collaboration are areas that the organization should strive to improve through the use of an Intranet. A workflow analysis should consider ways in which the Intranet can automate various organizational tasks and processes. For example, if a company has a geographically dispersed project team, the Intranet might be used to post and update project information as various tasks are completed. Other team members could then visit the Intranet page at any time to check the project status.

The following functionality checklist is helpful when developing a list of the functions that need to be supported by the Intranet:

- The user interface must be intuitive and tested
- The Intranet's design should support continuous updates
- The Intranet may need to be integrated with database management systems to allow users to access information (such as customer and product data)
- The Intranet should support existing (legacy) applications, as needed
- The Intranet should have built-in directories, such as corporate telephone numbers and e-mail addresses
- The Intranet should incorporate groupware applications
- Support (or future expansion) for online conferencing should be considered

- The Intranet should provide division-specific and corporatewide bulletin boards for electronic postings
- The Intranet should be designed with a document sharing and management process in mind
- The Intranet should foster teamwork and collaboration by enhancing channels of information distribution
- Search engines, which simplify a user's ability to locate and access information, should be made available
- The Intranet should support e-mail
- The Intranet should support (for future expansion) multimedia applications that use text, images, audio, and video
- Automated real-time Web-page generation should be encouraged
- The Intranet should be designed so it can interface, at least potentially, with factory equipment, other manufacturing devices, or other critical legacy systems
- The Intranet should support the automation of organization workflows

2.1.5.6 Content Management

Many organizations struggle with the tasks of information creation, management, and dissemination. They are time consuming and difficult to control. The Intranet alone cannot solve information management problems unless specific Intranet solutions are implemented that directly address the need for document management. The following list identifies content-management tasks that should be considered in the Intranet plan:

- Users must have the ability to easily add or update content on a regular basis
- Users must have the ability to protect their content from changes by other users
- A content-approval process should be defined and in place. This process should encompass ways to manage and control document revisions, especially changes to shared documents

As policies and procedures relating to content management are formulated, it is important to designate responsibilities to specific individuals to ensure that they are put into place and followed. An Intranet style guide should be developed that provides page layout, design elements, and HTML/XML code guidelines. The style guide will help the organization to maintain a consistent look and feel throughout the Intranet's Web pages. The style guide should contain information on where to obtain standard icons, buttons, and graphics, as well as guidelines on page dimensions and how to link to other pages. As part of the style guide, it is helpful to create Web page templates. These templates consist of HTML/XML files, and are used to provide a starting point for anyone interested in developing Web pages or content for the Intranet. Although it is very easy to create a working Web page and to publish it for mass viewing, the real challenge is in producing a well-conceived Web page.

2.1.5.7 Training and Support

After the Intranet is up and running, efforts should be focused on how to maintain the information content and on employee training. Part of the document-management strategy should encompass the selection of content stakeholders. Content stakeholders are individuals in different departments or work groups who are responsible for the creation and maintenance of specific content. Stakeholders can be department managers, team leaders, or content authors and publishers.

Some organizations create a position called a *Webmaster*. This position is responsible for maintaining and supporting the content published on the Intranet. A good Webmaster should have the following skills:

- Basic Internet skills, including an understanding of e-mail, FTP, and Telnet
- A thorough understanding of HTML/XML document creation
- Experience with CGI programming
- Programming experience with languages such as Perl, C/C++, and Java

- Experience with content creation and the conversion of text and images
- Knowledge of client–server processing
- Experience with server setup and maintenance
- Knowledge of your organization's structure and inner workings
- Organizational and training skills

It is possible that the organization may choose to decentralize the maintenance of the information content. In this case, individuals from various departments might be selected to maintain the content relating to their respective department. These individuals should be trained to handle a variety of maintenance issues. A decentralized approach depends on having more than one individual with the necessary skills available to maintain the Web pages. A decentralized support structure gives authors and content owners direct control and responsibility for publishing and maintaining information. This can help prevent bottlenecks in making information available in a timely fashion.

Training for stakeholders, Webmasters, and Intranet users is an important part of an Intranet strategy. Intranet customers and content stakeholders should be trained to understand the Intranet and how it will improve the organization and the way the company does business. They should also be given training on how to create, utilize, and maintain content on the Web page(s). Companies that invest in the education and training of their employees will have a better chance of creating and maintaining a successful Intranet.

2.1.6 Intranet Deployment

Since Intranets are easy to set up, many companies do not realize what the true resource requirements are to maintain the Intranet with up-to-date information. The goal of this section is to provide a realistic perspective on how organizations are most likely to achieve long-lasting benefits from the Intranet.

Some companies invest much more than their competitors in information technology, such as an Intranet, but still fail to effectively compete in the marketplace. Computers alone do not, and cannot, create successful companies. A good start is to empower all employees to contribute to the Intranet. As is true for any collaborative effort, every member is responsible for the overall success of the team.

2.1.6.1 Technological Considerations

The major technological challenges facing the organization after the initial implementation of an Intranet include:

- Converting existing paper documents into electronic documents that employees can access electronically via the Intranet.
- Connecting existing databases to the Intranet so they are accessible by a wide range of computing platforms (such as Windows- and Mac-based systems).
- Coordinating the use of multiple servers used across departmental lines.
- Continuously enhancing the Intranet's features and capabilities to keep employees motivated to use the Intranet.
- Installing security features within the Intranet to prevent unauthorized access to confidential or sensitive information.

Intranet technology, and information technology in general, is changing so fast that keeping up with the latest software and hardware solutions requires a substantial ongoing organizational commitment.

2.1.6.1.1 Conversion of Paper Documents into Electronic Form

The first issue facing companies after the initial Intranet release is how to convert large numbers of existing paper documents into electronic format ready for distribution on an Intranet. There are many tools, such as HTML Transit, that can be used to convert documents from most electronic formats to HTML

TABl E 2.1.1 Intranet Document Tracking Information

Data for Tracking Intranet Documents
Name of document
Document description
Page owner
Type of document (i.e., official, unofficial, personal)
Confidentiality status (i.e., confidential, nonconfidential, etc.)
Original publish date
Date document last modified
Frequency of update (i.e., daily, weekly, monthly, etc.)

or XML format. Microsoft's Internet Assistant for Microsoft Word can also be used to easily convert existing Word documents into HTML or XML documents. After paper documents have been converted to HTML or XML and placed on the Intranet, the next challenge is to keep the documents up to date.

Obsolete information can frustrate Intranet users and may encourage them to revert to old ways of information gathering (i.e., calling people, walking to various offices, and writing memos). One way to minimize this problem is to create a database containing the document title, date of last change, and frequency of update in a database. Other useful information that can be used to track the status and nature of documents on the Intranet is shown in Table 2.1.1. A program can then be written to search the Intranet for documents that have not been updated recently. The program can then issue e-mail to the document owner to request an update.

2.1.6.1.2 Interface to Legacy Database(s)

Connecting databases to the Intranet is not an easy task, and may require additional staff or reassignment of current programming staff. Legacy database vendors are currently working on various Intranet solutions to facilitate the implementation of this requirement.

Companies may need to connect the Intranet to legacy databases in order to access:

- Financial reports (regarding project costs, product costs, the overall financial health of the enterprise, etc.)
- Document-management systems
- Human resources information (e.g., so employees can review details on health care and benefits)

2.1.6.1.3 Use of Multiple Servers

As the Intranet becomes more complex, multiple servers will be needed. This is especially true for companies that have a large number of divisions and business units using the Intranet. For example, a product-development group may need to provide team members the ability to search project-specific databases, submit forms to various databases, and to use a private online discussion group. The Webmaster may find it impossible to support these service needs in a timely manner. When this happens, companies frequently relegate the task of server maintenance to each respective department.

Over the next few years, installing and using a Web server will become as easy as installing and using word processor software. Web servers will probably become part of the Windows NT server operating system. When each department is responsible for maintaining their own Web server, it is particularly important to choose server software that is easy to install and maintain. A Pentium-class machine running Windows NT server software and Microsoft's IIS is a good choice for small departments. Another way to provide departments with their own domain name and disk space is to use a virtual domain name. Companies use virtual servers to reduce hardware costs. In the case of the Web, an HTTP-based server runs on a server computer. For example, a company may need two types of Web servers, one that allows easy access and one that requires usernames and passwords. In the past, the company would have

purchased two different computers to run the Web server software. Today, however, the company can run both servers on the same system—as virtual servers.

2.1.6.1.4 Standardizing Hardware and Software

To avoid supporting multiple hardware and software components, it is important to standardize the server software, hardware, HTML and XML editing tools, and browser software. This will help to minimize the potential for unexpected network errors and incompatibilities.

2.1.6.2 Maintaining the Information Content on the Intranet

One of the major challenges organizations must face is how to transition from paper-based systems to computer-based systems, while keeping information up to date.

2.1.6.2.1 Automating HTML/XML Authoring

After establishing a policy for the distribution of Intranet documents, it is advisable to develop a set of guidelines that clearly specifies who is responsible for keeping them current. Inaccurate information greatly reduces the effectiveness of the Intranet. If employees lose confidence in the accuracy of the online information, they will revert to calling people to find information. Unfortunately, many people tend to ignore the need to update information, irrespective of its form (i.e., electronic or print).

In some cases, the Intranet will contain information that employees must update daily, weekly, or monthly. Spreadsheets can be used to capture highly time-sensitive data. Macros can be written (typically, by a staff programmer) that automatically convert the spreadsheet data into HTML or XML format.

2.1.6.2.2 Managing Document Links

In a traditional document-management system, documents often reference one another. In most cases, authors list the applicable references at the top of each new document. Intranets, unfortunately, create a situation where organizations cannot easily control the accuracy of links in documents.

HTML or XML document developers can use links freely and, in many cases, without checking the accuracy of those links. Even if employees test the initial accuracy of their document links, it is difficult to maintain and check the accuracy of those links after the document is released. If you have ever encountered a "broken" link when surfing the Web, you know that it can be frustrating. People depend on links within a Web document to find information. Today, however, there are a few mechanisms available to assure the accuracy of document links. Employees must understand that other people may link to their pages, and that they should not freely move the location of their documents. Employees must view their needs in the total organizational context.

2.1.6.3 Centralized vs. Distributed Control

The implementation of an Intranet is a major change for any organization. Although change is not easy, people are more inclined to modify their behavior when leaders have a clear sense of direction, involve employees in developing that direction, and are able to demonstrate how the Intranet will positively affect the employees' well being. Managers should work with their employees to show that Intranets can free them from the routine aspects of their job. This, in turn, will allow employees to spend more time learning and developing new ideas for the corporation.

Some of the benefits that can be obtained using a distributed model of Intranet control are:

- Employees can tap into the knowledge of everyone in the organization, making everyone a part of a solution.
- The power of any one Webmaster to dictate the Intranet's form and function is limited.
- It empowers departments to create their own information databases and to work with outside customers and vendors.

2.1.7 Intranet Security Issues

By their very nature, Intranets encourage a free flow of information. This means that it is also very easy for information to flow directly from the Intranet to the desktops of those who might seek to gain access to information they should not have. To guard against this situation, adequate security measures should be in place when the Intranet is deployed. In the discussion that follows, we review various security techniques to protect an Intranet from unauthorized external and internal use.

2.1.7.1 Firewalls

The Internet was designed to be resistant to network attacks in the form of equipment breakdowns, broken cabling, and power outages. Unfortunately, the Internet today needs additional technology to prevent attacks against user privacy and company security. Luckily, a variety of hardware and software solutions exist to help protect an Intranet. The term *firewall* is a basic component of network security. A firewall is a collection of hardware and software that interconnects two or more networks and, at the same time, provides a central location for managing security. It is essentially a computer specifically fortified to withstand various network attacks. Network designers place firewalls on a network as a first line of network defense. It becomes a "choke point" for all communications that lead in and out of an Intranet. By centralizing access through one computer (known as a *firewall-bastion host*), it is easier to manage the network security and to configure appropriate software on one machine. The bastion host is also sometimes referred to as a *server*.

The firewall is a system that controls access between two networks. Normally, installing a firewall between an Intranet and the Internet is a way to prevent the rest of the world from accessing a private Intranet. Many companies provide their employees with access to the Internet long before they give them access to an Intranet. Thus, by the time the Intranet is deployed, the company has typically already installed a connection through a firewall. Besides protecting an Intranet from Internet users, the company may also need to protect or isolate various departments within the Intranet from one another, particularly when sensitive information is being accessed via the Intranet. A firewall can protect the organization from both internal and external security threats.

Most firewalls support some level of encryption, which means data can be sent from the Intranet, through the firewall, encrypted, and sent to the Internet. Likewise, encrypted data can come in from the Internet, and the firewall can decrypt the data before it reaches the Intranet. By using encryption, geographically dispersed Intranets can be connected through the Internet without worrying about someone intercepting and reading the data. Also, a company's mobile employees can use encryption when they dial into your system (perhaps via the Internet) to access private Intranet files.

In addition to firewalls, a router can be used to filter out data packets based on specific selection criteria. Thus, the router can allow certain packets into the network while rejecting others.

One way to prevent outsiders from gaining access to an Intranet is to physically isolate it from the Internet. The simplest way to isolate an Intranet is to not physically connect it to the Internet. Another method is to connect two sets of cables, one for the Intranet and the other for the Internet.

Even without a connection to the Internet, an organization is susceptible to unauthorized access. To reduce the opportunity for intrusions, a policy should be implemented that requires frequent password changes and keeping that information confidential. For example, disgruntled employees, including those who have been recently laid off, can be a serious security threat. Such employees might want to leak anything from source code to company strategies to the outside. In addition, casual business conversations, overheard in a restaurant or other public place, may lead to a compromise in security. Unfortunately, a firewall cannot solve all these specific security risks.

It should be noted that a firewall cannot keep viruses out of a network. Viruses are a growing and very serious security threat. It is necessary to prevent viruses from entering an Intranet from the Internet by users who upload files. To protect the network, everyone should run antivirus software on a regular basis.

The need for a firewall implies a connection to the outside world. By assessing the types of communications expected to cross between an Intranet and the Internet, one can formulate a specific firewall design. Some of the questions that should be asked when designing a firewall strategy include:

- Will Internet-based users be allowed to upload or download files to or from the company server?
- Are there particular users (such as competitors) that should be denied all access?
- Will the company publish a Web page?
- Will the site provide Telnet support to Internet users?
- Should the company's Intranet users have unrestricted Web access?
- Are statistics needed with respect to who is trying to access the system through the firewall?
- Will a dedicated staff be implemented to monitor firewall security?
- What is the worst-case scenario if an attacker were to break into the Intranet? What can be done to limit the scope and impact of this type of scenario?
- Do users need to connect to geographically dispersed Intranets?

There are three main types of firewalls: network level, application level, and circuit level. Each type of firewall provides a somewhat different method of protecting the Intranet. Firewall selection should be based on the organization's security needs.

2.1.7.1.1 Network, Application, and Circuit-Level Firewalls

2.1.7.1.1.1 Network-Level Firewall A *network-level firewall* is typically a router or special computer that examines packet addresses, and then decides whether to pass the packet through or to block it from entering the Intranet. The packets contain the sender and recipient IP address, and other packet information. The network-level router recognizes and performs specific actions for various predefined requests. Normally, the router (firewall) will examine the following information when deciding whether to allow a packet on the network:

- Source address from which the data is coming
- Destination address to which the data is going
- Session protocol such as TCP, UDP, or ICMP
- Source and destination application port for the desired service
- Whether the packet is the start of a connection request

If properly installed and configured, a network-level firewall will be fast and transparent to users.

2.1.7.1.1.2 Application-Level Firewall An *application-level firewall* is normally a host computer running software known as a *proxy server*. A proxy server is an application that controls the traffic between two networks. When using an application-level firewall, the Intranet and the Internet are not physically connected. Thus, the traffic that flows on one network never mixes with the traffic of the other because the two network cables are not connected. The proxy server transfers copies of packets from one network to the other. This type of firewall effectively masks the origin of the initiating connection and protects the Intranet from Internet users.

Because proxy servers understand network protocols, they can be configured to control the services performed on the network. For example, a proxy server might allow FTP file downloads, while disallowing FTP file uploads. When implementing an application-level proxy server, users must use client programs that support proxy operations.

Application-level firewalls also provide the ability to audit the type and amount of traffic to and from a particular site. Because application-level firewalls make a distinct physical separation between an Intranet and the Internet, they are a good choice for networks with high security requirements. However, due to the software needed to analyze the packets and to make decisions about access control, application-level firewalls tend to reduce the network performance.

2.1.7.1.1.3 Circuit-Level Firewalls A *circuit-level firewall* is similar to an application-level firewall in that it, too, is a proxy server. The difference is that a circuit-level firewall does not require special proxy–client applications. As discussed in the previous section, application-level firewalls require special proxy software for each service, such as FTP, Telnet, and HTTP.

In contrast, a circuit-level firewall creates a circuit between a client and server without needing to know anything about the service required. The advantage of a circuit-level firewall is that it provides service for a wide variety of protocols, whereas an application-level firewall requires an application-level proxy for each and every service. For example, if a circuit-level firewall is used for HTTP, FTP, or Telnet, the applications do not need to be changed. You simply run existing software. Another benefit of circuit-level firewalls is that they work with only a single proxy server, making it easier to manage, log, and control a single server than multiple servers.

2.1.7.1.2 Firewall Architectures

Combining the use of both a router and a proxy server into the firewall can maximize the Intranet's security. The three most popular firewall architectures are the *dual-homed host firewall*, the *screened host firewall*, and the *screened subnet firewall*. The screened host and screened subnet firewalls use a combination of routers and proxy servers.

2.1.7.1.2.1 Dual-Homed Host Firewalls A *dual-homed host firewall* is a simple, yet very secure configuration in which one host computer is dedicated as the dividing line between the Intranet and the Internet. The host computer uses two separate network cards to connect to each network. When using a dual-home host firewall, the computer routing capabilities should be disabled, so the two networks do not accidentally become connected. One of the drawbacks of this configuration is that it is easy to inadvertently enable internal routing.

Dual-homed host firewalls use either an application-level or a circuit-level proxy. Proxy software controls the packet flow from one network to another. Because the host computer is dual homed (i.e., it is connected to both networks), the host firewall can examine packets on both networks. It then uses proxy software to control the traffic between the networks.

2.1.7.1.2.2 Screened Host Firewalls Many network designers consider *screened host firewalls* more secure than a dual-homed host firewall. This approach involves adding a router and placing the host computer away from the Internet. This is a very effective and easy-to-maintain firewall. A router connects the Internet to your Intranet and, at the same time, filters packets allowed on the network. The router can be configured so that it sees only one host computer on the Intranet network. Users on the network who want to connect to the Internet must do so through this host computer. Thus, internal users appear to have direct access to the Internet, but the host computer restricts access by external users.

2.1.7.1.2.3 Screened Subnet Firewalls A *screened subnet firewall* architecture further isolates the Intranet from the Internet by incorporating an intermediate perimeter network. In a screened subnet firewall, a host computer is placed on a perimeter network that users can access through two separate routers. One router controls Intranet traffic and the second controls the Internet traffic. A screened subnet firewall provides a formidable defense against attack. The firewall isolates the host computer on a separate network, thereby reducing the impact of an attack to the host computer. This minimizes the scope and chance of a network attack.

2.1.7.2 CGI Scripting

Web sites that provide two-way communications use CGI (common gateway scripting). For example, if you fill in a form and click your mouse on the form's Submit button, your browser requests that the server computer run a special program, typically a CGI script, to process the form's content. The CGI script runs on the server computer, which processes the form. The server then returns the output to the browser for display.

From a security perspective, the danger of CGI scripts is that they give users the power to make a server perform a task. Normally, the CGI process works well, providing an easy way for users to access information. Unfortunately, it is also possible to use CGI scripts in ways for which they were never intended. In some cases, attackers can shut down a server by sending potentially damaging data through the use of CGI scripts. From a security perspective, it is important to make sure that users cannot use CGI scripts to execute potentially damaging commands on a server.

2.1.7.3 Encryption

Encryption prevents others from reading your documents by "jumbling" the contents of your file in such a way that it becomes unintelligible to anyone who views it. You must have a special key to decrypt the file so its contents can be read. A key is a special number, much like the combination of a padlock, which the encryption hardware or software uses to encrypt and decrypt files. Just as padlock numbers have a certain number of digits, so do encryption keys. When people talk about 40-bit or 128-bit keys, they are simply referring to the number of binary digits in the encryption key. The more bits in the key, the more secure the encryption and less likely an attacker can guess your key and unlock the file. However, attackers have already found ways to crack 40-bit keys.

Several forms of encryption can be used to secure the network, including: link encryption, document encryption, Secure Sockets Layer (SSL), and secure HTTP (S-HTTP). The following sections describe these encryption methods in more detail.

2.1.7.3.1 Public-Key Encryption

Public-key encryption uses two separate keys: a public key and a private key. A user gives his/her public key to other users so anyone may send them encrypted files. The user activates his/her private key to decrypt the files (which were encrypted with a public key).

A public key only allows people to encrypt files, not to decrypt them. The private user key (designed to work in conjunction with a particular public key) is the only key that can decrypt the file. Therefore, the only person that can decrypt a message is the person holding the private key.

2.1.7.3.2 Digital Signatures

A digital signature is used to validate the identity of the file sender. A digital signature prevents clever programmers from forging e-mail messages. For example, a programmer who is familiar with e-mail protocols can build and send an e-mail using anyone's e-mail address, such as BillGates@microsoft.com.

When using public-key encryption, a sender encrypts a document using a public key, and the recipient decodes the document using a private key. With a digital signature, the reverse occurs. The sender uses a private key to encrypt a signature, and the recipient decodes the signature using a public key. Because the sender is the only person who can encrypt his or her signature, only the sender can authenticate messages. To obtain a personal digital signature, you must register a private key with a certificate authority (CA), which can attest that you are on record as the only person with that key.

2.1.7.3.3 Link Encryption

Link encryption is used to encrypt transmissions between two distant sites. It requires that both sites agree on the encryption keys that will be used. It is commonly used by parties who need to communicate with each other frequently. Link encryption requires a dedicated line and special encryption software. It is an expensive way to encrypt data. As an alternative to this, many routers have convenient built-in encryption options. The most common protocols used for link encryption are PAP (Password Authentication Protocol) and CHAP (Challenge Handshake Authentication Protocol). Authentication occurs at the data link layer and is transparent to end users.

2.1.7.3.4 Document Encryption

Document encryption is a process by which a sender encrypts documents that the recipient(s) must later decrypt. Document encryption places the burden of security directly on those involved in the communication. The major weakness of document encryption is that it adds a step to the process by which a sender and receiver exchange and receive documents. Because of this extra step, many users prefer to save time by skipping the encryption. The primary advantage of document encryption is that anyone with an e-mail account can use document encryption. Many document encryption systems are available free or for little cost on the Internet.

2.1.7.3.5 Pretty Good Privacy (PGP)

Pretty good privacy (PGP) is a free (for personal use) e-mail security program developed in 1991 to support public-key encryption, digital signatures, and data compression. PGP is based on a 128-bit key. Before sending an e-mail message, PGP is used to encrypt the document. The recipient also uses PGP to decrypt the document. PGP also offers a document compression option. Besides making a document smaller, compression enhances the file security because compressed files are more difficult to decode without the appropriate key. According to the PGP documentation, it would take 300 billion years for someone to use brute force methods to decode a PGP-encrypted, compressed message.

2.1.7.3.6 Secure Socket Layer (SSL)

The Secure Socket Layer (SSL) was developed by Netscape Communications to encrypt TCP/IP communications between two host computers. SSL can be used to encrypt any TCP/IP protocol, such as HTTP, Telnet, and FTP. SSL works at the system level. Therefore, any user can take advantage of SSL because the SSL software automatically encrypts messages before they are put onto the network. At the recipient's end, SSL software automatically converts messages into a readable document.

SSL is based on public-key encryption and works in two steps. First, the two computers wishing to communicate must obtain a special session key (the key is valid only for the duration of the current communication session). One computer encrypts the session key and transmits the key to the other computer. Second, after both sides know the session key, the transmitting computer uses the session key to encrypt messages. After the document transfer is complete, the recipient uses the same session key to decrypt the document.

2.1.7.3.7 Secure HTTP (S-HTTP)

Secure HTTP is a protocol developed by the CommerceNet coalition. It operates at the level of the HTTP protocol. S-HTTP is less widely supported than Netscape's Secure Socket Layer. Because S-HTTP works only with HTTP, it does not address security concerns for other popular protocols, such as FTP and Telnet.

S-HTTP works similarly to SSL in that it requires both the sender and receiver to negotiate and use a secure key. Both SSL and S-HTTP require special server and browser software to perform the encryption.

2.1.7.4 Intranet Security Threats

This section examines additional network threats that should be considered when implementing Intranet security policies.

2.1.7.4.1 Source-Routed Traffic

As discussed earlier, packet address information is contained in the packet header. When source routing is used, an explicit routing path for the communication can be chosen. For example, a sender could map a route that sends packets from one specific computer to another, through a specific set of network nodes. The road map information contained in the packet header is called *source routing*, and it is used mainly to debug network problems. It is also used in some specialized applications. Unfortunately, clever programmers can also use source routing to gain (unauthorized) access into a network. If a source-routed

packet is modified so that it appears to be from a computer within your network, a router will obediently perform the packet routing instructions, permitting the packet to enter the network, *unless special precautions are taken*. One way to combat such attacks is simply to direct your firewall to block all source-routed packets. Most commercial routers provide an option to ignore source-routed packets.

2.1.7.4.2 Protecting against ICMP Redirects (Spoofing)

ICMP stands for Internet Control Message Protocol. ICMP defines the rules routers use to exchange routing information. After a router sends a packet to another router, it waits to verify that the packet actually arrived at the specified router. Occasionally, a router may become overloaded or may malfunction. In such cases, the sending router might receive an ICMP-redirect message that indicates which new path the sending router should use for transmission.

It is fairly easy for knowledgeable hackers to forge ICMP-redirect messages to reroute communication traffic to some other destination. The term *spoofing* is used to describe the process of tricking a router into rerouting messages in this way. To prevent this type of unauthorized access, it may be necessary to implement a firewall that will screen ICMP traffic.

2.1.8 Internet Security

John Braun

The Internet offers the ability for anyone, from an individual computer owner to a major multinational corporation, to make information and computing resources available to the world. Alas, since the Internet was initially designed with ease of communications, rather than security, in mind, care must be taken to ensure that sensitive information is not made available to the world as well. This section will discuss several aspects of protection information that one makes available via a network or the Internet.

2.1.8.1 Physical Security

Although not as trendy or exciting as some of the exotic attacks that can be made against a network on the protocol level, physical security is nonetheless important. You can have the best security software in the world installed on your network, but it may do you little good if an attacker can waltz up to a key piece of network or computing equipment and disable it!

Key pieces of network hardware, such as routers, firewalls, and servers, should be stored in a secure room with some sort of access control such as a traditional or electronic lock, card reader, or other means that can limit access to authorized individuals. Access to especially sensitive devices should be further restricted by placing them in a locked cabinet. Also, access to buildings that contain these rooms should also be controlled with security guards or other access control so that visitors can't wander about.

Exposed network cables, especially those connected to routers, hubs, and other devices that carry traffic for several other users, should also be physically protected. If an attacker has access to these cables, they could physically cut them and insert equipment that could monitor and even generate network traffic. For maximum protection, network cables should be placed inside of pressurized pipes with sensors placed on various locations along the piping. Network monitoring tools should also be used so that a break in a cable can be identified quickly.

2.1.8.2 Modems

Modems present two security threats. First, modems offer a channel for data to leave your premises, circumventing security and auditing measures that may be in place for the rest of the network. A review of services that are accessed by modem should be made, and if possible, this access should be rerouted over a secure internal network. Second, modems offer a potential method for unauthorized individuals to access your network from the outside. Since there may be a need for users who are on the road or working from home to access a network remotely, additional security measures need to be taken for

these connections. One measure is a dial-back modem, which will call back the user at a predetermined number before allowing access to the network. Another is a system where each user is provided with an electronic card that displays a random number every few minutes. A similar device that performs the same calculation to produce this number is located on the network one wishes to access. Without the card, and the ability to produce this number, the remote user is denied access.

2.1.8.3 Data Security

There are many aspects to data security. One is to prevent files from being viewed by someone other than the owner. On a multiuser system, this concern can be addressed by proper system administration. Users should not be allowed access to directories or files that do not belong to them. On a single-user system, the ability to share files over the network should be closely monitored, so that users don't inadvertently allow access of their hard drive contents to anyone who cares to look. Another aspect of data security is to protect the contents of file or network data via encryption. This can be especially important for data traveling over a network, since the data can be broadcast to other terminals or network devices that are not the intended recipients of the data. By using encryption, data that falls into the wrong hands will be unusable unless an encryption key or password is also known. Although security at the TCP/IP level is still a ways off, several third-party products provide the ability to protect and encrypt data.

2.1.8.4 Passwords

Since passwords usually comprise an initial layer of defense against an attack, they should be chosen and implemented with care. A written policy and/or enforcement by the operating system can help. Passwords should not be dictionary words, should be as long as possible, contain a series of letters, numbers, and other characters, and should be changed on a regular basis. When deciding on a policy, care should be taken to balance security needs versus ease of use. If a policy is too tedious to follow, users may end up writing their passwords down somewhere near their terminal, eliminating any sort of benefit the password policy could have offered.

2.1.8.5 Workstation Security

Unattended workstations could be a great danger to the entire system, and a security system could be completely wasted if an unauthorized person could access someone else's logged-in workstation. For that reason, users need to be aware of this danger and be properly trained how to secure an unattended workstation, either by logging off or by using a screen saver or screen lock that activates after a short amount of inactivity.

2.1.8.6 TCP/IP Security

Since the current version of TCP/IP was designed to provide a robust, standard method of moving data on a network, rather than security, one should be aware of several attacks that could compromise network security or availability.

2.1.8.6.1 IP Spoofing

Spoofing is the act of altering the contents of a TCP or IP packet header in order to trick the remote system into thinking the packet is valid. One trick is to change the source IP address of a packet to an address that is valid on a network behind a firewall or router. Older equipment that would have otherwise blocked the packet will allow it to go through since it appears to be coming from a friendly network. There are attacks where a connection can be hijacked or terminated by combining IP address spoofing with the spoofing of the SEQ and ACK fields in a TCP header. The SEQ and ACK fields help synchronize traffic between two hosts. If these fields are modified by attackers, the attackers can take over connections, while legitimate hosts lose the connection since their packets now appear to be out of order. Additional fields could be activated so that the connection is terminated prematurely.

2.1.8.6.2 Denial of Service (DoS)

Many DoS attacks take advantage of nuances in the method used to establish a TCP/IP connection. Since connections may take a while to establish, portions of the TCP/IP establishment process include timeouts so that slow equipment or busy networks will not cause a connection attempt to fail. However, a program that intentionally completes only a portion of this negotiation will result in a host waiting for a connection to complete, when it never will. While the host is waiting for the connection attempt to time out, system resources are being used. If enough of these bogus attempts are made, the host will run out of resources, and future connection attempts will be refused.

Another major type of attack involves the sending of single packets, whose contents have been modified to some unexpected or invalid data. This can result in the remote system crashing with a nasty Blue Screen of Death, system bomb, or core dump. One attack sends an ICMP (a special type of IP packet, with no TCP) echo request, also known as a ping, whose data payload is very large. Since a ping packet normally has no data associated with it, some implementations that don't expect this data will grind to a halt when receiving this type of packet. Another attack interferes with the data offset field in the TCP header, so that the remote host is tricked into trying to read packet data where none exists, also causing a crash.

2.1.9 Summary

Intranets are being used to improve the overall productivity of the organization. Important Intranet concepts covered in this chapter are summarized below:

- TCP/IP was created because of the need for reliable networks that could span the globe. Because of its reliability and ease of implementation, TCP/IP has become the standard language (or protocol) of the Internet. TCP/IP defines how programs exchange information over the Internet.
- An Intranet is based on Internet technology. It consists of two types of computers: a client and a server. A client asks for and uses information that the server stores and manages.
- Telnet, FTP, and gopher are widely used network programs that help users connect to specific computers and to transfer and exchange files.
- The World Wide Web (or Web) is a collection of interlinked documents that users can access and view as pages using a special software program called a browser. The two most popular browser programs are Netscape Navigator and Microsoft Internet Explorer.
- HTML (Hypertext Markup Language) and XML (Extended Markup Language) are used to describe the layout and contents of pages on the Web.
- Java is a new computer programming language that allows users to execute special programs (called applets) while accessing and viewing a Web page.
- A network computer is a low-cost, specialized computer designed to work in conjunction with the Internet and Java application programs.

To be effective, the Intranet must deliver quality information content. To ensure this, management must be in a proactive role in assigning staff who will keep the corporate information reservoirs on the Intranet current and relevant. The following is a checklist of some of the ways to encourage the development of a high-quality Intranet:

- Give users access to document management systems and various corporate databases.
- Distribute the responsibility of maintaining the Intranet to increase the number of staff involved in developing and enhancing Intranet content.
- Create a corporate culture based on information sharing.
- Place employee training at the center of an Intranet deployment strategy.
- Design and implement appropriate security measures as soon as possible.
- Use firewalls to control access to the network.
- Use antivirus software.

- Implement a security plan that controls the access that employees and outsiders have to the network.
- Design and implement CGI scripts with security in mind.
- Encourage users to encrypt files before sending confidential data across the Internet/Intranet.

2.2 Virtual Private Networking Solutions

Endre Sara

Today's large corporate networks are geographically distributed and clients or employees need access to corporate information from different locations. The cost of a long-distance dialup session is very high, and it is also not efficient to deploy point-to-point connection between each possible location.

Virtual private network (VPN) is a concept of securely transferring sensitive corporate information among various geographically dispersed sites over a public network, such as the Internet. The market numbers for these services tell a success story and a win/win situation for both providers and users. But while the market numbers are good, numerous concerns remain. They are:

- There are not enough integration services to help users deploy VPNs.
- The products are not yet interoperable.
- Security standards are not yet unique.
- There are different protocol standards
- Many users are not yet fully comfortable using Internet technologies in the mainline business.

Typical users of VPNs are driving the implementation of VPN services. The most important benefits and points to consider are summarized in Table 2.2.1. Weights are not included in this consideration.

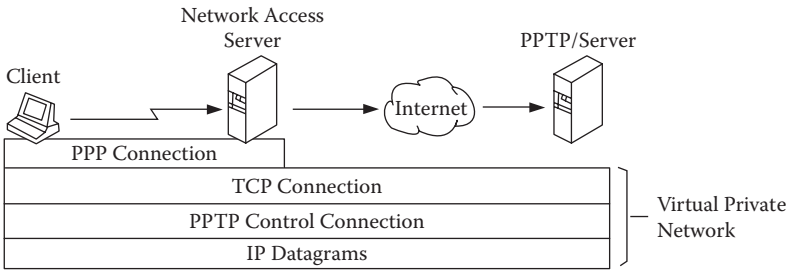
This document describes and compares the available standards and products for VPN solutions. The VPN is a network that uses a private address space that operates over another network infrastructure. It means that the VPN will use the same physical cabling, switches, bridges, and routers, but it uses a different address space. This is accomplished by encapsulating the VPN traffic (which doesn't have to be IP) into secure protocols. The emerging standards concentrate around Layer 2 and Layer 3 protocols.

2.2.1 Layer 2 Protocols

Layer 2 protocols enable the transfer of data from a remote client to the private network of an enterprise by creating a virtual private network most often across a TCP/IP-based data network. Layer 2 protocols

TABl E 2.2.1 VPN Benefits and Concerns

VPN Applications	Benefits	Points to Consider
Dial access for remote users	Outsource modems reduce dial-in costs	Client software
	Eliminate access lines	Are appropriate tunneling protocols supported in client software?
Connecting branch offices	Reduces number of dedicated lines	Encryption performance issues
	Lets IT managers consolidate central-site WAN equipment	Does VPN access-control system integrate with existing user access privileges?
Extranet	Gives trading partners and customers access to Intranet	Does system scale well?
	Makes collaborating with contractors and consultants much easier	Are there tools to handle the administrative burden of adding new users?
New business	Can create just-in-time networks for short-term projects	Interoperability of different VPN equipment
	Can give worldwide sites access much sooner than waiting for leased lines	Management of mixed equipment environment is not easy



FIGu RE 2.2.1 The PPTP tunnel.

support on-demand, multiprotocol virtual private networking over public networks such as the Internet. Internet access is provided by an Internet service provider (ISP), who wishes to offer services other than traditional registered IP address–based service to dialup users of the network.

This architecture is transparent to the end systems. In case of connecting two distant local area networks (LANs) through a VPN, the users will notice no difference while their traffic is being encapsulated in IP packets and transmitted to the remote VPN access server, which puts them back to the remote LAN. If a remote user wants to connect to the private network of the enterprise through a VPN connection, his/her computer has to support the implemented VPN protocol to be able to encapsulate the traffic. Although this encapsulation provides some security against intercepting the actual data, additional encryption should be implemented to provide secure communication.

2.2.1.1 Point-to-Point Tunneling Protocol (PPTP)

The PPTP networking technology is defined as an extension to the remote access Point-to-Point Protocol (PPP; RFC 1171). PPTP is a network protocol that encapsulates PPP packets into IP datagrams for transmission over the Internet or other public TCP/IP-based networks.

After the client has made the initial PPP connection to the ISP, a second dialup networking call is made over the existing PPP connection. Data sent using this second connection is in the form of IP datagrams that contain PPP packets, referred to as *encapsulated PPP packets*.

The second call creates the VPN connection to a PPTP server on the private enterprise LAN, referred to as a *tunnel*. This is shown in Figure 2.2.1.

The secure communication using the PPTP protocol typically involves three processes, each of which requires successful completion of the previous process:

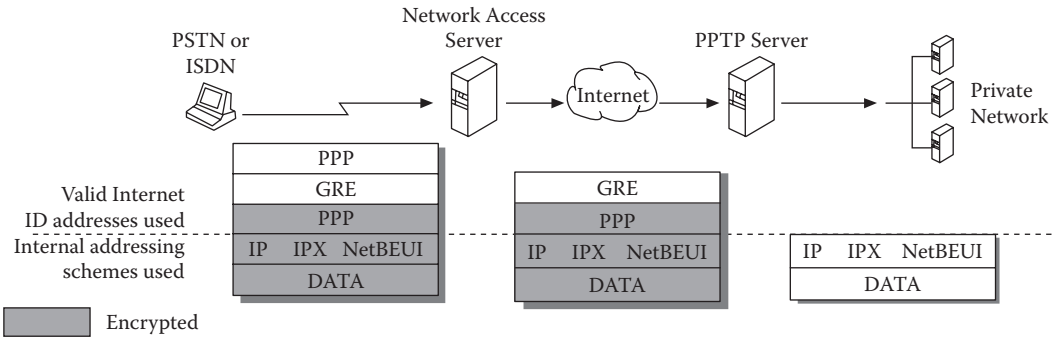
PPP Connection and Communication: The PPTP client uses PPP to connect to an ISP by using a standard phone line or ISDN line. This connection uses the PPP protocol to establish the connection and encrypt data packets.

PPTP Control Connection: Using the connection to the Internet established by the PPP protocol, the PPTP protocol creates a control connection from the PPTP client to the PPTP server over the Internet. This connection uses TCP to establish the connection and is called a PPTP tunnel.

PPTP Data Tunneling: Finally, the PPTP protocol creates IP datagrams containing encrypted PPP packets, which are sent through the PPTP tunnel to the PPTP server. The PPTP server disassembles the IP datagrams and decrypts the PPP packets, and then routes the decrypted packets to the private network.

Note that the encapsulated PPP packet can contain multiprotocol data such as TCP/IP, IPX, or NetBEUI protocols. Because the PPTP server is configured to communicate across the private network by using private network protocols, it is able to read multiprotocol packets.

Figure 2.2.2 illustrates the multiprotocol support built into PPTP. A packet sent from the PPTP client to the PPTP server passes through the PPTP tunnel to a destination computer on the private network.



FIGu RE 2.2.2 Connecting a dial-up networking PPTP client to the private network.

PPTP encapsulates the encrypted and compressed PPP packets into IP datagrams for transmission over the Internet. These IP datagrams are routed over the Internet until they reach the PPTP server that is connected to the Internet and the private network. The PPTP server disassembles the IP datagram into a PPP packet and then decrypts the PPP packet using the network protocol of the private network. As mentioned earlier, the network protocols on the private network that are supported by PPTP are IPX, NetBEUI, or TCP/IP.

An ISP's network access server may require initial dial-in authentication. If this authentication is required, it is strictly to log on to the ISP network; it is not related to the PPTP server authentication.

There are different options to provide data encryption between the PPTP client and the server. Microsoft uses the RAS (Remote Access Service) "shared secret" encryption process. It is referred as a shared secret because both ends of the connection share the encryption key. Under Microsoft's implementation of RAS, the shared secret is the user password. Other encryption methods base the encryption on some key available in public; this method is known as *public key encryption*. Microsoft's PPTP uses the PPP encryption and PPP compression schemes called Microsoft Point-to-Point Encryption (MPPE). The Compression Control Protocol (CCP) used by PPP is used to negotiate encryption. The encryption key is derived from the hashed password stored at both the client and the server. The RAS RC4 standard is used to create the 40-bit session key based on the client password. This key is used to encrypt all data that is passed over the Internet, keeping the connection private and secure.

PPTP is aimed primarily at Internet-based remote access. The main advantages are multiprotocol support and simplicity, because it functions on the Layer 2 level. This can be a preferred solution in a multiprotocol environment, but data security is a concern. The proposed standard does not provide a data encryption solution, although Microsoft has a vendor-specific solution as discussed earlier. The other difference is that compared to Layer 3 solutions, PPTP only provides a single point-to-point connection. This means that there can be no simultaneous Internet access while using a VPN connection. With multipoint tunneling, such as a Layer 3 solution discussed later, a user could have an Internet session at the same time as several VPN connections. This is also an inherent consequence from the PPTP architecture being a solution based on a client–server model, while Layer 3 solutions are based on a more general host-to-host model.

2.2.1.2 Layer 2 Forwarding (L2F)

L2F achieves private network access through a public system by building a secure "tunnel" across the public infrastructure that connects directly to a user's home gateway. Multiple corporate networks can use a single local telephone number terminated on a service provider's dialup switch or access server. The access server establishes identity, sets up a private tunnel to the user's home gateway router, and tunnels clients to that gateway. The gateway is responsible for authentication of the remote user, thereby ensuring client control of access security and addressing.

IP/UDP	L2F	PPP (Data)
Carrier Protocol	Encapsulator Protocol	Passenger Protocol

FIGu RE 2.2.3 Tunneling packet format.

A key component of the virtual dialup service is tunneling, a vehicle for encapsulating packets inside a protocol that is understood at the entry and exit points of a given network. These entry and exit points are defined as a tunnel interfaces. The tunnel interface itself is similar to a hardware interface, but is configured in software.

Figure 2.2.3 shows the format in which a packet would traverse the network within a tunnel.

Tunneling involves the following three types of protocols:

- The passenger protocol is the protocol being encapsulated; in a dialup scenario, this protocol could be PPP, SLIP (Serial Line Internet Protocol), or text dialog.
- The encapsulating protocol is used to create, maintain, and tear down the tunnel. Cisco supports several encapsulating protocols including the L2F protocol, which is used for virtual dialup services.
- The carrier protocol is used to carry the encapsulated protocol; IP will be the first carrier protocol used by the L2F protocol, because of IP's robust routing capabilities, ubiquitous support across different medias, and deployment within the Internet.

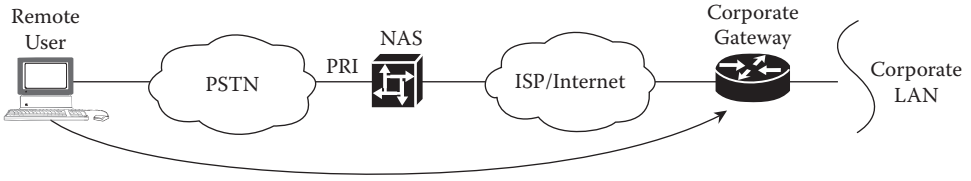
No dependency exists between the L2F protocol and IP. In subsequent releases of the L2F functionality, Frame Relay, X.25 VCs, and asynchronous transfer mode (ATM) switched virtual circuits (SVCs) could be used as a direct Layer 2 carrier protocol for the tunnel.

Cisco's L2F implementation provides several management features. End-system transparency ensures that neither the remote end system nor its corporate hosts should require any special software to use this service. Authentication is provided by dialup PPP, Challenge Handshake Authentication Protocol (CHAP), or Password Authentication Protocol (PAP), including Terminal Access Controller Access Control System Plus (TACACS+) and Remote Authentication Dial-In User Service (RADIUS) solutions, as well as support for smart cards and one-time passwords; the authentication will be manageable by the user independent of the ISP. Addressing will be as manageable as dedicated dialup solutions; the address will be assigned by the remote user's respective corporation, and not the ISP. Authorization will be managed by the corporation's remote users, as it would be in a direct dialup solution. Accounting will be performed both by the ISP (for billing purposes) and by the user (for charge back and auditing).

Figure 2.2.4 illustrates the topology of a virtual private connection using Cisco's L2F.

In a traditional dialup scenario, the ISP using the Network Access Server (NAS) in conjunction with a security server follows an authentication process by challenging the remote user for both the username and password. If the remote user passes this phase, the authorization phase can begin.

For the virtual dialup service, the ISP pursues authentication to the extent required to discover the user's apparent identity (and by implication, their desired corporate gateway). No password interaction is performed at this point. As soon as the corporate gateway is determined, a connection



FIGu RE 2.2.4 Virtual dialup topology.

is initiated with the authentication information gathered by the ISP. The corporate gateway completes the authentication by either accepting or rejecting the connection. (For example, the connection is rejected in a PAP request in which the username or password are found to be incorrect.) Once the connection is accepted, the corporate gateway can pursue another phase of authentication at the PPP layer. These additional authentication activities are outside the scope of the specification, but might include proprietary PPP extensions or textual challenges carried within a TCP/IP Telnet session.

For each L2F tunnel established, L2F tunnel security generates a unique random key to resist spoofing attacks. Within the L2F tunnel, each multiplexed session maintains a sequence number to prevent the duplication of packets.

Cisco provides the flexibility of allowing users to implement compression at the client end. In addition, encryption on the tunnel can be done using IP security (IPSec).

The advantages of L2F are similar to those of PPTP because the two solutions are similar, and they are being merged to a common standard called L2TP. L2F also has a nice advantage of connecting multiprotocol networks because of its Layer 2 functionality. But again there is no support in the proposed standard for VPN data encryption. Cisco refers to the IPSec standard as a possible encryption method for IP traffic carried with L2F. For non-IP traffic, L2F lacks the solution for security. In comparison with Layer 3 solutions, L2F provides only a single point-to-point connection, which makes parallel Internet access impossible while being connected to the VPN.

2.2.1.3 L2TP

The Internet Engineering Task Force (IETF) draft for PPTP, which is titled "Point-to-Point Tunneling Protocol" (draft-ietf-pptp-00.txt) was submitted to the IETF in June 1996 by the companies of the PPTP Forum, which includes Microsoft Corporation, Ascend Communications, 3Com/Primary Access, ECI Telematics, and US Robotics. Cisco's proposal for L2F was submitted to the IETF for approval as a proposed standard. Northern Telecom, Inc. and Shiva Corporation have announced their support for L2F. At the June 1996 IETF meeting in Montreal, the IETF PPP Extensions working group agreed to combine Cisco's proposal with PPTP proposed by Microsoft Corporation. The emerging proposed standard, Layer 2 Tunneling Protocol (L2TP), is currently drafted by Cisco Systems, Microsoft, Ascend, 3Com, and US Robotics. The L2TP draft finalized in 1999 has been supplemented with the L2TP Multicast Extension Draft (2001) and comparable extensions.

A typical connection scenario would start with the remote user initiating a PPP connection to an ISP via either the PSTN or ISDN. The ISP's access point, the L2TP Access Concentrator (LAC), accepts the connection and the PPP link is established. The ISP may now undertake a partial authentication of the end system/user. Only the username field would be interpreted to determine whether the user requires a virtual dialup service. It is expected, but not required, that usernames be structured (e.g., username@company.com). Alternatively, the ISP may maintain a database mapping users to services. In the case of virtual dialup, the mapping will name a specific endpoint, the L2TP Network Server (LNS). If a virtual dialup service is not required, standard access to the Internet may be provided.

If no tunnel connection currently exists to the desired LNS, one is initiated. L2TP is designed to be largely insulated from the details of the media over which the tunnel is established; L2TP requires only that the tunnel media provide packet-oriented point-to-point connectivity. Obvious examples of such media are user data protocol (UDP), Frame Relay PVCs, or X.25 VCs. Once the tunnel exists, an unused slot within the tunnel, a *Call ID*, is allocated, and a connect indication is sent to notify the LNS of this new dialup session. The LNS either accepts the connection, or rejects it. The initial connect notification may include the authentication information required to allow the LNS to authenticate the user and decide to accept or decline the connection. In the case of CHAP, the setup packet includes the challenge, username, and raw response. For PAP or text dialog, it includes the username and clear text password. The LNS may choose to use this information to complete its authentication, avoiding an additional cycle of authentication.

If the LNS accepts the connection, it creates a *virtual interface* for PPP in a manner analogous to what it would use for a direct-dialed connection. With this virtual interface in place, link layer frames may now pass over this tunnel in both directions. Frames from the remote user are received at the POP, stripped of CRC, link framing, and transparency bytes, encapsulated in L2TP, and forwarded over the appropriate tunnel. The LNS accepts these frames, strips L2TP, and processes them as normal incoming frames for the appropriate interface and protocol. The virtual interface behaves very much like a hardware interface, with the exception that the hardware in this case is physically located at the ISP POP. The other direction behaves analogously, with the LNS encapsulating the packet in L2TP, and the LAC stripping L2TP before transmitting it out via the physical interface to the remote user.

At this point, the connectivity is a point-to-point PPP session whose endpoints are the remote user's networking application on one end and the termination of this connectivity into the LNS's PPP support on the other. Because the remote user has become simply another dialup client of the LNS, client connectivity can now be managed using traditional mechanisms with respect to further authorization, protocol access, and packet filtering.

Accounting can be performed at both the L2TP Access Concentrator (LAC) as well as the LNS. This document illustrates some accounting techniques that are possible using L2TP, but the policies surrounding such accounting are outside the scope of this specification.

For the virtual dialup service, the ISP pursues authentication only to the extent required to discover the user's apparent identity (and by implication, their desired LNS). This may involve no more than detecting DNIS information when a call arrives, or may involve full LCP negotiation and initiation of PPP authentication. As soon as the apparent identity is determined, a call request to the LNS is initiated with any authentication information gathered by the ISP. The LNS completes the authentication by either accepting the call or rejecting it. The LNS may need to protect against attempts by third parties to establish tunnels to the LNS. Tunnel establishment can include authentication to protect against such attacks.

L2TP, like other Layer 2 protocols, does not provide any further security for data encryption, but rather refers to Layer 3 encryption techniques for IP traffic, such as IPSec.

Although L2TP seems to be the result of different Layer 2 tunneling initiatives, it still does not have the widest acceptance in the industry. Its predecessor PPTP is popular because of the large number of Windows NT users, but IPSec, a general initiative to add security to the IP protocol, has the strongest support by manufacturers and suppliers. In a multiprotocol environment there will still be a need for Layer 2 tunneling with the addition of encryption for IP traffic (using IPSec). But in an IP-only environment, Layer 3 solutions are more effective. Unless augmented with IPSec these Layer 2 solutions cannot support extranets, because extranets require keys and key management.

In comparison with other Layer 2 protocols, L2TP has better features, such as an addition to PPTP and L2F, and the support for ATM or SONET as an underlying transmission medium, which is only planned for the other two protocols.

2.2.2 Layer 3 Tunneling Protocols

In the previously discussed architecture the PPP connection begins at the remote client and terminates at the corporate network's L2TP server, going through the ISP access point and the corporate gateway.

Layer 3 tunneling proposes a different scenario initiating the PPP connection from the remote client, but terminating it at the ISP. This requires the reencapsulation of the PPP frame and transmitting the Layer 3 information only to the corporate access router. This access router does not need to support any additional standard, but it acts only as a simple router. The difference from traditional dialup services is that the ISP's IP gateway will provide the IP address to the client. It can roam with this address as long as the IP gateway does the reframing of the IP PPP packet and sends it to the corporate gateway as a standard IP packet.

The advantage of the latter scheme is that there is no need to support L2TP at either the remote client end or at the corporate gateway end. The remote client only needs a standard IP stack, and the corporate

gateway acts as a simple IP router. In this case, if the remote node is a router connecting a subnetwork to the corporate network, the packets can be routed just like any other traffic trough the ISP network. In either case an additional functionality is needed to provide security on the IP Layer (network layer). Although L2TP can encrypt the PPP packets on Layer 2, it does not claim to be secure against denial of service attacks or man-in-the-middle attacks (someone modifying the PPP frames in the tunnel).

2.2.2.1 IPSec

IPSec is a protocol suite defined by the IETF working group on IP security to secure communication at the Layer 3 (network layer) between communicating peers. The goal of the IPSec protocol suite is to provide secure tunneled transport of IP data only. Essentially, it takes private IP packets, performs data security functions such as encryption, authentication, and integrity, then wraps these secured packets in other IP packets for transport across the Internet. Key management functions also will be a part of the IPSec protocol suite. The IETF has issued five requests for comments—RFC 1825 through 1829. An interesting note is that if IPv6 succeeds in replacing IPv4, IPSec will be the automatic Internet VPN standard since it is integrated into the IPv6 specifications.

Like the Layer 2 VPN protocols, IPSec works as a LAN-to-LAN and dialup-to-LAN solution. It is designed to support multiple encryption protocols, a feature that allows users to choose a desired amount of data privacy. Obviously, IPSec will only be of value to companies that want to tunnel IP exclusively since it does not support other data protocols.

There are several different scenarios where IPSec can be used. In case two hosts A and B want to communicate with each other through a firewall, the host A can tunnel packets to the firewall, the firewall can decrypt/authenticate the packets, and send them to B based on its rules. In a different setup there can be a secure tunnel between host A and host B, where the firewall is authorized to act as a key management proxy, and has the capability to decrypt the packets and apply its packet filtering policy. A third setup is a combination of the previous two, where the inner payload is secured from host A to B, and the outer payload is secured and tunneled through the firewall. The advantage of this scheme is that the firewall is able to authenticate packets and decide whether to allow the packet without applying its filtering rules. This is typical of what happens today, where an employee gets into the network via dialup PPP.

There is a different scheme, where packets have to be secured while traveling the Internet. In this case IPSec will secure packets between two or more border routers of a topologically distributed organization. In this case since security associations are set up between the border routers, any traffic should go through these routers. All packets between the two routers must contain valid IPSec, otherwise they will be dropped.

A growing number of VPN, security, and major network companies either support or plan to support IPSec. It is also strongly supported by a user group consisting of manufacturers and suppliers. Although it deals with IP-only traffic, it is the most often recommended or chosen solution to ensure privacy in VPN communication. It can be implemented as a single Layer 3 solution, but it can be implemented over a Layer 2 solution to provide data encryption for IP traffic. It is capable of maintaining multiple tunnels, including simultaneous VPN and public access connection inherently from its general host-to-host model. It can also support extranets with its built-in key management functionality, which is missing in other Layer 2 solutions.

2.2.2.2 Mobile IP

Mobile IP is intended to enable nodes to move from one IP subnet to another. It is just as suitable for mobility across homogeneous media as it is for mobility across heterogeneous media. That is, Mobile IP facilitates node movement from one Ethernet segment to another as well as it accommodates node movement from an Ethernet segment to a wireless LAN, as long as the mobile node's IP address remains the same after such a movement.

Mobile IP introduces the following new functional entities:

- **Mobile node:** A host or router that changes its point of attachment from one network or subnetwork to another. A mobile node may change its location without changing its IP address; it may continue to communicate with other Internet nodes at any location using its (constant) IP address, assuming link-layer connectivity to a point of attachment is available.
- **Home agent:** A router on a mobile node's home network that tunnels datagrams for delivery to the mobile node when it is away from home, and maintains current location information for the mobile node.
- **Foreign agent:** A router on a mobile node's visited network that provides routing services to the mobile node while registered. The foreign agent detunnels and delivers datagrams to the mobile node that were tunneled by the mobile node's home agent. For datagrams sent by a mobile node, the foreign agent may serve as a default router for registered mobile nodes.

A mobile node is given a long-term IP address on a home network. This home address is administered in the same way as a permanent IP address is provided to a stationary host. When away from its home network, a care-of address is associated with the mobile node and reflects the mobile node's current point of attachment. In this case Mobile IP uses protocol tunneling to hide a mobile node's home address from intervening routers between its home network and its current location. The tunnel terminates at the mobile node's care-of address. The care-of address must be an address to which datagrams can be delivered via conventional IP routing. At the care-of address, the original datagram is removed from the tunnel and delivered to the mobile node.

The Mobile IP standard contains a very strong authentication between the mobile node and the home agent to authenticate themselves. The default algorithm is keyed MD5, with a key size of 128 bits. This will result in the mobile node's traffic being tunneled to its care-of address. But the standard does not provide privacy protection; it rather refers to other IP encryption standards, such as IPSec.

The Mobile IP standard, as with other Layer 3 standards, has the advantage of scalability, security, and reliability, but they are more complex to develop and, inherent in their Layer 3 functionality, they only support a specific protocol, which is IP in this case.

2.2.3 Frame Relay

Traditionally, VPNs were provided in the form of broadband packet switched services, such as Frame Relay, X.25, or ATM. Now, with growth of the Internet as a viable service infrastructure, it is possible to run VPNs over an alternative protocol. As it can be seen from the previously discussed standards, these latter solutions tend to be more popular. The reason is fairly simple: while traditional VPNs are very useful for fixed LAN-to-LAN connectivity, they do not easily accommodate individual users whose only access to the outside world is in the form of their PC, a modem, and the public switched telephone network. VPNs that run over IP are easily accessed by these users.

With any choice of the above-mentioned protocols, virtual circuit connections can be defined between remote locations, and the LAN traffic can be bridged over these circuits. This usually provides an emulated Layer 2 network for the users, which can be used to transmit multiprotocol traffic. Although proper authentication and encryption also has to be taken care of in these solutions, these networks are not as sensitive to security threats, as they are usually a private ATM, Frame Relay backbone of the provider, but in any case less public than the Internet in the previous solutions.

The advantages of this solution are the built-in quality of service (QoS) guarantees that are part of the virtual circuit definitions. Where the bandwidth availability could be argued in the past for an Internet-based VPN solution compared to these solutions, nowadays the ISPs can also provide high-speed Internet connections, especially when they utilize only their backbone network to provide VPN services. The disadvantage of the Frame Relay, ATM, or X.25-based VPN services is that these means of access should be available at each location where the VPN service has to be used. It usually means special equipment, wiring, and additional management needs. The user mobility is not solved with these solutions.

2.2.4 Layer 2 or Layer 3 Comparison

The goal of Layer 2 tunneling protocols is to transport Layer 3 protocols such as AppleTalk, IP, and IPX across the Internet. To achieve this, the architects of PPTP and L2F leveraged the existing Layer 2 PPP standard, which is designed to transport different Layer 3 protocols across serial links. In these schemes, Layer 3 packets are encased in PPP frames, which are then encased in IP packets for transport across the Internet.

From a security standpoint, Layer 2 tunneling protocols are insufficient to be secure VPN solutions on their own. None of these protocols provide the data encryption, authentication, or integrity functions that are critical to maintaining VPN privacy. The L2TP specification disclaims any data security functions and refers IP data security to IPSec, but no serious security provisions or references are made for the other Layer 2 protocols. In addition, none of these protocols provide a mechanism for key management, which limits their scalability.

PPTP and L2F are vendor-specific, proprietary protocols, so interoperability is limited to products from supporting vendors. In contrast, L2TP is a multivendor effort, so interoperability is not as much of a problem. It is important to note that when utilizing tunneling protocols besides IP, users will have to rely on vendor-specific data security features. On the upside, PPTP, L2F, and L2TP can transport multiple protocols. They also function both in LAN-to-LAN and dialup-to-LAN tunneling modes, allowing them to cover the applications most desired for VPN.

In case of Layer 3 tunneling there is no need for a globally unique address space, which is a requirement with L2TP for remote client address assignments. This global address space doesn't have to be registered, since it is seen from the corporate network but not visible from the public network (Internet).

Another difference is that the Layer 2 tunneling causes an additional overhead as opposed to Layer 3 tunneling, which only sends regular IP packets after the ISP's IP gateway to the corporate network. (The tunneling takes place only between the remote client and the IP gateway.) This might make Layer 3 solutions more scalable, but with the loss of the additional features, such as the freedom of network protocols that can be used over the PPP layer.

References

Bay Networks: http://www.baynetworks.com/Solutions/vpn/
Cisco: http://www.cisco.com/warp/public/779/servpro/solutions/vpn/
Microsoft: http://www.microsoft.com/ntserver/nts/commserv/exec/feature/VPNFeatures.asp
Shiva: http://www.shiva.com/remote/vpn.html
3Com: http://www.3com.com/enterprise/vpn/

2.3 Web-Enabled Data Warehousing

Dermot Murray

2.3.1 Introduction

Since the early 1980s, business analysts have identified the inherent value in analyzing the huge amounts of data generated in production online transaction processing (OLTP) systems. Hidden deep inside such data repositories is key information that can make a product or service more marketable, a customer more profitable, and processes more efficient. This process of analyzing production data in order to unearth that critical business intelligence is known as decision support systems (DSS). The problem has always been getting at that information in such a way that it adds value to the business as a whole. William H. Inmon promoted the concept of data warehousing in the 1980s as a means of separating operational systems from DSS in order to extract the data necessary for business intelligence without

impacting mission-critical processing systems. Since then, there has been a huge growth in the market for data warehousing and decision support tools, with all of the major database vendors such as Oracle, IBM, and Sybase developing products to satisfy the demand.

This paper examines the next logical step in the evolution of data warehousing technology; i.e., Web-enabling the applications that provide access to this key business data. With the growth in the use of intranets, extranets, and the Internet in general, such Web-enabled data warehousing products have revolutionized the way business analysts generate the reports and charts they need in order to analyze the trends and patterns in the operational data. We will explore the concepts behind Web-enabled data warehousing and look at the technology that makes it all happen.

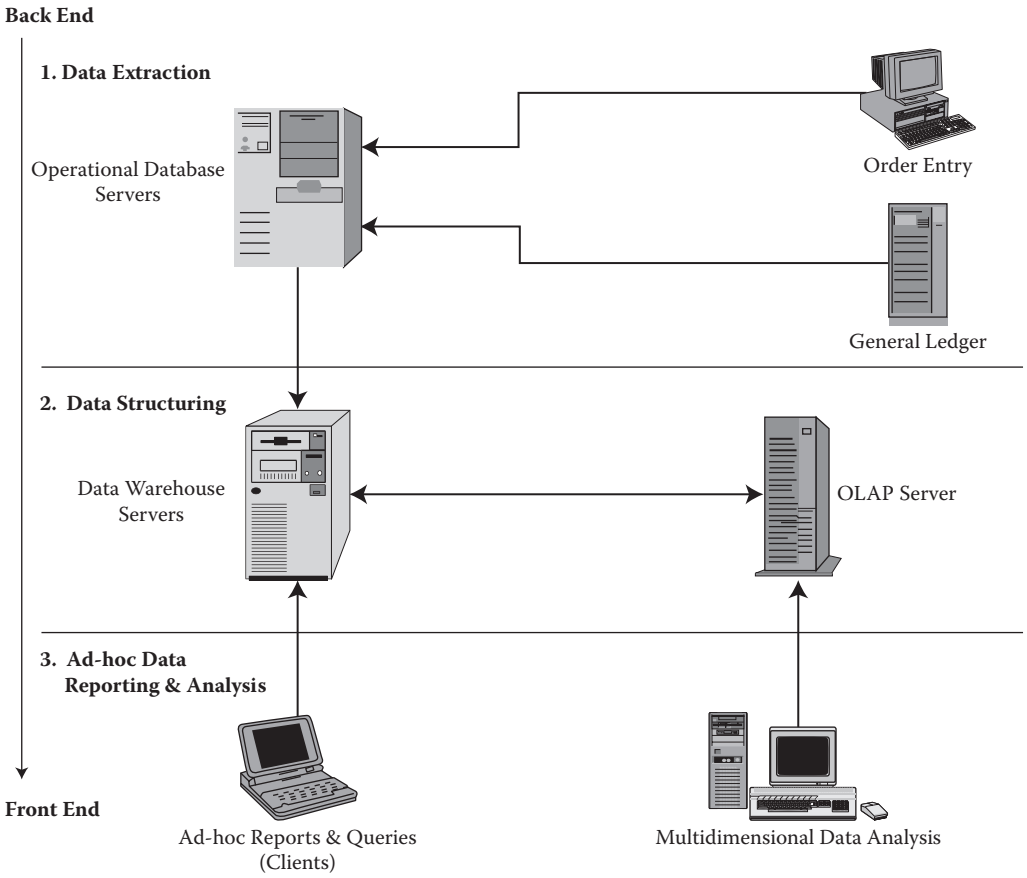
2.3.2 Data Warehousing Overview

2.3.2.1 Concepts

The concept behind data warehousing first emerged in the client–server platform environment of the 1980s. Although Inmon first started writing about data warehousing in 1981, it wasn't until the early 1990s that industry giants such as IBM, Oracle, and Sybase started taking the technology seriously. Today, the worldwide market for data warehousing products approaches $8 billion with growth expected. Driving the popularity of data warehousing is the need for executives and business analysts to gain competitive advantage from analyzing trends and patterns hidden within the mountains of data that companies generate in their day-to-day transactions. Data warehousing is the process of extracting data from various OLTP systems into a centralized format that can be analyzed using DSS and executive information system (EIS) tools, which provide more business-specific and powerful queries for higher-level managers and executives. Collectively, these tools are known as online analytical processing (OLAP) or multidimensional analysis (MDA). The data warehouse itself can either be a single or distributed specialized database management system (DBMS) that contains replicated data from different sources within the organization. This data is usually extracted from internal data production sources such as OLTP and enterprise resource planning (ERP) systems, but is increasingly from external sources such as Dow Jones, Reuters, and even the Internet itself. The data is typically cleansed and transformed to a format that can be analyzed by DSS tools. Information about the content and format of the data is stored as *metadata* in information directories that can be accessed by both business analysts and database administrators (DBA) alike. Figure 2.3.1 demonstrates the various steps involved in data warehousing.

2.3.2.2 Advantages

The main advantage of a data warehouse is that it provides executives with up-to-date data that can be queried for reporting and analysis in order to assist them in making strategic decisions about the operations of their organizations. The SQL (Structured Query Language) queries generated by OLAP tools are typically very sophisticated and can contain multiple criteria such as sales by region, state, and customer. The results of these queries are displayed as reports or charts that can then be presented in a concise form to executives who typically don't want all of the detail that is contained in the original production data. Due to their multiple-criteria nature, these queries are usually long-lived and may even result in lost or stray processes that could jeopardize the data integrity of a real-time production environment such as OLTP or ERP. Therefore, data warehouses are usually separate DBMSs that store replicated and transformed copies of the production data and are not required to provide the same fast response times that are necessary for mission-critical OLTP systems. However, the new genre of operational data stores provides faster response to DSS/EIS queries by using near-real-time transactional data for the most current query and analysis. In addition, DSS/EIS tools provide drill-down capabilities that allow executives to get more detailed information on a particular trend or subject matter by generating more detailed queries on that subject. Data warehouses also support data mining tools that provide analysts with the ability to discover unexpected patterns in data by using fuzzy logic searches.



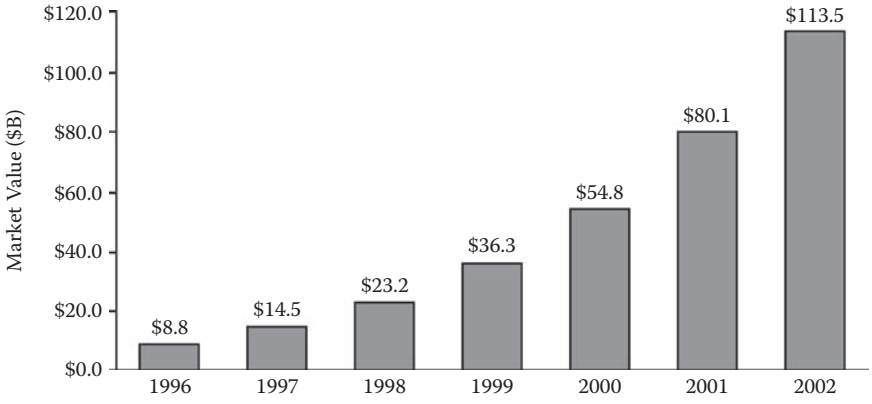
FIGu RE 2.3.1 Data warehouse/OLAP environment.

2.3.2.3 Data Marts

A very popular form of data warehousing is the data mart, which is typically a subset of the data warehouse that contains data of specific interest to a particular department, such as marketing, sales, or human resources. These data marts are less expensive to build and maintain than enterprisewide data warehouses and offer immediate business value to the departments that they service. However, the very independence that data marts provide tends to act as an obstacle to true enterprise integration in information reporting and analysis. While the debate between centralized enterprisewide data warehousing and decentralized data marts has consumed industry advocates for the last ten years, a compromise concept of dependent data marts has potentially bridged the gap. This hybrid solution provides for the extraction of data from the centralized data warehouse to the departmental data marts so that each department is working off the same enterprise data, albeit formatted to their individual needs.

2.3.2.4 Future Growth

In spite of the technical and organizational challenges that data warehouse implementations pose, the ultimate benefits that the technology offers has led to a huge growth in its adoption. Most global firms have implemented data warehouses. The future of data warehousing also seems secure with market estimates expecting growth to over $100 billion by 2002 (Figure 2.3.2). This growth has been realized and continues globally. One of the biggest factors in the expected growth of data warehouse implementation is the ease of access to timely reports and queries that Web-based data warehouse tools provide.



FIGu RE 2.3.2 Worldwide data warehousing solutions market (Source: Palo Alto Management Group).

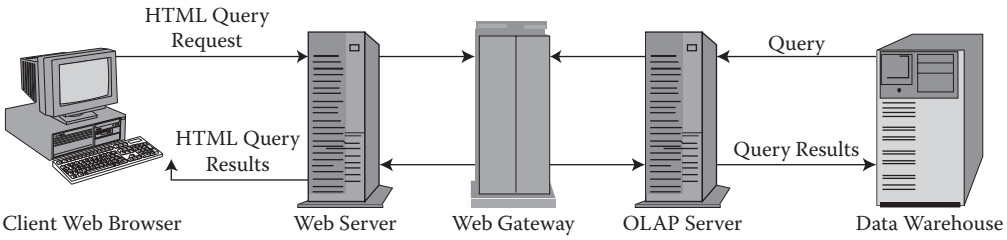
2.3.3 Web-Enabled Data Warehousing

2.3.3.1 Benefits

Perhaps the most perplexing aspects of implementing a data warehouse strategy have been the escalating costs of installing and maintaining the GUI-based clients that are used to generate the queries. The average cost of implementing a data warehouse project can take up perhaps as much as 20% of the overall IT budget. One solution to minimizing the costs associated with data warehouse implementations is Web-enabled data warehousing, a concept that has grown in popularity since it first appeared in the mid 1990s. These tools have given IT departments a more cost-effective option for allowing access to the enterprisewide data warehouses for a larger group of users. By allowing users to generate queries and reports through their Web browsers, IT departments can roll out data warehouse installations in a much shorter period of time by simply allowing the users to have the appropriate level of access to the DSS application server. With the advent of virtual private networking (VPN) and subsequent extranet technology, various levels of data warehouse access can even be provided to external users such as suppliers and customers. This is particularly critical in today's supply chain environment where companies have to work closely with their strategic partners, i.e., suppliers and customers, and therefore have to provide a certain level of access to their enterprise data warehouses for reporting and analysis. This Web-enabled approach has had the effect of reducing the costs associated with hardware and software by using the "thin" client approach, as opposed to the "fat" client requirements of the previous client/server model. This server-centric model also has the effect of reducing software licensing costs by switching from a per-seat basis to a more cost-effective server-based licensing model. The move away from dependence on fat desktop clients to the thin browser-based client model may also allow the new breed of Internet appliances, such as mobile phones and network computers, to be able to access information from data warehouses over the Internet. For instance, the advent of "push" technology could allow a sales executive to be paged if sales figures for a certain region fall below a predefined threshold.

2.3.3.2 Making It Happen

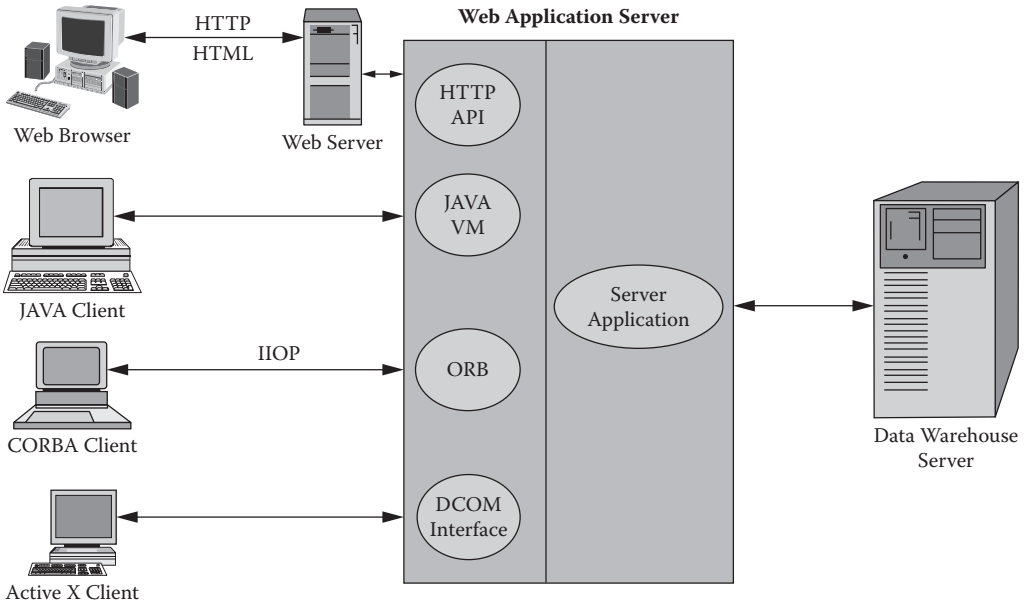
Web-enabled data warehousing involves a combination of HTML, XML, HTTP, and mobile component-based technology, typically Java or ActiveX. In this model, a user can generate an SQL query using a HTML/XML form that embeds the control information for the search criteria into a HTML/XML query and sends it to the remote Web server, which in turn passes the request to a Web gateway server. The gateway server converts the HTML/XML query into an OLAP-specific request and passes it to the OLAP server, which executes the query directly against the data warehouse. Typically the Web gateway and OLAP server are bundled into the vendor-supplied DSS server application package. When



FIGu RE 2.3.3 Web-based client–server architecture.

that application retrieves the resulting data, it sends the data embedded in an HTML/XML file back to the Web server, which forwards it back to the Web client. Some DSS servers also store predefined DSS objects such as reports and queries that can be executed by the user to generate more standardized reports. This type of interaction is similar to three-tier client–server architecture, and in fact most industry analysts would argue that Web-enabling traditional client–server applications are just the next step in the evolution of the client–server application—the intergalactic model (see Figure 2.3.3).

The interaction between the DSS server application and data warehouse is provided using distributed-object-computing protocols, such as CORBA, IIOP, or DCOM, which provide the client–server state management unlike HTTP, which is a stateless protocol. In some cases, the DSS server downloads client-based applications, such as Java applets or ActiveX controls, which actually run within the Web browser's memory space and interact directly with the application server. The purpose of this is to provide an improved graphical user interface to the end user in situations where plain old HTML is not sufficient for multidimensional data visualization, such as graphs and charts. These applets also provide more robust object-based communications between client and application server by using CORBA Internet Inter-ORB Protocol (IIOP) and Java remote method invocation (RMI), with HTTP only being used for downloading the applets to the client. Figure 2.3.4 demonstrates the different levels of interaction among the client, application server, and data warehouse server using both the Web and object-based protocols such as CORBA and DCOM.



FIGu RE 2.3.4 Object-based architecture (Source: Adapted from *DB2 Magazine*).

The advent of Web-enabled data warehousing has spawned the growth of information delivery, whereby specific business information is "pushed" to information consumers at predefined intervals. In effect, consumers subscribe to information by searching the metadata stored in the information directories using Java/ActiveX agents or Web-based search engines in order to locate the information that is of interest to them. The information delivery itself can either be schedule driven, meaning that the appropriate decision-support objects are executed at user-defined intervals and the resulting information delivered to a Web server, or event driven when particular business events occur. These user-defined events, also known as triggers, could also indicate that new data has arrived from the source DBMS. "Push" technology is also being incorporated into the channel method of delivering information, whereby a business analyst can subscribe to a channel, which in turn delivers the most up-to-date information on a particular subject.

2.3.3.3 Obstacles and Limitations

Although Web-enabled data warehousing has revolutionized the distribution of enterprisewide business information, it has not been the panacea for all data warehouse access problems. There are limitations to what "power users" can do with Web-enabled DSS tools, and in those cases, the traditional desktop approach is still being used. Such functionality as drill-down analysis and multidimensional queries are still not as effective with the Web-based tools, and users who need this functionality still require the more mature client–server tools to perform these analyses. As Web-enabled tools become more sophisticated, this problem will be rectified, but in the meantime, IS managers must deal with the reality that both Web-enabled and desktop data warehouse access must coexist for the foreseeable future.

Ironically, the ActiveX and Java components that are designed to make user interaction with the data warehouse environment possible can cause incompatibility problems if the end user's Web browser cannot support them. This might be the case if the browser is an older version or if the particular downloadable component only works with either Netscape or Internet Explorer. Of course, the solution to this problem is to standardize Web browsers and plug-in components—a daunting task for organizations with potentially thousands of users, but nevertheless manageable.

Security has always been a problem with enabling external users to access corporate information. As mentioned earlier, the growth in supply chain management has dictated that companies work in tandem with their suppliers and customers in order to develop their products and services. This trend has forced companies to open up their corporate information systems, including data warehouses, to these external users. Of course, this has led to fears that such mission-critical and confidential information could be compromised in transit between users across a public network such as the Internet. Another potential security hole is the remote access required by mobile users, such as sales agents, who need to pull up reports using dial-up connections over the Internet. Improvements in corporate firewall and VPN technology, coupled with better encryption algorithms and public key infrastructure, have eased these fears. However, organizations implementing Web-enabled data warehousing must constantly be vigilant not only for potential hackers from the outside but also for internal users who should only have the level of access to reports and data that they require. One potential solution to this problem that has been promoted since Web-enabled data warehousing began is to distribute the data among data marts, thereby limiting the users' access to only the data stored in those data marts.

Like every new technology, Web-enabled data warehousing has its downside as illustrated above. However, with vendors addressing these issues all the time, the usage of this technology has had a phenomenal growth rate and while traditional client–server warehouse access is not going away any time soon, Web-enabled DSS tools are making strong inroads into the marketplace.

2.3.4 Vendors

This burgeoning sector of the industry has produced a number of leading products from both established and startup vendors. While most of the major database vendors have been very proactive in

providing Web-based access to their data warehouse products, the market for packaged applications has come mostly from start-up vendors like MicroStrategy. As there are too many vendors to mention in this market segment, I will discuss a sample of the products being offered by these different categories of vendors.

2.3.4.1 Major Database Vendors

Oracle Corp., arguably the biggest name in the relational database market, has its presence thanks to the Oracle Express suite, which includes the Express Server and Express Web Agent modules. This product was actually acquired from Information Resources in 1995. The Oracle Express Server is the actual back-end OLAP engine that performs the end-user queries. This is the component that provides the interaction between the Express OLAP Server and the Web server. This module takes advantage of the Network Computing Architecture model, which is Oracle's blueprint for a three-tiered thin client environment, and the Express Stored Procedure Language cartridge, which provides the communication between the Web server and the Express Server. The Express Agent Developer's Toolkit, which comes with the Express Web Agent, supports the development of both HTML pages and components such as Java and ActiveX to produce customizable reports and analyses for the end user.

For their part, IBM provides Web access to the DB2 OLAP Server through the use of its Net.Data development platform. This product supports standard SQL statements as well as C++ and Java enablers that allow developers to write macros that automate SQL queries. On the server side, it supports FastCGI, which is a high-performance Web server interface that provides better performance for Net.Data applications. In addition, IBM integrated its OLAP Server with Hyperion's Essbase Web Gateway, which is a Web application server component similar to Oracle's Web Agent.

As for the other major database vendors, it appears that they have been slow to fully integrate their OLAP products with Web access, including Microsoft. For their part, Informix has tried to buy its way into the Web access market by acquiring Red Brick Systems, which had teamed up with Web application specialists Caribou Lake to deliver Web-based OLAP products. Therefore, this is still a market segment where the niche DSS vendors have been able to take the lead, at least for now.

2.3.4.2 Start-up DSS Vendors

MicroStrategy, Inc., once the market leader in Web-based OLAP products, developed a suite of DSS applications that interacts with DBMS platforms such as Oracle, Informix, and DB2 among others. This suite includes the DSS Web Server and DSS Broadcaster products, which use the Internet and the World Wide Web as the communications medium. DSS Web Server is a Web-based interface that works with the company's DSS Server OLAP engine to allow users to generate reports and analyses over the Web. Because the company uses ActiveX and Java components in the DSS Web Server they have been able to provide features such as drill-down analysis and report pivoting.

InfoSpace, Inc., is the developer of SpaceOLAP, a fully Java-compliant solution that provides client/server-like interfaces for reporting and analysis. It includes a Java Application Server that resides on the Web server and supports data extracts from Oracle Express and other OLAP servers.

Information Builders services the Web-enabled data warehousing market through its WebFOCUS product line. WebFOCUS provides reporting and publishing via the Web from both legacy databases and data in ERP applications. The company takes advantage of Java applets, which provides for customized reporting and a Java Developers Workbench for designing customized reports from a Web browser.

2.3.4.3 Established DSS Vendors

The more established data warehousing companies such as Prism Solutions and Brio Technology have also developed Web-based interfaces in their OLAP products. The Brio Enterprise Server suite provides two modules that allow the user to extract information from data warehouses via the Web—the OnDemand Server and the Broadcast Server—competitor products to MicroStrategy's DSS Web Server and DSS Broadcaster, respectively. Prism Solutions added Web-enabled functionality to its Warehouse

TABl E 2.3.1 DSS Vendors and Products

Vendor	Product(s)	Technologies Supported
Oracle Corp.	Express Server (Web Agent)	HTML, Java, ActiveX
IBM	Net.Data, DB2 OLAP (Hyperion Essbase)	SQL, C++, Java, CGI
MicroStrategy	DSS Server OLAP, DSS Web Server, DSS Broadcaster	HTML, Java, ActiveX
InfoSpace	SpaceOLAP	Java, HTML
Information Builders	WebFOCUS	Java
Brio Technology	Enterprise Server (OnDemand Server, Broadcast Server)	CGI, Java Runtime Environment (JRE)
Prism Solutions	Warehouse Directory (Web Access module)	Java, XML

Directory product when it introduced the Web Access module, which allows users to view and query the Directory from their Web browsers. Having been acquired by Ardent Software, Inc., this module has since been integrated as a standard feature of the Warehouse Directory package and Extensible Markup Language (XML) is understood to be the conduit for sharing access to corporate data warehouses.

Table 2.3.1 summarizes the list of vendors and products mentioned above.

2.3.5 Future Trends

The trend to add more functionality to Web-based DSS tools will continue to grow as more companies realize the benefits of implementing Web-based access to their corporate data warehouses. However, don't expect to see the demise of pure client/server-based DSS tools any time soon as analysts predict that a coexistence of both forms of access will prevail for a time.* As mentioned earlier, the main reason is that there are still features and functionality available with mature client/server-based tools and not with Web access tools. DSS vendors are working hard to add that functionality to their Web-based offerings so that in the future, even power users will be able to get the reports and analyses they need via the Web. Overall, the emphasis will shift away from desktop access and more toward the centralized "intergalactic" Web-based model as data warehouse implementations become larger and more users require access.

2.3.5.1 Web Farming

One area of interaction between data warehouses and the Web that is set to grow steadily in the next few years is the concept of Web farming. This is defined as the search for "systematic business intelligence by farming the information resources of the Web, so as to enhance the contents of a data warehousing system."† Essentially, it is the process of using the Web not as a means of distributing data warehouse information to the end user but as a means of obtaining the raw data that goes into the warehouse itself. Such external data sources include commercial databases, such as the IBM database of patents, and public databases, such as the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) operated by the U.S. Securities and Exchange Commission to track publicly traded companies.‡

Similar to extracting production data from OLTP systems into data warehouses, Web data must be refined to be suitable for use by the warehouse applications. This process, known as *acquisition*, can vary depending on the sources of the Web content and is crucial in ensuring the usability and reliability of the data content as it becomes part of the decision-support structure. While this sector is still in its infancy, it won't be long before the major vendors turn their expertise to developing products that will make this form of data collection more reliable and efficient.

---

* "Self-Storage; Lower Data Warehouse-Management Costs with Web Access," Communications News, August 1998.

† Richard D. Hackathorn, Web Farming for the Data Warehouse (San Francisco, CA: Morgan Kaufmann Publishers, 1998).

‡ "Routing the Web for Your Data Warehouse," DBMS, August 1998.

2.3.6 Conclusion

The value of using the World Wide Web to disseminate decision-support information is evident from the growth in the use of Web-based DSS and OLAP tools. By providing access to corporate data warehouses over the Web, companies are empowering a larger user base with the necessary tools to analyze the business data. This in turn has opened up the decision-making process to more people within the organization. The lower costs of rolling out and maintaining Web-based data warehouse access has made this medium very attractive to companies that wish to gain the most value from their data warehouses. In addition, the trend to open up information systems to the customers and suppliers who are part of the supply chain has made easier access to data warehouse information more critical to these external users.

While this technology has grown in leaps and bounds and looks set to take over as the predominant form of data warehouse access, IS departments will still have to grapple with coexistence between Web-based and client/server-based access. But with the improving feature set of the various products being offered by the many vendors in this arena, Web-enabled data warehouse access will become the *de facto* access medium for this powerful business information.

References

Carrickhoff, R. 1997. SpaceOLAP. *DBMS* (August).

ComputerWire PLC. 1998. End-User Query and Reporting Tools. London: ComputerWire, PLC (March).

Datamation. 1997. *Data warehousing management/productivity tools*. White paper. Datamation.

Davis, B. 1999. MicroStrategy's Revenue, profit soars. *InformationWeek* (April).

Fournier, R. 1998. *A Methodology for Client/Server and Web Application Development*. Upper Saddle River, NJ: Yourdon Press.

Hackathorn, R. 1998. *Web Farming for the Data Warehouse*. The Morgan Kaufmann Series in Data Management. San Francisco, CS: Morgan Kaufman.

Hackathorn, R. 1998. Reaping the Web for your data warehouse. *DBMS* (August) 11(8): p. 36.

Koch, C. 1999. The middle ground. *CIO Magazine* (January).

Orfali, R., D. Harkey, and J. Edwards. 1996. *The essential client/server survival guide*. New York: John Wiley & Sons, Inc.

Palo Alto Management Group, Inc. 1998. *Database solutions white paper* (July).

PR Newswire Association, Inc. 1998. Red Brick and Caribou Lake Software team up (December).

Reinauer, R. 1998. Self-storage: Lower data warehouse-management costs with Web access. *Communications News* (August).

Row, H. 1996. Just browsing thanks. *CIO Magazine* (October).

Schroeck, M. 1998. Data warehousing is worth the investment. *InternetWeek* (June).

Schwartz, S. 1997. Warehousing and the 'Net—Marriage made in heaven. *Insurance & Technology* (July).

Scott, J. 1998. Warehousing over the Web. Association for Computing Machinery, *Communications of the ACM* (September).

Vizard, M. 1999. Ardent's Peter Fiore is passionate about data warehousing. *InfoWorld* (May).

White, C. 1998. Warehouses Webicum: Evolution of a species. *DB2 Magazine* (April).

White, C. 1998. Building Web information systems. *Byte Magazine* (July).

2.4 Web Performance Management

Kornel Terplan

The Web plays a significant role for both enterprises and private consumers. In the years to come, Web services and Web applications will dominate how enterprises do business in both business-to-business

(B2B) and business-to-consumer (B2C) domains. The business impacts of Web performance are the focus of more attention than ever. The big question is how many resource enterprises and customers will be able to control.

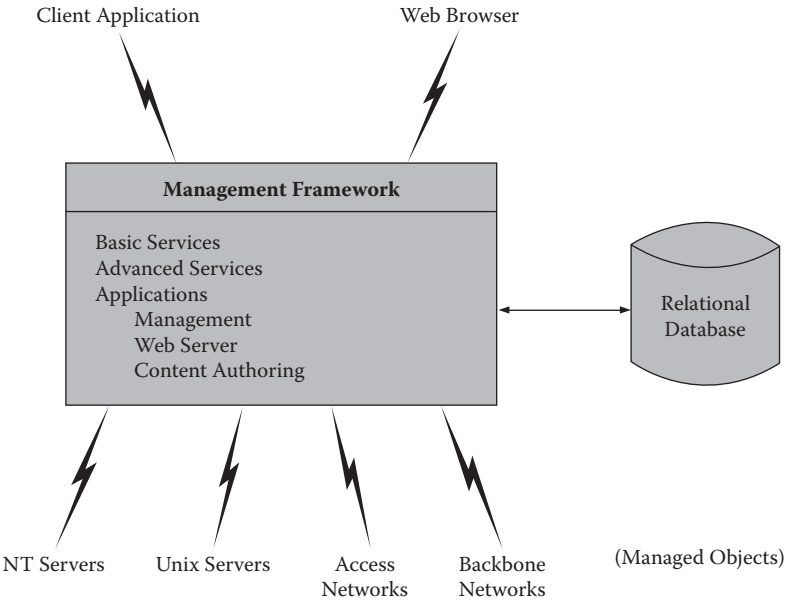
After outlining generic and specific challenges associated with managing intranets, this section focuses on emerging measurement and management tools such as log file analyzers, traffic monitors, load distributors, traffic shapers, and application performance management solutions. The section ends with a discussion of opportunities to integrate these new tools with existing management platforms and applications.

2.4.1 Internet, Intranets, and Extranets

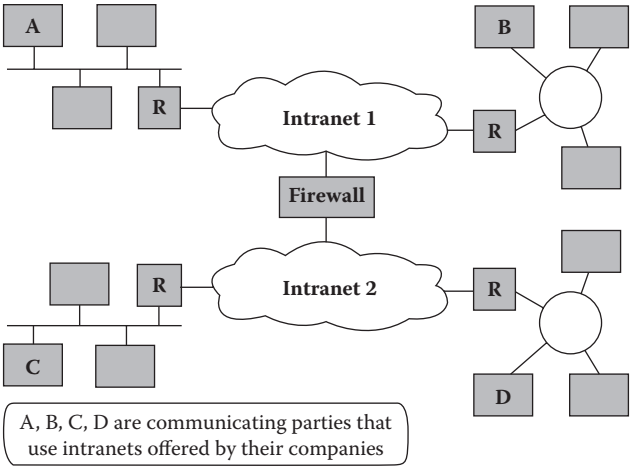
Intranet management refers to deploying and coordinating resources in order to design, plan, administer, analyze, operate, and expand intranets to meet service-level objectives at a reasonable cost and with optimal resource capacity. The experiences amassed over the last twenty-five years with managing data networks can be applied to intranet management; existing management concepts are still valid. Critical success factors are applicable as well. With respect to managing intranets, success factors include the following:

- Management processes that can be grouped around fault, configuration, performance, security, and accounting management
- Management tools, usually assigned to human resources, that support management processes
- Management team human resources, with their skills and network management experiences

There are similarities between intranet management and management of other networks. The intranet management architecture is shown in Figure 2.4.1. The management framework (at center) consolidates, processes, displays, and distributes information to authorized individuals. The framework should be equipped with Web capabilities meeting most user expectations, meaning that views and reports are converted into Hypertext Markup Language (HTML) pages and are accessible from universal browsers. Management applications consist of a mix of well-known applications, such as trouble ticketing, asset management, and change management, and new applications dealing with log file analysis, load



FIGu RE 2.4.1 Intranet management framework.



FIGu RE 2.4.2 Use of intranets.

balancing, packet shaping, content authoring, and application performance monitoring. Specific challenges associated with intranet management are addressed throughout the remainder of this section.

The Internet is a network used by millions of people every day. At the same time, Internet is a generic term for a bundle of technologies available under the Internet umbrella. The Internet exhibits a number of similarities with the global telephone system. Anyone who is a subscriber can be reached if the correct country code, area code, and phone number are dialed. In the case of the Internet, visitors type in the Universal Resource Locator (URL) to access the necessary information. The billing process is also similar; the longer the talk or surfing, the higher the bill.

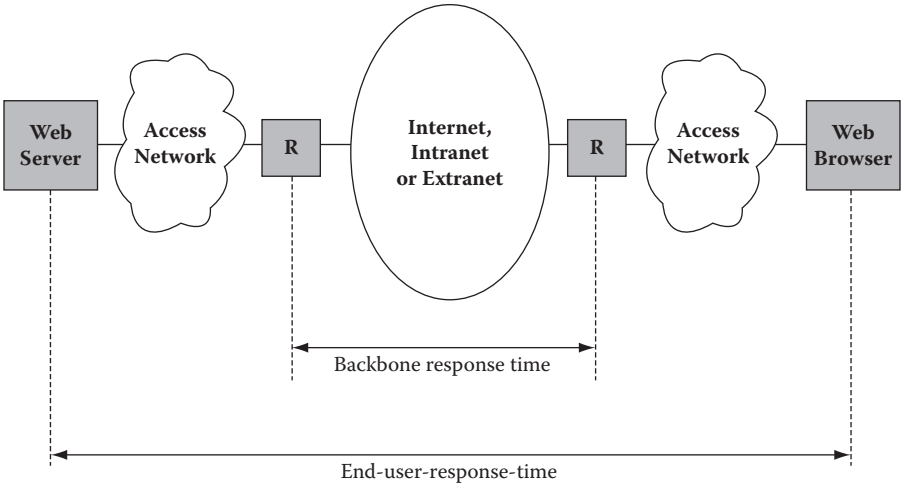
The ownership is not as clear with the Internet as it is with public telephone systems. The multiple owners of the Internet physical backbone are hidden from users. Administration and management are becoming more important given the fast-growing number of subscribers. For example, the single administration issue of address management causes many headaches at present. Country institutions are coordinated by an independent U.S.-based company.

It is tempting to consider the Internet as the central switching point of corporate networking. However, performance and security considerations drive corporate network managers to use privately owned Internet-like networking segments called intranets. Examples of intranets are shown in Figure 2.4.2; A, B, C, and D are communicating parties using the intranet(s) offered by their company.

Intranets are company-internal networks that use Internet technology. In particular, Web technology is used for information distribution (e.g., unifying company documentation, making internal hiring procedures visible), and Web protocols are used for internal information transfer. The intranet backbone is based in Internet Protocol (IP) Layer 3. If interconnections to other networks or other companies are required, firewalls are deployed to protect the company-owned intranet. Firewalls are actually filters; packets without the necessary authorization code cannot pass the firewall.

If partnerships are the targets, the networking equipment of partnering companies can be connected. In such cases, the connected intranets are called extranets. Firewall requirements are much less stringent with extranets than with intranets. Typical applications of extranets are car manufacturers and their parts suppliers; airlines and travel agencies; groups of telcos attempting to complement local, long distance, and international services; and service providers and customers.

The Internet can still be utilized as part of intranets and extranets. For instance, virtual private networks (VPNs) offer secure Internet channels that can be used by communicating parties in intranets and extranets. In addition, a couple of technical solutions are available that are based on Layer 2 or Layer 3 technologies.



FIGu RE 2.4.3 Principal structure of systems and networking components.

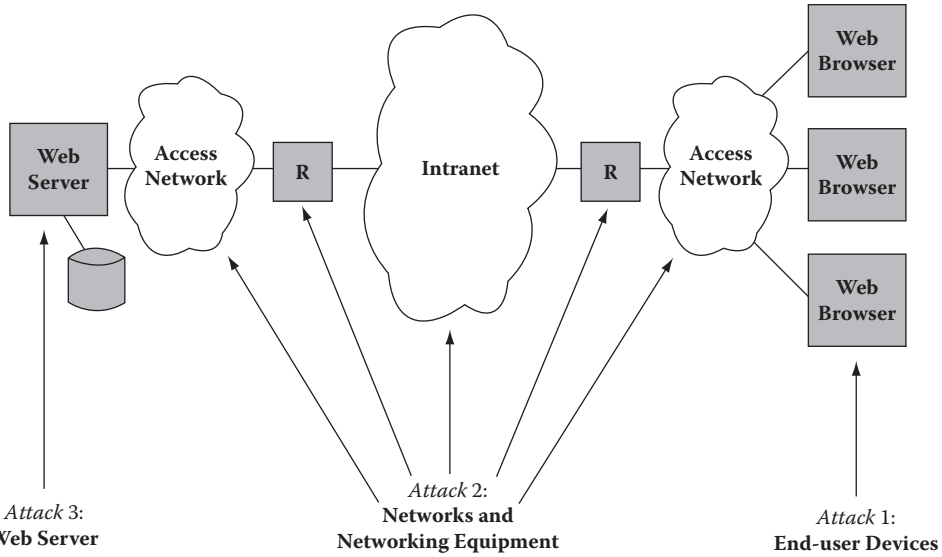
2.4.2 Generic Intranet Management Challenges

This section investigates how management functions can be reimplemented in intranets, as well as in supporting Web services. Also, challenges in each functional area will be highlighted.

2.4.2.1 Performance Management Challenges

Feasible network architectures for intranets and extranets are shown in Figures 2.4.3 and 2.4.4. The components of intranets are similar to those of other types of networks. Principal components include:

- Web servers that maintain home pages
- Web browsers that directly support users in viewing, downloading, and uploading information to and from Web servers



FIGu RE 2.4.4 Access points with security risks.

- Backbone that offers broader bandwidth for high data volumes
- Access network that offers narrower bandwidth for lower data volumes
- Networking components, including routers, switches, traffic shapers, and firewalls
- Communication protocols such as IP for backbones and higher-layer protocols such as Hypertext Transfer Protocol (HTTP), Simple Network Management Protocol (SNMP), and **File Transfer Protocol** (FTP) to support management applications

Figure 2.4.3 shows a typical arrangement in a simplified form. From the perspective of management, all of these components are managed objects. One additional type of managed object must be considered—the application running in a Web server.

According to Internet World Stats, 1.3 billion people were using the Internet as of December 30, 2007. In an article published in the *Harvard International Review*, philosopher N.J. Slabbert, a writer on policy issues for the Washington, DC–based Urban Land Institute, asserted that the Internet is fast becoming a basic feature of global civilization, with the result that what has traditionally been called "civil society" is now becoming identical to an information technology society defined by Internet use.

According to a 2001 study, there were more than 550 billion documents on the Web, mostly in the "invisible Web," or deep Web (i.e., Web pages that do not exist until they are created dynamically as the result of a specific search) (http://en.wikipedia.org/wiki/World_Wide_Web#Statistics). A 2002 survey of 2,024 million Web pages determined that by far the most Web content was in English: 56.4%. Next were pages in German (7.7%), French (5.6%), and Japanese (4.9%). A more recent study in which Web searches in 75 different languages were used to sample the Web determined that there were more than 11.5 billion Web pages in the publicly indexable Web as of the end of January 2005.

More than 100 billion Web sites were operating as of March 2008 (http://en.wikipedia.org/wiki/World_Wide_Web#Statistics), of which 74% were commercial or other sites operating in the dot-com generic top-level domain. With respect to services paid for by advertising, Yahoo! was able to collect the most data about commercial Web users, approximately 2,500 bits of information per month on each typical user of its site and its affiliated advertising network sites. Yahoo! was followed by MySpace, with about half that amount, and then by AOL-TimeWarner, Google, Facebook, Microsoft, and eBay. Approximately 26% of Web sites operated outside of dot-com addresses.

As a result of these patterns, performance metrics are extremely important. From a technical viewpoint, everything can be measured. From a practical viewpoint, however, only a few indicators are of prime interest. Two in particular are considered in every enterprise: response time and resource utilization.

Not only resource-level response time but also user-level response time should be measured. There are now several types of tools to choose from—some of them measure throughput rates, some simulate network traffic and tally the results, some gauge performance by running within the applications themselves, and some rely on a combination of these techniques. Altogether, there are four approaches:

- Monitors or packet analyzers
- Synthetic workload tools
- Application agents
- Application response measurement (ARM) management information bases (MIBs)

End-user response time is helpful in service-level agreements. If performance is to be optimized, more details are needed about contributors such as networks, systems, and applications. When response time segments are known, resources can be optimized through proper capacity planning.

Use of resources has a direct impact on response times. Payload is always an issue with resource utilization. Operating systems place a load on servers; protocol control characters mean additional bytes to be transferred. Both represent overhead, but they cannot be avoided completely. The same is true with monitors and the transfer of monitored data for further processing. Overhead can be controlled, however, and in this instance productive operations are not affected. Further details on performance-related metrics in intranets are provided in other chapters.

In summary, tuning and optimizing intranets and Web services may be very different than is the case with traditional networks. User behavior, application performance, unusual traffic patterns, asynchronous resource demand, and additional protocols result in performance management challenges.

2.4.2.2 Security Management Challenges

Opening networks, connecting partners, and using a public domain such as the Internet lead to considerable increases in security risks. VPNs are a possible method of combining existing infrastructure with acceptable levels of protection. Security expectations may be different in different industries, but generic security management procedures are identical or at least very similar. Security management enables intranet managers to protect sensitive information by (1) limiting access to Web servers and network devices among users both inside and outside of the enterprise and (2) notifying the security manager of attempted or actual violations of security.

Intranet security management consists of the following:

- Identifying the sensitive information to be protected
- Locating the vulnerable access points to sensitive information
- Securing these access points
- Maintaining secure access points

Identifying sensitive information means classifying information. Most organizations have well-defined policies regarding what information qualifies as sensitive; often it includes financial, accounting, engineering, and employee information. In addition, however, certain environments involve sensitive information that is unique to them. The main purpose of intranets is to improve documentation and communication within enterprises. Web servers are the focal point of information maintenance. Depending on individual responsibilities, access rights to information sources can be relatively easily structured and implemented. In summary, sensitive information is found on home pages, with particular content residing on Web servers.

Once the Webmaster and network managers know what information is sensitive and its location, they must determine how users can access it. This often time-consuming process will usually require that Webmasters and network managers examine each piece of hardware and software offering a service to users. In this respect, intranets are not different from any other complex networks. Generic sensitive access points are as follows (Figure 2.4.4):

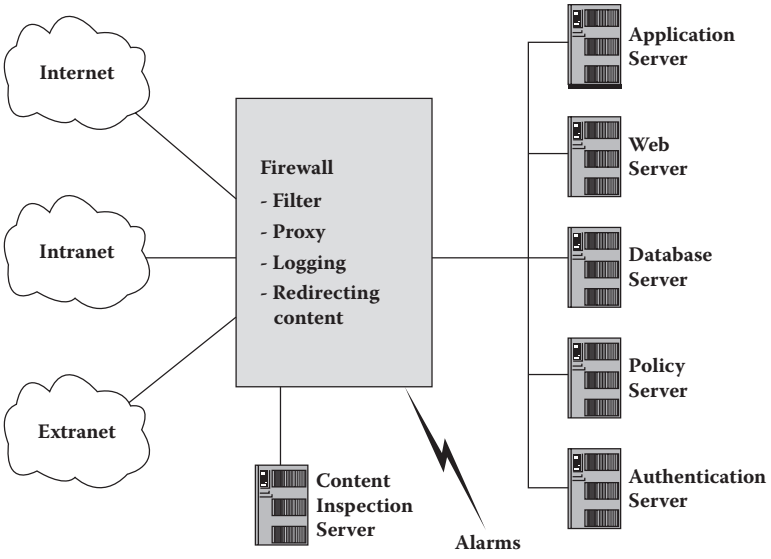
- End-user devices such as browsers
- Access and backbone networks
- Web servers maintaining sensitive information

The next step in security management is to apply the necessary security techniques. Sensitive access points dictate how protection should be deployed through a combination of policies, procedures, and tools. In this respect, the following security techniques must be considered:

- End-user devices such as universal browsers: use of chip cards or chip keys
- Access and backbone networks: use of encryption, authentication, and firewalls
- Web servers: use of server protection, operating systems protection, special tools, and virus protection

The final step in effectively securing access points in intranets is maintenance. The key to maintenance is locating potential or actual security breaches, which requires ongoing stress testing of intranets, assignment of tasks to outside professional security companies, reviews of security violation case studies, and evaluations of new security management techniques and tools.

Firewalls (Figure 2.4.5)—devices that control the flow of communication between internal and external networks such as the Internet—play a significant role in intranet security management. Firewalls serve several functions. First, they act as filters for inbound Internet traffic to enterprise servers,



FIGu RE 2.4.5 Firewall architecture.

preventing unnecessary network packets from reaching Web and application servers. Second, they provide proxy outbound connections to the Internet, maintaining authentication of internal Internet users. Third, they log traffic, providing an audit trail for usage reports and various planning purposes.

Firewalls are not without risks. Many companies assume that once they have installed a firewall, they have reduced all their network security risks. Typically, firewalls are difficult to penetrate, but when they are broken the internal network is practically open to the intruder.

Furthermore, a firewall does not address internal network compromise. Approximately 70% of all network security breaches occur from within the corporation, that is, by individuals already past a firewall. A modem dialup established by the company or by an engineer for remote access is one easy way past a firewall. Also, misconfigured firewalls may cause problems. Firewalls are highly susceptible to human error. In a dynamically changing environment, system managers routinely reconfigure firewalls without regard to security implications. Access control lists on a firewall can be numerous and confusing. Intranet managers should ensure that firewalls have been set up correctly and that they are performing well.

A network-based intrusion detection system is required to protect the intranet perimeter network from hacker attacks. Network-based intrusion detection systems may be deployed as probes or agents running on servers. Probes are the most effective method of providing network-based intrusion detection. Probes minimize the effects on existing systems by serving as passive listeners reporting back to a centralized console without interruption. Intrusion detection will comprise the following functions at the network device level:

- Inspection of data streams as they pass through the network, along with identification of and action on the signatures associated with an unauthorized activity
- Activation of an alarm immediately upon detection of the event
- Notification of the appropriate security personnel and triggering of an automated response

In addition to intrusion detection, a transmission control protocol (TCP) proxy aggregator may be considered. This will tighten security through the firewall by limiting exposed ports. It also provides an offload for session/connection management and a more robust technical implementation in terms of the number of port permutations supported.

Tunneling and encryption are used to deploy networks that need to appear point to point but consist of various routes to an endpoint, providing data integrity and confidentiality. Usually, tunneling

protocols, such as the Layer 2 Tunneling Protocol (L2TP), Point-to-Point Tunneling Protocol (PPTP), and Internet Protocol Security (IPSec), are used, along with encryption standards such as DES, MD5, Triple DES, and others.

Mobile code programs, such as Java and ActiveX, pose increasing security threats. Content inspection software should:

- Provide full control over Java, ActiveX, and other mobile code activity
- Prevent undetected, costly mobile code attacks, such as industrial espionage and data modification
- Enable safe Internet/intranet/extranet surfing while taking full advantage of Java and ActiveX technologies

A content inspection server accepts mobile content redirected from a firewall in order to scan for attack signatures. If the scan detects a vulnerability, the content will be blocked and the client prevented from downloading the mobile code. This denial will alert an appropriate administrator and notify the requesting client. If the scan does not detect any vulnerability, the mobile code is redirected to the firewall for routing to the client.

In summary, intranet security management challenges are increasing as a result of the high number of network access points and access technologies. New techniques and new tools are required in combination.

2.4.2.3 Accounting Management Challenges

Although there are no significant differences between intranet components and the components of other types of networks and systems, there are fundamental differences in terms of traffic patterns that may affect accounting strategies. Accounting management involves collecting data on resource usage in order to establish metrics, check thresholds, and, finally, bill users. Billing is a management decision, but usage measurements are a must in intranets. The principal accounting steps are as follows:

- Gathering data on use of Web servers, Web services and applications, and access and backbone networks
- Gathering data on the use of content provided by various content authors
- Setting usage quotas as part of service-level agreements with users
- Billing users for their use of resources through a combination of prepaid and postpaid alternatives

Proper instrumentation is necessary in gathering data on usage. Stand-alone monitors, log file analyzers, and built-in accounting agents are most commonly used. Accounting management requires continuous monitoring, but the amount of collected data is usually not critical in terms of overhead.

Service-level agreements may include an expected level of resource utilization by single users or user groups. Either time durations or byte volumes may be agreed upon. Services offered and/or prices charged may change when agreed-upon data volume quotas are exceeded. Service-level agreements and their continuous monitoring help enterprises plan for the appropriate amount of capacity.

Billing for these services is a new area, one that is not yet well understood. Users are often charged (1) a one-time installation fee followed by monthly fees or (2) fees based on the level of resources used. The first case is straightforward. The user is billed for installation of and access to the intranet and then is charged a standard fee for each month of use. When this method is used, accounting management is not necessary. Although this is the easiest system to implement, it can become difficult to justify why users with very different traffic patterns and volumes are billed the same amount.

The second case is more difficult and requires more engineering. Again, there are additional alternatives, including billing based on total number of visits, billing based on total number of packets sent or received, and billing based on total number of bytes sent or received.

Accounting and billing are more complicated when multiple suppliers are present in intranets. When this is the case, a clearinghouse must be used to gather and allocate usage data and then generate convergent bills to users. It is expected that users will receive only one bill for intranet services.

In summary, the accounting management process can be fundamentally different in intranets than in private enterprise wide area networks (WANs) and local area networks (LANs). In particular, usage-based data collection and convergent billing are the challenges faced in intranet accounting management.

2.4.2.4 Configuration Management Challenges

Configuration management is the process of identifying systems and network components and using that data to maintain the setup of all intranet resources. Configuration management consists of the following steps:

- Identifying the current intranet environment
- Modifying that environment through moves, adds, and changes
- Maintaining an up-to-date inventory of components and generating various reports

Identifying the intranet environment: The intranet environment can be identified manually (by engineers) or automatically (through intranet management systems). Intranets do not require special treatment; this discovery and mapping step is identical to that associated with other networks and systems.

SNMP-oriented platforms offer configuration and topology services in two different ways: The discovery function identifies all managed objects with valid IP addresses on the LAN or across LANs; the mapping function goes one step further and displays the actual topology of the LAN or across LANs. Both functions can be used successfully with intranets. Managed objects without IP addresses are not discovered. Discovery and mapping processes require time and may affect production. Careful consideration of periodicity is required. Many companies deploy intranet visualization tools instead of or in addition to discovery and mapping. These tools are typically very user friendly and easy to use, but they are independent from the actual network. Without synchronization of the tools' database with the actual network, these visualization tools are useless. However, a combination of the discovery feature of the management platform with a visualization application can be very successful.

Modifying the configuration environment*:* Move, add, and change (MAC) rates are higher than average in intranet environments. Moves, adds, and changes can be due to restructuring of buildings and infrastructures, deployment of new applications, and equipment changes. Moreover, change rates are not predictable when offering services to mobile users. Modification would probably be manual if the data collection method were manual and automatic if the data collection method were automatic. Stable procedures that can be implemented effectively are required. Intranets become a very important part of the IT infrastructure, requiring high availability and good performance. The MAC window is narrowing with requirements that MACs be prepared very carefully. Requesters are expected to complete forms detailing the nature of changes, their effects on other managed objects, fallback procedures, desired dates, priorities, and human resources requirements. Also, the MAC process should be carefully monitored. When problems occur, fallback procedures are expected to be triggered. After successful completion of the MAC process, all related files and databases must be updated accordingly.

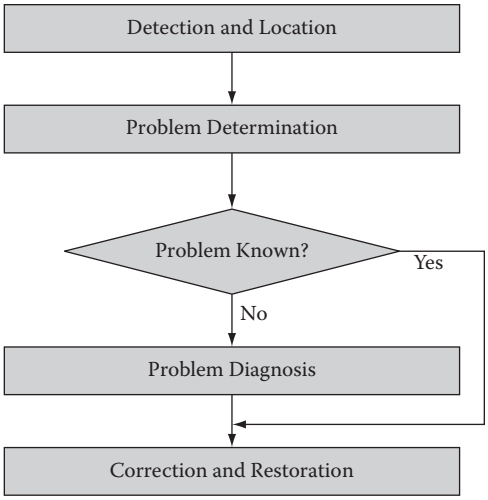
Maintaining the configuration*:* Asset and inventory management is one of the critical success factors in intranet management. Typically, relational databases are used to store and maintain technical and financial data on systems and network components. Access is usually via Structured Query Language (SQL); reporting is supported by standard or additional third-party reporting tools.

Asset management should work together with other management tools implemented in other management areas. In particular, links among asset management and the following elements are relevant in managing intranets: trouble ticketing, performance tuning, security violation traces, and accounting details.

In summary, managing intranet configurations does not introduce additional challenges to configuration management.

2.4.2.5 Fault Management Challenges

Fault management is the process of detecting, locating, isolating, diagnosing, and correcting intranet problems. Fault management consists of the following steps (Figure 2.4.6):



FIGu RE 2.4.6 Fault management functions.

- *Detecting and locating problems:* Intranet components generate a number of messages, events, and alarms. Meaningful filtering, combined with user input, helps to detect abnormal operations. Management platforms and applications are usually able to determine the locations of faults.
- *Determining the cause of the problem:* Information generated by element managers or correlational data provided by management platforms is used to ascertain the cause of the problem.
- *Diagnosing the root cause of the problem:* In-depth measurements, tests, and further correlation of messages, events, and alarms will help to determine the root cause of the problem.
- *Correcting the problem:* Various hardware and software techniques are used to repair or replace managed objects, and operations can return to normal.

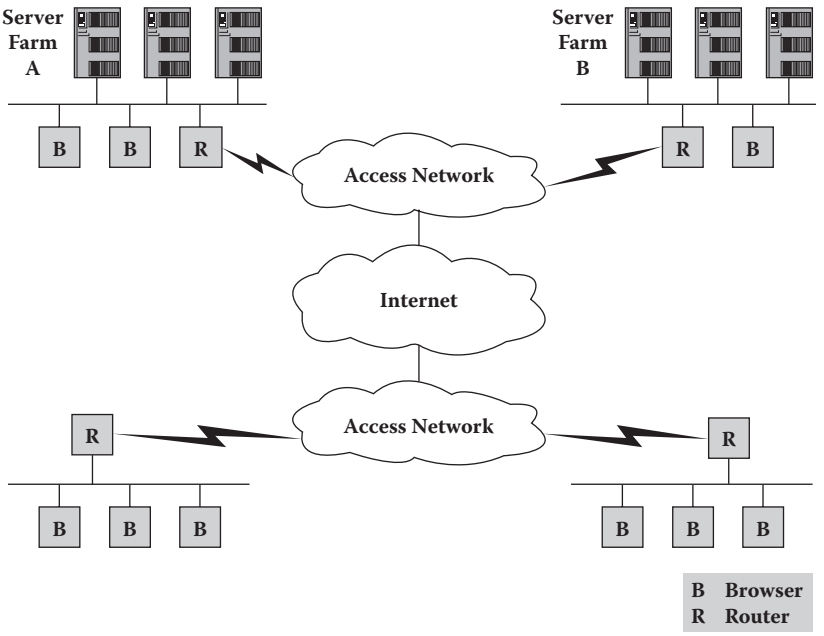
In summary, managing intranet faults does not introduce additional challenges to fault management.

2.4.3 Specific Challenges to Intranet Performance Management

The emergence of intranets has dramatically altered the way information is accessed within and outside the enterprise. Although there is a good amount of knowledge regarding intranet components such as servers, networks, and browsers, integrated management of these components produces several challenges for IT managers. Content, server, network, and browser management are all critical success factors. If IT managers do not focus sufficient attention on these factors, their intranets will fail. Figure 2.4.7 shows the components of intranets.

The emergence of Web computing is dramatically altering the way information is accessed. The heavy use and popularity of the World Wide Web is the most dramatic evidence. Looking at the enterprise, there is evidence that the Web browser has become the window of choice into corporate documentation and information. There are several important implications of this trend:

- All information can be viewed as Web content, accessible directly through a Web browser, a browser plug-in, or a dynamic piece of code (e.g., Java) downloaded automatically to the client. This content can be as varied as a static Web page, a Common Gateway Interface (CGI) script front-ending an existing database application, or new media such as streaming audio or video.
- The information access model has changed from one in which a client-specific configuration is required in order to access information to one in which access is always available unless policies are explicitly defined to prevent it.



FIGu RE 2.4.7 Components of intranets.

- Flash crowds, where certain content in the intranet generates significant unexpected traffic, are frequent, making traditional network design techniques based on measuring peak and average loads obsolete.
- Information accessed on or through Web servers represents the bulk of traffic on intranets (approximately 80%). Therefore, effective management of Web resources, bandwidth, and traffic is critical if there is to be an acceptable quality of service (QoS) in terms of Web-based computing.

2.4.3.1 Managing Content

All information can be viewed as content. The structure and arrangement of content will determine success or failure from the enterprise point of view. Page layouts may differ considerably depending on the content for targeted visitors. Not only the content of single pages but also their links to each other have a great impact on visitor satisfaction. Individual visitors expect:

- Easy-to-read layout combining text and graphics
- Easy navigation between pages
- Easy return to the home page
- Rapid painting of pages
- Efficient links to interactive services
- Up-to-date status of pages
- Visualization of site structure
- Sitewide change management of pages
- Ease in selecting pages to print or download

Goals and interests of companies offering information on home pages include the following:

- Rationalizing information distribution to internal customers
- Fully meeting content expectations of external visitors
- Managing intranet resources effectively
- Meeting performance expectations of external visitors

- Meeting business goals by using intranet technologies
- Providing an opportunity to deploy extranets to link business partners
- Meeting high security standards
- Monitoring visitors' behavior to allow rapid changes, increasing user satisfaction

Improvements in content management will have a positive impact on overall performance. While Web server performance improvements are part of a performance optimization solution, they must be accompanied by improvements in network and content management technology if they are to have a true impact on Web scaling and performance. Specifically, developments in the following three areas are critically important:

- *Content distribution and replication:* If content can be pushed closer to the access points where users are located, backbone bandwidth requirements can be reduced and response time to the user can be improved. Content can be proactively replicated in the network under operator control or dynamically replicated by network elements. Caching servers are examples of network elements that can facilitate dynamic replication of content. Other devices and models are likely to emerge over time.
- *Content request distribution:* When multiple instances of content exist in a network, network elements must cooperate to direct a content request to the "best-fit" server at any given moment. This requires increasing levels of content intelligence in the network elements themselves.
- *Content-driven Web farm resource measurement:* A server or cache in a server farm ultimately services a specific content request. Local server, switching, and uplink bandwidth are precious resources that need to be carefully managed to provide appropriate service levels for Web traffic.

2.4.3.2 Web Server Management

Web traffic poses a significant number of challenges to existing Internet and intranet infrastructures. Most Web sessions are short-lived. As such, they involve fewer TCP packets than batch mode operations such as file transfers. In addition, HTTP traffic tends to spike and fall radically. This creates instant demand for "hot" content that in turn causes network and server congestion. Web site traffic is highly mobile in that a unique event on a particular Web site can trigger a high hit rate within a short period of time. This is typical in cases with periodic management report distribution and major system and network outages.

Web traffic behavior is significantly different from the typical client–server paradigm. It involves the following unique characteristics:

- The amount of data sent from a server to a client is significantly larger (by a ratio of 5:1) than the amount of data sent from a client to a server. This suggests that optimization of server-to-client traffic has a more significant impact on intranets and that redirection of client requests to the best-fit server could have significant performance advantages in regard to Web traffic flows.
- The median transfer size for Web server documents is small. This implies that Web flows are mostly short-lived, and they are more likely to create instantaneous congestion as a result of their bursty nature. As a result, a resource management model will probably have to deal with short-lived flows in an appropriate manner. Even though HTTP supports persistent connections, it is unclear how widespread deployment will be, or how soon it will occur, owing to interoperability issues with existing network caches.
- The top 10% of Web server files are accessed 90% of the time and are accountable for 90% of the bytes transferred. This suggests that Web server selection, caching, and content replication schemes that focus on this top 10% will yield the greatest gain.
- Significant percentages (15%–40%) of the files and bytes accessed are accessed only once. That is, some small number of large files often consumes a disproportionate amount of total server and network bandwidth. In addition, servers suffer performance degradation when they are subject

to significant job size variations. This is due primarily to memory fragmentation, which occurs when data varying in size are buffered in fixed-length blocks. Furthermore, subjecting servers to workloads consisting of both hot and one-time requests will result in lower performance levels because of frequent cache invalidation of hot objects. Therefore, a server selection strategy that takes into account content, job size, and server cache coherency can significantly improve network and server resource allocation and performance. In addition, requests for large files may be good candidates for redirection to a server that has a shorter round-trip time to the client.

- Hosts on many networks access Web servers, but 10% of networks are responsible for more than 75% of this usage. This suggests that resource management strategies that focus on specific client populations may yield positive results in some cases.

Real-time traffic is beginning to represent an increasingly significant proportion of Web traffic. Web site resource management strategies must take into account an increasing demand for support of real-time applications such as voice, distance learning, and streaming media. To deal with both legacy and Web traffic as well as real-time Web traffic, these strategies will need to include admission control along with bandwidth and buffer allocation components.

The hardware for Web servers is practically identical to that used with other servers. The software is divided in most cases between Unix derivatives, Solaris, Windows Family, and open source software; industry analysts expect a clear shift in the future toward open source software for price reasons. Web server sizing should follow generic guidelines as well as criteria specified through analyses of Web traffic patterns. If resource demand is higher than server capacity, multiple servers can be combined into server farms. Although this solution may satisfy resource demand criteria, it requires careful attention to allocation and flow control.

2.4.3.2.1 Content-Smart Quality of Service and Resource Management

In the case of a typical Web site, as mentioned, the top 10% of Web server files are accessed 90% of the time and are accountable for 90% of the bytes transferred. Therefore, techniques that optimize performance for these files will have the most significant impact on overall Web site performance. This requires that the network itself be aware of which content is hot and which servers can provide it. Since content can be hot one instant and cold the next, content-smart switches must gain information on hot content by tracking content access history as content requests and responses are processed.

If Web site servers, networks, and bandwidth resources are to be effectively managed, something must also be known about the content size and quality of service requirements. These content attributes can be gleaned through processing of active flows, proactive probing of servers, or administrative definitions. In addition, it is important to track server performance relative to specific content. All of this information can be maintained in a content database that provides an analogous function to a routing table in a router or switch. Content-smart switches make a content-routing decision based on the information contained in the database, connecting a client to a best-fit server in either a local or remote server farm. This enables the emergence of a business model based on replicating content in distributed data centers, with overflow content delivery capacity and backup in the case of a partial communications failure. In addition, overflow content capacity intelligence minimizes the need to build out to handle flash crowds for highly requested content.

2.4.3.2.2 Content-Smart Flow Admission Control

Two factors often contribute to congestion in a server farm. One is that servers are not up to the task of handling the amount of incoming traffic. The other is that the link bandwidth from servers to the Internet is overwhelmed by the combination of inbound and outbound traffic; this is complicated by the fact that the amount of outbound traffic from servers is considerably higher than the amount of inbound traffic. As a result, a successful TCP/HTTP connection could fail because of an inability to allocate the server the necessary bandwidth to deliver the requested content. To make matters worse, some server

implementations come to a grinding halt when they encounter an excessive number of TCP/HTTP connections, sometimes requiring a hard reboot.

2.4.3.3 Load Distribution and Balancing

Bandwidth in backbone and access networks should be managed effectively to satisfy the high performance expectations of site visitors. Usually, servers are consolidated into server farms that use the infrastructure of LANs. It is very unlikely that LANs cause bottlenecks. Larger enterprises may use multiple server farms deployed at various locations. In order to optimize content allocations, traffic and page references should be monitored and evaluated. At different locations in the network, there is the expectation that hardware and software will be installed that intelligently analyze requests and direct traffic to the correct destination, which could be (1) the server farm destination with the requested content, (2) the server farm destination with the lightest load, or (3) the server farm destination with the closest location to the visitor. There can be no compromise with respect to Item 1, but there can be trade-offs between Items 2 and 3 depending on the networking traffic.

At present, the Internet can be described by using a model where local bandwidth is plentiful in the premise LAN located at the edge of the Internet. However, uplinks from LANs or remote users to the Internet are often severely constrained in terms of bandwidth by orders of magnitude. Although congestion can occur anywhere in the Internet path between a client and a server, the most frequent culprits are the WAN connection between the client and the Internet and the WAN connection between the Web farm and the Internet. Actions taken to ensure that this bandwidth is not overcommitted will help improve end-to-end performance.

Instantaneous bandwidth mismatches can occur with a network device that functions as the demarcation point between the public Internet and the Web farm. Examples are as follows:

- The incoming link of the traffic is a faster media type (e.g., Fast Ethernet) and the outgoing link is a slower type (e.g., T1 or T3).
- The instantaneous fan-in (i.e., the number of flows being sent at the same time to the same output port) can vary dynamically from one moment to the next.
- A number of traffic sources (e.g., outbound server traffic) may be sharing the bandwidth of a broadband pipe in a bursty manner over a very-high-speed switching fabric (e.g., 10 Gbps). This creates a need to regulate flow admission into a slower pipe from multiple higher-speed traffic sources.

Information about Web page use and users, frequency of access, resource utilization, and traffic volumes can also be collected in the network or at the interfaces of the network. In many cases, the borders between tools and techniques in the server and networking segments are not clear. Tools are different from each other; the differentiators are data collection technologies, performance metrics used, and reports offered.

Effective bandwidth management is a critical success factor for the Internet and intranets. The role of network planners will be redefined in the years to come. Real-time and near-real-time bandwidth allocation definitions are needed. Network managers agree that load balancers are needed.

Little progress has been made in standardizing load distribution performance metrics. However, the following metrics can be used successfully:

- Number of referrals to server farms
- Number of lost requests due to load situations
- Number of requests with unacceptable response times
- Number of broken connections due to network problems

2.4.3.3.1 Content-Smart Link Management

Content-smart link management can ensure that more flows are not admitted than can be handled on average through the switch or on the uplinks. It is still critical, however, to deal appropriately with

traffic bursts and temporary congestion on these links to ensure that Web flows receive the appropriate quality of service. Priority queuing provides a way to prioritize requests according to their precedence. Fair queuing and weighted queuing methods represent improvements over the priority queuing scheme in that they address the low-priority traffic starvation problem with a scheme that separates traffic into well-identified flows so that each receives a "fair" or "weighted fair" share of transmission bandwidth.

The class-based queuing (CBQ) model, in which traffic is categorized into hierarchical groups, was developed by the Network Research Group at Lawrence Berkeley Laboratory as an improvement upon these existing bandwidth management techniques. Flows inherit their characteristics from their parent flow class tree and can have local characteristics of their own. Flows are identified according to IP address and the inner attributes within the IP header and payload. CBQ provides more granular control of transmission bandwidth and distributes it to member flow classes in accordance with their allocation policies. The model itself is independent of the scheduling techniques that run underneath it; therefore, implementation details will vary according to the target architecture.

Content-smart link management borrows concepts from CBQ. However, while CBQ operates in a packet-by-packet configuration based on Layer 3 and Layer 4 classification techniques, content-smart link management classifies flows at admission time according to the content requested, its attributes, and configured policies. These policies support the enterprise and service provider models described earlier, facilitating the classification of flows in a two-level hierarchy that includes owners (or customers) and content. Actual scheduling of flows is managed by a hardware-based scheduler that supports guaranteed bandwidth flows, prioritized/weighted flows, and best effort flows. Hardware-based scheduling is critical in scaling the Web farm.

2.4.3.3.2 Content-Smart Load Balancing

Simple load-balancing techniques such as round robin, weighted round robin, and least connections are inadequate for Web traffic. For example, Web traffic load balancers must support "sticky" connections, connections that allow a particular server to be selected regardless of server load owing to content locality or transaction integrity. Because of the disproportionate ratio of hot content files to total content (1:10), it is highly desirable to support a content replication model that does not require that content be equally and fully mirrored among servers in a server farm. This means a load-balancing technique must be intelligent enough to recognize whether content is available on a particular server before making the selection decision.

Content-smart load balancing takes into account several factors that have a significant impact on the overall performance and cost of a Web server farm:

- Server cache hit rate: By directing requests for hot content to a server that has recently serviced that content, this technique ensures that cache hit rate is high, reducing disk access latency for the most frequently accessed content. Since a significant percentage (15%–40%) of files are accessed only once and 90% are accessed only once or not at all, it is important to keep infrequently accessed files from thrashing a server cache. That is, an infrequently accessed file should be invalidated in the server cache promptly to increase the chances that a more frequently accessed file can remain in the cache.
- Burst distribution: Short-lived, bursty flows can best be handled by distributing them among eligible servers provided that the servers have been performing below a defined threshold for a period of time.
- Web flow duration: Most Web flows are short-lived. However, a relatively small number of infrequent, long-lived flows can have a significant impact on overall bandwidth and server resource consumption. For that reason, long-lived flows should be separated from short-lived flows, and short-lived flows of similar quality of service (QoS) requirements should be aggregated to increase TCP flow intensity and reduce per-flow resource allocation overheads.

- Content-biased server performance measurement: Current server loading can best be measured by examining the request–response time interval of a server as it handles requests. This measurement is most accurate when the connection between the switch and the server is direct. In addition, server performance is not uniform across all content. For example, computer-intensive applications may perform better on one server than another. Other servers may perform better with other types of content. Server performance information needs to be qualified by content.

Network managers must decide whether they need hardware- or software-based load balancers and embedded or stand-alone solutions (or combinations of these options). In the first case, considering high traffic volumes, hardware solutions should be preferred. In critical load situations, software solutions may slow down processes and risk performance. At this time, there are no accurate guidelines for tolerable workload, but a level of up to 5% seems to be reasonable.

Switches, routers, and firewalls are almost everywhere in Internet access networks and in intranets. Embedding traffic control and sharing functions would save extra components, but as stated earlier, would generate additional load and might impair principal functions. Embedded solutions may also include the use of remote monitoring (RMON) capabilities for real-time load profiling. Stand-alone solutions are sensitive against a single point of failure but would offer overhead-free traffic and load management. The following attributes may play an important role when alternatives are being evaluated.

2.4.3.3.2.1 Load-Balancing Switches

Benefits:

- Load balancing performed in a device already needed in the network
- Centralized management
- Good opportunity to control and guarantee QoS

Disadvantages:

- Possible effect on performance by management functions
- Single point of failure for both switch and management functions

2.4.3.3.2.2 Load-Balancing Firewall

Benefits:

- Load balancing performed in a device already needed in most networks
- Centralized management
- Includes special functions and services, such as traffic management and application-based load balancing

Disadvantages:

- Switches still needed
- Single point of failure for both firewall and management functions
- Dependence on hardware and operating system configuration

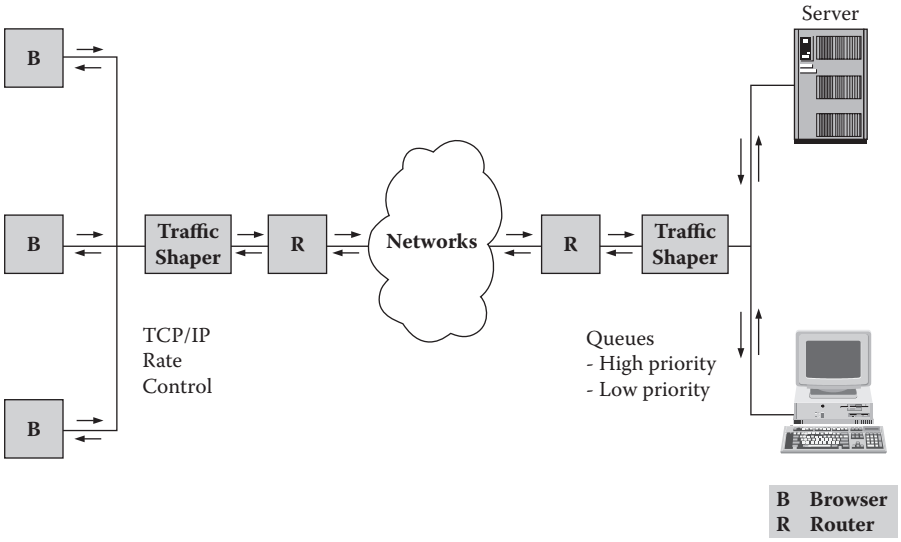
2.4.3.3.2.3 Load-Balancing Traffic Shapers (Figure 2.4.8)

Benefits:

- Load balancing performed by a device most likely already present in the network
- Centralized management
- Traffic shaping and balancing for Internet or intranet access in addition to server access

Disadvantages:

- In most cases, switches and firewalls needed in addition to these devices



FIGu RE 2.4.8 Load-balancing packet shapers.

TABl E 2.4.1 Comparison of Technologies for Access Networks

Criterion	T1, T3	ISDN	Frame	ATM	Cable	XDSL	Fixed Wireless	Wireless
Suitability	Medium	Medium	Good	Excellent	Excellent	Good	Good	Good
Maturity	High	High	High	High	Medium	High	Medium	Medium
Scalability	Good	Low	Medium	Excellent	Medium	Good	Medium	Medium
Distance limitations	None	None	None	None	Some	Some	Some	None
Cost	High	Low	Low	Medium	Low	Low	High	Medium

- Single point of failure for both traffic shaping and load balancing
- Little experience to date with performance and scalability

2.4.3.4 Technologies of Access Networks

There are many alternatives with respect to designing and deploying access networks. The basic technology selected has a significant impact on overall performance. Table 2.4.1 presents an evaluation of the most important technological choices according to the criteria of suitability, maturity, scalability, distance limitations, and cost.

The appropriate choice of access network technology must be seen in connection with content-smart control of the bandwidth provided in access networks. Traffic shapers and load balancers may also be used as generic terms; others such as adware and accelerator are frequently used in reference to WAN optimization.

2.4.4 Log File Analysis

Web site activity reporting involves analyses of:

- Basic traffic statistics (hits, page views, visits)
- Navigation patterns (referrers, next-click, entrance and exit pages)
- Content requested (top pages, directories, images, downloaded files)
- Visitor information (domains, browsers, platforms)
- Fulfillment of the Web site's objective (purchases, downloads, subscriptions)

Clearly, this last characteristic is the reason that Web site activity analysis has become an enterprise-critical priority for organizations investing massive amounts of time and money in their Web presence. How well a Web site is performing relative to its objective is what justifies continued investment. The easiest way to quantify return on investment is with meaningful Web activity reports.

Reporting is also essential in making decisions about content. Web site activity reports, by providing statistics about the most popular pages or files, give an organization quantifiable measurements as to what types of content appeal to its audience. Without reliable, comprehensive reports, the design of a Web site's content is based on educated guesses by the design team or editorial staff.

Similarly, Web site activity analysis reports provide an organization with information on its visitors: Where are they coming from, how do they get to the Web site, and what type of browser or platform are they using? When a corporation decides to deploy a Web site, it usually has an idea about who its audience will be. Does the actual audience resemble the predicted one? How does the audience change over time? What type of content improves visitor retention and session depth?

2.4.4.1 Usage Analysis

Web server monitors and management tools concentrate on how the Web server is utilized and how performance goals can be met. In addition to these tools, other tools are required that are able to continue the analysis by using log files filled by special features of the server operating system. This segment is devoted to log file analyzer tools that are able to provide the necessary data for an in-depth usage analysis.

Usage analysis is a means of understanding what is happening on an Internet or intranet server such as a Web server. Usage analysis tools piece together data fragments to create a coherent picture of server activity. Usage analysis can answer the following questions:

- How many individual users visited the site on a particular day?
- What day of the week is the site busiest?
- How many visitors are from a certain country?
- How long do visitors remain on the site?
- How many errors do visitors encounter?
- Where do visitors enter and leave the site?
- How long did it take most visitors to view the home page?
- Which links on other sites send the most visitors to this site?
- Which search engines send the most visitors to this site?

Reports can span any length of time, making it possible to see trends. They can also display any degree of granularity, allowing users to see both broad-ranging reports and detailed reports. Usage analysis is most frequently thought of in terms of Web servers. The reports created by usage analysis tools can be used throughout organizations to help people make informed decisions. Examples are as follows:

- Web developers use these tools to gauge the effects of site design changes. Using this information, they can make further refinements to the design of the site to maximize its effectiveness.
- Marketers use these tools to analyze the effectiveness of marketing programs and online ads.
- Site administrators can spot Web pages that are causing errors, determine future server hardware needs, and track FTP and proxy server activity.
- Salespeople can gather information about prospects, including their geographic location, how many pages they viewed, and how they found the site in the first place.
- Executives use the intelligence gathered with log analyzers as a resource when making a broad range of decisions.

Each time a visitor accesses a resource on a Web server—whether it is an image, an HTML file, or a script—the activity is typically recorded as a line in a text file associated with the Web server. This text file is known as the Web server log file.

Most Web servers write out log files in the combined log format, which differs from older log formats in that it contains browser and referral information. Referral information is important to determine what sites are sending the most traffic to the target address and what sites may have out-of-date links pointing to specific user sites. Referral information is also critical in gauging the effectiveness of online ads. Other information that can be included in a log file includes:

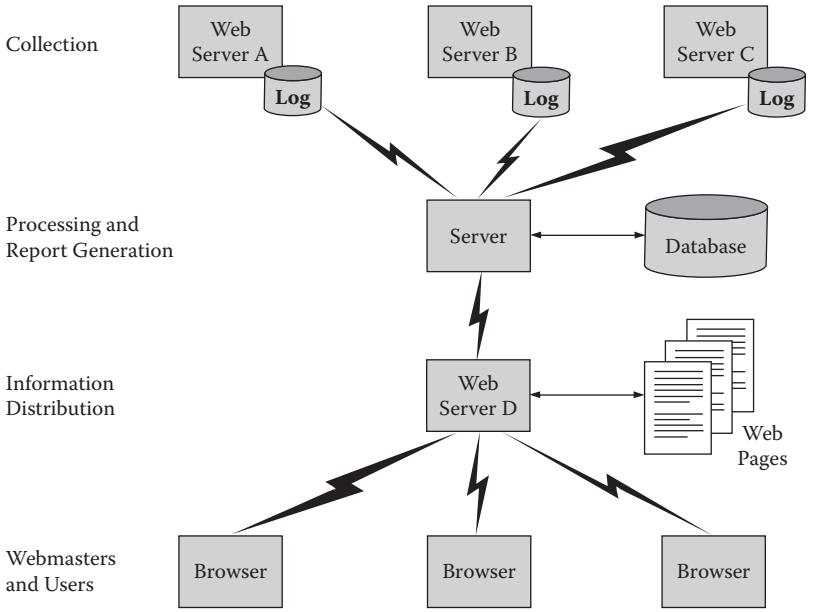
- Cookies: Persistent identification codes assigned to a user that allow the user to be tracked across several visits
- Session identifiers: Tools that track each visitor for the length of the visit only
- Amount of time the request required to fulfill: Enables server performance reporting

There are two basic types of usage analyzer tools: software-based tools and on-the-wire collectors. On the high end of usage, analysis tools are packet sniffers that offer on-the-wire reporting by installing an agent against the kernel of the operating system of the Web server. They run as the root in the kernel of the operating system on the Web server. Furthermore, they require a network to run in promiscuous mode in order to expose network traffic to the agent. Usually, there are very few reports packet sniffers can create and log file analyzers cannot. Log file analyzers can create reports on the usage of secure/encrypted communications, while packet sniffers cannot. Packet sniffers are more expensive, offer fewer reports, and offer only a few report distribution capabilities.

2.4.4.2 Issues in Log File Analysis

When selecting products, there are a number of criteria that must be carefully evaluated. While the market is large, a relatively small number of products are available. The criteria considered in product selection are also important when Webmasters want to position log file analysis within their IT administration or want to deploy this functionality within their organization.

The design of a product determines whether the product can support a distributed architecture or not. Distribution means that collecting, processing, reporting, and distributing data can be supported in various processors and at different locations. Figure 2.4.9 shows these functions with a distributed solution.



FIGu RE 2.4.9 Generic product architecture for log file analysis.

In Figure 2.4.9, Web servers A, B, and C can be very different types of servers, such as Netscape Navigator or Microsoft Explorer. Of course, many different Web server types are expected to be supported. Also, the hardware and operating system may be a differentiator for products. It is assumed that the Web server hardware has a reduced impact on log file analysis. The role of operating systems is more significant; the product should involve knowledge of exactly how log files are initiated and maintained. No problems are expected with leading Web server solutions based on Unix derivatives, Solaris, Windows Family, or open source software.

The data capturing technique is absolutely essential with log file analysis. The first question is where the logs are located. Figure 2.4.9 indicates that they are located in the Web servers. However, more precise information is required here:

- What memory area is used?
- What auxiliary storage area is used?
- What is the size of those areas?
- What types of log files are supported?

If log files are not processed in real time or near real time, it is important to know where they are stored until they are downloaded for processing. Log file analysis deals with very large data volumes, and these volumes depend on visitor traffic.

Log files are typically downloaded for processing. It is important to know how downloads are organized and how rapidly they are executed. As indicated in Figure 2.4.9, WANs sometimes involve limited bandwidth. Bandwidth is usually shared with other applications, and the result is potential traffic congestion. Bandwidth-on-demand solutions are rare with log file analysis. When transmission is arranged for low traffic periods, the actuality of log file analysis results may suffer. In such cases, local storage requirements increase, and processing, report generation, and information distribution are delayed by several hours or even by days.

Two solutions may help. The first solution is using intelligent profiling at the source of data collection. Redundant data are removed from logs during collection. Data volumes decrease along with local storage requirements, but processing requirements in Web servers increase considerably. In the second solution, data compression or data compaction may be used with the same results and impact as with the first solution.

Overhead is a critical issue with large data volumes. Data capturing is expected to introduce little overhead when logs are stored away immediately. If local processing is taking place, overhead must be very carefully quantified; if resource demand is high, overall Web server performance may be affected. Data transmission overhead can be heavy when everything is transmitted to the site where processing is taking place. It is expensive to dedicate WAN bandwidth only to log file analysis. If bandwidth is shared with other applications, priorities must be set higher for business applications than for transmitting raw log file data.

In the case of server farms, a local mediation device could help. LANs are used in connecting the mediation device; bandwidth is not as critical in LANs as it is in WANs. Processing and report generation remain at a special server that consolidates all data from mediation devices.

It is absolutely necessary to capture all data that are necessary to conduct a detailed Web site analysis of visitors or groups of visitors:

- Who is the visitor?
- What is the purpose of the visit?
- Where is the visitor coming from?
- When has the visit taken place?
- What key words have brought the visitor to the site?
- What search machines helped to access the site?
- How long was the visit?

Data losses cannot be completely avoided. Logging functions of Web servers, storage devices, or components of the transmission may fail; in such cases, there will be gaps in the sequence of events. Backup capabilities may be investigated, but IT budgets will not usually allow a large amount of money to be spent backing up high volumes of log file data. In the worst case, time windows are missing in reporting and in statistics. These gaps can be filled with extrapolated data.

Also, management capabilities are very important. One of the functions here includes automatic log cycling. Multiple logs can be used so that data are not lost. When one of the logs is full, the next log seamlessly takes over. Another function is the translation of domain name service (DNS). Its speed is absolutely important for real-time information distribution. If more meaningful reports are to be generated, results of log file analyzers must be correlated with other data sources. These other data can be maintained in other databases. Ad hoc database links should be established and maintained to allow correlation. Management of logs in any log file analyzer can be taken over by the Web server operating system. At present, only basic services are supported; additional services may follow. In the case of server farms and many individual Web servers, coordination of log transfers and processing is not a trivial task. Event schedulers may help in this respect.

Cookie support is important to speed up work initiated by visitors. In this logical connection between Web sites and browsers, a persistent identification code is assigned to a user that allows the user to be tracked across several visits.

Because of high data volumes, databases should be considered to maintain raw and/or processed log file data. Database managers would then offer a number of built-in features to maintain log files. Clustering visitors may be deployed from various perspectives, such as geography, common applications, common interests on home pages, and times of visits. Automatic log cycling can also be supported by database managers. Open database connectivity (ODBC) support helps to exchange data between different databases and to correlate data from various databases. In addition to log files, other data sources can be maintained in the same data warehouse. Above and beyond routine log file analysis with concrete targeted reports, a special type of analysis may also occasionally be conducted. This special analysis, called data mining, can determine traffic patterns and user/visitor behaviors. Both are important to sizing systems and networking resources.

One of the most vital questions is how log file analysis performs when data volumes increase. Volume increases can result when more pages are offered on more Web servers and when there are more visitors, longer visits, and extensive use of page links. In any case, collection and processing capabilities must be estimated prior to decisions on procedures and products.

In order to reduce log file processing and transmission loads, redundant data should be filtered as close as possible to data capturing locations. Filters can help avoid storing redundant data. Filters can also be very useful in the report generation process. Again, unnecessary data must not be processed for reports. Powerful filters can help to streamline reporting.

Not everything can be automated with log file analysis. The user interface is still one of the most important selection criteria for products. Graphical user interfaces are likely candidates, but simple products still work with textual interfaces. When log file analyzers are integrated with management platforms, this request is automatically met by such platforms.

Reporting is the tool used to distribute log file analysis results. Predefined reports and report elements as well as templates help to speed up the report design and generation process. Periodic reports can be automatically generated and distributed for both single Web servers and Web server farms. In the case of many Web servers, report generation must be carefully synchronized and scheduled. Flexible formatting helps to customize reports to special user needs.

Many report output alternatives are available. The most frequently used solutions include Word, Excel, HTML, and ASCII. Also, multiple choices are available with respect to distribution of reports:

- Reports can be stored on Web servers to be accessed by authorized users equipped with universal browsers.

- Reports can be uploaded into special servers or even pushed to selected users.
- Reports can be distributed as attachments to e-mail messages.
- Reports can be generated at remote sites; this alternative may save bandwidth when preprocessed data instead of completely formatted reports are sent to certain remote locations.
- Documentation can assume various forms. An integrated online manual is very helpful in providing immediate answers. Paper-based manuals are still useful for detailed answers and analysis. This role, however, will be taken over by Web-based documentation systems. In critical cases, a hotline can help with operational problems.
- Log file analysis is actually another management application. If management platforms are used, this application can be integrated into these platforms. There are many ways to integrate; in most cases, a command line interface (CLI) will be deployed.

2.4.4.3 Drawbacks of Pure Log File Analyzers

Log file analysis can provide a good entry-level summary of activities in and around Web servers. However, this technology involves several major problems.

The first major problem is traffic volumes. As traffic levels quickly reached exponential growth rates, nightly log file downloads soon became afternoon-and-evening and then even hourly downloads, since server disk drives would quickly fill with log file data. Compounding this problem was the fact that higher-traffic sites needed to load balance across several servers and physical machines, and thus log file downloads needed to be done not only many times a day but also across several machines each time. The quick fix to this problem was typically an automated script that would download log files on a preset schedule. However, this strategy failed to account for unexpected spikes in traffic and clogged internal networks with huge log files transmitted across the network several times a day.

The second major problem is data processing speed. Even if there were an easy way to continuously transfer log file data to a consolidated area, there was still the problem of how to process the gigabytes of log files into database tables in an efficient, continuous, and robust manner. Batch processing of log file data involved a considerable amount of time. In addition, the human resources demand for log file collection, processing support, and report compilation has exceeded expectations.

The third major problem involves incomplete data. In addition to log files, there are significant alternate sources of site activity data that contain more information than even the longest, most complex custom log file format can provide. A log file–only approach cannot guarantee a complete view of Web activities. A good example of missing data is network-level data that the Web server and the server's log file never see. For instance, a visitor requests a page that turns out to be too slow to download and decides to hit the browser Stop button, press the Back button, or otherwise terminate the request in mid-download. In this case, the network layer will log that action but will not notify the Web server about it. Similarly, there are significant amounts of data that are seen by Web servers but never written to the log file. Therefore, any measurement approach based solely on log files will occasionally miss critical information about user activity on the Web site.

The fourth major problem with the log file approach is flexibility. As sites become more sophisticated, one of the first obvious enhancements is to add dynamically generated content. Regardless of the type of content management system used, dynamic content typically results in URLs that are very difficult, if not impossible, for a human reader to decipher. Since log files are simply transaction records, dump reporting systems merely pass the nonsensical URLs through to the end-user report as the page that was requested, resulting in an unintelligible report with meaningless page names and URLs. The ideal solution would be to interpose an intelligent classification system between the raw activity data and end-user report. In practice, however, the reality of gigabytes of raw log files often leaves an in-house analysis team with few human resources to add even more complexity to an already slow log-based process. The inflexibility of log files with respect to handling the tracking of new technologies has been observed not only with dynamic content but also with personalization

applications, applet-based multimedia technologies, and a host of other new capabilities the log file approach was never designed to handle.

In summary, although log files were a convenient approach to measurement in the early days of the Web, they rapidly highlighted problems of:

- Labor intensity
- Slow data processing speeds and turnaround times measured in weeks
- Incomplete data, missing server- and network-level data
- Ineffective tracking of new feature enhancements such as dynamic content, personalization, and applet-based multimedia

In response to these problems, hybrid products have been developed and deployed.

2.4.4.4 Log File Analysis Tools

There are numerous log file analysis tools. Their depth and functionality are very different; some are complex in nature and offer more than simply log file analysis.

With the large number of mergers and acquisitions during the last few years, names of vendors and products are changing frequently. However, solutions can always be found with IBM, HP, CA, BMC, and Microsoft. Point products and new open source solutions are the targets of enterprise and service provider solution sets. A few examples are as follows.

Commercial analyzers:

Absolute Log Analyzer
Access Watch
Advanced Log Analyzer
Deep Log Analyzer
LogMiner
NetTracker
OpenWebScope
Sawmill
Visitors
Web Log Expert

Web log analyzers:

NetTracker
Opentracker.net
OpenWebScope
Sawmill
VisitorVille
Web Content Analysis
Web Log Analyzer Analysis
Web Log Expert
Web-Scope
Website Reporter
WebTrends

As can be seen, a few point products are represented in both groups.

2.4.5 Wire Monitors and Network Analyzers

Log files are not the only source of information for analyzing Web sites. Other tools reside "on the wire" or on LANs and collect information on performance and traffic metrics. Information depth and

overhead are significant indicators that may differentiate between log file analyzers and these products. In certain environments, the most effective results can be achieved only when both types of tools are deployed in combination.

Over the past several years, companies have adopted distributed multitier network infrastructures and moved business operations from traditional client–server applications to distributed Web-based applications. However, as more and more users come to depend on Web servers and TCP-based services, IT organizations are discovering that their current infrastructures are unable to offer the performance and availability expected by users; nor do they provide the management and monitoring capabilities required by IT organizations themselves.

2.4.5.1 Changes in Networking Infrastructures

The distributed multitier infrastructures that are being implemented by large corporations typically include four levels:

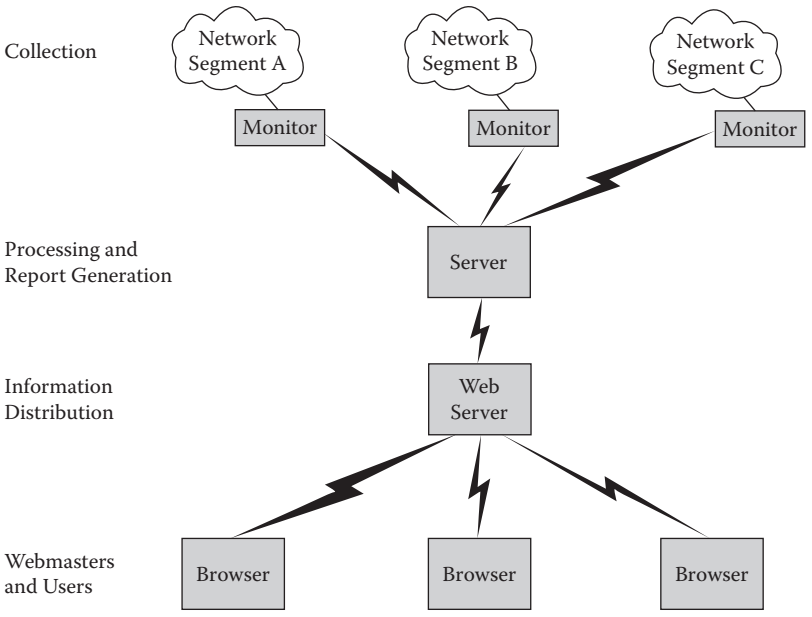
- WAN level: enables communication across multiple points of presence (POPs) or customer premises equipment (CPE)
- Web level: supports server farms providing a wide range of TCP-based services, including HTTP, FTP, **Simple Mail Transfer Protocol** (SMTP), and Telnet
- Application level: supports farms of application servers that offload computation from Web servers to increase overall site performance
- Aggregation network level: streamlines traffic and network admission to POPs and CPEs

IT organizations are deploying newly distributed, Web-based applications to take advantage of this new enterprise infrastructure. In place of fat software clients and centralized application servers, corporations are deploying Web browsers on every desktop, Web servers in departments and divisions, and application servers residing at multiple locations.

The new Web-centric model offers several advantages over the client–server model it replaces. IT departments can deploy Web browsers quickly and affordably to every desktop platform. Basic Web skills can be learned quickly and are popular with users. If an application requires modification to reflect changing business practices, IT departments need only modify the application itself, not the complex clients that formerly worked with the application. Most important, distributed, Web-based infrastructures move content and applications closer to users and provide improved reliability and availability. Employees can leverage this new infrastructure to improve internal business practices, communication with partners and suppliers, and services for customers.

While distributed multitier infrastuctures offer considerable advantages over earlier network architectures, as mentioned, they still do not offer the performance and availability expected by end users or the management and monitoring capabilities expected by IT organizations. Multitier architectures are physically well connected but not logically well connected. Standard network equipment enables traffic to flow, but not necessarily to the server best suited to respond. IT departments deploying these networks need traffic management solutions that intelligently direct TCP traffic to optimal resources at each tier of the enterprise infrastructure. An optimal traffic management solution requires communication between tiers. For example, there is little point in a DNS server directing traffic to a local server if that server is down or overloaded while another server is available with processing cycles to spare. To perform its job optimally, the DNS server needs availability and load information from the servers to which it directs requests.

The multitier model itself, when implemented with the standard software products available today, does not monitor services for system failures or spikes. Nor does it provide other capabilities that IT departments require to manage busy, distributed networks effectively. Specifically, it does not provide (1) policies for scheduling TCP traffic based on specific centralized events or (2) remote management reporting integration with standard network management tools. IT organizations need integrated soft-



FIGu RE 2.4.10 Generic product architecture for processing traffic measurement data.

ware systems that can be layered on top of the existing infrastructure to provide intelligent scheduling of requests and information.

2.4.5.2 Issues in Data Collection

The targeted metrics in data collection are the same as with log file analyzers, but the data source is different. When selecting products, there are a number of criteria, such as information depth, overhead, and reporting capabilities, that must be carefully evaluated. The market potential is good, with only a few vendors available. The criteria considered are also important when Webmasters want to position traffic measurements within their IT administration or deploy this functionality within their organization.

A product's design determines whether it can support a distributed architecture. Distribution refers to the fact that collecting, processing, reporting, and distributing data can be supported in various processors and at different locations. Figure 2.4.10 shows these functions with a distributed solution.

The monitors passively measure traffic in the network segments. They are actually microcomputers with ever-increasing intelligence. Their operating systems are either proprietary or open. Usually, they are programmed to interpret many protocols. TCP/IP, UDP/IP, and HTTP are high on the priority list of vendors.

The data capturing technique is essential with traffic measurement tools. Measurement probes are attached to the digital interfaces of the communication channels. They can reside directly on the network (stand-alone probes) or be co-located with networking equipment. In this case, the probe is used as a plug-in. Even software probes can be used and implemented in networking components or end-user devices. Hardware or software probes usually include event scheduling, which means determining polling cycles and time periods when downloading of measurement data is intended. Transmission should be scheduled for low traffic periods. Probes are expected to deal with large data volumes. These volumes depend, to a large degree, on visitor traffic in networking segments. Probes have limited storage capabilities; implementation examples show capabilities up to 24 hours. When this limit is exceeded, measurement data are overwritten by new data. Usually, measurement data are downloaded for further processing. It is important to know how downloads are organized and how rapidly they can be executed. As indicated in Figure 2.4.10, WANs are involved that may exhibit bandwidth limitations. Bandwidth is

usually shared with other applications, leading to potential traffic congestion. Bandwidth-on-demand solutions are rare with measurement probes. When transmission is arranged for low traffic periods, the actuality of measurement results may suffer. In such cases, local storage requirements increase, and processing, report generation, and information distribution are delayed by several hours or even by days.

Two solutions may help. The first is using intelligent filtering during and shortly after data collection. Redundant data are removed from captured packets during collection. Data volumes decrease, as do local storage requirements, but probe processing requirements increase. In the second solution, data compression or data compaction may be used with the same results and impact observed with the first solution.

Overhead is a critical issue with large data volumes. Data capturing is not expected to introduce any overhead in the case of hardware-based probes. Overhead is minimal with software-based probes. It is assumed that measurement data are stored away immediately after collection. If local processing is taking place, overhead must be critically quantified. If resource demand is high, probes must be upgraded properly. Data transmission overhead can be heavy when everything is transmitted to the site where processing takes place. It is too expensive to dedicate bandwidth to measurement and management functions only. If bandwidth is shared with other applications, priorities must be set higher for business applications than for transmitting measurement data.

It is absolutely necessary to capture all data that are necessary to conduct a detailed Web site analysis of visitors or groups of visitors:

- Who is the visitor?
- What is the purpose of the visit?
- Where is the visitor coming from?
- When has the visit taken place?
- What key words have brought the visitor to the site?
- What search machines helped to access the site?
- How long was the visit?

Data losses cannot be completely avoided. Probes, monitors, networking devices, user workstations, or transmission equipment may fail; in such cases, there will be gaps in the sequence of events. Backup capabilities can be investigated, but IT budgets will not usually allow a large amount of money to be spent on backing up high volumes of log file data. In the worst case, time windows are missing in reporting and in statistics. These gaps can be filled with extrapolated data.

As a result of high data volumes, databases should be considered to maintain raw and/or processed data. Database managers would then offer a number of built-in features to maintain data. Clustering visitors may be deployed from various perspectives, such as geography, common applications, common interests on home pages, and times of visits. Automatic log cycling can also be supported by database managers. ODBC support helps to exchange data between different databases and to correlate data from various databases. In addition to measurement data, other data sources can also be maintained in the same data warehouse. As an extension of routine log file analysis with concrete targeted reports, a special type of analysis—data mining—may also occasionally be conducted. Data mining can uncover traffic patterns and user/visitor behaviors. Both are important in sizing systems and networking resources.

One of the most important issues is how measurement data analysis performs when data volumes increase. Volume increases can occur when more pages are offered on more Web servers and when there are more visitors, longer visits, and extensive use of page links. In any case, collection and processing capabilities must be estimated and quantified prior to decisions on procedures and products.

In order to reduce processing and transmission load of measurement data, redundant data should be filtered as close as possible to data capturing locations. Filters can help avoid storing redundant data. Filters can also be very useful in the report generation process. Again, unnecessary data must not be processed for reports. Powerful filters can help to streamline reporting.

Not everything can be automated with measurement data analysis. The user interface is still one of the most important product selection criteria. Graphical user interfaces are likely candidates, but simple

products still work with textual interfaces. When measurement data are integrated with management platforms, this request is automatically met by such platforms.

Reporting is the tool used to distribute wire monitor analysis results. Predefined reports, report elements, and templates help to speed up the report design and generation process. Periodic reports can be automatically generated and distributed for both single Web servers and Web server farms. In the case of many Web servers, report generation must be carefully synchronized and scheduled. Flexible formatting helps to customize reports to special user needs.

Many report output alternatives are available. The most frequently used solutions include Word, Excel, HTML, and ASCII. Also, several choices are available with respect to distribution of reports:

- Reports can be stored on Web servers to be accessed by authorized users equipped with universal browsers.
- Reports can be uploaded into special servers or even pushed to selected users.
- Reports can be distributed as attachments to e-mail messages.
- Reports can be generated at remote sites; this alternative may save bandwidth when preprocessed data instead of completely formatted reports are sent to certain remote locations.

Documentation can assume various forms. An integrated online manual is very helpful in providing immediate answers. Paper-based manuals are still useful for detailed answers and analysis. This role, however, will be taken over by Web-based documentation systems. In critical cases, a hotline can help with operational problems.

Measurement data analysis is actually another management application. If management platforms are used, this application can be integrated into these platforms. There are many ways to integrate; in most cases, a CLI will be deployed.

2.4.5.3 Traffic Monitoring Tools

Only a few tools that support traffic monitoring are available. The large numbers of mergers and acquisitions during the past few years have led to frequent changes in the names of vendors and products. However, IBM, HP, CA, BMC, and Microsoft continue to provide solutions. Point products and new open source solutions are the targets of enterprise and service provider solution sets. Other wire monitoring tools are offered by Agilent, Datakom, Network General, NetScout, Omicron, Rohde & Schwarz, and Auritsu.

2.4.6 Load Balancing

In order to help IT managers track IP performance and to optimize bandwidth usage across WANs, several new vendors offer hardware- and software-based load-balancing products. Load balancers typically reside at the edges of corporate networks and determine traffic priorities. They apply a policy that defines different traffic types and determine what happens to each. A very simple policy may call for priorities for a specific sender. Other criteria may be TCP port numbers, URLs, and DNS. Traffic shaping may be supported by queuing or via TCP rate control. There are products available for both categories.

Optimization is accomplished by controlling enterprise traffic flows at the boundary between the LAN and the WAN. Because these products assign priority to traffic according to application type or even individual users, they allow IT managers to take the first steps toward policy-based QoS in their networks. These products are a logical evolution from the passive probes that allowed users a certain level of visibility for fault operations monitoring but no actual control over traffic. These products go further and can manipulate traffic. IT managers expect that this new class of traffic-shaping tools will ease bandwidth congestion without forcing the purchase of more and larger physical transmission lines. This segment introduces a couple of innovative solutions provided by start-ups and established flow-control companies.

2.4.6.1 Need for Bandwidth, Service Quality, and Granularity

For several reasons—the move to Internet/intranet-based business, the need for guaranteed bandwidth, the need for service-level agreements (SLAs), and the need for granularity—bandwidth management is rapidly becoming a must for Internet service providers (ISPs) as well as corporations running their own global intranets.

2.4.6.1.1 Move to Internet/Intranet-Based Business

Corporate networks are rapidly evolving from a classic client–server paradigm toward an intranet-based model based on information sharing and Web navigation. The number of enterprises transitioning their businesses to the Web, as well as the number of consumers, is growing almost exponentially. The result is a demand for significantly more bandwidth. Adding more channels and adding more bandwidth to each channel will not guarantee availability and performance where it is needed most. A Web-based model implies the following factors:

- Changing patterns of network use and unpredictable demands for bandwidth: Global users access the network 24 hours a day, 7 days a week. As information appears and disappears on Web sites, access patterns change and saturation moves around the network.
- Demand for increased amounts of bandwidth: People may stay on the link for extended periods of time and download large amounts of data.
- Demand for guaranteed QoS in terms of bandwidth and minimum delay: Emerging Internet applications are both bandwidth intensive and time sensitive, often requiring support for voice, video, and multimedia applications across the network infrastructure.
- Lack of control by IT staff: Workgroups and departments generally create their Web sites without IT approval, generating increased traffic without necessarily having the infrastructure to handle it. This often results in excessive traffic at the fringes of the network where Web sites are situated, generating traffic precisely where there is least provision.
- Changes in user attitudes: Users expect instant access to information without delays or restrictions, especially if that information is critical to their work.

2.4.6.1.2 Need for Guaranteed Bandwidth

Current networking technology has two major limitations:

The bandwidth available on a link at any given moment cannot be predicted in terms of quantity or quality. Bandwidth management is needed to allow applications that require a specific quality of service, in terms of bandwidth and delay (such as desktop video conferencing), to reserve the necessary bandwidth quality of service.

It is difficult to control which applications or users receive a share of the available bandwidth. In some circumstances, an application or a user can take control of all available bandwidth, preventing other applications or users from accessing the network. To solve this problem, the user can either add extra capacity at an additional cost, resulting in an overprovisioned network that still does not guarantee equal access, or introduce bandwidth allocation.

2.4.6.1.3 Need for Service-Level Agreements

VPNs are a popular value-added Internet service that corporations are increasingly moving toward. Enterprise customers seeking a VPN provider are more likely to sign with an ISP that can offer a contractual SLA, one that guarantees quality of service.

While SLAs cannot guarantee end-to-end service across the public Internet, they can be implemented for transport over a single-vendor network or for Internet server hosting. In these areas, an SLA is an important differentiator for an ISP.

Generally, the customer subscribes to a particular class of service and signs an SLA accordingly. Packet throughput is monitored as part of the agreement. It is expected that value-added, triple-play,

quad-play, and any-play services will exhibit very high growth rates globally. ISPs that want to get a piece of this additional business clearly need to implement bandwidth management in order to meet SLAs that guarantee QoS. Only efficient bandwidth management can enable them to tune network behavior so that customers receive the quality of service for which they are charged.

The new paradigm is a service-driven network. This responsive, reliable, modular infrastructure, based on the latest generation of management technology, is built on dynamic, flexible management services. To respond to today's business needs, ISPs and large enterprises must deploy service-driven networks; these networks deliver innovative services, such as unified roaming, push browsers, multicast, online shopping, push-to-talk, and streaming video, to customers faster and at a lower cost than ever before.

2.4.6.1.4 Need for Granularity

Bandwidth allocation based simply on filtering by protocol is not sufficient to meet bandwidth management needs. One of the key issues in this area is the extensive and increasing use of HTML/HTTP systems for OLTP. Within the next few years, the volume of HTTP-based OLTP traffic is expected to exceed the volume of traditional OLTP traffic. A fine level of granularity is needed for bandwidth management to take into account more than simply the protocol when assessing the relative importance of network traffic. Bandwidth management must base allocation not only on protocol type but also on the application and users involved.

2.4.6.2 Issues in Deploying Load-Balancing Products

Load balancing helps to utilize resources more effectively. At the same time, end-user response times can be stabilized and improved. This is an emerging area with a number of innovative hardware- and software-based products. The hardware solution is faster; software offers more flexibility if changes are required.

The functionality of a load balancer can be deployed in a stand-alone device or embedded into existing networking components such as routers, switches, and firewalls. The stand-alone solution offers broad functionality without affecting any other routing, switching, or firewall functions. But it will add components into the network, and these components must be managed; moreover, it may add another vendor to be managed. The embedded solution is just the opposite: easier management at the price of conflicting functions with its host.

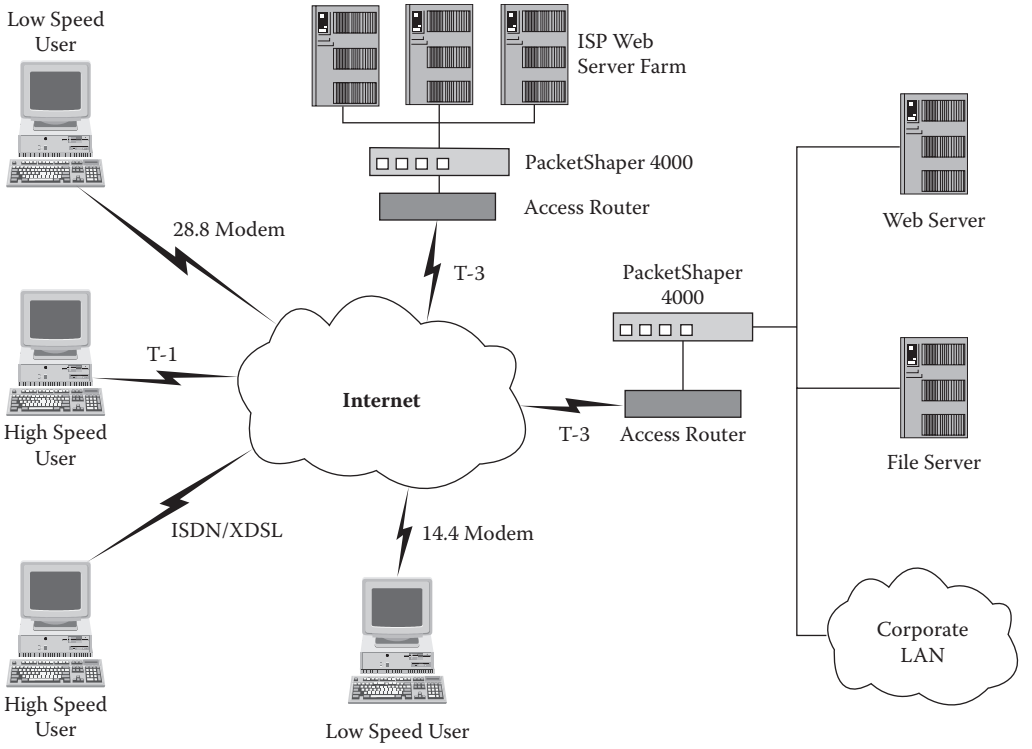
Load balancers are successful only when policy profiles can be implemented and used. Policy profiles are most likely based on supporting various transmission priorities. Priorities may be set by applications, users, or a combination of both. Solution technology may differ from case to case and from product to product, but most frequently the TCP flow is intercepted.

Load balancers are expected to support a number of services, such as quality control, resource management, flow control, link management, and actual load balancing. Advanced products support all of these services in dependency of page content. More work is required to gather the necessary information about content, but better services for high-priority content are available.

Functions in a narrower sense include traffic shaping, load balancing, monitoring, and baselining. Baselining, which refers to determining the optimal operational conditions for a certain environment, may be expressed through parameters such as resource utilization, availability, and response time. Load balancers should monitor these metrics and act on them. Traffic shaping and load balancing help restore normal conditions by splitting traffic, redirecting traffic to replicated servers, delaying payload transport, and so forth.

One of the most important questions involves the performance of load balancing when data volumes increase. Volume increases can result when more pages are offered on more Web servers and when there are more visitors, longer visits, and extensive use of page links. In any case, collection and processing capabilities must be estimated prior to decisions on procedures and products.

Load-balancing products can be managed by SNMP or Web-Based Enterprise Management (WBEM) agents. They are handled by managers as with any other kind of managed object. As before, various



FIGu RE 2.4.11 Packateer's PacketShaper in operation.

approaches may be taken for documentation and assistance to generate documentation. PacketShaper represents a group of products with sophisticated load-balancing capabilities. Figure 2.4.11 shows PacketShaper in operation.

Managing load balancers out of a management platform offers integration at the management application level. Baselining and monitoring may even be supported by other applications. When management intranets are employed, universal browsers can be used to view, extract, process, and distribute management information. The only prerequisites are that WBEM agents have been implemented and that Common Information Model (CIM) is supported for information exchange.

2.4.6.3 Load-Balancing Tools

Notwithstanding the changing names of vendors and products in the last few years resulting from mergers and acquisitions, solutions continue to be provided by IBM, HP, CA, BMC, and Microsoft. Point products and new open source solutions are the targets of enterprise and service provider solution sets. In addition, load-balancing and traffic-shaping tools are offered by such companies as 3Com, Cisco, Juniper Networks, F5, Foundry Networks, Nortel, Netscreen, Radware, and Microsoft.

2.4.7 Application Performance Management

Web application requirements have moved from nonexistent to mission critical within a very short period of time. However, the available tools have not kept up. In a business environment where "connections failed" means the same thing as "closed for business," IS/IT professionals are left to struggle with the challenges of building a highly available, high-performance server infrastructure. There are many interacting problems in this area:

- The majority of Web sites, both Internet and intranet, use single Unix derivatives, Solaris, Windows Family, or open source software servers. Like the mainframe solutions of the past, these centralized servers have become single points of failure. Even minor system upgrades become major service problems for demanding users.
- As the demands of interactivity grow, the cost of WAN bandwidth becomes a major factor. System configurations that force all user access out across the WAN for each request stretch out retrieval times and increase users' frustration levels.
- The increasing complexity of Web applications adds even more overhead; electronic commerce and multitier content architectures that build pages on the fly out of applications and databases make high reliability an even more important—and costlier—goal.

The severe problem in addition to all of these issues is that the Web technology base is narrow. In other words, solutions that can be applied to these problems are expensive and not very effective. Adding WAN bandwidth and a larger server are just the first steps in a never-ending circle. Adding mirrored, distributed servers increases server costs significantly as well as the complexities and costs of content distribution. Hiring more Webmasters and Web administrators to reboot downed Web applications and servers is not the ultimate solution. And, in a world of increasingly dynamic content and transactions, how effective will server caches and load-balancing tools really be?

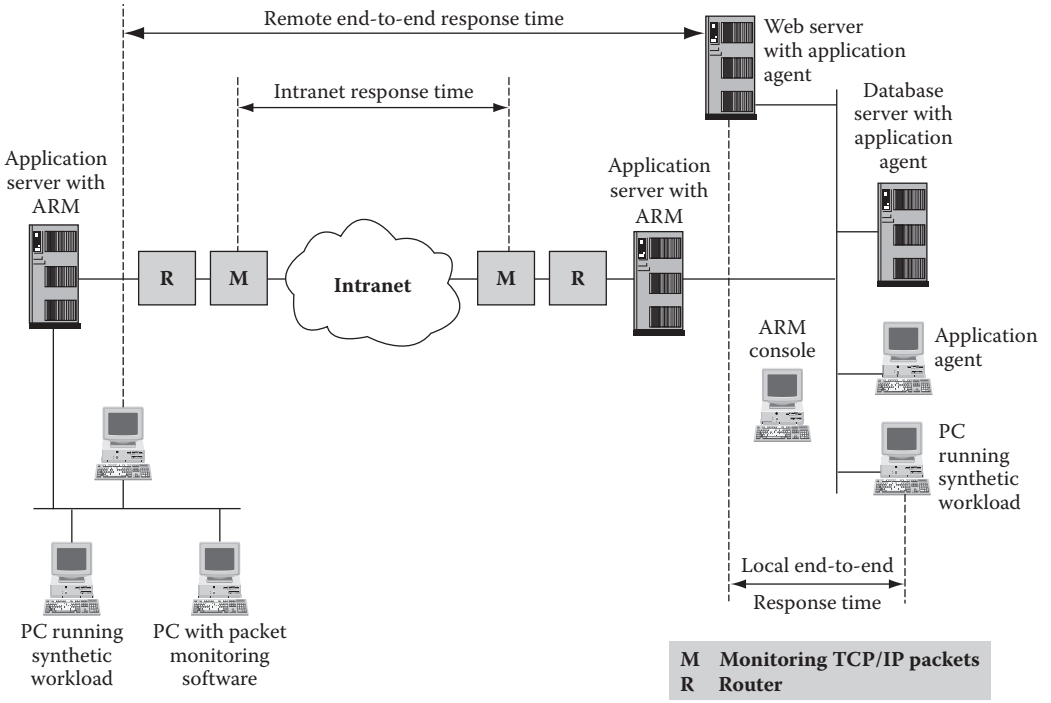
2.4.7.1 Response Time Measurements

Response time is one of the key SLA metrics. Its definition varies, but most users consider response time to be the duration between the sending of an inquiry and receipt of the full answer. There are two alternative response time measurements: (1) time elapsed until the first character of the response on the screen of the user and (2) time elapsed until the last character of the response on the screen of the user.

The second definition is better suited for the working cycle of users. The difference between RT1 and RT2 depends on many factors, such as the throughput of the backbone and access networks, servers in these networks, number of hops, and the hardware/software capabilities of the client's workstation or browser. Present measurement technology offers the following alternatives:

- Monitors and packet analyzers filter and interpret packets and draw inferences about application response times based on these results. These monitors passively listen to network traffic and calculate the time required for specific packets to move from source to destination. They can read the content of packages, revealing eventual application errors and inefficiency. However, they cannot measure response time end to end.
- Synthetic workload tools issue live traffic to allow a consistent measurement of response time on a particular connection in the intranet or for a given application. These tools are installed on servers, desktops, or both. They typically send TCP messages or SQL queries to servers and measure the time of the reply. Results from multiple sources are correlated to allow a more detailed view of intranet response times. Synthetic workload tools are very accurate with respect to end-to-end response time.
- Application agents work within or alongside applications using software that monitors keystrokes and commands to determine the length of time a specific transaction requires. They can run at both the client and the server, and they clock specific portions of the application at the server or the workstation. There is a need for customization in the use of application agents, along with a need to correlate many measurements to give users a performance estimate regarding their intranet.
- ARM MIBs define application programming interfaces (APIs), which allow programmers to write agents into an application so that network managers and Webmasters can monitor it for a range of performance metrics, including response time. This requires rewriting of existing code, which many companies are unwilling to do.

Figure 2.4.12 shows the locations of these tools and agents.



FIGu RE 2.4.12 Positioning response time measurement tools.

2.4.7.2 Application Performance Management Architectures

With most organizations relying heavily on the Web for customer interactions, online transactions are absolutely critical to business success. But performance issues are inevitable. The only way to ensure that service-level objectives are consistently met is through rigorous analyses of end-to-end transaction activities. However, traditional management tools are designed to provide a view of only a single point on the transaction chain. Organizations require an entirely new performance strategy, one that provides visibility into customer transactions and Web infrastructure performance.

With the advent of Web services and Web 2.0 technologies such as mashups, today's applications are too complex to manage with last-generation tools and technologies. New IT governance standards currently gaining acceptance among technology leaders also require that IT resources be managed more cohesively and proactively. What is needed is a holistic approach to application performance management (APM), one that employs systems that work across distributed enterprises. A key to getting application performance right is understanding the bigger-picture needs of the organization as defined through practices such as ITIL and CoBIT (GREE08).

The need for APM aligns closely with macro trends confronting IT. Application decomposition may well enable organizations to leverage information stored in previously inaccessible silos. Real-time Web services are required to make the necessary data available on demand. Management techniques should change given that some of the required data may be located outside IT's immediate control. "Webification" of existing enterprise applications often brings to light the need for new management systems and mindsets.

The APM architecture is built on three elements that enable testing and incident investigation capabilities: data collection, analysis engines, and reporting stations. These elements combine to build a suite of tools that proactively monitor systems and resolve application problems. In some cases, problems are diagnosed through active synthetic transaction monitors; in other cases, passive agent or agentless monitoring may be required.

Synthetic transaction monitors measure application performance by simulating user activity through predefined transactions. These tools can identify many user-perceived performance problems but often cannot determine where the actual problem is occurring. They require unique programming for each application monitored. Perhaps their most important use is in reporting user experience data, which can be trendable over long periods and through application revisions. Such data can be extremely useful in reporting on SLAs between an IT organization and its customers.

The other alternative is to capture application performance data passively by deploying software agents and/or hardware probes. While these provide a more detailed picture of the underlying application operation and infrastructure, they also can involve significant deployment and installation costs and require more operational attention. Such systems are likely to observe and record events that actually cause application performance bottlenecks, but locating those events and correlating them back to the business-impacting performance issue is an evolving science.

Hardware probes attach at key networking resources, such as switches and routers via their monitoring ports, and are normally passive. They also collect NetFlow statistics (a de facto standard for routers initiated by Cisco) to gain a more complete view of the IP infrastructure. The result is that these probes gather a large volume of data. To prevent these data from overloading the network—particularly WAN links—analysis engines must be deployed throughout the infrastructure. These systems aggregate and process the data from the various probes and, depending on the size of the organization, consolidate data from a number of sites. Redundant data must be deleted as soon as reasonable. Software probes operate similarly and are less expensive, but they may result in some overhead in the resource in which they are implemented and activated.

Finally, a monitoring station or management console that enables operators and managers to query the various components from a single location is required. Numerous functions and technologies are made available within the context to various managerial levels. Most important is the ability to analyze and correlate results from many locations such as the user's desktop, networking components, and the data center. The monitoring station should be capable of problem resolution functions and service-level monitoring and reporting.

2.4.7.3 APM Products and Product Suites

While many IT organizations strive for deep visibility into critical business application metrics, they are typically concerned about products that depend on manual, labor-intensive approaches to deploying, managing, and monitoring applications.

Holistic APM products are the answer. They not only identify problems but take corrective actions to resolve performance issues before they affect users and customers. Actions may include allocating additional bandwidth or server processing capacity and even rolling back configuration changes. When corrective-action capabilities are lacking, organizations will struggle with fixing problems manually, although fast identification and notification are important as well. Holistic products should locate and predict performance problems across all application components and support all facets of the application infrastructure, including system and network performance. Service modeling is important as well, including modeling critical application services and SLA compliance reports.

All APM products require application data. They may use proprietary agents, existing system agents, network packet capture, synthetic transactions, or adapters that pull data from other sources. Many products combine these mechanisms to capture data and aggregate them into meaningful information.

While synthetic transactions can detect issues and packet-capture tools can pinpoint the applications at fault, only agents can go deep and determine the specific root cause of the performance problem. To accomplish this, application agents and correlation are required. Products using synthetic transactions are effective in discovering performance issues, but alone they are unable to determine the cause. While network packet capture helps determine whether an application is occupying excessive bandwidth and causing performance issues, such products are best suited for troubleshooting more tactical issues owing to their network engineering approach to applications.

TABl E 2.4.2 Tools for Application Performance Management

Vendor	Product
BMC	ProactiveNet Analytics, Performance Manager
Compuware	ServerVantage, NetworkVantage, ClientVantage
Indicative	Service Director
NetIQ	AppManager
NetQoS	SuperAgent Appliance
Network General	Sniffer, IStream, AppIntelligence, Visualizer
NetScout	NGenius
Nimsoft	Nimbus
Quest Software	Foglight
Symantec	I3 Suite: Inform, Insight, Indepth
Computer Associates/Wily	Introscope
IBM	Tivoli suite
Hewlett Packard/Mercury	SiteScope

The leading products rely on application agents in combination with network packet and synthetic transaction data. While at times cumbersome, agents provide the best visibility into issues at the application layer. The bottom line is that there is no one-size-fits-all APM tool. Organizations that need flexibility and scale should expect to invest in substantial configuration and ongoing maintenance. Point products will provide value but may fall short for those seeking end-to-end insight. At present, no product is leading in terms of proactive corrective actions to resolve performance problems. While many can integrate with other automation tools, the ultimate holistic APM tool is not yet present.

Biddick (BIDD08) recommends the following criteria when evaluating products:

- Broad support for existing applications
- Ability to detect and report on problems as they occur
- Architecture to support distributed APMs
- Root-cause problem detection
- Ability to integrate with the surrounding environment
- Customization needs
- Overhead of transmitting measurement data
- Load increase due to synthetic workload
- Ability to collaborate with modeling tools

Table 2.4.2 lists tools that might be considered in application performance management.

2.4.7.4 Summary and Trends

The ultimate goal is to gain a holistic view that accounts for the unique characteristics of each device and system required by an application. With a coherent view of the application's end-to-end performance, managers can better understand the implications of infrastructure changes, support application lifecycle planning, and ultimately improve the ability to deliver what matters most—a satisfied user.

2.4.8 Web Performance Management Trends

Intranet management is an emerging area for Webmasters and Web administrators. It combines existing fault, performance, configuration, security, and accounting management processes with new management tools. Performance and security management are the two most challenging areas. Usage patterns, traffic peaks, unbalanced input/output streams from and to Web servers, server overload, and unstable

performance are areas in which Webmasters and network capacity planners face challenges. Security officers in corporations operating intranets must address proper partitioning of networking segments, selection and implementation of firewalls, stress testing of firewalls, and use of appropriate authentication techniques.

New intranet-related management tools—content authoring and auditing instruments, log file analyzers, traffic monitors, load balancers, and application monitors—can be used individually or in combination. It is expected that they will soon be integrated into systems and network management platforms.

References

ALDR99: Aldrich, S. Freshwater's Web application management, Patricia Seybold Group eBulletin, January 21, 1999.

BIDD08: Biddick, M. Cream of the APM crop, InformationWeek, pp. 46–47, February 18, 2008.

BOARD06a: Boardman, B. Cure the network hiccups, Network Computing, pp. 36–38, September 14, 2006.

BOARD06b: Boardman, B. Systems management, Network Computing, pp. 45–54, September 14, 2006.

BOBR98: Bobrock, C. Web developers follow old scripts, Interactive Week, p. 29, November 2, 1998.

BOCK98: Bock, G. E. Microsoft Site Server—Organizing and sharing the contents of a corporate intranet, Workgroup Computing Report, Patricia Seybold, August 1998.

BRUN99: Bruno, L. IP Balancing act: Sharing the load across servers, Data Communications, p. 29, February 1999.

GIBB98: Gibbs, M. Pinning down network problems, Network World, p. 43, March 2, 1998.

GREE08: Greenfield, D. Orchestrate flawless performance, InformationWeek, pp. 47–50, January 28, 2008.

HALL06: Hall, E. Take a byte out of management, Network Computing, pp. 40–42, September 14, 2006.

HERM98: Herman, J., and Forbath, T. Using Internet technology to integrate management tools and information, http://www.cisco.com/warp/public/734/partner/cmc/bmi_wi.htm.

HUNT96: Huntington-Lee, J., Terplan, K., and Gibson, J. HP OpenView—A manager's guide, McGraw-Hill, New York, 1996.

JAND98: Jander, M. Clock watchers, Data Communications, pp. 75–80, September 1998.

JAND99: Jander, M. Network management, Data Communications, p. 75, January 1999.

KAPO98: Kapoor, A., and Ryan, J. Reassessing networks for an IP architecture, Telecommunications, p. 48, October 1998.

LARS97: Larsen, A. K. All eyes on IP traffic, Data Communications, March 1997.

LEIN93: Leinwand, A., and Fang, K. Network management—A practical perspective, Addison-Wesley Publishing Company, New York, 1993.

MACV06: Macvittie, L. Tracking customers click by click, Network Computing, pp. 24–26, June 8, 2006.

POWE97B: Powell, T. An XML primer, InternetWeek, pp. 47–49, November 24, 1997.

REAR98: Reardon, M. Traffic shapers: IP in cruise control, Data Communications, p. 67, September 1998.

RUBI98: Rubinson, T., and Terplan, K. Network design—Management and technical perspectives, CRC Press, Boca Raton, 1998.

SANT97: Santalesa, R. Weaving the Web fantastic—Authoring tools, InternetWeek, November 17, 1997.

SCHU97: Schultz, K. Two tools for monitoring your Web site, InternetWeek, pp. 60–61, October 27, 1997.

STUR98: Sturm, R. Working with Unicenter TNG, QUE Publishing, Indianapolis, 1998.

TAYL96: Taylor, K. Internet access: Getting the whole picture, Data Communications, pp. 50–52, March 1996.

TERP96: Terplan, K. Effective management of local area networks, Second Edition, McGraw-Hill, New York, 1996.

TERP98a: Terplan, K. Web-based systems and network management, Xephon Briefing, London, October 14, 1998.

TERP98b: Terplan, K. Telecom operations management solutions with NetExpert, CRC Press, Boca Raton, 1998.
TERP99: Terplan, K. Web-based systems and network management, CRC Press, Boca Raton, 1999.
TERP05; Terplan, K. How to measure Web performance, CECMG Conference, Ulm, 2005.

2.5 Application Performance Management

Vadim Rosenberg

2.5.1 Introduction

Traditionally, communications service providers (CSPs) have relied on high-end, specialized hardware/software bundles to deliver a small, selective set of services. Today, with the industry's adoption of IP-based data transmission and call control protocols, the role of standards-based software platforms has increased dramatically. While more new revenue and subscriber retention depend on software-based value-added services and applications, the CSPs are not competing on the new features alone. Increasingly, rapid time-to-market, quick reaction to new requirements and business conditions, and top customer experience with the Next Generation (NG) services are becoming the key factors for CSP's success in this new environment. Any incremental improvement in the performance and availability of software-based services or applications results in a significant improvement in revenues, reduced customer churn, and improved reputation in the market. Thus, service assurance and customer experience management functions are increasingly becoming critical to the CSPs overall success.

Traditional network-based service assurance solutions can only show part of the picture. They are not capable of providing visibility into the application-level metrics and vital signs that contribute into the overall quality of service and end user experience. In order to adequately serve CSPs, service assurance solutions need to include application performance management (APM) in real time as the key component essential for the new software-based service delivery environments.

This segment examines the need for real-time APM within the CSP environment and how to meet the new requirements. It starts off by introducing general application performance management concepts and their evolution over time. After the modern-day APM standards and capabilities are explained, the section discusses the use cases, the essential requirements, the solution, and the future evolution of APM addressing the needs of the CSPs and Next Generation Networking (NGN)-based services.

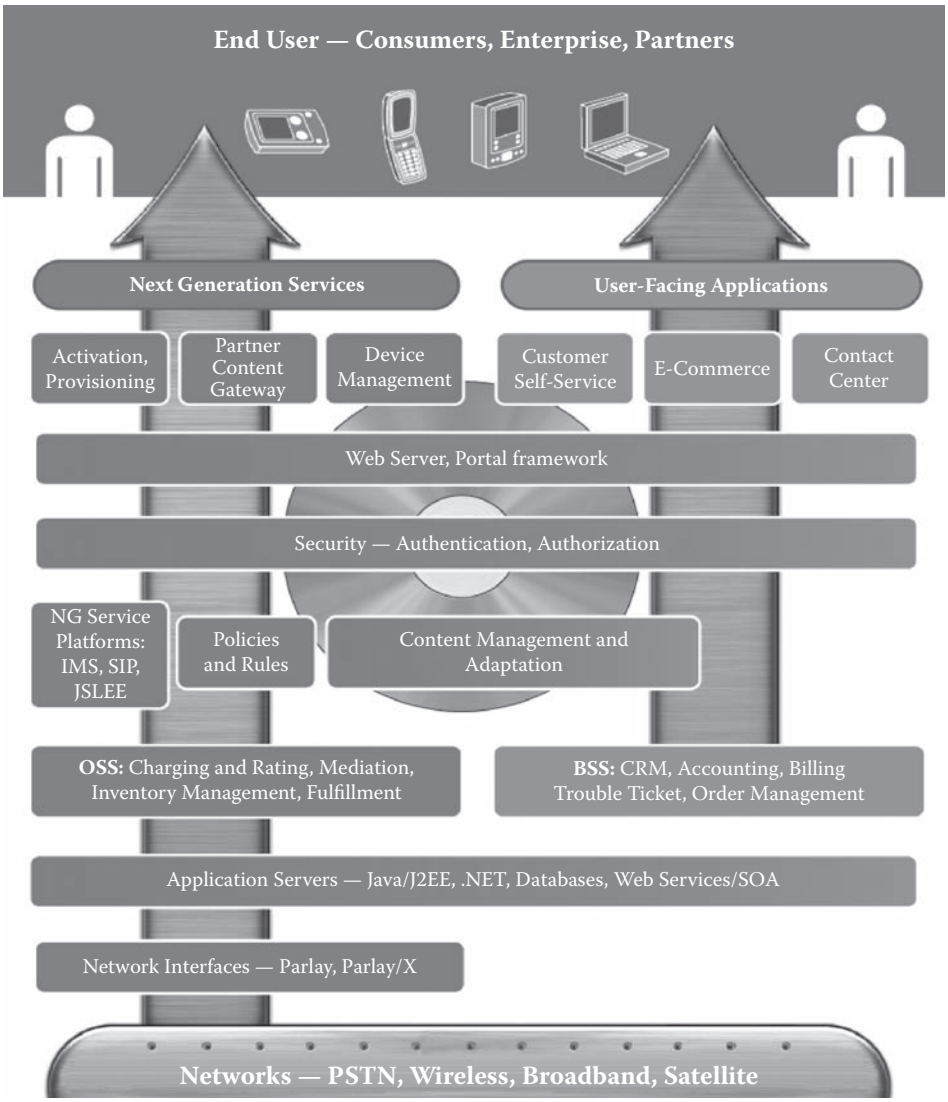
2.5.2 The Need for APM in Communication Services

The telecommunications industry is undergoing major changes both in the way business is done, as well as in the infrastructure that supports the services provided by the businesses. These changes call for new types of service assurance solutions based on APM technology.

2.5.2.1 New Possibilities Bring New Challenges

In the past, services like voice, 800 numbers, and caller ID were bundled with the large network switches and represented major investments depreciated over decades. With the emergence of Service Delivery Platforms (SDP), the NG services are ordered and delivered using generic (often even open source) Web and application servers, deployed on cheap, "commodity" hardware, delivered over IP networks—both private and public, and have an undetermined life span. Some services are rolled out "just in time" and then retired when revenues dry up—sometimes after days or even hours.

And what about OSS/BSS? Previously used internally and exclusively, billing, order management, Customer Relationship Management (CRM), and trouble ticket systems are now exposed to end users on a 24-hour, 7-day-per-week basis. Self-service provides customers with the same functions as a call center, but without waiting. For operators it is also a huge cost saving. The transition is happening



FIGu RE 2.5.1 In converged telco environments, layers of standards-based software deliver rich new services to end users. Managing this software presents a new critical challenge to telcos.

here as well: self-services are delivered via Internet technology using Web servers, portals, security, and messaging software, some of which wasn't even designed to handle telco availability and performance requirements. Previously billing or order management systems could be placed under scheduled maintenance or into batch mode at night with no significant impact on business, while today, customers expect round-the-clock availability. Online billing, support, plan configuration, and activation are not fancy nice-to-have features; they are considered essential and customers expect them to work. And this does not include services that can only be ordered or provisioned through self-service. A well-known voice over IP (VoIP) service provider sells their IP-enabled devices to be activated by customers online. If that feature does not work reliably, the vendor would have a serious problem. We are seeing a dramatic shift in telco. This shift makes standards-based software a critical, yet least understood, component in the otherwise familiar telco landscape shown in Figure 2.5.1. The new converged telco comes with new management challenges that may not be fully understood due to their complexity.

Another interesting side effect of this transition is the blurring of the line between your NOC and IT department. Who is now responsible for service delivery? Who owns the customer? Who delivers the revenue? How do you trace transactions across your NOC and IT infrastructure? How fast can you detect and analyze performance problems that affect new services?

In the past, network and element management tools provided all that was needed. With mostly network-based services, all we needed to know was if all network elements were healthy, if the bandwidth was fully utilized, and if all faults and errors were detected. These proprietary hardware-software bundles had their own management solutions; not to mention that the reliability of a typical Class 5 switch or similar Intelligent Network (IN) equipment was pretty high, to say the least. Users were getting their service directly from dedicated network elements, and therefore it was easy to understand the user experience from the metrics collected from the network itself.

Now, not only do you need to manage transactions running through layers of applications and interfaces, but you also have to derive the real customer experience out of it! And, if there is a problem, you need to detect and analyze it before your customers start calling or leaving.

2.5.2.2 The Rise of Software

We are not talking about the software that has been driving IN for years. We are referring to the software behind the Next Generation (NG) services like music, gaming, IPTV and mobile TV, navigation and location-based services, presence-based services, e-wallet, and m-commerce. This software is based on IT platforms (Java/J2EE, .Net, Web services/SOA), telco technologies (IMS, SIP, JSLEE, Parlay/X), and standard networks (TCP/IP), and delivers new customer-facing applications and value-added multimedia services. These new services must be as available as dial tone because whole new revenue models are based on their availability. They should be as user friendly as an old phone, a TV, or a fax machine. As we all know, telcos expect these services to make up for the declining revenues from legacy services.

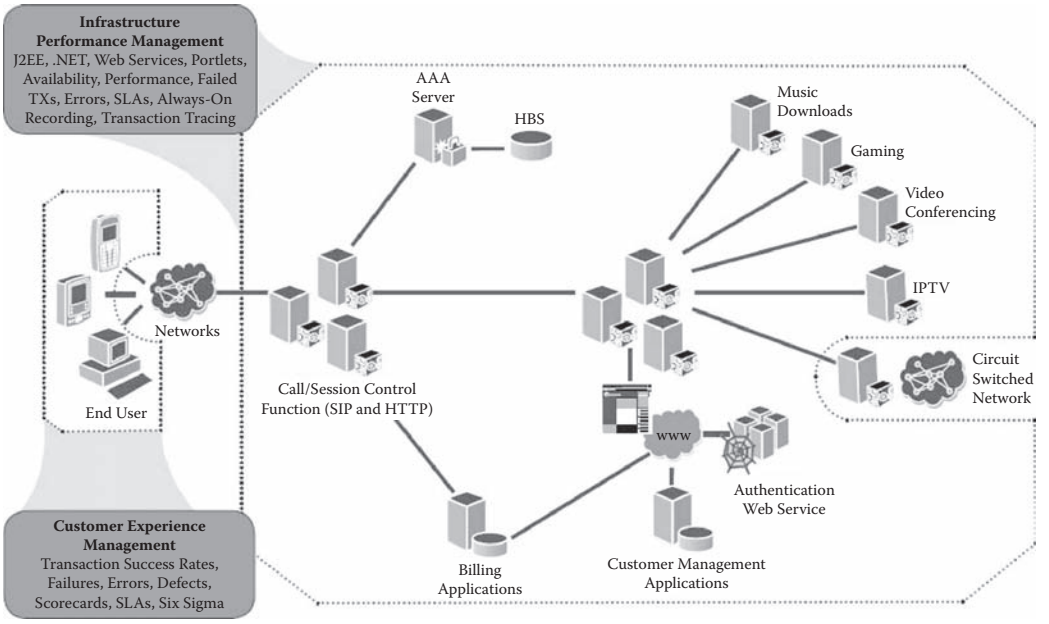
The communications services infrastructure is in the midst of a transformation from a proprietary, homogenous, circuit-based, hardware-centric telecommunications infrastructure to an open, heterogeneous, IP-based, software-centric one. The industry is in an evolution to NGN architectures such as IP Multimedia Subsystems (IMSs). An important part of this transformation is the adoption of standard protocols and interfaces, the underpinnings of the new services NGN introduces. Collectively these technologies enable a standard, repeatable way of building and deploying service applications, and this in turn has led to the advent of Service Delivery Platforms (SDPs).

With software-based SDPs, service silos (i.e., each service requiring its own stack of specific service components) are being replaced with a convergent platform utilizing common, reusable components to rapidly build and deliver rich value-added services. A CSP is no longer limited to the number and types of services it can offer, both home-grown and those of third-party vendors.

However, while these changes bring about many benefits, they also come with their problems. With new architectures, applications, protocols, and programming languages comes the need for new tools and ways to monitor and manage the quality of service of the systems serving the business need. The dramatic growth of new services further aggravates the management problem—more services to manage, with more boundaries and interdependencies. Increasingly, the innovative CSPs find themselves not being able to effectively monitor all the contributing components of their quality of service and the end user experience. When a problem is detected, especially retroactively, the blame game of which department is at fault starts. Problem management of the blame game escalates rapidly. Additionally, the real-time nature of the CSP business requires it to be able to quickly and proactively maintain the highest levels of performance and reliability which, though once the sole burden of hardware, is increasingly being shouldered by software.

2.5.2.3 The Acceleration of Business

On the business side, the pace of business is increasing as well as the demand for strong service level agreement (SLA) support. The breadth of a typical CSP's service offering has increased dramatically,



FIGu RE 2.5.2 Modern OSS/BSS and SDP environments leverage IT and Internet standards and technologies while providing next generation services to customers.

as well as the geographical reach. These factors are resulting in an increased customer base and subscription level. Any downtime means loss of revenue, damaged reputation, and increased customer churn.

The competitive climate mandates the ability to integrate and deploy new services rapidly, sometimes having a complete life cycle of just weeks. Increasing demands for service from the customer base has seen the once standard telco availability requirement of "five 9s" increase to "six 9s." Problems must be identified, triaged, diagnosed, and rectified before they start affecting end users. This indicates the mandatory requirement for both real-time APM and proactive problem identification and rectification.

The legacy services, such as basic voice and voice mail, have become commoditized. SMS/MMS has clearly become mainstream. CSPs are putting high hopes on the new value-added services based on content (ring tones, music, wallpaper), streaming media (video, IPTV), location-based services (using GPS), and miscellaneous services such as push-to-talk, gaming, dating, traffic info, etc. The availability and performance of the underlying applications, as well as of the Call Session Control Functions (CSCFs) based on Next Generation Network (NGN), are important issues to CSPs. They are seeing an increasing portion of revenues and revenue growth coming from many of these software applications. Every percentage improvement in the performance of service software applications usually infers even greater percentage improvement in revenues. Thus, performance management of these applications is increasingly becoming critical for communications service providers.

The uptake of successful services can be rapid, almost viral. As many operators are opening their interfaces for third-party content, such monitoring helps ensuring proper revenue sharing, auditing real-time bandwidth usage and policies fulfillment, and avoiding fines from SLA violations.

A similar transformation is taking place on the BSS side of the CSP environment. Previously used internally, billing, order management, CRM, and support systems are now exposed to end users 24x7 through self-service. Today, customers expect round-the-clock availability. A transition is happening here as well: self-services depend on Web servers, portals, security, and messaging software, some of which wasn't even designed to handle the availability and performance requirements of telecoms.

2.5.2.4 The Leading Performance Management Issues Today

Optimizing user-facing applications. The goal for customer inquiries, purchases, and support is first contact resolution—be it at the point-of-sale, call center, or in a self-service portal. Completing transactions efficiently requires gathering information in real time from multiple sources to create a single view of the customer, and then delivering it without delay. Any degradation in the performance or availability of the applications that must complete these functions will translate into lower productivity, decreased customer satisfaction, and in many cases, lost revenue.

Ensuring that next generation network software platforms and applications are carrier grade. With the shift toward software-based SDPs and IMS as part of Next Generation Networks (NGNs), operators and service providers need new tools to manage the complex software platforms within the network while IT professionals need to ensure that supporting OSS/BSS can scale to meet new real-time demands. They also need to effectively manage the performance and availability of services delivered on the network by third-party providers.

2.5.2.4.1 Optimizing OSS/BSS

OSS and BSS systems need to be ready for the real-time demands of Next Generation Networks, but they must also be ready to handle other immediate IT challenges. The rapid rollout of new billing models and provisioning requirements combined with continued growth of subscribers puts extra pressure on existing systems. Recent moves by many operators to rationalize their OSS/BSS systems promise to reduce costs and increase flexibility, but also introduce new complexities and management risks.

2.5.2.4.2 Ensuring Availability of Information Hubs and Busses

Increasingly, network operators are using integration platforms, SOA, and Web services to support growing transaction volumes and provide greater and deeper integration. This enables OSS/BSS to effectively share information, get a single view of the customer, and create interfaces with wholesale customers, content and service providers, and internal divisions. These systems require proactive management capabilities that enable proactive detection, isolation, and elimination of transaction performance issues.

2.5.2.4.3 Meeting Service-Level Agreements

Commercial relationships and regulatory demands place very strict SLAs on network operators. Failure to meet SLAs results in fines and loss of commercial opportunities. Without detailed systems monitoring and alerting, SLAs may be breached before proactive action can be taken to fix the issue.

2.5.2.5 The Need for New Type of Service Assurance Solutions

In the past, network and element management tools provided all that was needed to assure the high quality of service delivery. With mostly network-based services, all we needed to know was that all network elements were healthy, that the bandwidth was fully utilized, and that all faults and errors were detected. End users were getting their service directly from dedicated network elements, and therefore it was easy to understand the user experience from the metrics collected from the network itself.

Now CSPs need to manage transactions running through layers of applications and interfaces, and they also have to create the customers' experience through these applications and interfaces. Today, CSPs are competing less on features and are mostly fighting for the same customers in a highly disloyal environment. Customers are more often affected by the quality of their experience with one provider versus another than by the features offered by a particular provider. If there is a problem, CSPs need to detect and analyze it before customers start calling or leaving.

Customer satisfaction can rapidly improve by monitoring user experience in real time while correlating real response time to the end user with the events and processes happening in multiple components. This, in turn, increases customer retention and consequential revenues, and decreases costs of IT firefights.

The progressive, forward-looking CSPs are finding that:

1. Service assurance needs to be upgraded to include real-time proactive application performance management solutions.
2. For end user–facing environments and OSS/BSS applications, it is essential to get deep visibility into the presentation layer, the OSS/BSS applications and IT infrastructure, and all the way to the edge of the network.

Monitoring user experience in real time while correlating real response time to the end user with the events and processes happening in multiple components can help dramatically improve customer satisfaction, and in turn increase customer retention and consequently revenues, while decreasing costs of IT firefights.

For SDP, service assurance needs to be upgraded to include application performance monitoring solutions. If the new service is successful, its uptake can be rapid, almost viral. Real-time event and performance monitoring is essential to preempt problems and to support network planning. As many operators are opening their interfaces for third-party content, such monitoring helps to ensure proper revenue sharing, monitoring of real-time bandwidth usage and policies fulfillment, and avoidance of fines from SLA violations.

Doing this is not a simple task, but there are best practices that can be employed, including the following:

Deep diagnostics and visibility into the execution environment (portal, partner gateway, Java, or .NET application server, SIP/JSLEE, Web services and SOA), all the way to network interfaces like Parlay and Parlay/X. With deep visibility you can trace an individual user transaction as it touches different components and systems, and correlate it with other events to understand the root cause of a problem.

Monitor all data all the time to catch early symptoms as they happen. With probing, those symptoms may fall between the intervals and not manifest themselves with synthetic transactions. You need to monitor real transactions in real time. Probing still makes sense when real transactions are few and rare, but it will not tell you how the system would behave under the real load. Also, all data needs to be monitored for deep analysis, not just a few data points in synthetic transactions.

Management should be nonintrusive and have little or no impact on the managed environment itself. There are ways to optimize metrics collection so overall additional load on the system is minimal.

Role-based customizable reporting gives different units within your organization the appropriate view of the environment relevant to their function (business managers, NOC and IT administrators, or QA and software engineers).

2.5.2.6 Areas of Applicability

Communication service providers can achieve significant financial and qualitative benefits by proactively managing the availability and performance of their mission-critical OSS/BSS applications, SDP and deployed services, and network interfaces.

2.5.2.7 Increasing the Profitability and Efficiency of Service Delivery

In the context of service delivery, the link between performance and bottom-line results is especially pronounced:

- The ability to quickly identify and eliminate performance risks from the rollout of new services accelerates time to revenue and the return on investment (ROI) of new service creation.
- Improvements in availability of mission-critical, customer-facing systems prevent customer churn and preserve the integrity of the brand, in addition to directly increasing revenue.
- Acceleration of service ordering, activation, and provisioning enhances the revenue potential of every service and avoids customer support costs.

- Ensuring that performance threats are identified and isolated before they affect service quality helps meet SLAs and protect revenue.

2.5.2.8 Assuring High Quality of Customer Self-Service (CSS)

By preventing availability and performance issues from reaching the customer, operators can enjoy the full benefits of moving customer interactions to self-service channels. The financial benefits come from higher revenue through self-service channels, reduced operating costs, and more effective and efficient IT operations. High-quality self-service also drives qualitative benefits including improved customer satisfaction and retention, positive word-of-mouth, and enhancement of the company's brand image. The broader benefits achieved through proactive monitoring and SLA management of self-service applications are:

- Understanding of customers' actual experiences of interacting with the CSS system
- Ability to create early-warning flags about performance threats
- Presenting performance data in ways that are actionable for audiences ranging from Web application owners to middleware groups to customer care executives
- Ability to rapidly isolate and permanently resolve issues when they arise

2.5.2.9 Enhancing Contact Center Operations

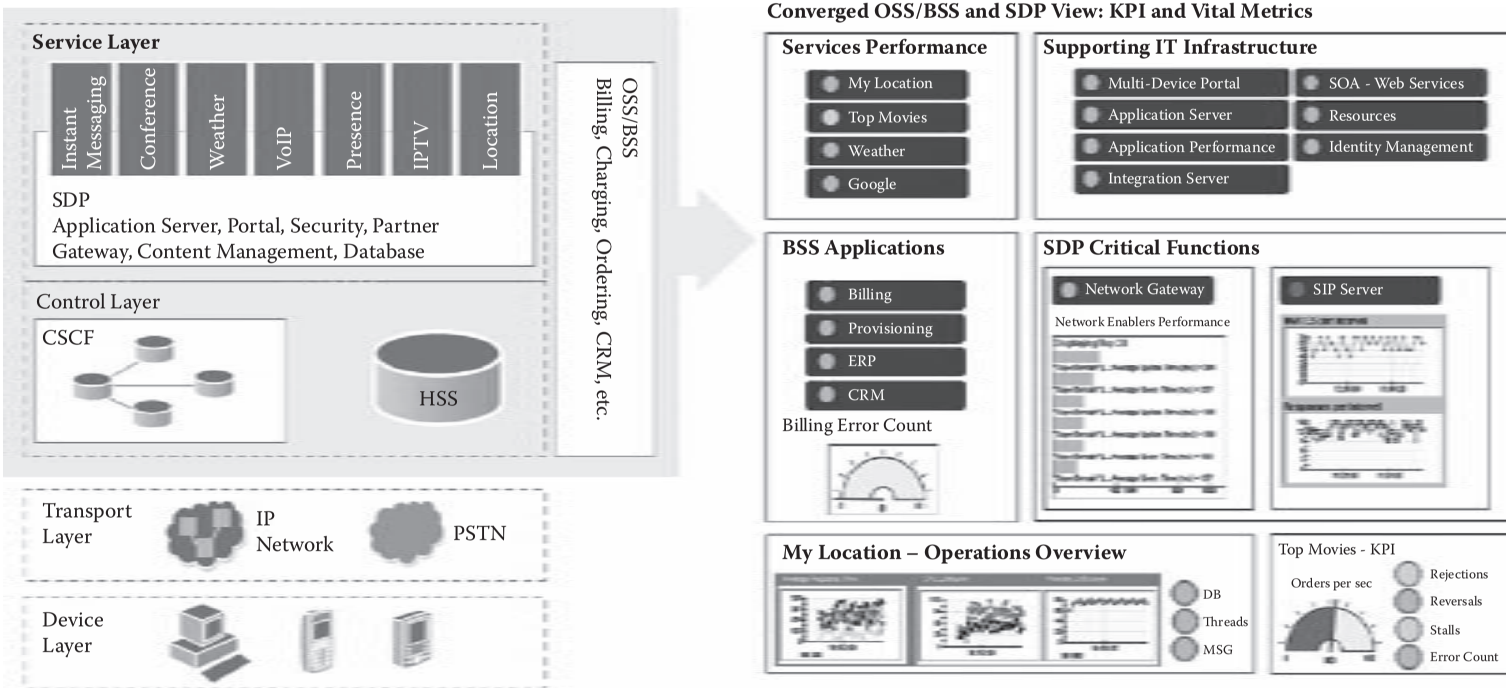
CA Wily Technology helps operators achieve improvements in contact center application performance and availability, which have a direct impact on almost every contact center success metric. Customers have realized significant cost savings and revenue gains from monitoring and managing their mission-critical applications with CA Wily Introscope and CA Wily Customer Experience Manager. By measuring key parameters, such as screen pop times or transaction success rates against customized thresholds, operators have been able to identify and resolve performance threats before they impacted end users. When a performance problem does occur, customers are able to quickly isolate the source, significantly reducing the length and impact of outages and performance issues. The specific benefits from proactively using performance monitoring products in a contact center environment are:

- A more available and responsive agent desktop shortens both handling time and after-call work time, which reduces the average cost per contact.
- Shortened handling time provides time for additional cross-sell or up-sell initiatives.
- Better-performing agent desktop and back-end connectors also ensure that agents always have the right information and recommendations, thus reducing queues, wait times, and abandonment, while raising revenue.
- Customer service representative (CSR) job satisfaction rises if the tools they use empower rather than frustrate them in accomplishing their goals.

2.5.2.10 Operational APM

The operations side of a CSP is responsible for the actual communications services delivery and thus has the clearest need for a service assurance function. For example, an IMS architecture defines three management planes: transport, control, and service. Within the control plane resides the core call session control function (CSCF) and user profile database (HSS), and within the service level reside the value-added services. Business transactions are overlaid over functional transactions such as Session Iniation Protocol (SIP) traffic flows between the CSCF and applications. Visibility of these transactions is the key to understanding the performance and health of the operational systems and preventing problems, and in diagnosing any problems that do occur.

From a use case perspective, the user of a CSP service interacts with either another user or with an application. These interactions occur as transactions over several systems and can involve several protocols and applications. In the past, there was a division of labor in monitoring these transactions—use protocol analyzers for protocol (network) flows, and application profilers for application traces, and then somehow the two were correlated. This is a tedious, time-consuming, and error-prone method. A



FIGu RE 2.5.3 New Generation services and OSS/BSS must be managed in a new way where performance and event metrics are collected from within the environment to correlate distributed transactions, heterogeneous components, infrastructure, and protocols.

significant improvement is to tap into the transaction flow end-to-end with a common toolset providing visibility into key areas such as SIP message processing within the CSCFs, HSS, and applications.

2.5.2.10.1 Use Case 1: Diagnostics of a Service Outage

Service outages can be quickly detected, analyzed, and corrected through deep diagnostics and transaction tracing before the majority of customers suffer.

When a critical revenue-generating service experiences a sudden performance problem, an executive responsible for the service can get a high-level alarm once certain KPIs or Key Performance Indicators are violated; for example, when "place order" transactions have an unusually high number of errors or slow performance. At the same time, IT staff gets a detailed view of the specific software areas of the environment. "Place order" transactions may fail because the database is reaching its capacity, so searches for customer information are slow and eventually time out. If the transaction involves a third-party content provider there may be a problem with the partner gateway, or even a licensing issue. If an operation on a network element is required, a Parlay interface to that element may be reporting errors. In all of these cases, real-time deep visibility into different layers of software can help quickly detect, troubleshoot, and resolve the problem.

2.5.2.10.2 Use Case 2: Detection of New Subscriber Activation Failures

This use case exemplifies APM for the diagnosis of a service activation problem. Delayed activation of a new subscriber delays revenues from that customer, and can even cause a loss of a customer. Such a situation can occur if a subscriber is signed up within the BSS but is not provisioned in the OSS. Often the CSP is first made aware of the problem when the new customer reports a complaint; for example, a customer complains of not being able to place a call though a VoIP terminal device that has not been properly configured.

The use of a real-time APM solution can diagnose this problem. One method would be to investigate for failed SIP registrar transactions, which indicate that a client was unable to register their device for access to the network. Once found, this transaction, which traverses asynchronously through the operational system, can be traced to the point of failure, which in this case would be the HSS failing to authenticate the user as the user profile was not configured in it (as one possibility).

2.5.2.10.3 Use Case 3: Detection of Overloaded Services

This use case exemplifies APM for prevention of a problem. An overloaded service means that one or more users are unable to use a service. In a pay-per-use environment, this equates to an impact on revenue. Often such occurrences are known to the CSP only when (and if) a disgruntled customer reports the problem. Being proactive is important as surveys have shown that a customer experiencing this frustration often switches to another service provider.

The use of an APM solution can reduce or eliminate this problem. One way to enable this is to monitor SIP errors. If a called SIP endpoint is unable to respond to a call request (i.e., is overloaded), the calling terminal is sent the "480 Service Not Available" error code. A spike in the issuance of these errors from service endpoints indicates an overloaded service case. Other detection criteria include stalls or time-outs of SIP invite messages, or an increase in SIP "503 Service Unavailable" errors. Provided there is visibility into SIP transactions, these situations can be detected and the afflicted service identified.

2.5.2.11 Business APM

CSPs are increasingly turning to customer self-service (CSS) channels to manage the increased number of customers they must service. The availability and reliability of these applications is vital to customer happiness. The following use case introduces the typical benefits an APM solution can provide and then describes how a business view of APM can benefit.

2.5.2.11.1 Use Case 4: Customer Self-Service

CSPs are increasing the use of Web-based applications (as the most common CSS channel today) to provide customers with all kinds of services, enabling both increased revenues via subscription of new or additional services and cost savings via reduced costs for customer care. CSS transactions can be complex as they span several systems from the customer to the back-end systems. Maintaining the transaction rate is important to maintaining revenues, and increasing the transaction rate correspondingly increases potential revenues.

A business transaction–aware, real-time APM toolset that can monitor end-to-end CSS transactions as they occur gives the business owner, in this case the customer care unit, visibility into the success of catering to customer needs. Identifying and eliminating the cause of defective transactions raises transaction volumes and increases customer satisfaction. The output metrics of real-time transactions monitoring also contributes to revenue assurance.

2.5.2.12 Benefit of Homogeny

As CSPs increasingly see the integration of more third-party applications into their operational environment, the benefit of a homogenous real-time APM solution becomes clear. Each third-party application is able to connect to the host environment's APM system and to provide its specific APM metrics to the common repository, and be used when required for the rapid triage of a problem, real or potential, as well as for real-time SLA assurance. CSP vendors will provide this capability out of the box as a competitive differentiator.

2.5.2.13 Doing More with APM Data

The need to monitor key performance indicators in order to ascertain whether business services are being delivered in an effective and timely manner is only one important value that APM can offer. Knowing that a problem exists is often not enough; a quick remedy is key.

We need to process the collected data to determine the service state and to determine what corrective actions need to be taken in order to return to a stable state, and where possible, take those corrective actions. This introduces the need for proactive APM that (a) allows definition of specific events of concern and the necessary management actions, (b) provides alerts for these events ("event occurred" "remedial action is being taken" "event has been handled"), and (c) enables automated agents to trigger a predescribed remedy.

2.5.2.13.1 Use Case 5: Dynamic Business Subscription Management

John Doe, an NG service entrepreneur, enters the business with a bandwidth-dependent application. John's CSP provides an applications hosting service on their SDP platform that is tiered (Gold/Silver/Bronze) and priced according to the resources to which the customer subscribes.

For John's application, each user incurs more bandwidth. John decides to start his offering on the service provider's basic hosting package (Bronze), which offers 1 GB transfer per month and 1 server CPU. He knows that if successful, his service will require more bandwidth in time and more CPU power, and thus he will have to continually monitor to see when he has to upgrade his service. Such an action will be beneficial to both John and his hosting provider.

Within the monitored data are metrics that reflect transactions rates (for example, user subscriptions or purchase orders) as well as CPU and bandwidth usage. An event trigger could be written to detect when historical trends show the need for the next service band. If the hosting service provided a customer self-service interface (say via Web services), then the triggered script could additionally issue the upgrade request, updating the contractual obligations between the CSP and the hosted service provider.

In a dynamic (pay as you go) scenario John's application could request resource increases and decreases as well.

2.5.3 Application Performance Management Overview

As an essential function of any mission-critical software-based environment, APM provides visibility into the key parameters of any service-offering application: availability (is the service available?), and performance (how well is it performing in terms of response time and throughput?). APM covers the actions of collecting performance measurement data (or metrics) from running software applications and infrastructure, analyzing this data to determine events and trends of interest, and supporting the decision-making process to ultimately improve the performance and availability of the running system.

Application performance problems are usually related to issues of resource availability or capacity (such as database, back-end systems, CPU utilization, memory availability, and bandwidth) or to an application's code behavior (business logic). Monitoring these metrics provides visibility into the health of a running application.

2.5.3.1 Interested Parties for APM

In general, there are two groups of users interested in the information provided by APM tools: technical and business. Each group has a different expertise level and different view into the common issue of performance management. Often it is difficult for the two groups of users to discuss the same issue as each has their own points of view concerning what they see.

Business owners are interested in APM as it pertains to the availability and performance of revenue-generating services. They view activity in terms of business transactions; how many are proceeding per unit time, how many have failed, what is the impact to the users.

The key questions of the business owner can be grouped into ongoing (frequent) and one-time (infrequent). Examples of ongoing questions include: "How well are my communication service applications performing?" "Are my customers needs being met?" "Do I need to add capacity?" Examples of ongoing one-time questions include: "Are any customers experiencing downtime (am I losing revenue)?" Any problems leading to these conditions need to be quickly identified and then remedied to avoid negative impact on revenues, customer satisfaction, and operational costs. Business users may be able to analyze and even resolve problems in cases where the cause is expressed in business terms (for example, the CSP's license with the third-party service provider has expired, or the contract with a partner needs to be renegotiated to address growing traffic).

Technical users are interested in APM as it pertains to the design and efficiency of the code and components needed to provide the services. They view activity in terms of object instantiations, procedure calls, method invocations, stack traces, database request statements, processing power, and memory utilization, etc. Technical users are generally expected to use APM tools to get to the root of a problem and find the most efficient way to resolve it.

2.5.3.2 Evolution of APM

APM's capabilities, as materialized in products and tools, underwent significant growth as it matured over time. In the early days, custom code was inserted into an application in a nonstandard way by programmers with the goal of logging diagnostics information at execution time. While this provided value in diagnosing a problem, it required much overhead both in terms of coding effort (which impacted the development time), application maintenance (APM had to be maintained for each new release), and run-time execution costs (which impacted execution and often interfered with the validity of the data gathered). At the time when software infrastructure and business code were blended together, and no standards prevailed, these efforts had to be repeated for every new application or platform.

Further refinements were made to address some of these drawbacks, such as compilers supporting directives to collect generic data (reducing coding overhead) and selective tuning of what data metrics were collected on a per-run basis (necessitated to reduce run-time overhead). But there were still drawbacks; generic compiler directives did not cover all situations, and reducing the amount of metrics gathered

during run time (in order to reduce run-time overhead) ran the risk of key diagnostic data being missing or lost (if the criteria for determining the subset of metrics to be enabled during run time were incorrect).

Run-time impact was a huge issue; APM was at a state were performance data was collected at run time and typically processed postcollection. The introduction of a client–server programming model with standard infrastructure services (like database access, etc.) later solved some of the issues by providing limited performance-monitoring functions as a part of standard platform. New problems surfaced, however: since business logic now resided on both client and server, the resulting performance was hard to correlate. And business code itself still had to be created at the time of development. Trying to debug an end-to-end client–server problem was difficult at best due to different tools, metric names, etc.

The introduction of Java and BCI (byte code insertion) into the mix brought some relief with code and framework standardization across the development divide and the ability to auto-instrument code. Industry's adoption of standard IP-based object-oriented platforms like J2EE and .Net helped to further improve APM through built-in instrumentation of the leading software infrastructure platforms, and automatic discovery and instrumentation of components of the applications deployed on these platforms.

However, a major breakthrough in APM occurred through changing the way in which performance data was used and visualized. Until this point, the data collected offered a functional view—stack traces, procedure or method calls, variable values, time intervals. This viewpoint was good for the technical group developing and maintaining the application code but not for the business group providing the services. To serve these needs, the ability to view performance data in terms of business transactions was introduced. Now business owners could get a real-time view into the execution of critical transactions and correlate that with the behavior of an application and the end user experience.

APM tools that were capable of supporting both views—business and technical—gave the added benefit of providing a common language and reference for both groups to communicate effectively.

2.5.3.3 Reactive and Proactive APM

The purpose of advanced APM tools is to collect essential data to report not only on current problems, but also to predict and warn on potential pending problems or trends. To date, however, any proactive measures typically required operator (human) intervention to assess what the impact would be and initiate remedial actions. This approach has its clear drawbacks resulting in late problem detection (often through end-user complaints) and long resolution time. The effectiveness and timeliness of remedial actions that can be taken is highly dependent on the available operator resources having the right skill level. The ultimate goal of proactive APM is to prevent problems, or detect and fix them before they affect end users.

Via the use of heuristic algorithms and an enablement engine, certain situations that fall under these categories can be defined and acted upon while minimizing the drawbacks of the operator approach such as the time to fix, ability to fix, and cost to fix a problem.

2.5.3.4 APM Tool Requirements

APM tools must meet several requirements to be effective. A primary requirement is that any tool that collects performance data from a running system must have as little impact as possible on the system itself while providing enough data to accurately reflect the real transactions execution in a production environment. Ideally, the APM tool's overhead should be as minimal and as nonintrusive as possible. On the other hand, APM tools should run in a production system on a 24/7 basis, collecting real production data to be used immediately in the event of a production problem and also providing in-production (run-time) views of the health of the monitored system.

The APM tools should be easy to use. A user's time should be spent in understanding and diagnosing performance issues, not fighting with the tools to get them to produce the information or visibility needed to determine the cause of a problem.

APM tools must provide real-time monitoring of real data and transactions. This requirement is based on the key differences between the legacy network-based services and the software-based next generation services. With physical networks, their topology, possible faults, and disruptive factors are well known at design time, and it is easy to develop probes to test each area of the architecture. The majority of traditional network-based service-assurance tools use probes and synthetic transactions sent at predefined intervals of time. At the network level, all transactions and data types are also well known in advance, which allows the probes to effectively provide the performance and fault information. With software, the number of permutations increases exponentially due to multiplicity of possible business logic paths, user interactions, changing resource capacity, and bugs introduced with new or upgraded applications, making it hard to impossible to adequately provide the real performance and customer experience information based on synthetic transactions. Also, the real-time requirement comes from the fact that human user sessions can span over longer periods of time than hardware-based sessions, making it essential to know the performance at each step of the process. Testing at certain periods has a high potential of missing some important events and conditions that may contribute into the resulting quality of service.

It is still important, however, to be able to generate an automatic load of "typical" transactions in situations when the real traffic is low, or during test cycles.

APM tools should be deep and very specific in their visibility; however they should be easily customizable for different roles within an organization. An APM solution should cater to different users of APM. The best practice is to offer each user a customized view. Personalized user profiles, triggered on user login, could support customized dashboards that present data meaningful for this specific user. For example, a support person debugging a dropped call would be interested in examining SIP transactions and inquiring as to the health of SIP proxies, registrars, etc., whereas the business owner of a service would be interested in knowing that application's health, transaction rates, and the number of logins per user type, etc.

An additional benefit of using roles-based views based on a common terminology is that they can serve as a common language between business and technical parties.

Due to the dynamic nature of technology and business, the tools should be extensible so that new capabilities can be added to the system. A common framework would eliminate the need for additional training as the new extensions reuse the existing methodologies.

In larger deployments, the ability to integrate with other management tools (e.g., IT and network management) to support a complete and holistic view adds value and the additional capability of being customizable and embeddable into a system environment further increases overall solution value. A documented API and SDK provided with the advanced APM tools allow them to be integrated with the broader service assurance solutions as well as built into custom applications. This feature is critical for network equipment manufacturers (NEMs) and independent software vendors (ISVs) who are increasingly using standards-based software platforms to offer next-generation service delivery solutions. For them, a built-in APM function provides critical competitive advantage through deep visibility into previously hidden areas of service performance and availability, the ability to track SLAs, and additional input for the service assurance function.

The data accessed by APM tools can sometimes be of a sensitive nature. In such cases the tools should provide access security. At a minimum, access should be limited to authorized users only.

Finally, the tools should be able to automate or perform the mundane or time-consuming steps and thus expedite the time taken to diagnose or predict the occurrence of an issue. In particular, a capability like automatic root-cause analysis of an issue is key to a quick recovery.

2.5.4 Performance Management for Next Generation Service Delivery Platforms

2.5.4.1 SDP Overview

Service Delivery Platforms (SDPs) and network migration to IP multimedia subsystems (IMSs) are strategic investments allowing operators to provide more innovative services to subscribers and businesses.

At the same time they are disruptive to the traditional network-centric management solutions based on clear association of service, server, and operational ownership. SDPs and IMSs separate management of delivery channels (the transport mechanisms for the service) from the management of the service value chain (the run-time relationships among the end user, operator, content publisher, and service providers), making service performance management a distributed task. Moreover, each subscriber application class (video streaming, messaging, or VoIP) has a different model and data sources for providing service-specific performance metrics.

To overcome limitations of managing performance and availability of SDPs with traditional network-centric management tools, visionary operators making such strategic investments are looking for new solutions that are able to correlate real-time customer experience with the performance of service execution and delivery. Only such solutions allow an operator to control cost of service performance versus revenue from subscriber usage, analyze business impact of missing service level objectives, and take immediate corrective action.

Today, operators begin to view assurance from a customer, in addition to a network, perspective. They are combining traditional network-facing assurance functions, such as fault management and performance management, with customer-facing assurance functions such as service quality management and customer care, enabling a more holistic view of the customer's experience of the service.

2.5.4.2 SDP Is Strategic Investment

SDPs are complex, distributed software solutions providing the infrastructure and framework services for the rapid deployment, provisioning, execution, management, and billing of value-added services independent of networks and devices. With SDPs each operator and service provider can use a commodity platform to launch and test as many combined services as possible, using flexible and attractive charging models supported by a shared investment and with little impact on the network.

IMS is another strategic investment, allowing operators to get closer to the Internet model and leverage network value into new, customer-centric services at lower operational costs. While influenced by IT architectural concepts like J2EE and .NET, open APIs, SOA, and business process management, any SDP solution is deployed as an overlay to core network elements and services. As such, SDP is becoming the responsibility of network operations teams. The adoption of standards-based software platforms in the network domain has already been advanced by network equipment providers who embed commodity application servers into their products to lower the cost while enhancing their features' programmability and integration capability.

In this new quest for providing value to their customers, the following questions arise:

- Are the network operations teams enabled with the right tools and methodologies to reduce the risk of SDP deployments into the network, launch new services, and monitor performance and availability of a complex, distributed, pure software infrastructure?
- Are they really in control of services run time and able to report to business operations on their customers' experience with the service or the level of usage of their network assets?
- If a problem arises, do service providers have the ability to triage the problem in real time and then do deep-dive diagnostics to facilitate fixing the problem?

2.5.4.3 Challenges of Controlling SDPs with Traditional Network Management Tools

To answer the above questions, IT or network operation teams use the common language of performance. They use performance metrics to make decisions to address customer satisfaction level and service delivery costs or to better understand the availability level of the platforms and services they operate.

Progressive organizations already understand that good performance measurements are the enablers of service quality, and are looking for standards to guide them. Unfortunately, they are finding that the deployment of SDPs, IMSs, and mobile services has outpaced what standards bodies can specify for performance measurement. SDPs move the service execution out of the network context. The service

execution happens inside the SDP and depends on subscriber-related information (subscriber profile including the contractual quality of service) as well as network-related information (network-connecting enablers that are available at the service execution time for the requested quality of service). The service execution on the software platform is what a subscriber really experiences. This is the only context in which an operator has the opportunity to make a difference.

If network operation teams want to understand what performance at service execution time is, and correlate that with business rules and network status, they need a performance monitoring solution that fits between these two layers (business processes and network resources) and is intimately integrated into an SDP's architecture.

Network-centric management solutions are not sufficient because they do not have the visibility into distributed software or service composition at run time. Monitoring the service anywhere on an SDP must happen at those component interfaces where the service is contracted, at execution time when the service is delivered, and from the inside of each component, to gain insight into resource usage for executing the service. Traditional network management frameworks based on active probes pinging for response time or collecting performance metrics files will treat services or applications as technology silos disconnected from the rest of the environment. They cannot "see" the functional components and the run-time behavior of a service (e.g., how do methods calls perform at component interfaces) or the dynamics of resource consumption (e.g., depletion of connection pools, constant increase in memory usage, sockets bandwidth, etc).

When something goes wrong on the SDP, what can a network manager do to identify the root cause of the problem before customers begin complaining about service degradation? Without visibility in the service layer, how can network operations teams stop this degradation?

The bottom line: Service and technology innovations require new operational solutions that assure the same high level of service quality as traditional networks offered.

The solution: Performance should be monitored at many levels in order to get the big picture—not just at the network-element level, but also at the transaction level. Ideally, operators will increasingly look to real-time tools in order to respond immediately to issues rather than having to clear them up later, once the damage to their reputation has already been done.

Service Delivery Platforms, just like other mission-critical distributed software in an operator's network such as billing engines, customer care systems, or core network services, need a performance monitoring solution that is able to:

- Collect performance metrics from the inside of the service delivery components
- Collect performance metrics of back-end resources while their usage is driven by real user transactions or business processes
- Automatically discover a service as it is deployed on a platform and start monitoring its performance when and if the service is invoked by its subscribers
- Automatically detect changes in software components when services are replaced with newer versions, thus being able to correlate performance degradation with such events
- Meet the top–down approach for service Key Quality Indicators (KQI) and Key Performance Indicators (KPI) modeling, with the bottom–up collection of granular metrics and their aggregation to meaningful KPIs
- Monitor customer experience as services are invoked from the device and be able to tie this data back to the service execution processes running inside of the software platform

This approach not only monitors service performance, but also captures information that can be fed into decision processes fast enough to influence a service during its run time.

For example, monitoring how close a content provider is to reaching its allocated quota of messages can signal automatic contract renegotiation and prevent service interruption for subscribers who are

currently downloading content from that provider. This results in revenue protection for operator and content provider as well as increased customer satisfaction.

2.5.5 Performance Management for Next Generation OSS/BSS and SDP: The Solution from CA Wily Technology

To meet the new requirements of managing OSS/BSS and SDP, CA Wily Technology proposes a solution that is based on the following core elements:

CA Wily Introscope® identifies when a problem occurs or is going to occur and provides the information necessary to fix the problem. It collects and aggregates data that is needed to assess performance and availability of services and resources of the SDP, and is compatible with the technology stacks chosen for various SDP implementations (Java/J2EE/JSLEE or .NET, SIP, XML/SOAP/Web Services, databases, and HTP servers, etc.). It creates and stores consistent sets of data (metrics and transaction traces) reflecting real-time SDP infrastructure and service performance as they are invoked by subscribers or supporting processes. From this consistent set of data, CA Wily serves business- or operations-focused dashboards, drill-downs into specific components performance and executes actions in support of business or network operations. Through custom or standard reference integration points, this solution can be integrated in any existing end-to-end service assurance or customer care and billing system extending their operational visibility to SDP services, a unique value proposition from CA Wily.

CA Wily Communications Service Provider Extension (CSPE) is designed to be embedded into next-generation solutions from NEMs and ISVs to enable deep visibility into the service execution and delivery processes and interfaces, and to ensure the top performance and availability of the overall system.

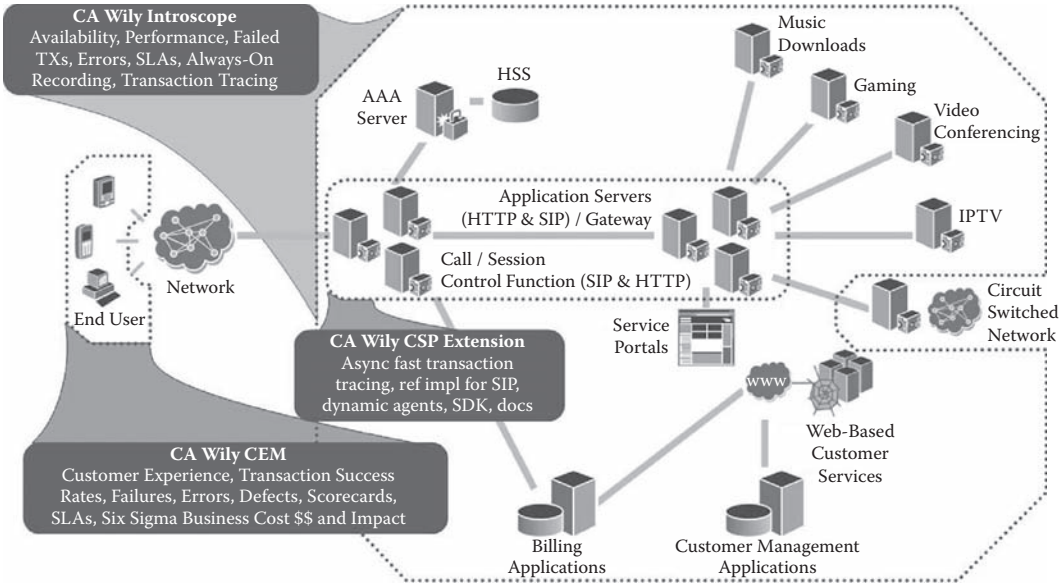
Built on the proven CA Wily Introscope platform, CA Wily CSPE is capable of deep monitoring of real user transactions in real time with very little overhead in production environments. It provides deep real-time monitoring of asynchronous fast-running transactions across multiple components, the underlying Java-based software infrastructure, the operating system, and the related back-end resources and OSS/BSS. Designed as a customizable component of third-party next generation service delivery solutions, CA Wily CSPE delivers risk management of new service creation and deployment, proactive monitoring of Key Performance Indicators (KPI), and rapid problem detection and analysis.

Complete with an SDK and a library of documented APIs, an out-of-the box reference implementation for the Session Initiation Protocol (SIP), and a comprehensive development guide with best practices and code samples, CA Wily CSPE is ideally suited to be customized for a broad range of solutions. With CA Wily CSPE, NEMs and ISVs can accomplish the following:

- Build deep visibility into service delivery and call-control supporting functions
- Create fully integrated dashboards, reports, and alarms
- Extend the existing network management tools into application software layers

CA Wily CSPE provides deep real-time monitoring of asynchronous fast-running transactions across multiple components, the underlying Java-based software infrastructure, the operating system, and the related back-end resources and OSS/BSS. Designed as a customizable component of third party next-generation service delivery solutions, CA Wily CSPE delivers risk management of new service creation and deployment, proactive monitoring of Key Performance Indicators (KPI), and rapid problem detection and analysis.

CA Wily Customer Experience Manager (CEM) detects and analyses the real-time experience of a subscriber, identified by its Mobile Station International ISDN Number (if present in the HTTP header) as transactions are executed through a Web interface (e.g., a mobile portal, self-service portal, etc.). Fully integrated with CA Wily Introscope, CEM can correlate these transactions with transactions traced and metrics collected from inside of the OSS/BSS or SDP execution environment. CEM can also detect and analyze business-to-business transactions carried through Web services invocations.



FIGu RE 2.5.4 CA Wily Solution for end-to-end OSS/BSS and SDP performance management.

Using this unified, cross-component, platformwide performance and availability monitoring solution, service providers benefit from:

- Real-time performance monitoring of OSS/BSS and SDP mission critical layers and subscriber services
- Customer experience monitoring and correlation with service performance
- Visibility into multiple distributed components added to service assurance and other OSS/BSS processes providing real-time alignment of critical customer-facing operations with business and network operations
- Risk management of new service creation and deployment
- Ability to integrate with network management, and broaden its management solution based on standards (JMX, SNMP)

CA Wily proposes that performance monitoring must become a core function of any OSS/BSS environment and that a performance monitoring enabler should be added to any SDP architecture. This solution reduces cost and complexity of the performance management by offering one enabler that can be integrated with many external management applications, services, and processes, making available the metrics collected from different functional components, from subscriber services to user-driven interactions with these services, as well as with the related OSS/BSS.

CA Wily's performance monitoring solution is specially designed for portals, Web services, and distributed and high-performance Web execution environments, which are typical for most SDP solutions. CA Wily is the acknowledged leader in Web application and server performance management in critical and complex deployments. CA Wily's performance instrumentation is standards-based and does not require application source code modification. It has the least impact on performance compared to any other solution for the same quality and depth of performance data it collects and analyzes.

CA Wily has helped solve many critical issues for telecommunications organizations worldwide, including:

- Application failures and delayed launches caused by rapid growth of subscribers and billing system complexities

- Self-service portal failures leading to lost prepay and ring tone transactions and unhappy partners
- Process failures in which the CTO hears about application problems from customers or the line of business rather than the IT staff
- Broken SLAs resulting in large regulatory fines
- Slowly responding call-center screens lowering CSR productivity
- Increased call-center load due to a malfunctioning IVR application
- Lost or mishandled orders due to failures in an order management or provisioning system
- Hidden performance issues in integration projects
- Finger pointing among network and operations teams, outsourced developers, and operations
- Inability to measure and manage SLAs with a third-party content and service provider

CA Wily Technology is the market-leading provider of APM solutions. CA Wily provides performance management software for converged service delivery platforms, subscriber services, business processes, and related OSS/BSS. By providing end-to-end visibility into customer transactions in real time, CA Wily's products enable operators to successfully manage health and availability of their critical applications, services, and software infrastructure. CA Wily's solutions ensure monitoring and delivery of SLAs, problem detection and incident handling, and successful customer experience in the increasingly complex, next generation of network-based, service-oriented platforms and applications.

2.5.6 Summary and Trends

As CSPs are increasingly using standards-based software for new applications and services, managing performance, availability, and the overall customer experience is becoming a new challenge. New application performance management tools, when used proactively as part of SDP and OSS/BSS architectures, can ensure high quality of service, low costs, and top customer satisfaction, ultimately making operators successful in this new environment.

The key requirements for APM are:

- **Deep diagnostics and visibility** into the execution environment (portal, partner gateway, Java or .NET application server, SIP/JSLEE, Web services, and SOA), and into network interfaces (Parlay and Parlay/X). The ability to trace an individual user transaction as it touches different components and systems, and correlate it with other events to understand the root cause of a problem.
- **Continuous monitoring of real transactions and real data** to catch early symptoms as they happen. With probing, those symptoms may fall between the intervals and not manifest themselves with synthetic transactions. Probing still makes sense when real transactions are few and rare, but it will not tell you how the system behaves under the real load. All data needs to be monitored for deep analysis, not just a few data points in synthetic transactions.
- **24x7 management of production environments** with little or no impact on the managed environment itself. There are ways to optimize metrics collection so overall additional load on the system is minimal.
- **Role-based customizable reporting** gives different roles within your organization the appropriate view of the environment relevant to their function (business managers, NOC and IT administrators, or QA and software engineers).

CA Wily offers a real-time APM solution fitting the above-stated requirements. It is based on the proven CA Wily Introscope monitoring tools used by 20 out of 30 Tier-1 CSPs worldwide, and by over 120 operators in 28 countries. The solution delivers risk management of new service creation and deployment, proactive monitoring of Key Performance Indicators (KPI), and rapid problem detection and analysis.

Acronyms

APM	Application Performance Management
CEM	Customer Experience Manager
CRM	Customer Relationship Management
CSCF	Call Session Control Function
CSPE	Customer Service Provider Extension
CSR	Customer Service Representative
CSS	Customer Self-Service
CSP	Communications Service Provider
HSS	Home Subscriber Server
IN	Intelligent Network
ISV	Independent Software Vendor
KPI	Key Performance Indicator
NEM	Network Equipment Manufacturer
NGN	Next Generation Network
NOC	Network Operating Center
QOS	Quality of Service
SIP	Session Initiation Protocol
SDP	Service Delivery Platform
SLA	Service-Level Agreement
SOA	Service-Oriented Architecture

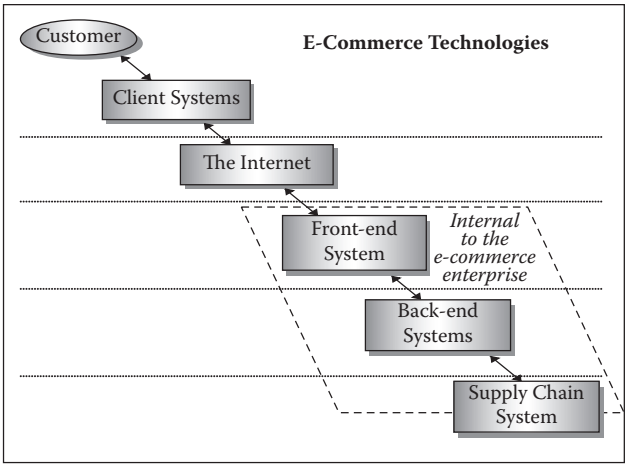
2.6 Electronic Technologies

Patricia Morreale and Mihir Parikh

Increasingly, our interactions as individuals and organizations are initiated by, and frequently restricted entirely to, electronic technologies. While in years past, human interactions that were not conducted in person were frequently governed by land-line wired telephone calls or postal mail correspondence, we now frequently engage in time-independent dialogues using electronic technologies such as e-messaging, e-commerce, or wireless communication in a variety of forms. These forms of communication are frequently asynchronous, rather than synchronous.

Asynchronous communication does not require that all parties involved in a communications exchange be available at the same time. Rather, a message prepared and sent by one party can be received and reviewed by another party or parties at a later time. E-mail and text messaging are examples of asynchronous communication. Synchronous communication, also known as direct communication, in contrast to asynchronous communication, does require that all parties be present at the same time in order to complete the communication. Traditional voice phone calls or instant messaging with immediate response are two examples of synchronous communication.

These interactions have been enabled by devices and services that have come into service recently and are characterized by being introduced into society and common use in a novel manner. In the past, new services and technologies were adopted in a top–down manner, by senior technology people able to afford and enjoy the newest technologies. The innovations filtered down into the general population over time, as initial price points were reduced and innovation became more commonly available to all. Now, electronic technology innovations are driven in a bottom–up manner, where the early adaptors are preteens, teens, and twenty-year-olds who are willing to try and share new technologies with their peers and elders.



FIGu RE 2.6.1 E-commerce technologies.

This bottom–up approach to technology innovation and adaptation is a first, and is indicative of the rate of change in electronic technologies. Approaches to gathering data and using data quickly become obsolete or require enhancement to continue to provide the information desired. In this section, broad areas of electronic technology are presented and discussed.

2.6.1 E-Commerce Technologies

Electronic commerce (e-commerce) is a result of sweeping changes over the past years in electronic communications and technology. In simple terms, e-commerce is defined as a business conducted over the Internet with the use of computer and communications technologies. There are three major types of e-commerce: business-to-business (B2B) e-commerce, business-to-consumer (B2C) e-commerce, and electronic markets. B2B e-commerce deals with one business providing products and services to another business typically as a part of the supply chain or as an enabler of business processes. Examples of B2B e-commerce include an auto-part manufacturer supplying an auto company, or a bank providing credit card payments and other financial services to a retailer. For many years, electronic data interchange (EDI) handled B2B transactions in many companies. Now, Web-based open system applications are replacing proprietary EDI systems. B2C e-commerce deals with a business providing products and services to consumers at the end of the supply chain. Examples include a book retailer selling books to a reader (Amazon.com) and a broker providing financial trade executions to an individual investor (E*Trade). Electronic markets provide marketspaces on the Internet, as opposed to marketplaces in the physical world, where buyers and sellers can meet and exchange products and services. Electronic markets are of two types: consumer markets and business hubs.

Annual worldwide e-commerce will increase to more than $1 trillion by 2012.

Information technologies are key enablers of e-commerce. A combination of many information technologies ranging from a Web browser to logistics management systems makes e-commerce possible. Identifying and implementing the right technologies to execute business processes is critical to succeed in the e-commerce space.

Figure 2.6.1 shows different categories of e-commerce technologies. All of these technologies are required and play significant roles in managing and conducting business over the Internet.

2.6.1.1 Client Systems

Client systems reside on customers' computers. Customers utilize these technologies to participate in e-commerce activities.

Web browser and cookies: The most common of all client systems is a Web browser. Web browsers provide an interface through which a user can view information on the Internet. In the last five years, Web browsers have evolved from a small software using simple hypertext markup language (HTML) to a very sophisticated software that uses technologies such as Java, XML, and different types of plug-ins and Web applications in addition to a more advanced version of HTML. Most Web browsers utilize cookies whereby they transmit basic information about a user to the Web server on the other end for identification purposes. Such identification can be used to personalize services for the user. Multiple levels of cookies are used. Low-level cookies simply provide username and password, while high-level cookies may include information about credit card, mailing address, previous purchase patterns, and browsing habits.

Communications software and hardware: Communications software and hardware help customers' computers to connect to the Internet via one of the multiple modes including modem, cable modem, satellite links, and local area networks.

Plug-ins: Plug-ins are independent software applications utilized to show special data files within a Web browser. Plug-ins enable a Web browser to show a multimedia presentation or to play a streamed audio piece. Shockwave Flash, RealPlayer, and Adobe Acrobat Reader are plug-ins.

Software agents: Agents are other types of independent software that can assist users in carrying out some specific activities. Such activities include filtering information, searching the Web to find the right information, and comparison shopping.

Biometric identification: Biometric identification is a technology that uses a measurable physical characteristic to recognize the identity, or verify the claimed identity, of an enrollee. It utilizes physical characteristics such as fingerprints, facial design, iris patterns, retina patterns, hand geometry, signature verification, and voice recognition instead of keys, passwords, and plastic cards. Several advantages, including reduction in fraud and never losing the identification, have prompted several large companies such as IBM and Motorola to invest heavily in biometric identification. Biometric identification is one of the fastest-growing areas in client systems and security.

Push technologies: These publish-and-subscribe technologies enable delivery of possibly useful information without the recipient asking for it. Push technology, also referred to as server push, utilize extensive user profiles containing preferences of each user to match and deliver information over the Internet. The request for a given transaction is initiated by the publisher or central server. This is in contrast to pull technology, where the receiver or client initiates the request for transmission of information.

2.6.1.2 The Internet

Many advanced telecommunications technologies have been utilized to create and operate the Internet. Most prominent are optical fiber, routers, digital switches, Synchronous Optical Network (SONET) rings, asynchronous transfer mode (ATM), frame relay, Transmission Control Protocol (TCP), and Internet Protocol (IP). These hardware and software technologies provide the backbone lines and the transmission rules to carry out information exchange over the Internet. As the Internet is growing, new technologies are emerging to increase speed and volume of data transmission. The continuing convergence of textual data with audio and video data over the Internet is prompting new technologies to support quality of service (QoS), which recognizes the differences between data types and assigns appropriate priority for transmission.

2.6.1.3 Front-End Systems

Front-end systems are the ones with which the customers interact. They provide a face for an e-commerce business.

Web pages: Web pages are the data files containing HTML-coded information. In the early days of e-commerce, Web pages were static and generated by human Web programmers. Now, in most e-commerce sites, Web pages are dynamic and are generated by Web page management systems. These systems work with other front-end systems and back-end systems to develop HTML-coded content that the customers receive.

Traffic management: Due to a special sales event or some top news, sometimes an unexpected number of visitors go to a Web site all at once. Such an unexpected load puts pressure on the Web server and dramatically reduces its speed and increases download time. If the load extends for a longer period of time, it crashes the system. To avoid such a mishap, traffic management tools are used. These reduce massive congestion by spreading traffic load on multiple servers and increase overall network efficiency.

Search engines: Several types of search engines are utilized by e-commerce companies. Some search engines provide capabilities to search a specific product based on description, features, or other information. Some search engines provide capabilities to search and locate Web content based on key words or phrases. Search engines can be used to search a specific Web site, a regional part of the Web, or the whole Web. Often directory engines are used along with search engines to automatically categorize Web content for future searches.

Site servers: Site servers are the most comprehensive tools for e-commerce. They provide support ranging from creating an electronic storefront to controlling and managing it. Site servers include support for site building, standard code sharing, code library development and management, dynamic Web page generation, product promotion, product cataloging, order taking and processing, securing transactions, and managing payment systems. Site servers help provide product information, dynamic pricing information, marketing and promotion, shopping cart services, tax calculations, shipping and handling calculations, and automated postsales follow-up. In addition, they capture market demographic information and coordinate with back-end systems. Some off-the-shelf products are available to support small business shopping services. However, a major e-commerce site requires custom application development to support the above-discussed activities.

Shopping engines: These types of software enable customers to find and compare different products or services on features and prices. Some shopping engines also include product reviews from independent agencies or with current users of the products to help others make more informed purchase decisions. It is a very useful tool in e-retailing.

Customer relationship management (CRM): As the search cost for every new customer is staggering, increasing customer satisfaction and loyalty is crucial to maintain the current customer base and for e-commerce success. They provide a database of frequently asked questions (FAQ), a searchable knowledge base, multiple ways to assist shoppers in real time (e-mail, Internet telephony, video conferencing, etc.), follow-up support, order tracking, return processing, after-sales services, and warranty processing. They also help maintain profiles of buyers that contain their shopping behaviors and preferences. The currently available CRM-related technologies include enterprise portal, mobile computing, net telephony, desktop video conferencing, speech recognition, call center systems, and data warehousing.

Personalization: As the number of e-commerce businesses increases, a key differentiator would be how well an e-commerce business customizes its storefront for each customer. Knowing the customer and his or her preferences will improve customer service and retention. This personalization or mass customization requires creating user profiles from purchase patterns and browsing patterns, and applying business rules and inference on the information collected in the user profiles to create new knowledge about the users. Key technologies used for personalization are databases, dynamic Web pages, business rules, inference engines, cookies, and push technologies. In some cases, data warehousing and mining technologies can also be used for personalization.

Security: More than half of B2C e-commerce transactions are paid with credit cards. Protecting transmission of credit card and other private, confidential information over the Internet and securing the information on merchants' Web servers are two of the most pressing issues in e-commerce. Encryption of data through Secure Sockets Layer (SSL) and private and public keys are the most commonly used technologies to secure Internet transmission of confidential data.

2.6.1.4 Back-End Systems

While the front-end systems manage the interface with customers, the back-end systems carry out operations and manage e-commerce organizations.

Enterprise resource planning (ERP): ERP systems are commercial software packages utilized to integrate information flowing through different functions of an organization, such as financial and accounting, human resources, sales and customer service, supply chain, etc. ERP systems coordinated with front-end systems can capture orders, provide order confirmation, accept payment, check credit cards for approval, process coupons and other promotions, handle billing and invoicing, control inventory and procurement, integrate with payment systems, and coordinate with fulfillment systems for order execution. Before ERP systems, these processes were handled by various, independent information systems indigenous to different functional divisions in the organization. Information systems from one functional division were often not compatible with the information systems in other functional divisions. This brought inefficiencies in business processes and overall higher costs. ERP systems promise seamless integration and easy information flow among different functional divisions. However, successful implementation of an ERP system has been one of the major critical issues in utilizing ERP systems.

Databases, data warehouses, and data mining: Databases and data warehouses are at the center of running a business in this information economy, especially an e-commerce business. They provide repositories for information collected through business processes. This information is the lifeblood of organizations and its optimum use is very important. Several emerging database technologies, such as multidimensional databases, provide a holistic perspective and better understanding of the information. When used in conjunction with data mining technologies, e-commerce businesses can find and exploit hidden relationships and buying behaviors to increase market share and sales.

2.6.1.5 Supply Chain Systems

These systems work with business alliance partners and other enablers. They provide smooth transfer of information with partners to carry out outsourced business processes. Some of these systems are external to e-commerce organizations and implemented by the partners.

Supply chain management: Most emerging e-commerce companies are not vertically integrated. They depend on many partners and intermediaries on both sides of the supply chain (a value addition sequence through which raw material flows to become a finished product). Supply chain management systems help businesses coordinate their processes with suppliers, manufacturers, raw material providers, shippers, distributors, and associated retailers. In e-commerce businesses, greater efficiencies are achieved by moving information rather than actual products along the supply chain. Actual products are generally delivered directly to the consumer by the product manufacturers without any supply chain intermediaries handling or storing the products. Supply chain management systems help control product life cycle, forecast demand, arrange advanced scheduling, plan manufacturing and distribution, and enable order promising and processing.

Payment systems: The majority of payments over the Internet are conducted through credit cards and checks. Payment systems help e-commerce businesses coordinate with banks and credit card companies to approve credit card purchases and clear checks. Some e-commerce companies also utilize these systems to work with soft cash providers such as PayPal and electronic money.

Fulfillment and logistics management: These systems coordinate with logistics partners such as FedEx, UPS, and independent warehouse operators. These systems work with back-end and front-end systems to help determine shipping and handling charges, delivery terms, delivery schedule, order tracking, freight management, customs and excise duty clearance, and other fulfillment issues.

2.6.2 Web Service Delivery Challenges

2.6.2.1 High Cost of Small Errors

In e-commerce, there is little room for error. In most cases, you do not get a second chance. Integrating right strategies with right technologies and continuously improving competitive position are important

for survival and success. Any error in technology or strategy implementation can lead to a serious loss in market position and the ability to carry out business in the future. Once, eBay's stock price dipped more than 50% largely due to recurring, unexpected shutdowns of its Web site. Most of these shutdowns were not more than a few hours long. However, persistence and a strong commitment by the company's top management to improve technological infrastructure led to a rebound in the stock price. E-commerce businesses pay a very high cost for small errors. As an e-commerce business expands, the probability of making such errors increases. While most managers and leaders assume that being there first is the key to success, an early study found that many pioneers fail and most current leaders are not pioneers.* The study found that five factors (vision, persistence, commitment, innovation, and asset leverage) are critical to success. These factors often lead to making fewer errors and quickly correcting the errors when they are made.

2.6.2.2 Building Relationships

Technology is a double-edged knife. While it provides unprecedented advantages to you, it also provides the same benefits to your competitors and future competitors. It enables your existing competitors to quickly react to your moves. It also reduces the barriers to entry for new competitors. Assuming that if you build it, they will come and stay is one of the fallacies of e-commerce. To avoid this, you have to build a strong customer base and retain it. Building a strong customer base requires building relationships and providing useful services and content to your customers.

2.6.2.3 Speed

The speed of doing business has increased tremendously with the Internet. Speed is required in growth, in decision making, in adapting to the changing conditions, and in supporting and servicing customers. An e-commerce company has to continuously innovate and improve its business processes and Web-based storefront. It is always vulnerable to the quick imitation of its processes and its innovative shopping features by its competitors. Often, not only the Web site designs but also business models are copied overnight by competitors. This also requires building flexibility in the front-end and back-end systems to adapt to the continuously changing conditions and stay ahead of competitors. Therefore, industry experts recommend an open and adaptive architecture for enterprise information systems.

Operationally, when a customer visits an electronic storefront, performance of the Web site becomes an important issue. First of all, the customer wants quick downloading of the Web pages and immediate response to any search queries. If the Web site is slow, the customer will very likely move on to another competing store. This invariably happens when an unexpected number of customers come to the Web site at the same time. Several technologies can help improve the speed of the Web site. Use of traffic management tools can balance the load on a Web server and increase the speed of interaction. Scalability of the hardware and software provide quick integration of additional resources. The use of better search engines and shopping engines can also increase the speed of searching the requested product from millions of product profiles stored in the databases. Good shopping engines can also help identify related, complementary products for cross-selling and up-selling.

2.6.2.4 Security

With the growth of e-commerce, more and more business processes and databases are put on the Internet. This is required to improve customer service and increase organizational efficiency. However, this also makes the processes and databases vulnerable to malicious forces including business spies, computer hackers, and disgruntled former employees. Having complete control over who gets to see what and who gets to change what is extremely important. Security management software, virus protection software, and intrusion detection software can help increase security of the Web site.

---

* Gerard Tellis and Peter Golder, "First to Market, First to Fail? Real Causes of Enduring Market Leadership," Sloan Management Review, 32(2): 1996, 65–75.

2.6.2.5 Technology Evolution

E-commerce technologies are in a constant state of flux. E-commerce businesses have to continuously evaluate emerging technologies and adapt them quickly to stay competitive. As the technologies are constantly evolving, few standards exist. The ability to choose a right technology with a strong future becomes a critical skill in managing e-commerce.

2.6.3 Management of Web Services

> What is my ROI (return on investment) on ecommerce? Are you crazy? This is Columbus in the New World. What was his ROI?
>
> **—Andy Grove, Intel Chairman**

New technologies are emerging every day. Some of these technologies may become a killer app for e-commerce. While it is almost impossible to predict them, using standard measures (such as ROI) to evaluate may fail, too. However, several underlying trends may help to identify, evaluate, and manage the right technologies.

The drop in the cost of computing continues. Storage, processing, and distribution cost of information is decreasing to a level where the cost is less than the value of information. This has led to the development of new enterprises that provide free products and services in return for information and loyalty. More and more e-commerce businesses will continue to rise, making it difficult for the current businesses to compete. Rather than the traditional model of paying for access to a network connection, the newer model emerging is that of a consumer or enterprise paying for access to a digital library or other information service offering.

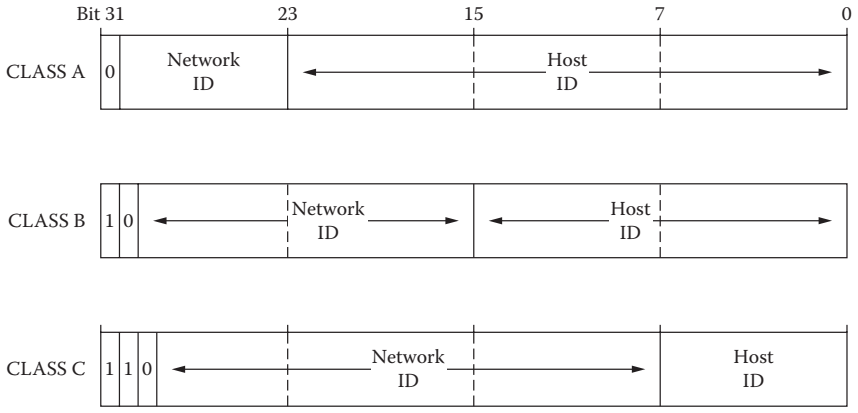
While the power of computer processors is going up, their size is decreasing. In addition, computer processors are now used in many devices and products (such as autos, refrigerators, washing machines, dishwashers, etc.). As computing technologies continue to expand to household products and appliances, communications technologies will follow. These products and appliances, when connected to the Internet, will create new e-commerce opportunities. For example, a refrigerator in the future may be able to automatically buy groceries for you. The container of milk is put in the refrigerator in one location where there is a sensor that notices how much milk is left. As soon as milk reaches the reorder level (determined based on your consumption pattern), the refrigerator will automatically connect with a grocery store on the Web (maybe NetGrocer or PeaPod) and place an order for milk. Comparable examples exist for the delivery of movies (NetFlix) and music (iTunes).

Management of existing and emerging Web services is problematic. In addition to the 24/7 expectation of availability, there is a fine line between anticipating a customer's need for service, through data mining and anticipatory sales, and maintaining customer privacy and security. Additionally, seasonal and timely spikes in Web service demand can create unanticipated delays and bottlenecks for customers seeking to obtain access to Web services via electronic technologies. New software applications are emerging every day. Each one makes e-commerce business processes more efficient and effective. This will enable more and more people to go online for their shopping, entertainment, business, and home management needs. New e-commerce models will rise to support these changes in customer behavior.

2.7 Internet Protocols

John Braun

The purpose of this section is to describe concepts that are essential to understanding how Internet protocols work. Basic addressing at the device or client level will be described, followed by a discussion



FIGu RE 2.7.1 IP address formats for Class A, B, and C.

of the protocols that are used to exchange information among these devices. Although the term *device* may seem vague at first, the types of devices that communicate via the Internet have matured from simple text-based information, to the use of audio, video, animation, and other forms of communication. Once the basics of communication are laid out, some of the specific protocols and applications that utilize these basics will be discussed. Security aspects of these protocols and applications will be covered. Useful search tools that can help locate information on the Internet will be covered, and a discussion of some of the major industry players will conclude this section.

2.7.1 Addressing for Internet

All devices on a network that supports Internet Protocol (IP) have a unique numeric address, 32 bits in length.

The most common way of representing a device's IP address is by using a "dotted quad"—four decimal numbers ranging from 0 to 255, separated by periods. This results in a theoretical range of 0.0.0.0 to 255.255.255.255.

There are currently five classes of IP addresses. They are presented in Figure 2.7.1, and Classes D and E are listed below.

A Class D address has the following format:
1110 MMMMMMMMMMMMMMMMMMMMMMMMMMMM(28)
A class E address has the following format:
1111 XXXXXXXXXXXXXXXXXXXXXXXXXXXX(28)
N = Network portion of address
L = Local portion of address
M = Multicast address
X = Undefined

This results in the following valid ranges:

Class	Begin	End
A	0.0.0.0	127.255.255.255
B	128.0.0.0	191.255.255.255
C	192.0.0.0	223.255.255.255
D	224.0.0.0	239.255.255.255
E	240.0.0.0	255.255.255.255

Class D and Class E addresses, which have the first (high-order) bits of the address set to 111, are classified as part of an undefined extended addressing mode.

There are a few addresses that have special purposes and should not be assigned to a device.

If the network ID portion of an address is set to all zeros, it means "this network." For example, address 0.0.0.42 would mean host 42 on this network.

If the host ID portion of an address is set to all ones, it means "all hosts." For example, the address 199.27.24.255 would mean all hosts on the 199.27.24 network.

The Class A address with a network ID of 127 is defined as a loopback address, where any packet sent by the host will be returned to the host without being sent across the network. This can be used for testing purposes, or as a sanity check to determine if one's TCP/IP implementation is working properly. The typical value used for this purpose is 127.0.0.1.

2.7.1.1 DNS

The domain name system (DNS) is a global network of computers that can translate a numerical IP address to a human-readable name, and also translate a name to the corresponding IP address. This makes navigation of a TCP/IP network much easier. For example, www.crcpress.com is much easier to remember than 199.29.24.3, the IP address which corresponds to this address.

Before DNS, each computer on a network would have to maintain a large file (typically called hosts) with all known IP address and name pairs. This is obviously impossible to do now with the sheer number of hosts on the Internet, but can still be useful for small, private TCP/IP networks that are not directly connected to the Internet.

When configuring a client to access DNS services, multiple DNS servers should be specified, if available. Modern TCP/IP implementations are smart enough to try another DNS server if the initial one is unavailable.

A DNS client will submit a DNS request to a server, and receive one of three types of reply. The client will be told that (1) the lookup was successful and be given the name, (2) that the server couldn't perform the lookup but knows another server that may, or (3) that the lookup failed.

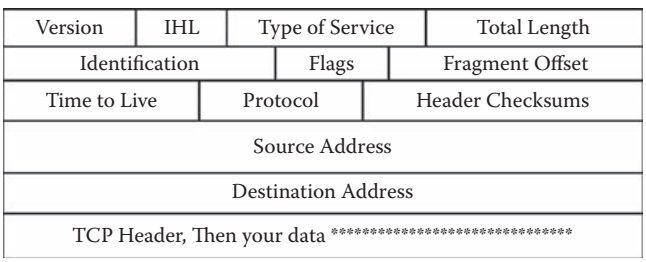
2.7.2 Communication Protocols in Internet

There are many protocols used on the Internet. Most are classified at either the lower layers (Layer 3, Network and Layer 4, Transport) of the OSI model, or at the application level (Layer 9).

2.7.2.1 IP (RFC 791)

The "IP" in TCP/IP refers to the Internet Protocol used at the network layer. The basic purpose of this protocol is to try to deliver packets. It does not offer such services as acknowledgment, retransmission, error correction, flow control, or guarantee of order of delivery. This is the job of higher-level protocols. The advantage of IP is that it offers a common framework for devices to communicate.

An IP header looks like the one in Figure 2.7.2.



FIGu RE 2.7.2 An IP header.

Source Port		Destination Port
Sequence Number		
Acknowledgment Number		
Data Offset	Reserved	Windows
Checksum		Urgent Point
Your Data *** next 100 octets **********		

FIGu RE 2.7.3 A TCP header.

2.7.2.2 UDP (RFC 768)

User Datagram Protocol (UDP) is a connectionless protocol that provides a means to send data with a low protocol overhead. Only the source port, destination port, length, and checksum are added to the raw data. However, it does not guarantee delivery or protection from duplicates. For applications where performance is critical, a small loss of data is not critical, and the link is known to be reliable, UDP may be used. Streaming audio or video are examples of applications where high performance is more important than a possible, but potentially correctable, loss of data.

2.7.2.3 TCP (RFC 793)

Transmission Control Protocol (TCP) offers a more robust, connection-oriented method for reliably sending data. Flow control and multiplexing are supported. Unlike UDP, TCP supports the concept of a continuous stream of data between two hosts. Unlike IP, TCP will make multiple attempts to deliver data if the initial attempt fails. If the integrity of data is critical, TCP should be used.

A TCP header looks like the one in Figure 2.7.3.

2.7.3 Information Transfer in Internet

A socket is a virtual communications channel that is established between two hosts. It can use either UDP or TCP for the transport protocol. Each socket has a unique descriptor, and multiple sockets can reside on the same port. This allows multiple clients to take advantage of a service on a single port without the server getting confused. In order to prevent congestion on a single port, many application-level protocols have a control connection on a known port, and then negotiate another port for subsequent data transfer.

2.7.4 Types of Internet Access

There are two basic types of Internet access. One is a full-time direct connection; the other is via a telephone line with a modem. The disadvantage of a direct connection is mostly cost in both dollars and network maintenance. The advantage is speed, where T1 (1.5 Mb/sec) or T3 (45 Mb/sec) rates are common measures. The disadvantage of a modem connection is reliability (line noise), availability (busy signals), and speed. Certain schemes such as v.90 can achieve up to 56-kbps download speeds. The advantage of a modem connection is the cost of both equipment and service, with 56-kbps modems available for under $300 U.S., and unlimited service for around $20 U.S. per month.

Integrated services digital network (ISDN) can provide 128 kb/sec transfer rates. It never seemed to catch on due to the difficulty in configuring the equipment and inconsistent pricing plans across the country. Many providers charge a flat rate, with some adding charges for each unit of time or unit of data sent, whereas a local telephone line typically allows unlimited usage.

Asymmetric digital subscriber line (ADSL) is a relative newcomer that takes advantage of existing twisted-pair wiring, and can reach speeds of 6 Mb/sec for downloading, and 640 kb/sec for

uploading. It is being deployed in major cities, but it remains to be seen how widespread the service will become.

Cable modems are slowly becoming available, and offer up to 10 Mb/sec transfer rates. Typically, the upstream connection consists of high-speed fiber, with the final connection to the cable modem being made with coax.

There are hybrid solutions, such as a satellite (with speeds of 400 kb/sec and higher), which uses a phone line for uploading data and a satellite for data download. This can be a good solution for scenarios such as surfing the Web, where the amount of data sent to request information is much less than bandwidth-intensive data types, such as graphics and sound, which comprise the data.

There are two major protocols used for establishing a TCP/IP connection over phone lines. Point-to-Point Protocol (PPP) is the more modern method, and PPP software is included with nearly every major operating system. Serial Link Internet Protocol (SLIP) is an older standard that is being phased out in favor of PPP.

2.7.5 Internet E-Mail

There are a few standards for sending and receiving Internet e-mail. The most popular protocol for receiving e-mail is Post Office Protocol (POP), which defaults to TCP port 110. A newer protocol, Internet Message Access Protocol (IMAP4), used for receiving e-mail, resides on TCP port 143.

The most common protocol used for sending e-mail is Simple Mail Transfer Protocol (SMTP), which defaults to TCP port 25. Note that POP can also be used for sending e-mail, but this feature is an optional extension of the POP protocol, and is not supported by many e-mail clients and servers.

2.7.5.1 SMTP

The SMTP service allows the sending of e-mail by providing relevant information, in a specific order, to a SMTP server. The conversation between the client (sender) and the server consists of human-readable text commands that are assigned four-letter codes, and three-digit code numbers as responses. A typical exchange may look like that shown in Table 2.7.1.

This system works well enough, with servers cooperating when mail needs to be forwarded to a domain outside of their own. Messages can be delivered in minutes, if all the servers in a message's path respond in a timely manner.

Although standard SMTP does not have any provision for confirmation of receipt (there are proposed standards, but they are not widely implemented yet), problems encountered along the path of the message are usually reported to the sender. If a server finds that another server is unavailable, it will usually

TABl E 2.7.1 A Typical E-Mail Exchange via SMTP

C: HELO megacorp.com	(sender identification)
S: 250 smtp.conhugeco.com	(server identification)
C: MAIL FROM:<johnbraun@megacorp.com>	(who is this from?)
S: 250 OK	
C: RCPT TO:<dilbert@anotherdomain.net>	(who is the recipient?)
S: 250 OK	
C: DATA	(ready for data?)
S: 354 Go ahead…	(sure)
C: Hey Dilbert get to work!	
C: .	(end of input)
S: 250 OK	
C: QUIT	(sender all done)
S: 221 Bye…	(server says seeya)

try to send the message several times before giving up. If a server crashes while processing mail, the message may get lost forever. For critical documents, it may be wise to ask the receiver to confirm receipt.

One problem with current SMTP implementations is that they don't confirm the validity of the sender's address. This has led to massive abuse by junk e-mail senders, who really don't want you to respond via e-mail, anyway. An attempt to reply will result in getting a message saying that the (bogus) return address doesn't exist, or that it has already been shut down.

There are proposed methods of authenticating an entity wishing to use an SMTP server, in hopes of reducing spam and other evils. There are also solutions, mainly using public key encryption and digital signatures, to confirm the identity of the sender. These features are not available at the protocol level yet.

2.5.7.2 POP

The POP protocol is used to retrieve e-mail for a specific user. Session establishment can be done with a simple username and password scheme, or can optionally use more sophisticated means like APOP. APOP never sends the user's password over the connection, instead applying the MD5 algorithm to the password and other time-sensitive data. Not all POP servers support APOP, and will return an error message if this method of login is attempted but not supported.

After a message is retrieved, it is usually deleted. Some POP servers allow the user to keep their e-mail on the server after it has been retrieved, but this is not guaranteed. The IMAP4 protocol is better suited for keeping e-mail on a server.

Some POP servers offer the ability to send e-mail as well as receive it. The advantage of using POP for sending e-mail is that it requires users to identify themselves, thereby reducing the generation of spam and other unauthorized e-mail. It also eliminates the need to maintain a separate SMTP server. The disadvantage of this method is that not all e-mail clients and servers support sending via POP.

2.5.7.3 IMAP4

IMAP4 is a newer protocol for retrieving e-mail. It offers a richer set of features than POP, including keeping e-mail on a remote server, searching e-mail before retrieval, and a greater number of authentication schemes.

The option of keeping e-mail on a server helps reduce resource requirements on the client, but increases the need for more disk space, as well as regular backups to ensure the historical data is preserved. This scheme can benefit devices such as portable computers, network computers, and portable digital assistants, which may have limited storage capacity. It can also offer convenience to mobile users, since their mail can be stored on the server, rather than being spread among multiple clients.

2.7.6 Telnet in Internet—STD8

Telnet is a service that resides on TCP port 23 and offers terminal services to remote users. The client and server can negotiate features of the connection, which can range from a simple ASCII exchange, to enhanced services such as cursor control and styled text.

A telnet client can be used to interact with other TCP services that exchange text or binary data, provided that the telnet client allows one to specify the port one wants to connect to. For example, a telnet client can connect to an SMTP server on port 25 and send e-mail. This can be handy for debugging or investigative purposes.

2.7.7 File Transfer in Internet (FTP, RFC 959, RFC 990—Port number)

File Transfer Protocol (FTP) provides a reliable, standard way to transfer both text and binary files between hosts. There are two types of connections when using FTP, a control connection and a data connection. A control connection is used for the client and server to exchange commands and status

information. The default port for a control connection is 21. The default client data port is 21, and the default server data port is 20.

In most cases, a data port with a value outside the range of commonly known services is selected for security reasons. The client can request a specific data port via the PORT command, or can ask the server to select one with the PASV command.

Although FTP makes every attempt to deliver a file, circumstances beyond the user's control (modem disconnect, network failure) may cause the transfer to be interrupted. Most, but not all, FTP implementations support a restart mode where file transfer can begin at a point other than the beginning of the file, allowing data transfer to begin at the point of interruption and complete.

2.7.8 News and Usenet (NNTP, RFC 977)

The most common way of distributing news on the Internet is by using the Network News Transfer Protocol (NNTP). The system, originally created to exchange messages in a university environment, has evolved to a worldwide messaging system with thousands of distinct topics of interest, called newsgroups.

Each newsgroup name consists of several elements, with the general classification at the beginning of the name, and the specific subject matter at the end. For example, the group comp.sys.mac.hardware.video is a computer-related group about video cards for Mac systems.

The most popular hierarchies are sometimes referred to as the "big seven" and include comp (computing), misc (miscellaneous), news (newsgroup-related), rec (recreation), sci (science), soc (social), and talk (discussion). There are many other hierarchies for alternative and localized content.

Clients to allow one to read and post news are available as both stand-alone applications or integrated with other Internet clients such as a Web browser. Sophisticated clients can allow one to organize messages about a similar topic into threads, or filter messages based on certain criteria. Although most messages pertain to the topic of the newsgroup, there are those who "spam" the newsgroups with ads or other material unrelated to the group.

Care should be taken when configuring your news client where it asks for your e-mail address. Your address is normally added to any messages that you post, with the intent of making it easier for others to make a personal reply. Alas, there are those who scan the newsgroups for e-mail addresses, which are then used for the purpose of sending junk e-mail. A good strategy is to add some text to your address that a human would know to delete before making a personal reply, such as john@ihatespam.megacorp.com.

A dedicated news host is required before clients can post and retrieve messages. Typically, a news host will serve a large group of clients, and will exchange articles with another upstream host. The connection between hosts is referred to as a feed, and is meant for propagating articles and control messages, and not meant for direct client connections. If all hosts along a certain path agree to exchange the same group, a message posted in a particular group on one host will eventually propagate to all other news hosts. Hosts can also choose to restrict the groups they carry, which can help conserve disk space and bandwidth.

Care must be taken in deciding which groups a host should carry, and how long the articles should remain on the server before being deleted or expired. Inaccurate estimates can result in a dreaded disk full error, which causes many hosts to reject articles until someone clears some disk space.

2.7.9 Mailing Lists in Internet

A mailing list is a method of allowing Internet users to communicate about specific topics via e-mail. To join a mailing list, one sends an e-mail to a special address, and indicates a desire to join the list, as well as what e-mail address should receive further information from the list. The subscriber will then receive information on how to submit messages to the list, and how to perform administrative functions, such as being removed from the list. If the list is unmoderated, any message submitted to the list will be sent to all other subscribers. If the list is moderated, an administrator will decide which submissions should

be distributed to other members. Messages may be sent one by one, or grouped and sent in a digest. Choosing to receive messages in digest form is a good idea for busy lists, lest your mailbox gets filled with hundreds of messages.

http://www.lsoft.com/listserv-hist.stm
http://www.wrt.tcu.edu/wrt/wri/files/barnes.htm
listserv distribute protocol information — rfc1429.txt

2.7.10 Information Search in Internet

2.7.10.1 Web Crawlers

There are several tools, known as Web Crawlers and more commonly referred to as spiders, which "crawl" through a Web site, acquiring some or all of the information on each page. They then allow users to search through this information in hopes of finding a resource relating to the topic of interest.

Many spiders have evolved into portals, which not only allow you to find information drawn from the Internet, but also provide links to other commonly used services, from maps to news headlines to package tracking. Examples include www.excite.com, www.lycos.com, and www.webcrawler.com.

2.7.10.2 Category Indexes

Another method of finding what you are looking for is to use a tool where sites are scanned by actual humans, and then placed into one or more categories. This method of locating information is a good complement to using a spider, since a spider may return too much information to be useful. The most popular of these services is Yahoo at http://www.yahoo.com.

2.7.11 Netscape and Microsoft

Netscape (NASDAQ: NSCP) and Microsoft (NASDAQ: MSFT) are two of the major players in the commercial Internet client and server markets. Netscape, established in 1994, was the first company to offer a commercial Internet browser, followed by server products. Microsoft was late to the party, but now also offers a full suite of Internet client and server products. Whereas Microsoft server products run on Windows, and their client programs span Windows, Mac, and UNIX, Netscape offers server products on both the Windows and UNIX platforms, and client programs for Windows, Mac, and UNIX.

Netscape has evolved from being a browser-only company, to one that now depends on server software and access to their NetCenter site for revenue. This is no doubt due to Microsoft's giving away their Internet Explorer browser. Netscape eventually made the source code to their browser available, in an attempt to gain back some browser market share by opening up their product.

Although the tight integration of Microsoft's browser and server products with the Windows operating system can offer increased functionality, it also tends to lock one into Windows. This can be a problem when attempting to use Microsoft products with more standards-based solutions like those from Netscape. Proprietary technologies like ActiveX controls and VBScript don't always work well with Microsoft products.

In the fast-moving world of the Internet, it is hard to predict which, if either, company will win. The strength of Netscape is that they provide solutions for a wider variety of systems, especially in the UNIX realm, and that they don't lock you into a single environment. The strength of Microsoft is that their products are on almost all desktop systems. For a solution where both client and server products are guaranteed to come from Microsoft, the increased functionality of being closely linked to Windows can be worth it.

To put things in perspective, the most popular Web server application as of this writing is the freeware Apache Web server, with almost 50% market share.

Intranet References

The following Web sites provide comprehensive sources of information and links to other sites for reference material relating to the Internet:
http://www.cisco.com
http://www.intel.com
http://www.microsoft.com
http://www.sun.com

References

Mockapetris, P. V. Domain Names—Implementation and Specification, RFC 1035, November 1, 1987.
Postel, J., Internet Protocol, RFC 791, USC/Information Sciences Institute, September 1981.
Reynolds, J., and J. Postel, Assigned Numbers, RFC 943, USC/Information Sciences Institute, April 1985.
*** Latest is RFC 990

Glossary of Intranet Acronyms and Terms

ARPAnet: Earliest packet switched network; the progenitor to today's Internet.
ASCII: Plain text file, containing only regular keyboard characters.
BCP: Bridging Control Protocol, used to configure bridge protocol parameters on both ends of a point-to-point link.
C/C++: High-level programming language allowing program control of system hardware, and designed for code portability on all types of computers. C++ is a programming language that extends the object-oriented capabilities of C.
CA: A Certificate Authority: a third party that can attest that you are on record as the only person with the key associated with a personal digital signature.
CCITT: International Telegraph and Telephone Consultative Committee.
CHAP: Challenge Handshake Authentication Protocol: a commonly used protocol for link encryption. Authentication occurs at the data link layer and is transparent to end users.
CGI: Common Gateway Scripting: used to support two-way browser communication.
Digital Signature: Used to validate the identity of the file sender's e-mail.
DNS: Domain Name Server protocol, part of TCP/IP.
ECP: Encryption Control Protocol, part of the PPP suite.
Ethernet: LAN protocol as specified in the IEEE 802.3 standard.
Firewall: Collection of hardware and software that interconnects two or more networks and, at the same time, provides a central location for managing security.
FTP: The File Transfer Protocol: a standard protocol for transferring and copying files from one computer to another.
Kbps: 1,024 bits per second.
Home Page: The first page that users see when they access a particular Web site.
HTMl : HyperText Markup Language: used to describe the layout and contents of pages on the Web. The Hypertext Markup Language (HTML) is the language of the World Wide Web.
ICMP: Internet Control Message Protocol: defines the rules routers use to exchange routing information.
ICMPv6: Internet Control Message Protocol Version 6.
IGMP: The Internet Group Management Protocol is used by IP hosts to report host group clusters to neighboring multicast routers.
IGRP: Interior Gateway Routing Protocol, part of TCP/IP.
IPCP: IP Control Protocol, used to configure IP parameters on both ends of the PPP link.

IPX: Internetwork Packet Exchange: Novell's implementation of the Xerox Internet Datagram Protocol (IDP), used to define a set of rules for coordinating network communication between network components.

ISDN: Integrated Services Digital Network: provides a digital communications circuit that allows transmission of voice, data, video, and graphics at very high speeds (from 56 Kbps to 128 Kbps), over standard communication lines.

ISP: Internet Service Providers act as middlemen, renting time to other users who want to access the Internet.

Java: A computer programming language that allows users to execute special programs (called applets) while accessing and viewing a Web page. Java is designed for creating animated Web sites.

l AN: Local Area Network.

l CP: Link Control Protocol: configures and tests the data link connection, and is part of the PPP suite.

MARS: Multicast Address Resolution Server.

Mbps: 1,000,000 bytes per second.

MPEG: Motion Picture Experts Group defining standards for handling video and audio compression.

OS/2: A high-performance, multitasking workstation operating system.

OSPF: Open Shortest Path First: link-state routing protocol used for IP routing.

PAP: Password Authentication Protocol: used for transparent session authentication occurring at the data link layer.

PGP: Pretty Good Privacy is a free (for personal use) e-mail security program developed in 1991 to support public-key encryption, digital signatures, and data compression. PGP is based on a 128-bit key.

PPP: Point-to-Point Protocol is one of the major protocols used to connect to the Internet. PPP is newer and faster than SLIP.

PKE: Public Key Encryption allows a sender to encrypt a document using a public key, which the recipient decodes using a private key.

POP3: Post Office Protocol version 3, allowing dynamic workstation access to mail drops on a server host.

PPTP: Point-to-Point Tunneling Protocol.

RIP: Routing Information Protocol: maintains network exchange and topology information.

RFC: Request for Comments: discussion notes, recommendations, and specifications for the Internet.

Router: This hardware device can be used to filter out data packets–based specific selection criteria. Thus, the router can allow certain packets into the network while rejecting others.

S-HTTP: Secure HTTP: a protocol developed by the CommerceNet coalition that operates at the level of the HTTP protocol.

Sl IP: Serial Line Internet Protocol: one of the major protocols used to connect to the Internet. It predates PPP.

SMTP: Simple Mail Transfer Protocol: specifies the format and delivery handling of electronic messages.

SPX: The Sequenced Packet Exchange protocol defines a set of rules for coordinating network communication between network components.

SSI: SSI software is used by Web servers to display and/or capture dynamic (changing) information on an HTML page.

SSl : The Secure Socket Layer (SSL) was developed by Netscape Communications to encrypt TCP/IP communications between two host computers.

Subnet: Partition on an IP network based on class, involving use and movement of a subnet mask.

Surfing: Term used to describe accessing (through a Web browser) a chain of documents through a series of links on the Web.

T1: A leased line that can support transmission speed to 1.54 Mbps.

TCP/IP: The Transmission Control Protocol/Internet Protocol: specifies the rules for the exchange of information within the Internet or an Intranet, allowing packets from many different types of networks to be sent over the same network.

TQM: Total quality management involves creating systems and workflows that promote superior products and services.

Telnet: This service allows connection to a remote Internet host so that programs can be executed from a remote computer.

u DP: User Datagram Protocol provides message service for TCP/IP.

u NIX: UNIX is an operating system well suited to the Internet's open system model.

u Rl : Millions of documents that are distinguished by a unique name called a URL (Uniform Resource Locator), or more simply, a Web address. The URL is used by Web browsers to access Internet information.

u u CP: Unix to Unix Copy: allows files to be copied from the Unix system to another.

XMl : Extended Markup Language: designed to provide a self-descriptive, platform-independent mechanism for exchanging management information between applications.

Web Browser: Allows you to traverse and view documents on the World Wide Web.

Windows NT: Provides a high-performance, multitasking workstation operating system.

WWW: The World Wide Web, or Web, is a collection of seamlessly interlinked documents that reside on Internet servers. The Web is so named because it links documents to form a web of information across computers worldwide.

2.8 Role of Open Source Software

Tivadar Szemethy

2.8.1 Introduction: The Nature of Open Source

Note: In this chapter, the abbreviation OSS stands for open source software, not to be confused with operations support system, which is the more common use of the acronym in the telecommunications industry. Other commonly used abbreviations for the open source concept are FOSS (free OSS) and FLOSS (Free/Libre OSS).

Open source products' source code is made available for end users. Furthermore, in a practical sense, software is considered open source if it has a license allowing the use of, changes in, and redistribution of the product regardless of modifications. Most OSS is also free (of charge) as well, since freedom with respect to change and redistribution is difficult when software from the legacy legal framework is used. Software companies pursuing the OSS model typically realize a profit by providing professional support and services rather than selling licenses and collecting royalties. Open source software is often developed in a loose collaborative manner, via Internet collaboration techniques.

Many varieties of OSS follow the bazaar methodology suggested by Eric S. Raymond in his landmark 1997 essay "The Cathedral and the Bazaar." There Raymond likened traditional software development to the building of a cathedral: Highly skilled craftsmen with individual responsibilities follow a master plan set by a venerable architect who has planned everything in detail and has precise oversight of the process. Roles (architect, manager, implementer) are clearly defined. In the bazaar, however, software emerges as the work of a swarm of members from "a great babbling bazaar of differing agendas and approaches." This model is characterized by the patterns outlined by Robles (2004), as follows.

- Users should be treated as co-developers: The project can reap important benefits by expecting more of the end-user community (as opposed to the traditional model). According to Linus's law, "Given enough eyeballs, all bugs are shallow." End users with access to the source code can assist

in finding and fixing problems. End users are not expected to be excellent programmers, but their sheer numbers can offset their lack of skill. In addition, each user machine serves as an extra test bed. In addition, allowing users to have their say in the development process ensures that the end product meets their requirements—a frequent flaw in traditional software systems.

- Early releases: Subsequent versions of the software should be released as early as possible (alpha stage). This increases the chances of locating testers and co-developers. End users are aware of potential problems and free to choose which version to use.
- Frequent integration: New code is integrated as quickly as possible. This alleviates lengthy testing cycles but leads to version proliferation. Again, end users choose the version they want to use, and stable versions emerge through evolution. Some OSS projects have nightly builds with automated integration supported by a large-scale automated test system.
- Several versions and branches: As a result of the two previous points, software versions proliferate. Thus, most OSS projects retain (at least) two "current versions": a development branch in which the latest features are integrated but not yet thoroughly tested and a stable branch that is in general use for a longer period of time. The latter version contains fewer features, but its performance has been proven during its longer lifetime. Users evaluate the benefits of using newer software with advanced features but greater risk versus an older, proven version that offers fewer features but more stability.
- High modularization: The structure of OSS is highly modular, allowing for customization and parallel development.
- Dynamic decision-making structure: Strategic decisions are often made informally. Project splits are common in response to the different needs of different user groups. Some OSS projects (typically those supported by foundations and large companies) have a well-defined decision-making leadership structure, while some are entirely informal and decisions are made after email or Web forum discussions.

In addition, although this is not one of the patterns characterized by Robles (though it is very important in the network management context), the software tends to be standards based. OSS typically follows open standards (if available), as they form natural cohesion points on which diverse developer groups can agree.

As a result of these factors, open source software develops more rapidly and is available in more varieties than its traditional, closed source counterpart. Users also have to accept the greater risk associated with using software from multiple, often hard-to-verify sources. Thus, the choice of an OSS solution should always be preceded by a careful evaluation and risk assessment.

Treating end users as co-developers also means that, in the case of OSS, more "initiative" is expected on the behalf of operators. Most OSS requires skilled operators who are more or less familiar with software development terms not in widespread use in their own particular trade (such as network management).

The most important advantage of OSS is that it is free of charge. Additional benefits include transparency and complete control over the application. Development of extensions or custom modules is easier.

2.8.2 Commercial and Noncommercial Applications and Tools

Open source software is typically commercialized through one of the following means:

- A dual-license model in which the code base is formed via a traditional open source–licensed product and a commercially licensed value-added extension is provided simultaneously. Vendors typically charge a perpetual license fee for additional closed source features, supplementary documentation, testing, and quality, as well as intellectual property indemnification to protect the purchaser from legal liability.
- Functional encapsulation, in which an open source framework or library is installed separately from the commercial product and the commercial product relies on the open source

functionality. Commercial vendors of such software argue that the commercial software was sold and shipped without the open source library, even though it uses it. Similar to the dual-license model, vendors typically charge a perpetual license fee for the functionality they provide under closed source.

- A software as a service model in which the vendor charges for the services rendered rather than for the software itself. In this arrangement, users access the service hosted on the vendor's computers; it is not installed on the users' systems. For the end user, the software implementing the service is often opaque and its exact license is irrelevant. Vendors typically charge a monthly subscription fee for use of their hosted applications.
- Charging only for additional services, such as support, training, and consulting services that assist in the deployment and use of OSS. Vendors typically charge an annual fee for support, per-student fees for training, and per-project fees for consulting engagements.

End users have the choice between different levels of professional support versus the complete do-it-yourself approach. All kinds of arrangements are common; for example, with complex systems users may decide to develop significant in-house expertise and opt for the DIY approach. For everyday software, such as a reporting tool or a productivity application, IT departments may simply contract an OSS support company to perform the installation and provide support for end users.

In rare cases, vendors of proprietary software agree to provide the source code to end users in a form of "source as documentation" without granting permission to modify and redistribute the code. These arrangements are not considered to be open source.

2.8.3 Opportunities and Vulnerabilities

This section summarizes the most important opportunities or business drivers behind the proliferation of OSS systems. For each topic, the business opportunity or benefit is discussed along with the potential pitfalls for OSS.

No-cost acquisition, modification, and (re)distribution: The most important benefit is the ability to obtain industry-grade software without paying a penny to a software vendor. Most OSS comes with no licensing fees. In the case of commercial OSS products (typically prepackaged software such as Red Hat Linux), this cost is usually a fraction of that of a similar vendor product (e.g., SCO Unix).

The ability to modify the product without legal hurdles and internal or external (re)distribution is also an important factor. Several companies offer value-added services for FOSS products.

In contrast, all OSS licenses strongly disclaim any form of warranty or reimbursement in the case of damage suffered from use of the software. This is no different than with commercial software, for which most vendors also disclaim reimbursement if the software fails. Some vendors offer reimbursements up to the purchasing cost of the software if customers' requirements are not met. This is not much of a help, as usually this amount is significantly less than the potential damage the software's failure can cause.

Reduced cost and time for development and integration: Quite often, OSS libraries or other basic infrastructure elements (such as the JBOSS Java Application Server) form the backbone of a customized solution. The availability of a proven, standards-based, well-documented OSS solution is a major stepping stone in such projects. OSS software by its very nature is standard compliant or standard setting, and thus its integration is easier than that of a proprietary solution.

Should the company choose to redistribute to the public the modules it has developed within an open source project, it can benefit from the co-developers' peer reviews, tests, and even bug fixes, fast and free of charge.

Integrating an OSS solution with an existing proprietary system might lead to additional development requirements. Although usually the proprietary system is to blame (e.g., owing to a lack of standard interfaces), both the OSS chosen for integration and the legacy system need to be scrutinized for interoperability in the presence of proprietary systems.

No vendor lock-in: A company using a standards-based OSS solution will not be bound to use proprietary solutions or pay recurring licensing fees. Publicly developed OSS systems are less likely to become "dead-end" solutions, as often occurs with pioneering vendors' systems. In contrast, companies relying on OSS systems should be on the watch for project or community splits, which are common in the OSS world. In this case, during the early stages of the diverging new project(s), migration paths are provided, as no development branch wants to be without a user community.

Commercial support for OSS software is usually available from multiple competing companies that specialize in such support. Competition also means varying quality; thus, the business partner must be evaluated beforehand. Fortunately, replacing an OSS support company is easy thanks to the strong competition.

Increased quality, reliability, safety, and integrity: Open source projects are born and evolve under the scrutiny of the public eye. In the case of less mature projects, each user is a tester and co-developer. If a user base is sufficiently large, it will discover and help mitigate quality and reliability problems. The same can be said for safety and security. As for the latter, companies relying on OSS solutions also need to be aware that any information about newly discovered vulnerabilities spreads fast within the OSS community and the general public. Thus, companies need to monitor such information constantly and be ready to act instantly according to an appropriate contingency plan. It also should be mentioned that with the source code available, an OSS can be audited at an arbitrary depth, as opposed to trusting the word of a vendor whose development and testing methodology is not disclosed.

Increased transparency and customizability: Related to the previous topic, OSS software tends to have a better architecture and to be more transparent. OSS solutions, engineered for a large user base, also tend to be more generic. They are easier to customize, but they require more in-depth customizations from their prospective users. Advanced skills and knowledge (over that of a vendor's solution) are needed from the company, and this expertise must be acquired either in-house or by hiring an OSS support company to do the customization.

Training, which requires advanced skills for operating OSS solutions, is typically available through self-learning (e.g., Web tutorials and extensive documentation). OSS support companies also offer educational services. External training from competing support companies tends to be cheaper than that of a software vendor, for whom training is just another source of revenue from an already committed customer.

Increased standards adoption and synchronization with technology trends: As mentioned earlier, OSS is standard based or standard setting (sometimes both). Thus, building on OSS solutions automatically ensures compliance with standards, both internally and externally. The same can be said about technology trends: Public OSS, shaped by its large and active user base, tends to shape as well as follow trends. OSS projects cover the entire breadth of the technology landscape, from established solutions to the "cutting edge."

For this very reason, the maturity of OSS solutions is not uniform. Many projects are academic or in an experimental stage. To avoid relying on immature systems, the OSS project's developer and user community needs to be evaluated. If a well-known, widely used solution cannot be found, smaller OSS projects need to be scrutinized for maturity.

Best indicators include development activity, frequency of releases, activity of online forums, and bug trackers or mailing lists. Positive signs are the presence of up-to-date, well-maintained documentation; "stable" and "development" project branches; and the availability of different packaging or distribution options. Although confusing, the latter are still a good sign of development activity and a sufficiently large user community to sustain different varieties of the same software.

As mentioned earlier, use of OSS typically requires a deeper understanding on the part of its operator. Although this might seem to be a burden with respect to operating costs, it is also an important driving factor for the operator team to be "on the edge" and to have a better overall understanding of their trade and the field. Good management and leadership will translate this into competitive advantage.

Synchronization with technology trends is a two-way process in the OSS world: Taking an active role in OSS projects and submitting source code enables the company to shape the evolving project. With

good strategy and effort allocation, influencing the entire field of technology through a high-impact OSS project (such as a reference implementation) is possible at relatively low cost. Contributing to a high-visibility OSS project has significant marketing value as well, and major contributions may establish the company as an expert or major player of the field, bringing strategic advantage.

Potential OSS vulnerabilities: The following is a brief list of potential problems OSS users need to be aware of.

- Lack of professional support: There is always the chance of running into a problem no one has faced before and no one is willing to fix.
- No fixed timetable (development, problem solving), volatile road map: Problems are solved sooner or later, but typically without deadlines. Road maps and development plans are subject to change.
- Low-quality code: Many OSS developers are not professionals. OSS projects often include academic code not suitable for mission-critical applications. Popular OSS projects evolve away from this stage with time, but the danger of running into a piece of code submitted by an overconfident student co-developer cannot be neglected.

2.8.4 Licensing

As with any other intellectual product, computer software is covered by copyright law. Thus, open source software is covered under a license, usually set explicitly by the author or the project's owner. OSS licenses are primarily concerned with (1) guaranteeing that free open source software remains free and open source through redistribution (i.e., no one should unfairly benefit from it) and (2) setting rules for the modification and inclusion of the software into other products. Thus, in the case of end users who do not plan to redistribute or modify the software, the actual license does not matter.

OSS licenses can be classified into two fundamental categories: copylefting and non-copylefting licenses. A program released under a copylefting license allows anyone to change the program, but those changes must be provided to recipients under exactly the same conditions as the original. A FLOSS program released under a copylefting license cannot be later transformed into a proprietary program by a third party. Most FLOSS software is released under copylefting licenses, such as the General Public License (GPL) and the Lesser/Library General Public License (LGPL). GPL is a "strict" license banning the integration of GPL code with non-GPL code; LGPL is more permissive but still requires the full source code to be disclosed to end users. Copylefting licenses allow proprietary vendors to easily incorporate the code into their products and modify it any way they like.

Non-copylefting licenses are often used when the goal is to promote the adoption of a standard. An example of this approach is the implementation of the Internet's suite of standards (often called the TCP/IP standards); the key code was licensed under a Berkeley Software Division (BSD)-style license, and it has become ubiquitous.

2.8.4.1 Public Domain

Releasing software to the public domain means abandoning all copyrights. The public domain principle is applicable only within certain legal environments. Public domain software is largely developed with government support at universities or research institutes. Taking the software into the commercial domain is allowed. The major types of open source licenses are summarized in Table 2.8.1.

2.8.5 Cost Considerations (CapEx, OpEx)

First and foremost, OSS promises significant savings on capital expenditures (CapEx) by eliminating acquisition and license fees. This is offset by the following initial expenditures: (1) cost of evaluation (quality and performance) for lesser-known systems and (2) increased customization and training costs for more generic OSS solutions.

TABl E 2.8.1 Types of Open Source Licenses

Property 6 ➧ / License 4 ➧		GPL	LGPL	BSD & MIT	APACHE	Public Domain	Microsoft MIT4 EULA
A	Can be stored on disk with other license types	✓	✓	✓	✓	✓	Bans FLOSS[5]
B	Can be executed in parallel with other license types	✓	✓	✓	✓	✓	
C	Can be executed on top of other license types	✓	✓	✓	✓	✓	
D	Can be executed underneath other license types	✓[1]	✓	✓	✓	✓	
E	Source can be integrated with other license types		✓	✓	✓	✓	
F	User decides whether and when to publish derived code	✓[2]	✓	✓	✓	✓	✓
G	Software can be sold for a profit	✓	✓	✓	✓	✓	✓
H	Binary code can be replicated by users as desired	✓	✓	✓	✓	✓	✓
I	Binary code can be redistributed as desired	✓[3]	✓	✓	✓	✓	
J	Binary code can be employed as desired by users	✓	✓	✓	✓	✓	
K	New users always receive source code of derived works	✓	✓[6]				
L	New users receive full source modification rights for derived works	✓	✓[6]				
M	New users receive full redistribution rights for derived works	✓	✓[6]				
N	Binary code can be released without source code			✓	✓	✓	✓
O	Derived code can have a different type of license		—[7]			✓	
P	Original source can be incorporated into closed source products					✓	

Note: Properties A through E refer to the ability of a license to coexist with other types of software. In this category, the most exclusive license is the Microsoft MIT EULA license. The GPL takes a very distant second place for exclusivity, since it forbids design-time incorporation of GPL source code into non-GPL source code. However, unlike the Microsoft MIT EULA, the GPL places no constraints on software simply running on the same system. The GPL even allows non-GPL software to use GPL software as long as the two programs are not inextricably linked to each other (i.e., they can both be used independently in other contexts). The LGPL allows software to be directly incorporated into non-FLOSS software. The BSD and Apache licenses are more accommodating still by allowing distribution in binary form only. The most permissive category is public domain software, which allows essentially any use. Properties K through M represent the flip side of the somewhat restrictive nature of the GPL: its ability to ensure that later generations of users will inherit exactly the same rights to use, change, and redistribute GPL software as the first generation of users. Properties N to P deal with further aspects of redistribution and derivative works: (L)GPL strictly requires software to be released together with its source (as the "ultimate documentation"), while other license types are more relaxed. All formal license types prohibit changing licensing through derivative work (O,P).

[1] Provided that both programs are fully and independently usable in other unrelated contexts.

[2] Provided that the binary code has not been previously released to the public.

[3] Provided that source code is always redistributed along with the binary code.

[4] The proprietary Microsoft MIT EULA is not related to the similarly named MIT (X/MIT) license.

[5] Specifically bans the use of GPL, LGPL, Artistic, Perl, Mozilla, Netscape, Sun Community, and Sun Industry Standards.

[6] The rights granted by LGPL do not necessarily extend to the applications linked to an LGPL library.

[7] The LGPL permits relicensing under GPL as a special case but does not allow relicensing under any other license type.

With few exceptions (where OSS solutions dominate the market, such as the Apache Web Server), adoption of OSS may increase operational expenditures (OpEx).

This is mainly due to the fact that OSS users need to assume a more proactive role. Activities such as following user community forums (e.g., monitoring community mailing lists) are much more important than with commercial products.

When they encounter a problem (e.g., a software bug), OSS users are expected to submit a developer-quality report if they want their problem solved by the community. These activities require alert and skilled operators.

Introduction of the first large OSS-based system might bring a "culture shock" as the company needs to adjust to the different ways of working with OSS. During this introductory period, it may be helpful to hire an OSS support company, which can provide support for internal users. Although this may increase initial expenditures, it results in a smoother transition. These challenges can be appropriately addressed with a proper OSS adoption policy.

The quick release cycle of OSS also means that the company has to continuously evaluate upcoming releases and decide when to upgrade. Following successive software releases takes effort and requires advanced skills on the part of users. On the other hand, given a large enough user base, maintenance of older versions will continue, since with OSS the user base performs the maintenance. If everything else fails, the company can take the initiative and hire or allocate resources to do the maintenance in-house as source code is available.

Generally speaking, the gains associated with CapEx reductions significantly outweigh the increased OpEx and introduction costs. Nevertheless, careful cost analyses with a special emphasis on OSS characteristics are necessary, as would be the case with any business process.

Competition among OSS support companies will keep prices low, and thus operational expenditures such as support costs will remain consistent. For the same reason, there is no sole intellectual property owner who can monopolize the solution, making long-term risk management easier.

Open source adopters are frequently warned about the lack of OSS warranties. In the telecommunications/management area, similar to the general software realm, it is extremely rare that software systems are shipped with explicit warranties. The vast majority of commercial software also comes with no warranty.

2.8.6 Concrete Examples

2.8.6.1 Web Servers

The Apache server family is a good example of a mature OSS product. Developed under the leadership of the Apache Software Foundation, a nonprofit organization, the project coordinates the work of hundreds of volunteers.

Development of the Apache HTTPS started in 1994, and the software played a key role in the growth of the World Wide Web. As such, Apache serves as a de facto reference implementation of the HTTP protocol and a technology leader.

With a large user base, Apache emphasizes stability and reliability as well as quality documentation. Support for older versions is also strong.

The Apache Foundation coordinates several other OSS projects as well, with the HTTP server being the flagship product. According to a Netcraft report*, the Apache HTTP server had a 48% share of the global Web server market as of October 2007.

2.8.6.2 Browsers

Mozilla Firefox is a mainstream browser that, according to a Net Applications (http://netapplications.com) report, had a 15% market share as of October 2007. The browser is the flagship product of the

* Netcraft report available at http://news.netcraft.com/archives/2007/10/11October_2007_web_server_survey.html.

Mozilla Foundation, whose mission is similar to that of the Apache Foundation. The HTML engine, the core of Firefox, is made available separately (via modularization), and there are many other derivative and branded browsers based on it. The Mozilla Foundation also coordinates many other, less prominent software projects such as the Mozilla Thunderbird mail client.

2.8.6.3 Database Management

The two major open source relational database engines are MySQL and PostgreSQL. MySQL development was coordinated by a small, for-profit company (MySQL AB in Sweden) until its acquisition by Sun Microsystems in early 2008. MySQL was the first major example of the business model in which a software company provides all product source code (and the rights to use it commercially) free of charge and derives profits only from services and support. MySQL is licensed in part under GPL and in part under the proprietary MySQL EULA.

MySQL had more than 11 million installations at the time of the Sun Microsystems acquisition (according to press releases), and it played a key enabling role in the proliferation of interactive, database-backed Web applications (such as e-commerce shops). Although it is not able to compete with industry heavyweights such as Oracle and DB2 in terms of robustness, scalability, and features available, MySQL is lightweight, fast, and dependable and serves a popular market demand.

PostgreSQL, the other significant OSS database management system, has a long history of evolution (it began in 1985 based on an older project called Ingres). The software is available under a BSD-style license and supports many interesting features such as concurrency and certain object-oriented concepts (such as inheritance).

According to the Annual Java Use & Awareness Study (www.bzmedia.com/bzresearch/x7672.htm) in December 2006, Apache Tomcat and Geronimo accounted for 64% and 12%, respectively, of the Java Application Server market at that time.

2.8.6.4 Languages and Development Environments

The key enabling factor in the proliferation of OSS was the groundbreaking work performed by the Free Software Foundation in providing FLOSS compilers and development tools, such as the GNU Compiler (GCC) and the GNU development tool chain. Their importance in the open source movement cannot be overestimated.

The most important open source development environment is Eclipse. The development is coordinated by the Eclipse Foundation, which was established by a consortium of influential software vendors (most prominently IBM but also BEA, Intel, CA, Sybase, Borland, and others). Thus, unlike many typical varieties of OSS, Eclipse is not the result of grassroots development. Eclipse is mainly oriented toward Java development, although it also provides C/C++ tooling. Eclipse is highly modular (built around the Eclipse Rich Client Platform).

2.8.6.5 Security Software

The open source community has long been a strong opponent of the security through obscurity method, according to which a solution maintains security by not revealing its algorithms, internal workings, and implementation. The argument has been that true security can be achieved only through open, verifiable, and peer-reviewed methods and software. In this area, the definitive example is the OpenSSH platform, which implements public-key cryptography for remote access and file transfer. OpenSSH was developed by the OpenBSD project and thus has a BSD license.

2.8.6.6 Application Servers

In the J2EE application server market, the two prominent open source solutions are JBoss AS (application server) and GlassFish. JBoss, developed by a company of the same name, was acquired by the Linux and Open Source company RedHat in 2007. JBoss has a very large and active user community and is a viable alternative to commercial competitors such as BEA WebLogic and IBM WebSphere.

GlassFish was developed by Sun Microsystems (with community contributions) and licensed under GPL and Sun's own OSS license, CDDL (Common Development and Distribution License). GlassFish is included with the latest official Java Enterprise Edition downloads and thus is rapidly gaining in popularity.

2.8.6.7 E-Mail Servers

According to an August 2007 Simple Mail Transfer Protocol (SMTP) survey (available at http://smtpsurvey.stillhq.com) conducted by E-Soft Inc., four SMTP servers had more than a 10% share of the installed base at that time: Sendmail, Microsoft Exchange, Exim, and Postfix. Of these servers, only MS Exchange is not open source (the leader, Sendmail, had a 29% share). SMTP servers tend to be very sophisticated software with complex functionality and strict security and scalability requirements. An important factor in open source software's success in this area is the community's ability to respond quickly to security problems. With the e-mail systems globally connected, malware (worms, trojans, viri) is able to spread fast and overwhelm the worldwide e-mail system in a matter of hours. Critical security patches are available much faster for OSS servers than for commercial e-mail systems such as MS Exchange.

2.8.6.8 Office Suites

The office suite market has long been dominated by closed source proprietary solutions, most notably Microsoft Office. In 1999, Sun Microsystems acquired StarDivision, the developer of the StarOffice suite and a small competitor of the MS suite at the time. In 2000, Sun created the GPL-licensed OpenOffice.org project based on the StarOffice code with the specific intent of reducing Microsoft's market share. After opening up of the source code, community-based development has intensified quickly, and today OpenOffice.org is a viable alternative to Microsoft Office in many respects. OpenOffice.org emphasizes strong cross-platform support (Windows/Unix/Mac) as well as the ability to import/export MS Office document formats.

2.8.6.9 Wikis

Web-based lightweight collaboration and information-sharing systems are becoming increasingly important—in both the business and private spheres—in our globalized world. Wikis are the best examples of these systems: They are enablers as well as products of the open source era. The most prominent example is MediaWiki, the engine behind Wikipedia, the online open content and multilingual encyclopedia that included more than 10 million articles as of March 2008. Wikimedia, coded with the open source PHP Web scripting language, stores its database in either MySQL or PostgreSQL (these systems are, again, open source). Apart from Wikimedia, there are dozens of simpler open source Wiki systems available, serving the needs of various communities.

2.8.6.10 Content Management Systems

Content management systems (CMSs) provide frameworks that allow people who are not Web or computer experts to create (typically collaborative) Web content. CMSs provide user-friendly content editors and manage users with various roles and privileges (e.g., authors, proofreaders, content managers). They also control the workflow of content publishing and discussions and provide further enhancements such as personalized content viewers and support for multiple presentation formats (e.g., HTML, PDF, Flash animation).

The most prominent open source CMS systems are Drupal and Joomla! Both systems are licensed under GPL and support multiple languages and internationalization (a very important feature), operate on multiple platforms, and use open, standard database interfaces.

2.8.6.11 Virtualization

Virtualization, an emerging technology trend, has long been the territory for proprietary solutions from vendors such as VMWare and Citrix. Virtualization allows the abstraction of hardware resources to the extent that a single computer is able to execute several instances of operation systems independent of

each other (with the obvious limitations associated with the physical resources of the host system). The first and most important open source virtualization platform is Xen, licensed under GPL. Development of the Xen codebase is overseen by the Xen Advisory Board. The most prominent industry players (e.g., IBM, Intel, HP, Sun) have joined the advisory board, which is headed by Citrix.

2.8.6.11 CRM and ERP

There is only one significant open source product in this area, SugarCRM. Adoption of OSS is somewhat difficult for systems such as customer relationship management (CRM). The relatively small user base (CRM administrators) demands stable solutions; core OSS ideas such as early and frequent releases and treating users as co-developers are not a good match for CRM systems. The developer (SugarCRM Inc.) of SugarCRM provides the core system (Sugar Community Edition) under a GPL license. The other two bundles (Sugar Professional and Sugar Enterprise) are sold for an annual subscription fee, sharing about 85% of their modules with the community edition. The company realizes profits mostly from software-related services (implementation/deployment and support) and from professional services such as CRM strategy development, analysis, and consultation; another area of profit is custom module development.

2.8.6.12 Business Process Management

Standard automated business process management (BPM) is a fresh direction within enterprise IT systems. There are several open source BPM engines and solutions centered around the BPMN/BPEL (Business Process Modeling Notation/Execution Language) standards.

One of the leading OSS BPM solutions is Intalio's BPMS (Business Process Management System). The software comes with its own license, which restricts its usage but makes the source code available for the end user. Intalio has also donated code (parts of the BPMS) to the Apache Foundation and Eclipse Foundation.

The other major open source product in this area is jBPM from JBoss, which is an LGPL-licensed BPEL environment written in Java.

2.8.7 Summary and Trends

The fear, uncertainty, and doubt that have surrounded open source since its inception seem to be dissipating. IT operators—especially in the telecommunications area, where compliance with standards is vital—have recognized that "riding the wave" is important. Furthermore, using OSS leads not only to cost reductions but also to a more efficient and better skilled operator team. Open source users need to be alert and proactive; in exchange, they gain the power to shape emerging solutions by participating actively in the open source user community. To promote technological innovation while controlling CapEx and OpEx, enterprises are increasingly considering the deployment of open source–based solutions.

There are a number of commonly recognized barriers to the adoption of open source software by enterprises. These barriers include the perception that open source licenses are viral, lack of formal support and training, velocity of change, and lack of a long-term road map. The majority of these barriers are risk-related. Many business models exist around open source software to provide a "whole product" to help reduce these risks. The whole product typically includes support, professional services, training, certification, partner programs, references, and use cases. These business models range from services-only organizations that do not participate in the development of the software to models in which the majority of the software is created by full-time committers employed by a central organization. Thus, in the long term the commercial support offered by competing OSS companies takes the place of vendors' support for their proprietary solutions. The competition of different OSS companies supporting the same free software supersedes the competition of vendors offering their own solutions, ultimately placing the customer in a better position.

Customers (OSS users) need to adopt the necessary risk management attitudes and methodologies. Most of these practices are not fundamentally different from those associated with traditional pro-

prietary software. However, OSS adoption and usage necessitates certain distinctive risk management practices in order to evaluate the opportunities and vulnerabilities related to this new paradigm.

References

Raymond, E. The cathedral and the bazaar, http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/. Published 2000.

Robles, G. 2004. A software engineering approach to Libre Software, in Robert A. Gehring and Bernd Lutterbeck, *Open Source Jahrbuch*.

Terplan, K. 2008. Open source for system, network, and service management. Presented at CECMG Conference, Hamburg.

Wiener, J., and Bruce, G., eds. 2005. Open source for Next Generation OSS: Issues and challenges, EURESCOM Study Report.

Acronyms

AAL	ATM Adaption Layer
BSB	Berkeley Software Distribution
CRM	Customer Relationship Management
ERP	Enterprise Resource Planning
ITU-T	International Telecommunications Union
GCC	GNU-C Compiler
OAM&P	Operations, Administration, Maintenance, and Provisioning
QoS	Quality of Service
SMTP	Simple Mail Transfer Protocol

Summary and Trends

Patricia Morreale and Kornel Terplan

Intranets and the associated network solutions and applications, such as virtual private networking and e-commerce have been one of the greatest areas of growth in modern telecommunications. With this growth comes potential problems and defensive solutions must be adopted. The Internet and intranet management solutions presented here provide an outstanding discussion of opportunities and pitfalls to avoid.

It is not only the performance the underlying network, but also the network applications that create the performance perception on the part of the user and the associated enterprise. With this in mind, Web and application performance management also require consideration in the design and deployment of network solutions.

Open computing and networking standards may help companies respond more rapidly to customer needs. Open computing is the philosophical principle that describes architecture and technology procurement policies and practices that align IT with the first principles of openness by ensuring interoperability with open standards. Open standards are specifications for APIs, protocols, data, and file formats that are openly documented and published without restrictions that limit implementations. Open standards such as HTTP, HTML, TCP/IP, XML, and SQL have evolved collaboratively, typically from various IT or software companies that collaborate under the auspices of standards organizations such as W3C, OASIS, OMA, ISO, and IETF. Open standards are implemented through offerings available in the market. Open architecture is a flexible architectural approach that allows for the loose binding of application functionality through standardized interfaces. Open architectures provide the necessary independence to isolate and distribute work to the most effective teams within and outside the organization.

Open source is software whose source code is published and made available to the public, enabling anyone to copy, modify, and redistribute the code without paying royalties or fees. Open source software usually evolves through community developers composed of individual programmers as well as large companies. Community innovation is the common thread that runs across each of the three open computing components. All three require active support and collaboration among individual developers, IT vendors, standards organizations, governments, and universities to accelerate innovation and promote the advancement of critical standards. Open computing accelerates the process from idea generation to market distribution through a standardized information technology platform. As a result of the alignment of technology and business strategies made possible through open computing, businesses have more technology choices and greater flexibility to solve business problems more efficiently.

3

Network Management and Administration

Introduction **3**-2

3.1 Management Concepts **3**-4
Management and Data Communications • Management Requirements • Management Paradigms • Open Management Systems • Distributed Management Systems • Management Systems Topological Frameworks • Management Systems Evolution • Summary and Trends

3.2 Management of Emerging Technologies **3**-19
Introduction • Foundation Concepts for Networking Technologies • Management Solutions for Emerged Technologies • Discussion of Select Technologies • Next-Generation Wireless Access Technologies • Next-Generation Telecommunications Application Technologies • Telco and Web 2.0 Applications • Summary and Trends

3.3 Management-Related Standards **3**-51
Manager–Agent Relationship • Simple Network Management Protocol (SNMPv1, SNMPv2, SNMPv3) • Remote Monitoring (RMON1 and RMON2) • Desktop Management Interface • TR-069 CPE WAN Management Protocol • Telecommunications Management Network • TOM and eTOM • OSS-J • Transitioning from IPv4 to IPv6 • Web Service Technologies and SOA: SOAP, UDDI, BPEL • Summary and Trends

3.4 Management Function **3**-84
Management Functions • In-Depth Considerations of Selected Management Functions • Summary and Trends

3.5 Support Systems for Service Providers **3**-104
Status, Definitions, and Markets for Support Systems • OSS/BSS Market Drivers • Future of OSS/BSS • Summary and Trends

3.6 Support Processes for Service Providers **3**-132
High-Level Breakdown of Support Processes • Enterprise Management • Customer Relationship Management (CRM) • Customer Billing and Collections Management • Service Management and Operations (SM&O) • Resource Management and Operations (RM&O) • Supplier/Partner Relationship Management (S/PRM) • Support Processes Taxonomy • eBusiness • IT Management Frameworks and the Information Technology Infrastructure Library (ITIL) • Summary and Trends

3.7 Management Frameworks and Applications **3**-169
Evolving Management Frameworks • Features and Attributes of Management Frameworks • Management Applications • Vendor Profiles • Consolidation of Support Systems • Results of Market Research • Summary and Trends

Árpád Bakay
Netvisor

Jim Frey
NetScout Systems

Joe Ghetie
Telcordia

Paul Hoffmann
Datakom

Norman Kincl
Hewlett-Packard

Tivadar Szemethy
Netvisor

Kornel Terplan
Industry Consultant and Professor

József Wiener
T-Com Hungary

3.8 Intelligence Support Systems 3-200
Positioning Lawful Intercepts and Surveillance • ISS Basics and Application Areas • ISS Positioning among Other Support and Security Systems • Basic Requirements for Lawful Intercepts • Principal Functions of Interception • Reference Models for Lawful Intercepts • Principles of Monitoring and Intercepts (Hardware and Software Probes) • Receiver Applications • Summary and Trends
3.9 Management of Sensor Networks 3-226
Objectives of Sensory Monitoring Systems • Challenges of Monitoring • Sensory Monitoring Technologies and Alternatives • Selecting and Applying Sensory Systems • Summary and Trends
3.10 Solution Architectures 3-241
Business Strategy Drives Technology • A Comprehensive Architectural Approach • Business Drivers Require a New Approach to OSS • The Job of an OSS • How to Build an OSS
Summary and Trends 3-258

Introduction

It is not enough for telecommunications service providers to develop, implement, and roll out new technologies. These technologies should be properly administered and managed. Over a five-year period, management and administration will account for up to 85% of operating expenses, with acquisition of the technology making up just 15%. This is a very important metric for justifying management and administration investments.

This chapter addresses administration and management issues. Management concepts outline the basics of managers and managed entities. Concepts include management models such as centralized and decentralized models, concentrated and distributed models, and the use of hierarchical schemes supported by umbrella managers. Also, open management is addressed, including the open systems conceptual model, associated systems concepts, and requirements for open management systems. The distribution of management processes and functions will play a key role in future management solutions. Managed entities must be connected with element managers and management platforms via in-band or out-of-band communication schemes. Examples of both alternatives are provided.

Administration and management are usually an afterthought when considering the deployment of innovative technologies. This chapter attempts to synchronize technology deployment and selection and implementation of management solutions. Innovative technologies such as asynchronous transfer mode (ATM), MPLS, WiMax, Wi-Fi, triple-play, quad-play, IPTV, voice over IP (VoIP), converged services, cable services, push-to-talk, unified messaging, presence services, and instant messaging are investigated with the goal of determining the extent to which management and administration solutions are available and implementable. In particular, the availability and structure of management information bases (MIBs) are analyzed. Typically, MIBs support most fault, configuration, performance, security, and accounting management functions. In combination with simple network management protocol (SNMP) managers, MIBs are useful in regard to data visualization, analysis, and reporting. State-of-the-art technologies require additional management tools and applications that help with real-time decision support.

Management and administration depend to a large extent on management standards. There are two principal groups of standards: standards for enterprise-level administration and management and standards for specific telecommunications environments. The section devoted to management standards focuses first on enterprise-level standards such as SNMP, RMON, and DMI. Components of telecommunications standards are also discussed in some depth (e.g., CMIP, CORBA, and DCOM). Readers are provided with a better understanding of management framework products, service delivery platforms, and management applications.

The telecommunications management network (TMN) is a simple model for streamlining management and administration. It uses four layers in addition to the network element's layer at the bottom. Management processes, functions, and tools can be categorized in accordance with these layers. The section focusing on TMN provides an in-depth discussion of various TMN models (information, functional, and physical), elements (operations system, workstation, mediation, Q adapter, and network element functions), and internal and external interfaces (Q3, Qx, X, F, and M), as well as the most appropriate use of data communication networks (DCNs).

Management-related technologies such as UDDI, LDAP, OASIS, BPEL, SAML, OSS/J, SID are discussed in some depth to illustrate to telecommunications service providers how these technologies can help them create and maintain services and acquire and retain customers. The use of Web 2.0 technologies for both service creation and management offers completely new opportunities in terms of solution architectures.

The TeleManagement Forum offers guidance for deploying and reengineering telecommunications business processes. This chapter adopts the basic business model of breaking down support processes into two dimensions: service life cycle (e.g., fulfillment, service assurance, billing processes) and service hierarchy (e.g., customer care processes, service development and operations processes, network and system management processes). NGOSS concepts based on eTOM and SID help service providers streamline their business processes and align them with IT support systems. Principal service fulfillment, service assurance, billing, and revenue assurance solutions are addressed in some depth.

Management frameworks, which consist of an application platform and management applications, are the heart of support systems for telecommunications providers. This chapter outlines principal attributes, such as architecture, application programming interfaces, protocol support, hardware and software platforms, graphical user interfaces, application programming interfaces, management functions supported, security modules, modeling capabilities, and internal systems services. Framework products are listed for both telecommunications and enterprise environments, and a few, such as Telcordia, Amdocs, Hewlett-Packard, IBM, and Oracle, are analyzed in some depth. Over the next couple of years, frameworks are expected to embed the best of suite management applications, with the result that full functionality will be available to implement operations, business, and marketing support systems.

Intelligence support systems (ISSs) focus on the expanded infrastructure requirements of service providers, which are basically no different than the requirements of operations support systems (OSSs) and business support systems (BSSs). Intelligence plays two principal roles. First, it provides surveillance by collecting information on illegal activities such as terrorism, criminality, fraud, and money laundering. Second, it provides basic data that improve the bottom line, such as revenue assurance, business intelligence, and fighting telecommunication fraud. In short, ISSs are software elements or units that interface with or are a part of billing, ordering, provisioning, and authentication systems, as well as interfacing with outside parties such as law enforcement agencies.

Monitoring plays a key role in managing large networks. One of this chapter's segments is devoted to technologies designed to collect, consolidate, process, and store large amounts of performance data. Both hardware- and software-based sensors are discussed in depth. Particular emphasis is placed on overhead and how monitoring tools integrate with enterprise management platforms and business intelligence solutions.

The support systems of telecommunications providers represent a complex but increasingly significant segment of the communications industry. After a discussion of the market drivers for support systems, such as network complexity, increased standards, high growth rates, deregulation, and convergence, the strategic benefits of such systems are examined. Support system suppliers, including consulting companies, computer manufacturers, equipment manufacturers, software companies, and outsourcers, are also described. The remaining sections focus on positioning products in terms of supporting markets (voice, data, Internet, cable, and wireless), supporting management areas (customer care and billing, provisioning and order processing, and network operations management), and compliance with TMN layers (e.g., business, service, network, and element management layers).

Traditional players, such as Telcordia, Amdocs, and Hewlett Packard, and new entrants, such as IBM and Oracle, attempt to consolidate and streamline their product portfolios and solution architectures through acquisitions. They are able to offer complete best-of-suite solutions for addressing practically all principal service provider support processes. The final segment of the chapter introduces solution architectures from Hewlett-Packard with a particular focus on the life cycles of service and resource management.

3.1 Management Concepts

Joe Ghetie

The last decade witnessed one of the most dramatic advancements of communications technologies and services in human history. Communication, as a way of conveying and exchanging management information, had found in the Internet one of the best examples of the explosive growth with a tremendous impact on the current and future abilities of humans to share information.

The dream of universal access to information, the dream of a giant village, the dream of fast, reliable, content-rich information exchange are today closer to reality than anybody has anticipated. Data communications, video communications, and both wired and wireless communications media have increased our ability to control large, global enterprises and businesses through communications.

Network and systems management are specialized systems targeting, monitoring, and controlling the vast array of network and computing systems resources used in communications, manufacturing, commerce, finance, banking, and education, as well as in research and development.

Management systems were born out of necessity to prevent, diagnose, configure, and solve problems raised by the size, complexity, and heterogeneity of multivendor, multiprotocol, and multitechnology environments that characterize the underlying network and computing systems.

Although management systems are value-added components to communications technologies, they are as vital as the transmission, switching, and operations systems in order to supervise and maintain the normal information exchange.

3.1.1 Management and Data Communications

Management systems aimed at monitoring and controlling communications systems represent conceptual design and associated infrastructure, which essentially resemble particular implementation of open systems.

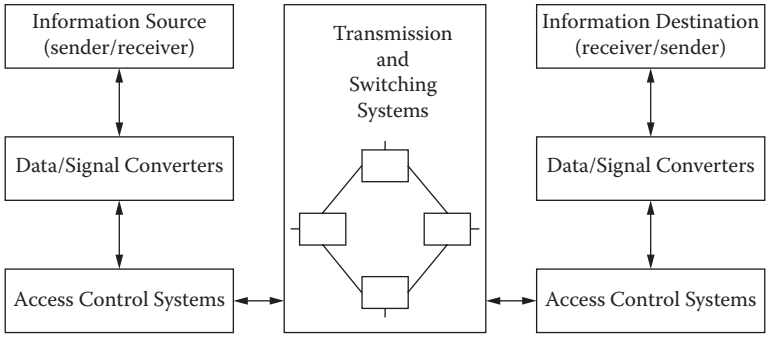
3.1.1.1 Communications General Model

A simplified view of any point-to-point communication assumes an information source (sending party) and the information destination (receiving party). The communication takes place over a transmission media, which can be a pair of copper wires, coaxial cable, fiber-optic, or a wireless media such as radio, microwave, satellite, or infrared rays (Figure 3.1.1).

The information source can be a telephone set, a computer, a TV pattern, a facsimile, or an instrumentation process. The information destination can be another telephone set or a computer, a TV set, a fax machine, or a control panel.

In order to be transmitted, the native information sources—voice, computer or instrumentation data, graphics, or video images—require successive data/signal conversions, according to adopted communications media and transmission technologies, and a rigorous security control of the access to shared

FIGu RE 3.1.1 Communication network conceptual model.



FIGu RE 3.1.2 Communication network conceptual model.

network resources. Therefore, new components such as data/signal converters and access control systems should be added to the communication model. This communication can be asymmetric, i.e., taking place only in one direction, or it can be symmetric, i.e., taking place bidirectionally (Figure 3.1.2).

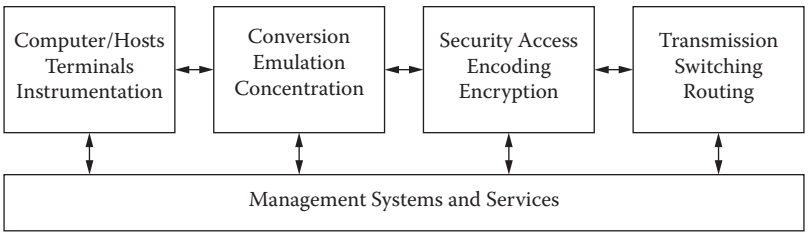
As a further consideration, the box representing the transmission media becomes more than a single conduit; a mixture of transmission/transport components and switching components in the form of circuits, links, nodes, routers, and switches participates in the design of a shared network environment.

3.1.1.2 Network Management General Model

The task of management, as derived from the general model of communications, is very clear: to be able to supervise, monitor, and control all the components that participate in the process of communications from the source to destination. That might include various computer hosts and terminals as sources/destinations of information, the devices performing data/signal conversions (protocol converters, emulators, concentrators, multiplexers), devices required to control the access to the network (security access, authorization, encoding, encryption), and all the components used in transmission, switching, and routing (Figure 3.1.3).

The task is not only clear but also quite challenging when the list of actual devices is spelled out. Many dozens of different technologies implemented in hundreds of different components, developed, designed, and manufactured by thousands of vendors, are all potential subjects of management systems, especially when it comes to the point of providing end-to-end, enterprisewide management services including monitoring, diagnostics, control, and reporting.

PCs, workstations, minicomputers, servers, mainframe computers, terminals, test equipment, phones, PBXs, TV sets, set-top boxes, cameras, modems, multiplexers, protocol converters, CSU/DSUs, statistical multiplexers, packet assembler dissemblers, ISDN adapters, NIC cards, codecs, data encoders, data compression devices, gateways, front-end processors, line trunks, repeaters, regenerators, matrix switches, DCS/DACs, bridges, routers, network access devices, accelerators, load balancers, wireless access points, sensors, and switches just begin a list of devices that should be or might be managed.



FIGu RE 3.1.3 Network management conceptual model.

The management picture is complete only if we consider, in addition to the management of network resources, the management of computing systems resources such as thousands of different businesses, users, systems applications, databases, and complex, specialized, large operations systems.

All the information collected and exchanged in conjunction with management operations is translated into management data, which is manipulated using techniques similar to those employed by data communications networks. However, substantial differences exist between data communication exchange and management information exchange to claim a specialized technical field, specialized communication protocols, information models, and specialized skills to design and operate management systems and interpret fault, performance, configuration, or security management information.

The following subsections will explain what is peculiar in management systems, the major requirements appended to management systems, and the management paradigms adopted in management.

3.1.2 Management Requirements

The diversity of managed resources, as found in traditionally distinct fields of communication such as voice, data, and video communications, generate different views on what should be the management functions and management requirements associated with management systems.

3.1.2.1 High-Level Management Functions

Regardless of the diverse management views, three high-level management functions top the list: monitoring, controlling, and reporting. Monitoring represents the continuous collection of management information about the status of management resources, delivered in the form of events and alarm notifications when the threshold attached to managed resource parameters is exceeded. Controlling is the targeted attempt of the manager or management application to change the status or configuration of selected managed resources. Reporting consists of delivering and displaying the management information in an accessible form for reading, viewing, searching, and ultimately interpreting the reported information.

In practice, several other functions are associated with management systems and management applications according to particular business needs such as provisioning, service activation, capacity planning, network/systems administration, inventory management, backup and recovery management, and management operations automation. Many of these complex functions include or are built on basic monitoring, controlling, and reporting.

3.1.2.2 High-Level Users Management Requirements

Based on the users' perspective on management, we can derive a set of high-level requirements associated with management, as listed below:

a. Ability to monitor and control end-to-end network and computing systems components.
b. Remote access and configuration of managed resources.
c. Ease of installation, operation, and maintenance of the management systems and their applications.
d. Secure management operations, user access, and secure transfer of management information.
e. Ability to report meaningful management-related information.
f. Real-time management and automation of routine management operations.
g. Flexibility regarding systems expansion and ability to accommodate various technologies.
h. Ability to back up and restore management information.

3.1.2.3 Driving Forces behind Management Technologies

Although the term *network management* gained a clear acceptance only in the mid-1980s with the advancement of IBM management tools (later incorporated into the IBM NetView family of management products), network management was equally driven by the development of telecommunications, data communications, and computing systems networking. For telecommunications and data communications, the

management technologies were concentrated on management of transmission and switching equipment (hardware devices, connections, circuits) along with conversion- and access-control devices. In the case of computing systems, the management technologies were concentrated on managing large computing system resources (hardware, interfaces, memory, data storage devices, etc.) and applications/databases.

With the convergence of telecommunications and computing systems, which embraces various technologies commonly known as *computer telephony integration* (voice over Internet is one of the most recent developments), the common point of these major fields becomes the network that connects these systems and the management of large data communications networks. This will be the dominant factor of the networks of the future.

3.1.2.4 Justifying Network Management Investment

It is well known that management systems are perceived as an overhead cost. However, the cost of not being able to prevent major network and systems problems or to quickly find and restore a system to normal functionality is even higher and can be crippling for many businesses relying on information exchange.

The following reasoning can be used in justifying the investment in network management. Some points can be quantified and used as a basis for a front-end analysis when selecting management systems.

a. Reducing downtime of critical components of networks and computing systems.
b. Controlling the corporate networks as strategic investment assets.
c. Controlling the performance, growth, and complexity of user applications.
d. Improving services in customer support and security of data transfer.
e. Controlling the cost of information technology deployment and operations.

3.1.3 Management Paradigms

Before analyzing the capabilities or the openness expected from management systems, we have to understand the fundamental paradigms used in management and the views associated with these paradigms.

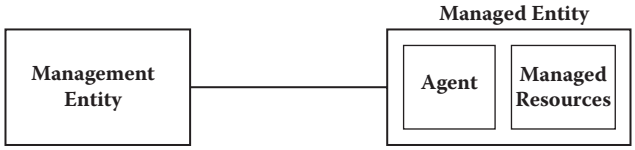
3.1.3.1 Management Basic Model

Conceptually, management systems are based on a simple model. In this model, management is the interaction/cooperation between two entities: the managing entity and the managed entity. The management entity represents a management system, a management platform, and/or a management application. The managed entity represents the managed resources. Looking at this simple model, it is important to note its similarity to the basic communication model presented at the beginning of this chapter (Figure 3.1.4).

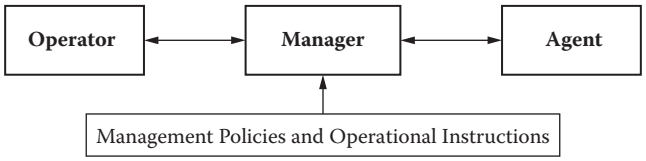
In order to communicate with the managed resources, which do not have any native mechanism to pass on management information, there is a need to create an intermediary component, the agent. The agent is also called the *management agent* or *managed agent*. The manager is the management entity, while the agent hides the interaction between the manager and the actual managed resources (Figure 3.1.5).

The manager–agent model is very common, used in describing the interaction between the management entity and the managed entity at a high level. This is the reason that all the paradigms natively created for management purposes closely follow the manager–agent model. In reality, the manager–agent model is more complex (Figure 3.1.6).

FIGu RE 3.1.4 Management basic model.



FIGu RE 3.1.5 Manager–agent model.



FIGu RE 3.1.6 Real manager–agent model.

The complexity becomes more evident when we consider the interactions between the manager or the management applications and the human operators. Other components, less visible but also very important because they shape the nature of interactions between managers and agents, are the management policies and the operational instructions given to the manager and implicitly to the operator.

There are other paradigms such as client–server and applications–object server that can be used for management information exchange. Natively, these paradigms have been conceived for building distributed applications or distributed object environments. Nevertheless, these general paradigms can be applied for management and there are products that use variations of these paradigms for management purposes.

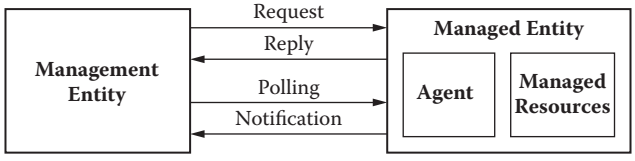
3.1.3.2 Management Views and Associated Models

Management assumes, as a primary function, the communication between the managing entity and managed entity. The management communication is based on the request–reply paradigm. The manager will request from the agent specific management information and the managed entity, through the agent, will reply with a message containing the information requested. If the request–reply communication is used continuously in order to reach each agent and the corresponding managed objects, the mechanism is called *polling* and it is primarily used in the management of Internet environments based on the Simple Network Management Protocol (SNMP) (Figure 3.1.7).

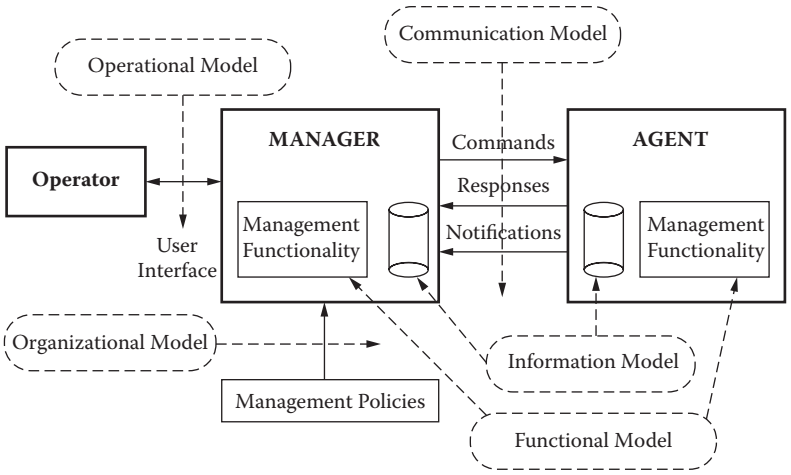
The request–reply mechanism is considered a synchronous communication mechanism, i.e., the manager expects an answer from the agent in a limited time frame before taking any action. If the reply is not received, a request for retransmission should be initiated by the manager.

There is an additional mechanism for communication between the manager and agents, called *notification*. The notification is an asynchronous mechanism initiated by the agent that communicates to the manager important changes in the status of managed resources that require either manager attention or intervention.

When building management systems, there are many aspects that should be taken into consideration. In addition to the communication model—responsible for exchanging commands, responses, and notifications between the manager and the agent—several other models are used in conjunction with the



FIGu RE 3.1.7 Manager–agent communication model.



FIGu RE 3.1.8 Manager–agent relationship models.

manager–agents relationship, as follows: architectural model, organizational model, functional model, and informational model (Figure 3.1.8).

The architectural model deals with the design and structure of the components participating in the management process, i.e., the manager or managers and the agents supplying management information according to the network topology. The manager can be designed as a management platform that consists of a management framework and a suite of management applications providing the actual management functionality such as configuration, fault, and performance management. More details will be provided in the sections dealing with management platforms.

The operational model deals with the operator's interface to the management system and specifies the nature and the type of interactions available to the user such as controlling managed objects, displaying and searching for specific events, dialog with the systems, and alerting the operator in case of critical alarms. Most of the operational specifications are included in the product's technical specifications such as user guide, administrative guide, etc.

The functional model refers to the structure of management functions performed by the management system through management applications. The functional model is considered a layered model where basic management functions such as configuration, fault, performance, security, and accounting management are the foundation of the functional model. Several other management functions such as trouble ticket administration, help desk, provisioning/service activation, and capacity planning consist of a combination of the basic management functions. At the pinnacle of the functional model, there are applications performing complex functions such as alarms/events correlation, expert systems, and management automation.

The organizational model is tightly linked to the overall management policies and operational procedures. This model specifies management domains, partition of management realm among the management operators, access of the user to the management systems, customer-based network management, interchangeability of the roles between managers and agents, and the overall cooperation between the manager and other managers or management applications.

The information model, although mentioned at the end of this list, is critical in handling all the management aspects. Given the variety of managed resources, in order to support their management in a common way, there is a need for an abstraction of managed resources in the form of a common information model, known by both manager and agents. The management information model establishes the basis for defining, naming, and registering the managed resources. Managed objects are considered abstractions of physical and logical managed resources. Therefore, the term *managed*

objects implies the use of an information model. Access to the managed resources is allowed only through the use of managed objects. The conceptual repository of management information is called the *management information base* (MIB). When we refer to a particular MIB, it means a collection of managed object definitions that describe a particular management domain or environment. The definition of managed objects is standardized, and on this basis a manager implementing a particular protocol and information model can communicate with distributed agents, which implement the same MIB.

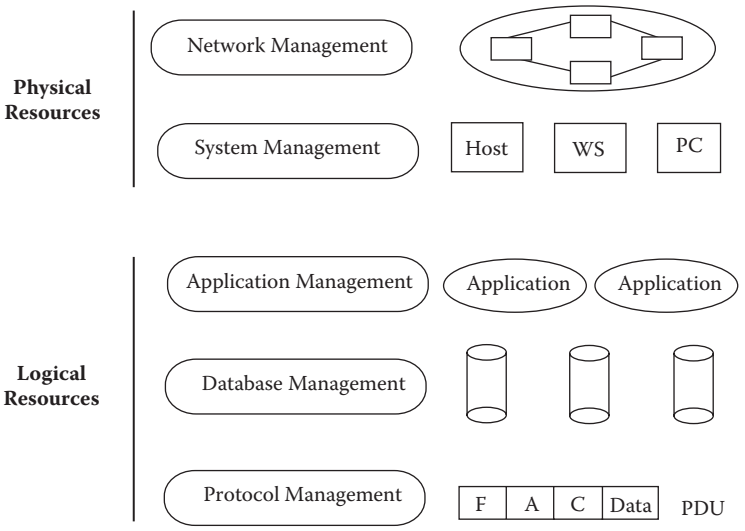
3.1.3.3 Management Domains

Historically, as we mentioned earlier, the notion of network management was launched by IBM. The IBM NetView products were in fact a combination of mainframe systems management and network management. Since then, the concept of management has evolved. At the beginning, the management products reflected the division, typical to most of the businesses, between network and computing systems management. With the advent of management platforms, the difference between network and systems management is blurring since the nature of the application and not the platform framework will determine the use of management systems.

Currently, it is commonly accepted that two major management domains can be considered when discussing the nature of managed resources: managing **physical resources** and **logical resources**. Physical resources are considered to be all the hardware components of converged (voice, data, and video) networks that participate in the process of exchanging information. This management domain is known as **network management**. The management of computing systems' physical resources such as processors, memory, input/output interfaces, and storage devices, are considered part of *systems management*.

The management of logical resources is built around **applications management** and **database management**, both associated with computing systems. Service management, user management, management of distributed transaction services, and data flow management are also considered systems management of logical resources (Figure 3.1.9).

There is a separate domain that deals with the management of specific logical resources, i.e., the protocols used in standards-based communications. Layered protocols, layered service primitives, and embedded management services are examples of protocol management. This type of management is applied to interfaces of particular technologies such as ATM, SONET, and WDM in the form



FIGu RE 3.1.9 Management domains classification.

of embedded channels or embedded **layer management entities** (LMEs). This type of management is conceptualized in the OSI Basic Reference Model, the foundation of standardized layered architecture and management.

3.1.4 Open Management Systems

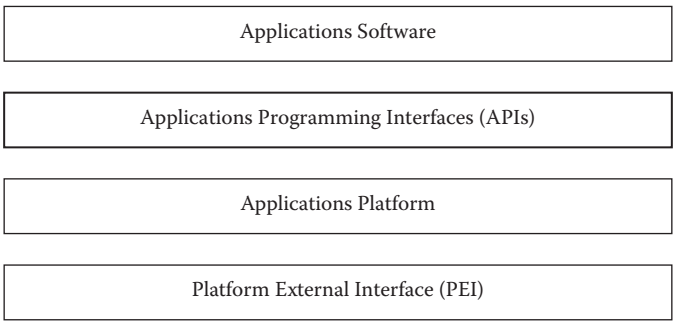
In order to evaluate the management systems, there is a need for a reference model. This reference model is the open system and its corresponding model, the open management system.

3.1.4.1 Open Systems Conceptual Model

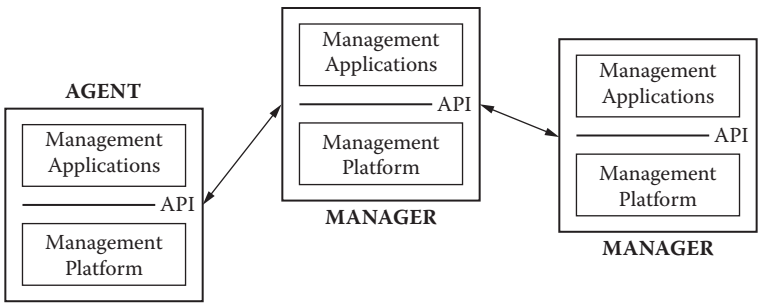
The open systems conceptual model assumes a design of systems modeled by the presence of four entities and by the relationship between these entities: **application platform, applications, application programming interface** (APIs), and **platform external interface** (PEI). This model can be applied to any computing system as part of the overall design and implementation. What makes any computing system (which runs software programs or applications) an open system is the separation of applications from the applications platform through APIs (Figure 3.1.10).

3.1.4.2 Open Management Systems Concept

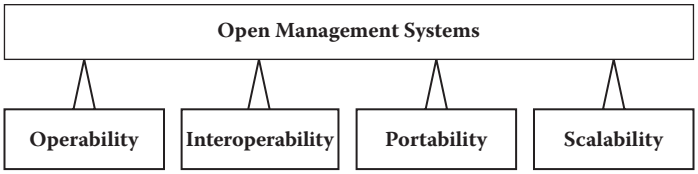
The open systems conceptual model can also be applied to management systems, i.e., to managers and managed agents. In this case, the applications will be specialized management applications providing fault, configuration, performance, security, and accounting management. The management platform is a management framework that consists of, in addition to the computing platform, specific management services such as event management services, communication services, graphical user interface services, or database services (Figure 3.1.11).



FIGu RE 3.1.10 Open systems conceptual model.



FIGu RE 3.1.11 Open management system.



FIGu RE 3.1.12 Open management system major requirements.

As mentioned earlier, key components to open systems are the APIs. In this case, the APIs are specific management APIs that allow the development of management applications by using specific management platform services. Last but not least, the management platforms are not isolated; they communicate with managed agents (as a minimum) or with other management systems, which may be modeled as open management systems. The platform external interface, in this case, will be an open standardized interface with well-defined management operations, services, and protocols.

3.1.4.3 Requirements for Open Management Systems

Four high-level requirements characterize open systems and open management systems: **operability**, **interoperability**, **portability**, and **scalabilty** (Figure 3.1.12).

Operability represents the ability of management systems to provide easy installation, operations, and maintenance, as well as adequate reliability and performance. Interoperability represents the ability of management platforms to transparently exchange management information with managed agents or peers' management systems. Portability expresses the ability of management platforms and/or management systems applications to be ported to a different environment (computing platform) with minimum changes or no changes. Scalabilty refers to the ability of management systems to be expanded in coverage, user domain, and management functions without the need to change the initial design.

3.1.5 Distributed Management Systems

Most of the computing systems, telecommunications, and data communications networks are distributed, i.e., interconnected by a communications network designed to transfer information and messages related to specific business needs. Management means managing various network and systems resources, which in most instances are physically separated. Therefore, by its very nature, the management is distributed.

A system is considered autonomous (it can be simple processor or multiple processor based) if the processes that constitute the system share the same memory. In contrast, distributed systems consist of interconnected autonomous systems with no shared memory. Since any networked computing environment is inherently a distributed system, the management of these systems is also inherently distributed.

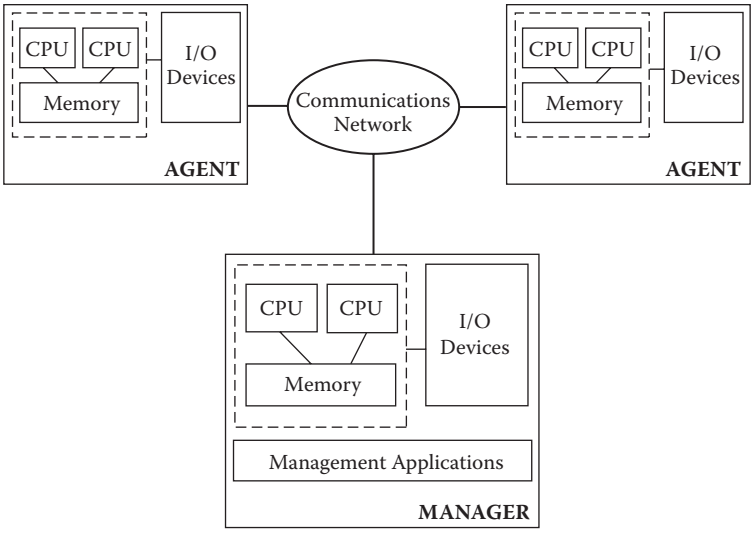
3.1.5.1 Distributed Network and Computing Systems

The true nature of management, as distributed or centralized, is determined not by the physical distribution of its components (managers and agents) but by the centralization and processing of management information (Figure 3.1.13).

If the system is designed to collect all the management information from all the agents (which constitute the management domain) in one point, we are dealing with a centralized type of management. If the collection of management information takes place in several interconnected processes and the information may be held in distributed databases, we are dealing with distributed management systems.

3.1.5.2 Distributed Management Systems Architectures

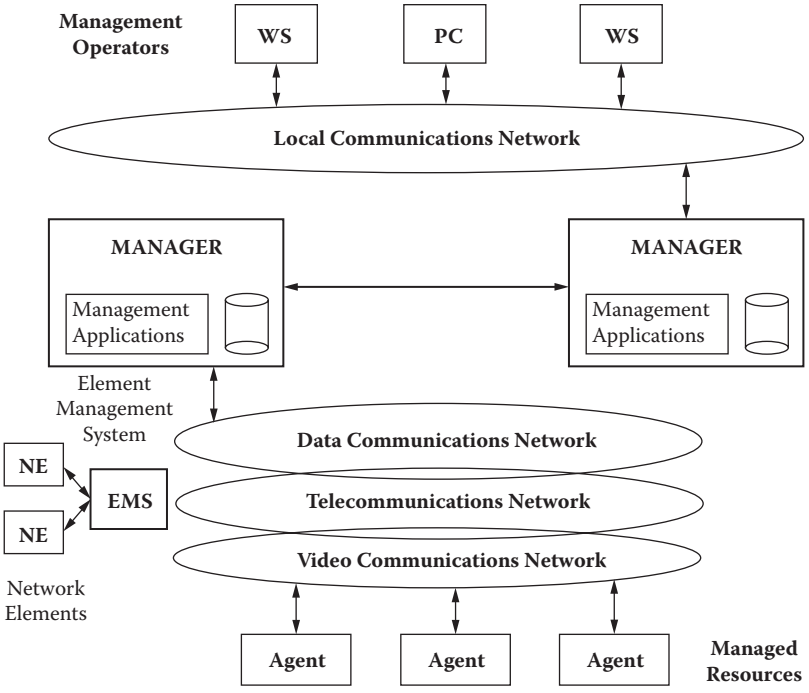
In a truly distributed management system, multiple management users or operators as management clients access the management server through a local or a wide area network. The actual manager runs



FIGu RE 3.1.13 Management of distributed systems.

the management applications and it is the holder of a MIB for a particular management domain. Each manager is responsible for the agents that are part of his/her domain.

The ability to exchange management information between servers (managers), keep in synchronization the shared MIB information, take over the management domain of a failed manager, and of the operators to interact with multiple managers, creates a truly distributed management system architecture (Figure 3.1.14).



FIGu RE 3.1.14 Distributed management systems architecture.

We have to emphasize that in all these examples we assume that the manager has a high degree of remotely accessing and configuring agents, with each agent acting as a management agent for a collection of managed objects and management processes.

3.1.5.3 In-Band and Out-of-Band Management Systems

All of the diagrams were presented in order to introduce the management concepts, and the properties of the management field included representation of interconnecting networks that carry management information. It is important to emphasize that these networks have rarely been designed as management-only infrastructures. Most of the management systems use for management information exchange the very network that carries the business-related data, voice, or video information. For the purpose of management, specialized protocols, operations, and application entities have been created and used. However, the management information is carried on the same physical infrastructure and on the same communication stack as the business information. In this case, we are dealing with the **in-band** type of management. This is a very cost-effective solution. However, there are some issues. By sharing the same service channels, the management information may take a significant chunk of the available bandwidth, and this may affect the overall performance of data exchange. That puts restrictions on how much and how often information is collected.

This is the reason that some management systems are built using **out-of-band channels**. The out-of-band management solutions may include unused bandwidth from a current channel allocation. A good example is the use of the low-band portion (50 Hz to 200 Hz) of the voice-grade channels as a dedicated data channel for management purposes. This solution is used for the management of the modems that share the same infrastructure with the voice communications. Other solutions consist of reserving a bit from the normal bit stream (for example, T1 multiplexer) to create a dedicated data channel for management purposes or assigning fields in each of the transmitted frames or cells for management purposes as it happens in the SONET and WDM technologies.

3.1.6 Management Systems Topological Frameworks

At a very high level, the architectural model of management systems is understood as the relationships among the main components of management systems, the managers, and agents. The accepted term for the architectural layout of the management network is called the *topology* and in most cases follows the business network infrastructure.

The topological view is the basis for the representation of the network and systems components using graphical user interfaces. An elaborate collection of graphical icons representing logical and physical resources, the links between these icons and the colors associated with the status of managed resources allow the management operators or the users to have a view of the management components along with their status.

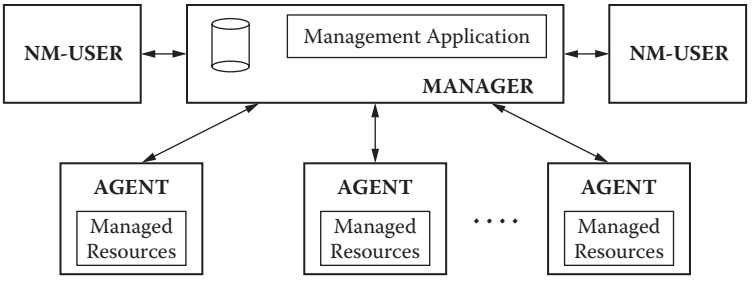
Three major topological frameworks are considered when designing management systems: **single manager, manager of managers,** and **network of managers**.

3.1.6.1 The Single Manager

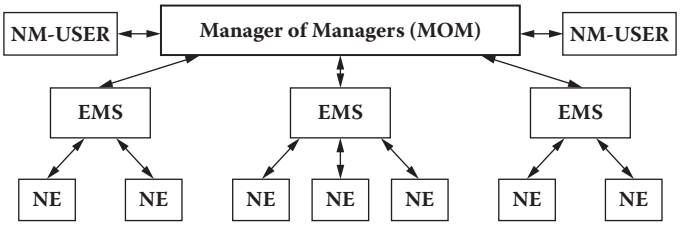
The single manager topology framework uses one management system that concentrates the collection and processing of management information from various managed resources such as routers, bridges, multiplexers, matrix switches, etc. Thus, the manager is the only point of exercising control over the network.

The system playing the role of the manager is usually a monolithic application that performs manage single manager topology is fully centralized. Historically, most of the management systems started as host-based, centralized systems. Today, they still represent the most common topological framework (Figure 3.1.15).

Regarding the single manager framework, we emphasize its weaknesses as follows: concentration of network management functions and applications in one point; limitation of the number of resources to



FIGu RE 3.1.15 The "single manager" topological framework.



FIGu RE 3.1.16 The "manager of managers" topological framework.

be managed (lack of scalability); and the high vulnerability of these systems when the manager fails. This topological framework is used for the management of small to medium-size networks and systems.

3.1.6.2 The Manager of Managers

The manager of managers (MOM) topology is a logically centralized framework with distributed control capabilities. The MOM acts as a single integration point for several distributed element management systems (EMSs) (Figure 3.1.16).

The actual management of managed resources/devices is provided by the EMSs that monitor and control a particular management domain, which may consist of a group of network components and associated applications. Usually, EMSs are designed to manage a family of similar products built around a particular technology. In other instances the management domain is determined by geographical, administrative, or jurisdictional considerations.

This topology is used for medium and large networks. Only vital, critical information such as alarms, security alerts, and capacity planning–related information is elevated to the level of MOM, which acts as a management integrator.

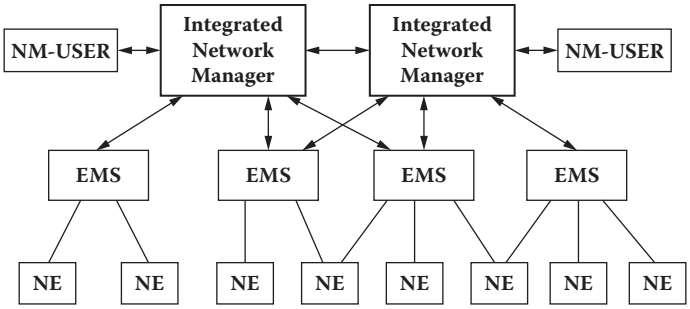
3.1.6.3 The Network of Managers

The network of managers topological framework provides fully distributed management based on cooperative management between integrated network managers (INMs).

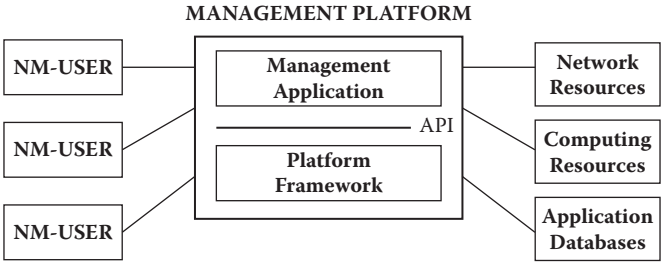
In this topological framework, management information can be exchanged between peer managers. Each INM is responsible for the management of its own grand domain. Cooperative links between INMs allow management information exchange. More than that, each INM can take over the management functions of an adjacent manager. Within each domain, the INM acts as the focal point of distributed management provided by several EMSs (Figure 3.1.17).

3.1.6.4 The Management Platforms

Management platforms do not represent a new topological framework; thus, they can be used in any of the topologies described in this section. The management platforms are designed as open management



FIGu RE 3.1.17 The "network of managers" topological framework.



FIGu RE 3.1.18 Management platform framework.

systems to allow the development and operation of portable distributed management applications. By employing, as part of the platform framework, advanced management services such as directory, security, and time services, in addition to basic communication, event management, graphical user interface, and database services, the management platforms can manage large, heterogeneous, multivendor, multitechnology, and multiprotocol environments.

Several management platform components, such as graphical user interfaces, management databases, and management applications, can be distributed among several computing platforms. Multiple management platforms can communicate to each other in order to manage large administrative domains (Figure 3.1.18).

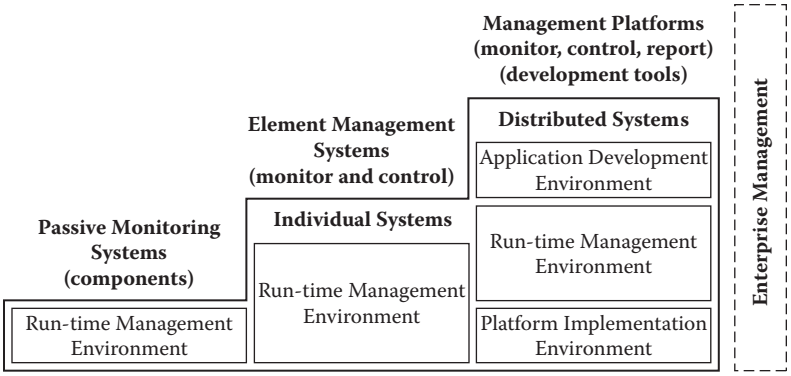
3.1.7 Management Systems Evolution

Although the management systems have been established as specialized systems to manage large and complex network and computing systems since the mid-1980s, distinct events and distinct phases can be identified in the technical and chronological evolution of management systems.

3.1.7.1 Management Systems Technical Evolution

The first phase in the management systems development is exemplified by **passive monitoring systems** targeted solely toward network components management and providing test, instrumentation, and protocol analysis results. This was characteristic for the management systems designed in the late 1970s and early 1980s.

The next major phase was the build-up of **element management systems** (EMSs), which provide monitoring and controlling capabilities of individual systems. Acting as stand-alone systems, the EMSs target the management families of network elements, equipment, and hosts. Generally, the EMSs contain a single management application bundled with the computing platform, forming the actual runtime operational management environment. These types of management systems were typical in the 1980s and they covered the management of modems, multiplexers, T1 multiplexers, matrix switches,



FIGu RE 3.1.19 Management systems evolution.

etc. In most cases, the management was limited to one type of equipment provided by a single vendor (Figure 3.1.19).

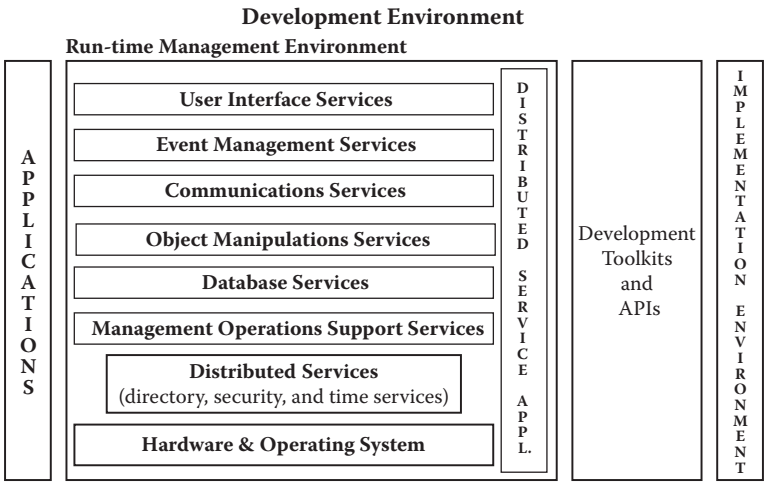
Management platforms are the third major generation in the history of management systems development that go beyond the run-time operational environment that characterized the earlier stages. By adding an application development environment with tools and APIs that allow multiple and portable applications to run on top of management platform framework, the management platforms have embraced many of the concepts of distributed open management systems.

The run-time environment is represented by common management services provided by the platform and reflects the overall operability aspects of a management platform. The development environment includes the run-time environment and provides portability for management applications and integration of these management applications with the platform services. In order to develop management applications that use platform management services, the development environment should include the run-time environment. In addition, the complexity of management platforms requires an implementation environment for testing and conformance to standards or vendor specifications. The implementation environment provides means to assess the management platforms' interoperability.

3.1.7.2 Management Platforms for Enterprisewide Management

In the previous sections we indicated the complexity and difficulties confronting the management field. We also indicated the shortcomings in the historical development of management systems. Management platforms do not solve these shortcomings overnight but they bring a flexible approach to the management of multivendor, multitechnology, multiprotocol networks and systems environments. This is why a close look at the design of management platforms is necessary.

Management platforms, either as autonomous or interworking management systems, should provide several basic management functions. First of all, a management platform has to communicate with the external world through platform external interfaces; that is, it has to provide communications services. Next, management information is exchanged in the form of management events that have to be stored, processed, and named. Therefore, there is a need for event management services. Furthermore, the events, as related to the management of network or computing systems resources, should be displayed (after the necessary processing) by using graphical user interfaces. Management information and all the components associated with management are organized using managed object models. Therefore, there is a need for a service that provides manipulation of managed objects. Ultimately, the management information about network configurations and managed resource status and parameters has to be stored in databases. Such database service may allow near real-time presentation of the status of the managed systems and components. Additional management operation services are needed to support all of the other platform services. In addition to these management services, there is a need for other



FIGu RE 3.1.20 Management platform architectural components.

distributed management services such as time service (synchronization), directory service (naming), and security services (Figure 3.1.20).

User interface services provide support for presentation of management information and support for interactions between users/human operators and distributed management applications used to manage network and computing systems. These services support both graphical user interfaces (GUIs) and asynchronous command line interfaces, by providing network and computing systems layout display based on visual icons, windows environment manipulation, online information, and general support for common applications development. The user interface is the most visible point of integration between various platform components and management applications.

Event management services provide common services to other platform management services and to the management applications running on top of management platforms. The events can be generated by network/computing systems, components state changes, systems errors, applications, and by users/operators. Common event operations include event collection, event logging, event filtering, and event administration.

Management communications services, either object-based or message-based, provide support for communications interfaces, management protocols, and communications stacks used to carry management information. Primarily, this support targets standardized management protocols such as Simple Network Management Protocol (SNMP), Common Management Information Protocol (CMIP) and Services (CMIS), and Remote Procedure Calls (RPCs).

Object manipulation services provide support for information exchange between objects as abstractions of physical and logical resources ranging from network devices and computing systems resources to applications and management services. Primarily, this support target object interfaces as defined by the OMG Common Object Request Broker Architecture (CORBA) and OMG Interface Definition Language (IDL). Object manipulation includes operations on the MIB, object support services providing location transparency for objects exchanging requests and responses, persistent storage for MIBs, and support for object-oriented applications development.

Database services provide support for management data storage and retrieval along with its integration with various platform services and management applications. Management information can include dynamic instance information related to configuration events, fault events, or performance events, to historical and archive information for security audit trail. The database services include database management systems (DBMSs), standardized database access and retrieval mechanisms such as structured query language (SQL), database concurrence mechanisms, and data backup mechanisms.

Management operations support provides common services to the management platform core and to the management applications running on top of the platforms. It includes management of the background processes associated with the platform hardware and operating system, and the handling of management applications.

Regarding management applications running on top of the management platforms, they can be classified in two major groups: *core management applications* and resource-specific applications. Core applications include functional management applications to manage specific management functional areas (SMFAs) such as configuration, fault, performance, security, and accounting management. Core applications also include compound applications, which provide cross-area functionalities, as required by user needs and business considerations (e.g., trouble ticketing). Resource-specific applications, as the name indicates, are built specifically to manage particular network devices or computing system components.

The management applications development environment mirrors the run-time environment and includes specific development tools for user interfaces, event management, communications, object manipulation, and database operations. The APIs for various platform management services are critical components in the development environment. The development environment includes the run-time environment since the newly developed management applications are tested against the run-time environment.

The management implementation environment mainly refers to the overall acceptance testing of management platforms as they are used in the management of real network and computing systems. Implementation testing is different from the testing done on various platform hardware and software components during their development and production. The implementation environment covers effective systems testing based on test criteria, test procedures, and test tools used to operate, maintain, and troubleshoot complex distributed systems, which are tailored to management platforms.

3.1.8 Summary and Trends

As we mentioned earlier, the management platform consists of management services and the actual management applications that run on top of management services. Both management services and management applications make use of the computing hardware and operating system. Although the design of the hardware and operating system is outside the scope of management platform design, it is important to understand the differences between various computing systems and to understand the alternatives in selecting hardware and operating systems. In a truly open system environment, the platform services and applications are supposed to be hardware and operating system independent.

3.2 Management of Emerging Technologies

Tivadar Szemethy

3.2.1 Introduction

Telecommunications services offered by domestic and international service providers are based on a mixture of emerged, emerging, and next-generation technologies. Emerged technologies include private leased lines on T/E-basis, ISDN, traditional voice networks, message switching, packet switching, and SS7-based signaling. Emerging technologies are the following: frame relay, asynchronous transfer mode (ATM), SDH/Sonet, digital subscriber line (DSL), cable, and Multiprotocol Label Switching (MPLS). Most of them will be addressed from the perspective of manageability. Each technology will be briefly introduced without details. Emphasis is on the availability of management information bases, management protocols, and management products.

In both emerged and emerging cases, management solutions are considered as mature. Simple Network Management Protocol (SNMP), Remote Network Monitoring (RMON), and Web-based management technologies have won over complicated CMIP-solutions. Solutions are built by service providers themselves or framework-based solutions from third parties have been implemented. The MIB and hardware/software agents play a very important role.

Next-generation wireless access technologies, WiFi and WiMax, will help to increase the mobility of enterprises and their employees. The technology is well proven, but implementation alternatives about affect the service quality and operational expenses. The move to IP as the common denominator is the future. All future services will be built on the basis of IP. Due to substantial quality improvements, all services will become acceptable not only to consumers, but also to enterprise customers.

This segment also addresses next-generation telecommunications application technologies, such as triple- and quad-play, Session Initiation Protocol (SIP)-based solutions, IPTV, and video on demand (VoD). Management is usually an afterthought for many service providers and suppliers. Most likely, strong future management solutions will be based on SIP.

Telco and enterprise Web 2.0 technologies will occasionally merge. Service providers will open their networking infrastructure for new resource-demanding applications provided by third-party content providers. But, being in the middle, service providers can earn considerable amounts by delivering and distributing content to customers and consumers.

3.2.2 Foundation Concepts for Networking Technologies

The majority of emerged and emerging technologies have a few basic foundation principles. These will be addressed in this section. The basics for this section can be found in more detail in BLAC94 and TERP98.

3.2.2.1 Connection-Oriented and Connectionless Communications

Communication systems that employ the concepts of circuits and virtual circuits are said to be connection oriented. Such systems maintain information about users, such as their addresses and their ongoing quality of service (QoS) needs. Often, these types of systems use state tables that contain rules governing the manner in which the user interacts with the network. These state tables, although clarifying the procedures between the user and the communication network, add overhead to the communication process.

In contrast, communication systems that do not employ circuits and virtual circuits are said to be *connectionless systems*. They are also known as *datagram networks* and are widely used throughout the industry. The principal difference between connection-oriented and connectionless systems is that connectionless protocols do not establish a virtual circuit for the end user communication process. Instead, traffic is presented to the service provider in a somewhat ad hoc fashion. Handshake arrangements and mutual confirmations are minimal and perhaps nonexistent. The network service points and the network switches maintain no ongoing knowledge about the traffic between the two end users. The state tables seen with connection-oriented solutions are not maintained. Therefore, datagram services provide no a priori or ongoing current knowledge of user traffic; however, they introduce less overhead.

3.2.2.2 Physical and Virtual Circuits

End users operating terminals, computers, and client equipment communicate with each other through a communication channel called the *physical circuit*. Physical circuits are also known by other names, such as *channels*, *links*, *lines*, and *trunks*. Physical circuits can be configured wherein two users communicate directly with each other through one circuit, and no one uses this circuit except these two users. They can operate the circuit, which is dedicated to users, in half duplex or full duplex. This concept is still widely used in simple networks without serious bandwidth limitations.

In more complex systems, such as networks, circuits are shared with more than one user pair. Within a network, physical circuits are terminated at intermediate points at machines that provide relay services on another circuit. These machines, which are known by such names as *switches*, *routers*, *bridges*,

and *gateways*, are responsible for relaying the traffic between the communicating users. Since many communication channels have the capacity to support more than one user session, the network device, such as the switch, router, or multiplexer, is responsible for sending and receiving multiple user traffic to or from a circuit.

In an ideal arrangement, a user is not aware that the physical circuits are being shared by other users. Indeed, the circuit provider attempts to make this shared operation transparent to all users. Moreover, in this ideal situation, the user perceives that the circuit directly connects only the two communicating parties. However, it is likely that the physical circuit is being shared by other users.

The term *virtual circuit* is used to describe a shared circuit wherein users are not aware of any sharing. The term was derived from computer architectures in which an end user perceives that a computer has more memory than actually exists. This other, additional virtual memory is actually stored on an external storage device. There are three types of virtual circuits, as follows.

- **Permanent virtual circuits (PVCs):** A virtual circuit may be provisioned to the user on a continuous basis. In this case, the user has the service of the network at any given time. A PVC is established by creating entries in tables in the network nodes' databases. These entries contain a unique identifier of the user payload, which is known by various names, such as a *logical channel number* (LCN), a *virtual channel identifier* (VCI), or a *virtual path identifier* (VPI). Network features such as throughput, delay, security, and performance indicators are also provisioned before the user initiates operations. If different types of services are desired, and if different destination endpoints must be reached, the user must submit a different PVC identifier with the appropriate user payload to the network. This PVC is provisioned to the different endpoint, perhaps with different services.
- **Switched virtual circuits (SVCs):** A switched virtual circuit is not preprovisioned. When a user wishes to obtain network services to communicate with another user, it must submit a connection request packet to the network. The address of the receiver must be provided, along with the virtual circuit number that is to be used during the session. SVCs entail some delay during the setup phase, but they are flexible in allowing the user to select dynamically the receiving party and the negotiation of networking parameters on a call-by-call basis.
- **Semipermanent virtual circuits (SPVCs):** With this approach, a user is preprovisioned, as in a regular PVC. Similar to the case with a PVC, the network node contains information about the communicating parties and the types of services desired. However, these types of virtual circuits do not guarantee that users will obtain their requested level of service. When networks are congested, users may be denied service. In a more likely scenario, continuation of a service is denied because the user has violated some rules of the communications. Examples are higher bandwidth demand and higher data rates than agreed on with the supplier.

3.2.2.3 Switching Technologies

Voice, video, and data signals are relayed in a network from one user to another through switches. This section provides an overview of prevalent switching technologies.

Circuit switching provides a direct connection between two networking components. Thus, the communicating partners can utilize the facility as they see it—within bandwidth and tariff limitations. Many telephone networks use circuit switching systems. Circuit switching provides clear channels; error checking, session establishment, frame flow control, frame formatting, selection of codes, and protocols are the responsibility of users. Today, the traffic between communicating parties is usually stored in fast queues in the switch and switched on to an appropriate output line with time division multiplexing (TDM) techniques. This technique is known as circuit emulation switching (CES). In summary, the attributes of CES are:

- Direct end-to-end connection
- No intermediate storage unless CES is used

- Few value-added functions
- Use of TDM to emulate circuit switching

Message switching, the dominant switching technology of the past two decades, is still widely used in certain applications (e.g., electronic mail) but is no longer employed in backbone networks. The switch is usually a specialized computer responsible for accepting traffic from attached terminals and computers. It examines the address in the header of the message and switches the traffic to the receiving station. As a result of the low number of switching computers, this technology suffers backup problems, performance bottlenecks, and lost messages due to congestion. In summary, the attributes of message switching are:

- Use of store-end-forward technology
- Disk serving as buffer
- Extensive value-added functions
- Star topology due to expense of switches

Packet switching relays small pieces of user information to destination nodes. Packet switching has become the prevalent switching technology of data communications networks. It is used in such diverse systems as private branch exchanges (PBXs), LANs, and even with multiplexers. Each packet occupies a transmission line only for the duration of the transmission; the lines are usually fully shared with other applications. This is an ideal technology for bursty traffic. Modern packet switching systems are designed to carry continuous, high-volume traffic as well as asynchronous, low-volume traffic, and each user is given an adequate bandwidth to meet service-level expectations.

The concepts of packet and cell switching are similar; each attempts to process the traffic in memory as quickly as possible. However, cell switching uses much smaller protocol data units (PDUs) than packet switching. PDU size is fixed with cell switching and may vary with packet switching. In summary, the attributes of packet and cell switching are:

- Hold-and-forward technology
- RAM serving as buffer
- Extensive value-added functions for packet but not many for cells

Switching will remain one of the dominant technologies in the telecommunications industry.

3.2.2.4 Routing Technologies

There are two techniques to route traffic within and between networks: source routing and non-source routing. The majority of emerging technologies use non-source routing.

Source routing derives its name from the fact that the transmitting device—the source—dictates the route of the PDU through a network or networks. The source places the addresses of the *hops* in the PDU. The hops are actually routers representing the internetworking units. Such an approach means that the internetworking units need not perform address maintenance; rather, they simply use an address in the PDU to determine the destination of the PDU.

In contrast, non-source routing requires that the interconnecting devices make decisions about the route. They do not rely on the PDU to contain information about the route. Non-source routing is usually associated with bridges and is quite prevalent in LANs. Most of the emerging new technologies implement this approach with the use of a VCI. This label is used by the network nodes to determine where to route traffic.

The manner in which a network stores its routing information varies. Typically, routing information is stored in a software table called a directory. This table contains a list of destination nodes. These destination nodes are identifiers with some type of network address. Along with the network address (or some type of label, such as a virtual circuit identifier), there is an entry describing how the router is to relay the traffic. In most implementations, this entry simply lists the next node that is to receive the traffic in order to relay it to its destination.

Small networks typically provide a full routing table at each routing node. In the case of large networks, full directories require too many entries and are expensive to maintain. In addition, exchange of routing table information can affect the available bandwidth for user payload. These networks are usually subdivided into areas called *domains*. Directories of routing information are kept separately in domains.

Broadcast networks contain no routing directories. Their approach is to send the traffic to all destinations.

Network routing control is usually categorized as *centralized* or *distributed*. Centralized routing control uses a network control center to determine routing of packets. Packet switches are limited in their functions. Centralized control is vulnerable; a backup is absolutely necessary, which increases operating expenses. Distributed routing control requires more intelligent switches, but they provide a more resilient solution. Each router makes its own routing decisions without regard to a centralized control center. Distributed routing is also more complex, but its advantages over the centralized approach have made it the preferred routing method in most communications networks.

3.2.2.5 Multiplexing Technologies

Most emerged and emerging technologies use some form of multiplexing. Multiplexers accept low-speed voice or data signals from terminals, telephones, PCs, and user applications and combine them into one higher-speed stream for transmission efficiency. A receiving multiplexer demultiplexes and converts the combined stream into the original lower-speed signals. There are various multiplexing techniques:

- **Frequency division multiplexing (FDM):** This approach divides the transmission frequency range into channels. The channels are lower frequency bands; each is capable of carrying communication traffic such as voice, data, or video. FDM is widely used in telephone systems, radio systems, and cable television applications. It is also used in microwave and satellite carrier systems. FDM decreases the total bandwidth available to each user, but even the narrower bandwidth is usually sufficient for users' applications. Isolating the bands from each other costs some bandwidth, but the simultaneous use outweighs this disadvantage.
- **Time division multiplexing (TDM):** This approach provides the full bandwidth to the user or application but divides the channel into time slots. Each user or application is given a slot, and slots are rotated among the attached devices. The TDM multiplexer cyclically scans the input signals from the entry points. TDMs work digitally. The slots are preassigned to users and applications. When there is no traffic at the entry points, the slots remain empty. This approach works well for constant bit rate applications but leads to wasted capacity for variable bit rate applications.
- **Statistical time division multiplexing (STDM):** This approach allocates time slots to each port on a s**tatistical time division multiplexer**. Consequently, idle terminal time does not waste the capacity of the bandwidth. It is not unusual for two to five times as much traffic to be accommodated on lines using s**tatistical time division multiplexers** in comparison to a TDM solution. This approach can accommodate bursty traffic very well but does not perform well with continuous, nonbursty traffic.
- **Wave division multiplexing (WDM):** WDM is the optical equivalent of FDM. Lasers operating at different frequencies are used in the same fiber, thereby deriving multiple communications channels from one physical path. There is a more advanced form of this technology—dense wave division multiplexing (DWDM)—with even better efficiency.

3.2.2.6 Address and Identification Schemes

In order for user traffic to be sent to the proper destination, it must be associated with a destination identifier. There are two typical techniques in use.

An explicit address has a location associated with it. It may not refer to a specific geographical location but rather a name of a network or a device attached to a network. For example, the Internet Protocol (IP) address has a structure that permits the identification of a network, a subnetwork attached to the

network, and a host device attached to the subnetwork. The ITU-T X.121 address has a structure that identifies the country, a network within that country, and a device within the network. Other entries are used with these addresses to identify protocols and applications running on the networks. Explicit addresses are used by switches, routers, and bridges as an entry into routing tables. These routing tables contain information about how to route the traffic to the destination nodes.

Another identifying scheme is known as a *label*, although other terms may be more widely used. Those terms are *logical channel number* (LCN) and *virtual circuit identifier* (VCI). A label contains no information about network identifiers or physical locations. It is simply a value that is assigned to identify each data unit of a user's traffic.

Almost all connectionless systems use explicit addresses, and the destination and source addresses must be provided with every PDU in order for it to be routed to the proper destination.

3.2.2.7 Control and Congestion Management

It is very important in communication networks to control the traffic at the ingress and egress points of the network. The operation by which user traffic is controlled by the network is called *flow control*. Flow control should ensure that traffic does not saturate the network or exceed the network's capacity. Thus, flow control is used to manage congestion. There are three flow control alternatives with emerged and emerging technologies, as follows:

- *Explicit flow control:* This technique limits how much user traffic can enter the network. If the network issues an explicit flow control message to the user, the user has no choice but to stop sending traffic or to reduce traffic. Traffic can be sent again after the network has notified the user about the release of the limitations.
- *Implicit flow control:* This technique does not absolutely restrict flows. Rather, it recommends that users reduce or stop the traffic they are sending to the network if network capacity situations require limitations. Typically, the implicit flow control message is a warning to the user that the user is violating its service-level agreement with the internal or external supplier regarding network congestion. In any case, if the user continues to send traffic, it risks having traffic discarded by the network.
- *No flow control:* Flow control may also be established by not controlling the flow at all. Generally, an absence of flow control means that the network can discard any traffic that is creating problems. While this approach certainly provides superior congestion management from the standpoint of the network, it may not meet the performance expectations of users.

3.2.3 Management Solutions for Emerged Technologies

The present status of emerged technologies, such as private leased lines, voice networks, SS7-based signaling techniques, and message and packet switching, can be summarized as follows:

- *Domination of proprietary solutions:* This means that the management protocols selected are controlled by the supplier of the equipment or facilities vendors.
- *SNMP reengineering:* Many equipment vendors include SNMP agents in their devices to meet the requirements of customers. These agents provide information on performance management and reporting, but they usually do not change the real-time processing of status data within the devices.
- *Heterogeneous management structures:* In most cases, these structures are hierarchical and include a manager of managers. This manager uses a proprietary architecture. Most of the interfaces to element managers and managed devices are proprietary.
- *Low penetration of the telecommunications management network (TMN):* Suppliers have recognized the need for a generic standard, but they are not willing to invest heavily in supporting it. Some providers go as far as supporting the Q3 interface.

- *Heavy operating support systems:* The legacy-type operations support systems (OSSs) support emerged technologies on behalf of suppliers well, but they aren't flexible enough to address future needs. They are lacking with respect to separating operations functionality from operations data, using flexible software, and separating network management from service management.

3.2.4 Discussion of Select Technologies

This section highlights a few emerged, mainstream technologies and discusses their management aspects. These technologies include frame relay, Fiber Distributed Data Interface (FDDI), Switched Multimegabit Data Service (SMDS), ATM, SONET/SDH, and mobile and wireless communications. The capabilities of these technologies are presented, including technology descriptions and evaluations of management capabilities.

3.2.4.1 Asynchronous Transfer Mode

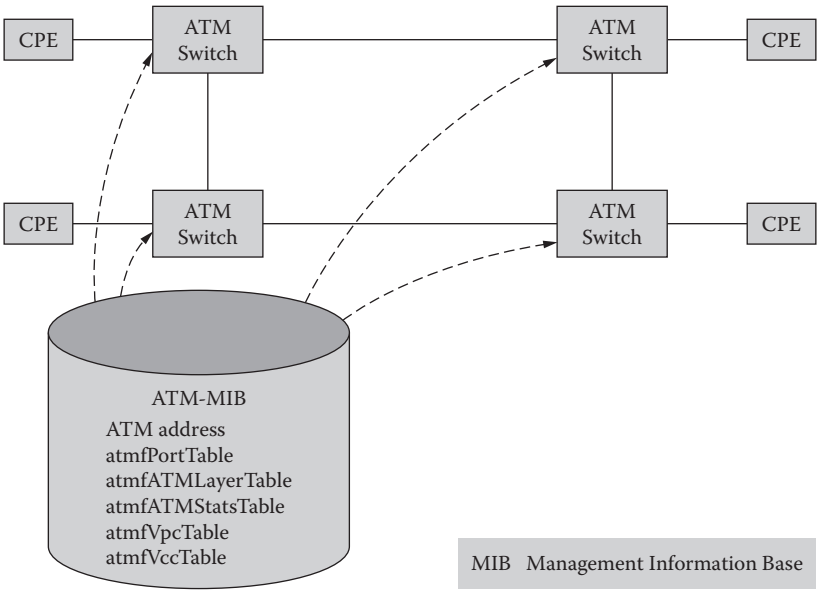
The purpose of ATM is to provide a high-speed, low-delay multiplexing and switching network to support any type of user traffic, including voice, data, and video applications. ATM is one of four fast relay services. ATM segments and multiplexes user traffic into small, fixed-length units called *cells*. The cell is 53 octets, with 5 octets reserved for the cell header. Each cell is identified with virtual circuit identifiers that are contained in the cell header. An ATM network uses these identifiers to relay traffic through high-speed switches from the sending customer premises equipment (CPE) to the receiving CPE.

ATM provides limited error detection operations. It provides no retransmission services, and few operations are performed on the small header. The intention of this approach—small cells with minimal services performed—is to implement a network that is fast enough to support multimegabit transfer rates.

ATM is a new technology designed to be the foundation for providing convergence, multiplexing, and switching operations. It resides on top of the physical layer.

3.2.4.1.1 *MIB Availability*

The ATM Forum has published an MIB as part of its Interim Local Management Interface Specification (ILMI; Figure 3.2.1). The ATM MIB is registered under the enterprise node of the standard Structure of



FIGu RE 3.2.1 ATM management information base.

Management Information (SMI) in accordance with the Internet. Management Information Base (MIB) objects are therefore prefixed with 1.3.6.1.4.1.353.

Each physical link (port) at the UNI has an MIB entry defined in the atmfPortTable. This table contains a unique value for each port, an address, the type of port (DS3, Sonet, etc.), media type (coaxial cable, fiber, etc.), status of port (in service, out of service, etc.), and other miscellaneous objects.

The atmfAtmLayerTable contains information about the User–Network Interface (UNI)'s physical interface. The table contains the port ID, maximum number of VCCs, VPCs supported and configured on the UNI, active VCI/VPI bits on the UNI, and a description of the public or private aspects of the UNI.

The atmVpcTable and atmVccTable contain similar entries for the VPCs and VCCs, respectively, on the UNI. These tables contain the port ID, VPI or VCI values for each connection, operational status (up, down, etc.), traffic shaping and policing descriptors (to describe the type of traffic management applicable to the traffic), and any QoS that is applicable to the VPI or VCI. The ATM Forum has defined two aspects of UNI network management: ATM layer management at the M plane and ILMI specifications.

- **M-plane management:** Most ATM M-plane management functions are performed with the SONET F1, F2, and F3 information flows. ATM is concerned with F3 and F4 information flows.
- **Il MI:** Because International Telecommunications Union-Standardization Section (ITU-T) and the American National Standards Institute (ANSI) have focused on C-plane and U-plane procedures, the ATM Forum has published an interim specification called ILMI. The major aspects of ILMI are the use of SNMP and an MIB. The ILMI stipulates the following procedures. Initially, each ATM device supports the ILMI and one ILMI MIB instance for each UNI. The ILMI communication protocol stack can be SNMP/UDP/IP/AAL over a VPI/VCI value. SNMP is employed to monitor ATM traffic and the UNI VCC/VPC connections based on the ATM MIB with the SNMP get, Get-Next, Set, and Trap operations.

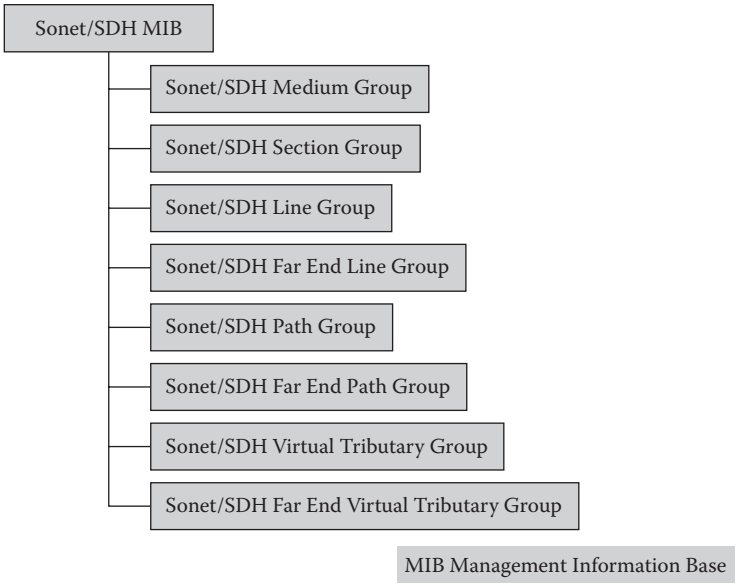
3.2.4.2 SONET and SDH

SONET/SDH is an optical-based carrier (transport) network utilizing synchronous operations between the network components. The term *SONET* is used in North America, and SDH is used in Europe and Japan. Attributes of this technology are:

- A transport technology that provides high availability with self-healing topologies
- A multivendor that allows connections without conversions between the vendors' systems
- A network that uses synchronous operations with powerful multiplexing and demultiplexing capabilities
- A system that provides extensive Operations, Administration Maintenance, and Provisioning (OAM&P) services to the network user and administrator

SONET/SDH provides a number of attractive features relative to other current technology. First, it is an integrated network standard on which all types of traffic can be transported. Second, the SONET/SDH standard is based on optical fiber technology, which provides superior performance in comparison to microwave and cable systems. Third, because SONET/SDH is a worldwide standard, it is now possible for different vendors to interface their equipment without conversion.

Fourth, SONET/SDH efficiently combines, consolidates, and segregates traffic from different locations through one facility. This concept, known as *grooming*, eliminates backhauling and other inefficient techniques currently being used in carrier networks. *Backhauling* is a technique in which user payload is carried past a switch that has a line to the user and sent to another endpoint. Then the traffic to the other user is dropped and the first user's payload is sent back to the switch and relayed back to the first user. In present configurations, grooming is eliminated, but expensive configurations, such as back-to-back multiplexers, which are connected with cables, panels, or electronic cross-connect equipment, are required. Fifth, SONET/SDH eliminates back-to-back multiplexing overhead by using new techniques in the grooming process. These techniques are implemented in a new form of equipment called an *add-drop multiplexer* (ADM).



FIGu RE 3.2.2 SONET/SDH management information base.

Sixth, the synchronous aspect of SONET/SDH results in more stable network operations. These types of networks experience fewer errors than the older asynchronous networks and provide better techniques for multiplexing and grooming payloads. Seventh, SONET/SDH has notably improved OAM&P features relative to current technology. Approximately 5% of bandwidth is devoted to management and maintenance. Finally, SONET/SDH employs digital transmission schemes. Thus, the traffic is relatively immune to noise and other impairments on the communications channel and the system can use efficient TDM operations.

SONET has been around a couple of years. The technology is not completely new, but its implementation is new.

3.2.4.2.1 MIB Availability

The SONET/SDH MIB consists of eight groups. Each of the groups described below includes two tables: the Current Table and the Interval Table (Figure 3.2.2).

The SONET/SDH XXX Current Table contains various statistics collected for the current 15-minute interval. The SONET/SDH XXX Interval Table contains various statistics collected by each system over a maximum of the previous 24 hours of operation. The past 24 hours can be broken into 96 completed 15-minute intervals. A system is required to store at least four completed 15-minute intervals. The default value is 32 intervals.

- *SONET/SDH Medium Group:* SONET/SDH interfaces for some applications may be electrical interfaces rather than optical interfaces. This group handles the configuration information for both optical SONET/SDH interfaces and electrical SONET/SDH interfaces (e.g., signal type, line coding, line type).
- *SONET/SDH Section Group:* This group consists of the SONET/SDH Section Current Table and the SONET/SDH Section Interval Table. These tables contain information on interface status, counters on errored seconds, severely errored seconds, severely errored framing seconds, and coding violations.
- *SONET/SDH Line Group:* This group also consists of the SONET/SDH Line Current Table and the SONET/SDH Line Interval Table. These tables contain information on line status, counters on errored seconds, severely errored seconds, severely errored framing seconds, and unavailable seconds.

- *SONET/SDH Far End Line Group:* This group can be implemented only in SONET/SDH systems that provide far end block error (FEBE) information at the SONET/SDH Line Layer. It consists of the SONET/SDH Far End Line Table and the SONET/SDH Far End Line Interval Table.
- *SONET/SDH Path Group:* This group consists of the SONET/SDH Path Current Table and the SONET/SDH Path Interval Table. These tables contain information on interface status, counters on errored seconds, severely errored seconds, severely errored framing seconds, and coding violations.
- *SONET/SDH Far End Path Group:* This group consists of the SONET/SDH Far End Path Current Table and the SONET/SDH Far End Path Interval Table.
- *SONET/SDH Virtual Tributary Group:* This group consists of the SONET/SDH VT Current Table and the SONET/SDH VT Interval Table.

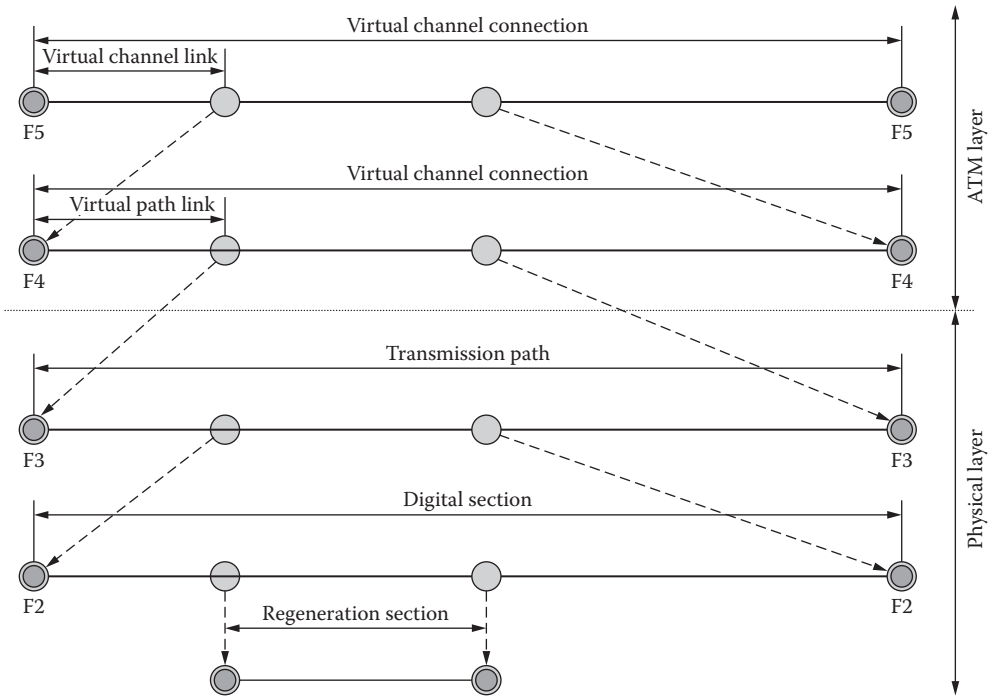
For Synchronous Digital Hierarchy (SDH) signals, virtual tributaries are called virtual circuits (VCs) instead of virtual tributaries (VTs).

VT1.5	VC11
VT2	VC12
VT3	none
VT6	VC3

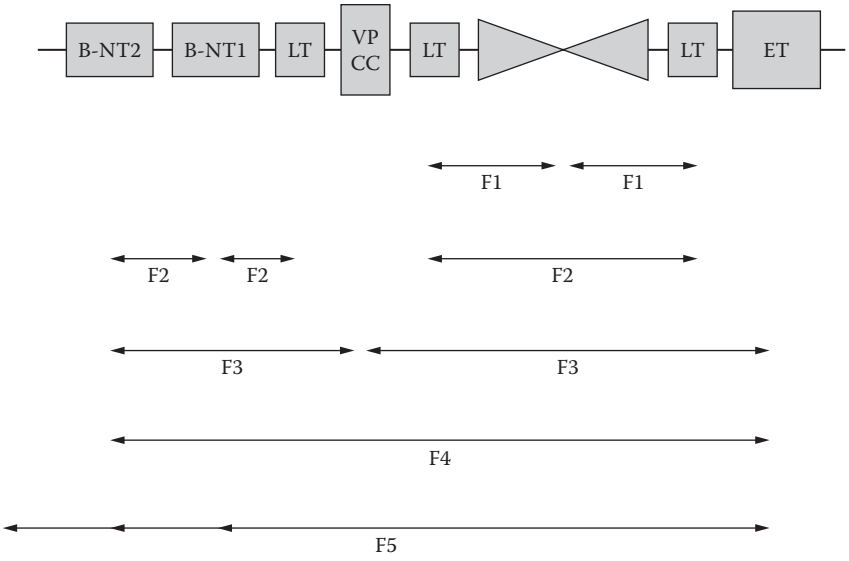
These tables contain information on virtual tributaries width and status, counters on errored seconds, severely errored seconds, severely errored framing seconds, and unavailable seconds.

- *SONET/SDH Far End VT Group:* This group consists of the SONET/SDH Far End VT Current Table and the SONET/SDH Far End VT Interval Table.

The operation, administration, and maintenance (OAM) functions are associated with the hierarchical, layered design of SONET/SDH. Figure 3.2.3 shows the five levels of the corresponding OAM operations, which are labeled F1, F2, F3, F4, and F5. The F1, F2, and F3 functions reside at the physical



FIGu RE 3.2.3 Relationships in ATM layers.



FIGu RE 3.2.4 Information flows in BISDN.

layer; F4 and F5 functions reside at the ATM layer. The Fn tags depict where the OAM information flows between two points, as shown in Figure 3.2.4. The five OAM flows are as follows.

- *F5:* OAM information flows between network elements performing VC functions. From the perspective of a BISDN configuration, F5 OAM operations are conducted between B-NT2/B-NT1 endpoints. F5 deals with degraded VC performance, such as late arriving cells, lost cells, and cell insertion problems.
- *F4:* OAM information flows between network elements performing virtual path (VP) functions. From the perspective of a BISDN configuration, F4 OAM flows between B-NT2 and ET. F4 OAM reports on an unavailable path or a VP that cannot be guaranteed.
- *F3:* OAM information flows between elements that perform the assembling and disassembling of payload, header, and control (HEC) operations and cell delineation. From the perspective of a BISDN configuration, F3 OAM flows between B-NT2 and VP cross connect and ET.
- *F2:* OAM information flows between elements that terminate section endpoints. It detects and reports on loss of frame synchronization and degraded error performance. From the perspective of a BISDN configuration, F2 OAM flows between B-NT2, B-NT1, and LT, as well as from LT to LT.
- *F1:* OAM information flows between regenerator sections. It detects and reports on loss of frame and degraded error performance. From the perspective of a Broadband ISDN (BISDN), F1 OAM flows between LT and regenerators.

3.2.4.3 Digital Subscriber Line Technologies

In this case the enabling technology is the digital subscriber line (xDSL), a scheme that allows the mixing of data, voice, and video over telephone lines. There are, however, different types of DSL to choose from, each suited for different applications. All DSL technologies run on existing copper phone lines and use special and sophisticated modulation to increase transmission rates (ABER97).

The asymmetric digital subscriber line (ADSL), the most publicized of the DSL schemes, is commonly used as a transport for linking branch offices and telecommuters in need of high-speed intranet and Internet access. The word *asymmetric* refers to the fact that it allows more bandwidth downstream (to the consumer) than upstream (from the consumer). Downstream, ADSL supports speeds of 1.5 to 8 Mbit/s, depending on line quality, distance, and wire gauge. Upstream rates range between 16 and

640 Kbit/s, again depending on line quality, distance, and wire gauge. For up to 18,000 feet, ADSL can move data at T1 using standard 24-gauge wire. At distances of 12,000 feet or less, the maximum speed is 8 Mbit/s.

ADSL delivers a couple of other principal benefits. First, ADSL equipment installed at carriers' central offices offloads overburdened voice switches by moving data traffic off the public switched telephone network and onto data networks, a critical problem resulting from Internet use. Second, the power for ADSL is sent by the carrier over the copper wire, with the result that the line works even when local power fails. This is an advantage over ISDN, which requires a local power supply and thus a separate phone line for comparable service guarantees. The third benefit is again over ISDN; ADSL furnishes three information channels—two for data and one for voice. Thus, data performance is not affected by voice calls. Rollout plans are very aggressive with this service. It is expected to be widely available by the end of this decade.

The rate-adaptive digital subscriber line (RADSL) has the same transmission limits as ADSL. However, as its name suggests, it adjusts transmission speed according to the length and quality of the local line. Connection speed is established when the line synchs up or is set by a signal from the central office. RADSL devices poll the line before transmitting; standards bodies are in the process of determining whether products will constrain line speed. RADSL applications are the same as with ADSL and include Internet, intranets, video on demand, database access, remote LAN access, and lifeline phone services.

High-bit-rate digital subscriber line (HDSL) technology is symmetric, meaning that it furnishes the same amount of bandwidth both upstream and downstream. The most mature of the xDSL approaches, HDSL has already been implemented in telco feeder plants—the lines that extend from the central office to remote nodes—as well as campus environments. Because of its speed—T1 over two twisted pairs of wiring and E1 over three—telcos commonly deploy HDSL as an alternative to T1/E1 with repeaters. At 15,000 feet, the HDSL operating distance is shorter than that of ADSL, but carriers can install signal repeaters to extend the useful range (typically by 3,000 to 4,000 feet). Its reliance on two or three wire pairs makes it ideal for connecting PBXs, interexchange carrier points of presence (POPs), Internet servers, and campus networks. In addition, carriers are beginning to offer HDSL to carry digital traffic in the local loop between two telco central offices and customer premises. HDSL's symmetry makes this an attractive option for high-bandwidth services such as multimedia, but availability is still very limited.

The single-line digital subscriber line (SDSL) is essentially the same as HDSL with two exceptions: It uses a single wire pair, and it has a maximum operating range of 10,000 feet. Since it is symmetric and needs only one twisted pair, SDSL is suitable for applications such as video teleconferencing or collaborative computing with identical downstream and upstream speeds. Standards for SDSL are still under development.

The very-high-bit-rate digital subscriber line (VDSL) is the fastest DSL technology. It delivers downstream rates of 13 to 52 Mbit/s and upstream rates of 1.5 to 2.3 Mbit/s over a single wire pair. However, the maximum operating distance is only 1,000 to 4,000 feet. In addition to supporting the same applications as ADSL, VDSL, with its extended bandwidth, could potentially enable carriers to deliver high-definition television (HDTV). VDSL is still in the developmental stage.

A number of critical issues must be resolved before DSL technologies achieve widespread commercial deployment. Standards are still under development. Modulation techniques such as carrierless amplitude phase (CAP) and discrete multitone (DMT) have been separated by standards bodies. Some other problems include interoperability and security issues, elimination of interference with ham radio signals, and lowering of power system requirements from the present 8 to 12 watts down to 2 to 3 watts. A nontechnical but important factor will be how well carriers can translate the successes they have realized in their xDSL technology trials to market trials and then to commercial deployments.

3.2.4.3.1 MIB Availability

MIB definitions are in the works with standard committees and with vendors. Present definitions are not complete but can be used as guidelines. The MIB is currently structured into eight groups:

- The *xdlsDevIfStats* group provides statistics specific to the xDSL link. Statistics are collected on a per-port basis at specific intervals. Hence, such a table is indexed by the xdslDevifStatsIfIndex and xdslDevIfStatsInterval. Also, statistics are grouped into remote and central statistics. *Remote* means that the statistics are collected by the device at the customer premises. *Central* means that the statistics are collected by the device located at the central office. The objects that are not divided into these two groups are related to both ends of the xDSL link.
- The *xdslDevIfConfig* group provides configuration information specific to an xDSL device or system. The table is indexed by an object that corresponds to ifIndex. These ifIndex entries themselves denote and identify specific xDSL interfaces on the board or module. Also, the configuration parameters are grouped into two broad categories, up (the upstream direction, from the customer premises to the central office) and down (the downstream direction, from the central office to the customer premises).
- The description of the *xdslRemoteSys* group is identical to that of the mib-2 system MIB.
- The *xdslRemoteDTEStatus* group provides status information about the data terminal equipment (DTE) ports of DSL remote terminal units (RTUs).
- The *xdslDevMvlIfConfig* group provides configuration information specific to an xDSL (MVL) device or system. The table is indexed by an object that corresponds to ifIndex. These ifIndex entries, themselves, denote and identify specific xDSL (MVL) interfaces on the board or module.
- The *xdslDevNAPCustomerAccount* group provides customer accounting information on each DSL port. Network access providers can accurately bill their end-station DSL customers according to amount of usage. The table is indexed by ifIndex and xdslDevNAPCustomerAccountInterval; ifIndex identifies the specific xDSL interface on the device, and xdslDevNAPCustomerAccountInterval specifies the accounting information for the current day or the previous day. Customer data excludes all traffic used for management purposes.
- The *xdslLinkUpDownInformation* group indicates the reason why the DSL link went down. This information is obtained when the link comes up.
- The *xdslRemoteInjection* group identifies the processes at the remote site, indicating injection types and various traps.

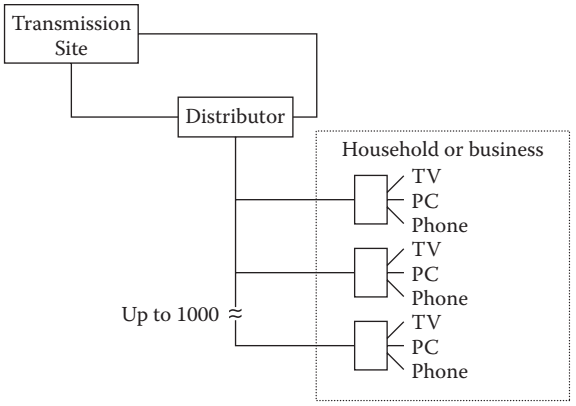
Management is more difficult than with other technologies because more information is processed in real time or near real time. In most cases, proprietary solutions dominate. Implementation of TMN and other standards is extremely slow.

3.2.4.4 Cable Technnology

Cable service providers can enter the competition for voice and data services. Depending on the country in question, there are millions of households and businesses with cable television connections. In the majority of cases, cable television is a distribution channel supporting only one-way communication. Through the use of cable modems, however, channels can be provided for two-way communications, allowing consumers to send back data or use the cable for telephone conversations.

Figure 3.2.5 shows the structure of such an arrangement. The structure is simple, requiring just a few additional networking devices. The components are as follows.

- The equipment at the transmission site generates television signals and houses switches that route phone calls and monitor the network.
- Parallel fiber-optic lines spread out over the area. If a line is out, traffic moves almost instantly to another route.
- Electronic nodes convert signals for transmission via coaxial cables. One node can serve approximately 500 to 1,000 homes.
- Coaxial cable runs into a box on the side of the home. Electronic equipment in the box splits signals, sending telephone traffic over regular phone lines and television signals over cable lines to the television set.



FIGu RE 3.2.5 Use of cable technology in subscriber environments.

- Phones work on the current wiring. Television can be interactive, allowing signals to flow back up the cable line. Personal computers also can be connected through cable modems, which transmit information up to 1,000 times faster than phone lines.

Transforming cable television systems into local telephone networks will not be easy or cheap. Plans are ambitious to compete with local service providers. A basic cable system starts out as copper coax that carries signals in a line from the headend—where TV signals are generated—to each consumer. Signals go one way; in the case of cable cuts or other damages, service to all consumers from that point is interrupted.

A system designed to handle phone calls and interactive TV looks much different. The main trunks of the network are set up in interconnecting rings of fiber-optic lines, which can carry thousands of times more information than copper lines. If a line is cut, traffic moves to another ring in a microsecond, and practically no one loses service. Because phone calls are two way, a phone-cable system must be two way and must contain the sophisticated switches necessary to send calls to the right places.

About 40% of today's cable TV systems are still copper. The other 60% have fiber in their trunk lines, but most of them need upgrades. Only a few are set up in rings and have switches.

Power could cause another problem. Telephone lines have just enough electricity running through them to power the phone and the network, independent of the power grid. If a storm knocks out the main electrical grid, most of the time the phone still works. Cable lines do not carry power. If electricity goes out, so does cable. So would a phone hooked to an unpowered cable network. Cable lines, however, are capable of carrying power. But adding power from a node over the coaxial cable could add noise. There is debate today in the industry about whether power should be obtained from the cable-phone network or whether another solution should be used, such as attaching backup batteries to the sides of houses.

Once the pieces of the network come together, cable companies will face other issues. One issue is number portability—allowing consumers to keep their phone numbers even if they change phone service from the local phone company to the cable company. At present, if customers change carriers, their number must be changed. State or federal regulations are expected to challenge this situation. Cable-phone systems also will have to prove to a skeptical public that they can be as reliable as current phone systems, which almost never break down. In addition, cable systems must overcome their reputation for overly frequent service interruptions.

From a technology standpoint, the competition in the high-speed data services market will initially be between a dedicated architecture and a shared architecture. Both ISDN and other dedicated solutions, such as xDSLs and cable-LAN services, require infrastructure upgrades, and that means significant investments on the part of service providers. While most performance comparisons focus on peak bandwidth, other aspects of network usage, including customer density and average session time, can also affect cost and quality of service.

The type of upgrade that is needed to deliver high-speed data over cable networks should be fairly obvious: Cable television transmission is typically one way, but a data and phone network must permit two-way traffic. Limited forms of data service are possible if the telephone line is used as a return route, but this is not a long-term solution. Some cable networks already have migrated to a hybrid fiber-coax infrastructure, but many operators are still immersed in the upgrading process. Once the networks are tailored for two-way traffic, broadband LAN technology will be incorporated so that digital data can be transmitted over a separate channel. Most of the vendors that supply technology to cable operators will use an Ethernet-like approach in which the consumer's computer will have to be fitted with a network interface card and a cable modem for accessing the cable LAN. Access speed will depend upon the peak rate of the cable modem and the volume of traffic on the cable LAN. Subscribers will likely experience connection speeds that vary according to such factors as usage.

Cable LANs will operate full duplex with two channels, each channel sending data in a different direction. For customers using Internet or intranet applications, one of these channels would connect to an Internet or intranet POP router, which would then forward data packets from all users on the neighborhood LAN to and from all other systems on the Internet or on intranets. The other channel would receive data from the Internet or intranet service. Cable LANs are unlike ISDN-based services in that they furnish full-time connections rather than switched or dial-up links. The obvious advantage is that the need for dedicated transmission and switching resources would be eliminated. Users can access the cable LAN when needed, and only the cable network makes use of the Internet or intranet point of presence.

A shared cable LAN requires only a single connection to the Internet or intranet provider. With ISDN, the lines of many individually served subscribers have to be multiplexed and concentrated, and the number of subscribers online at any given time is limited by the number of connections between the provider and the telecommunications network.

In terms of geographical coverage, cable LANs can be quite large. The original intention was that they would be implemented for residential broadband services. The future target, however, is corporate internetworking.

Currently, there are no standards governing transmission of data over cable LANs, but this has not slowed equipment suppliers. Suppliers are bringing proprietary products to the market as rapidly as possible.

The cable modem, which is used to connect the consumer's PC with a coax drop linked to the two-way, broadband cable LAN, is specifically designed for high-speed data communications. When the modem is in operation, cable television programming is still available. But the return path that the cable modem uses is limited and must be shared among all digital services, including interactive television, video on demand, telephony, videoconferencing, and data services. One problem with cable modems is the immaturity of existing devices. The majority of manufacturers are targeting peak bandwidth between 128 Kbps and 30 Mbps.

Competing with cable technology are, first of all, ISDN and emerging technologies such as ADSL, RADSL, HDSL, SDSL, and VDSL, each of which uses wire already in place with corporations and residential customers. Because ISDN delivers both voice and data, it makes use of circuit-switching and packet-switching technology. For high-speed data services, the local telephone plant must be upgraded so that two-way digital transmission—over the existing copper pairs that ISDN also relies upon—is possible. In addition, ISDN-capable digital switches must be installed at the telecommunications central office. ISDN may offer faster data transmission than analog modems, but the dial-up access model is the same. From the data transmission perspective, this is an inefficient approach because circuit switching requires the service provider to dedicate resources to a customer at all times. In other words, it is not consistent with the bursty nature of data services. The result is wasted bandwidth.

Implementing an ISDN network could end up being more costly than deploying cable LANs. In telecommunications networks, for example, an assessment of the incremental costs of providing ISDN Internet or intranet access includes the cost of initializing a subscriber's ISDN circuit at the central office; the cost of multiplexers, concentrators, and terminal servers; and possibly the cost of T1 lines from the telco to the Internet or intranet provider.

Cable-LAN access might prove to be less expensive for data and online services. It is assumed that cable television providers are facing less financial risk than telecommunications providers when deploying data services. Lower costs translate into consumer savings. The cost per bit of peak bandwidth (ROGE95) in providing Internet or intranet access is significantly lower for hybrid fiber-coax networks than it is for ISDN—about 60 cents for a 4-Mbps residential service as opposed to $16 for a 128-Kbps service. Shared fiber-coax networks also compete well against dedicated ISDNs with respect to average bandwidth and peak bandwidth. A 4-Mbps residential cable service, for example, can provide the same average bandwidth and about 32 times the peak bandwidth of a 128-Kbps ISDN service for 40% less money.

While deployment costs for both technologies continue to decline, ISDN deployment still runs several hundred dollars more per subscriber than cable-LAN deployment. This difference, in addition to the higher performance of cable LANs, will be an important factor as Internet, intranet, and online service access becomes a commodity product. Yet the cable-LAN approach and the cable services industry have their shortcomings as well. Cable-LAN modems and access products are still proprietary, so once an operator has selected a vendor, it is likely to remain locked in. Also, shared networks function properly only when subscribers' usage habits are well known. The more subscribers deviate from an assumed usage profile, the more performance is likely to deteriorate. Large changes in usage could affect the cost of providing acceptable service levels over neighborhood cable LANs.

Finally, the cable industry itself is greatly fragmented in many countries. This inhibits coordinated efforts, which are vital to the rapid deployment of services. Although telecommunications providers are in competition against each other, they will work better as a group if they act quickly in deploying the resources needed to make ISDN universally available.

In order to avoid these bottlenecks and other fault- and performance-related problems, cable service providers need a powerful management solution. This requires a complete rethinking of presently used solutions. The management solution should address high volumes of events, alarms, and messages; fail-safe operations; traffic control and management; service quality assurance; and collection of accounting data.

3.2.4.4.1 MIB Availability

Again, the MIB definitions in the works with standard committees and vendors can be used as guidelines. At present, the MIB is structured into six groups:

- The *docsDevBase* group extends the MIB-II "system" group with objects needed for cable device system management. It includes the device role, date and time, serial number, and reset conditions.
- The *docsDevNmAccess* group provides a minimum level of SNMP access security to the device by network management stations. Access is also constrained by community strings and any vendor-specific security. The management station should have read-write permission for cable modems.
- The *docsDevSoftware* group provides information for network downloadable software upgrades. It includes file identification, administration status, operational status, and current version.
- The *docsDevServer* group provides information about the progress of the interaction between the cable modem (CM) and cable modem termination system (CMTS) and various provisioning servers. It includes boot state, Dynamic Host Configuration Protocol (DHCP), server time, configuration file, and Trivial File Transfer Protocol (TFTP) configuration parameters.
- The *docsDevEvent* group provides control and logging for event reporting. This group offers very detailed information on control parameters, syslog details, throttle conditions, severity codes, and priorities. It also offers entries for vendor-specific data.
- The *docsDevFilter* group configures filters at the link layer and IP layer for bridged data traffic. This group consists of a link-layer filter table, docsDevFilterLLCTable, which is used to manage the processing and forwarding of non-IP traffic; an IP packet classifier table, docsDevFilterIpTable, which is used to map classes of packets to specific policy actions; and a policy table, docsDevFilterPolicyTable, which maps zero or more policy action tables. At this time, the MIB specifies only one policy action table, docsDevFilterTosTable, which allows manipulation of the types of service

bits in an IP packet based on matching certain criteria. The working group may include additional policy types and action tables in the future, for example, to allow QoS to modem service identifier assignment based on destination.

Management is more difficult than with other technologies as a result of the lack of management capabilities in managed devices. In most cases, proprietary solutions dominate. Implementation of TMN and other standards is extremely slow.

3.2.4.5 Multiprotocol Label Switching

Multiprotocol label switching (MPLS) is a packet-forwarding (switching) technology that attaches labels to the packets traveling through the network and makes routing decisions based on these labels. MPLS labels are *prepended* to the packets, and the label value (ID) is placed at the very beginning of an MPLS packet. This enables MPLS devices to use wire-speed packet switching technology (where label/destination lookup is performed in hardware) instead of the CPU-intensive destination address-based routing table lookup performed by IP routers.

Advances in router technology have reduced the speed advantage of MPLS, but its unique flexibility and low-level QoS-handling characteristics have made MPLS the dominant technology in provider networks. MPLS also offers resilience in the form of MPLS local protection, also known as MPLS Fast ReRoute, which is much faster than the case with pure IP routing protocols. According to the Open Systems Interconnected (OSI) classification, MPLS operates between Layers 2 (data link) and 3 (network); therefore, MPLS is often referred to as a Layer 2.5 protocol, and it provides the best of both worlds. It is in close interaction with data link characteristics (unlike a traditional IP routing protocol), thus providing good QoS; at the same time, MPLS is aware of and supports higher-level network conditions and topology.

MPLS uses IP addresses (either v4 or v6) to identify endpoints and, as a result, is naturally IP compatible. For this reason, MPLS networks are most often used to carry IP traffic, although MPLS can be used to carry many native frame formats such as Ethernet, ATM, or SONET as well.

In hierarchical MPLS, multiple MPLS labels can be prefixed to the same payload packet; thus, the MPLS header is referred to as the *label stack*. An MPLS label stack entry consists of the following fields:

- 20-bit label ID
- 3-bit QoS value (traffic class)
- 1-bit "bottom of stack" flag indicating the last entry on the label stack
- 8-bit TTL (time to live) field

Network entry and exit (ingress and egress in MPLS parlance) routers are called LER (label edge router) devices. These either push (prepend) or pop (remove) one or more entries to/from the label stack.

Using labels, MPLS networks form Label Switched Paths (LSPs) within the network. Ingress routers use the destination address of the payload packet to determine which LSP to use within the network and write the appropriate value into the label stack entry. Inside the network, Label Switched Routers (LSRs) distribute packets just by their top label information, which enables a very high performance and dependable core infrastructure.

MPLS employs two protocols for establishing network paths (LSPs). A standard IP routing protocol (such as Open Shortest Path First [OSPF] or Intermedtiat System-to-Intermediate System [IS-IS]) is used to distribute network topology information. The fact that MPLS networks are typically within a single administrative domain enables the use of simpler interior gateway protocols. In addition, MPLS defines traffic engineering (TE) extensions that enable the distribution of QoS and Shared Risk Link Group (SRLG) information so that routers can establish LSPs with guaranteed QoS and define independent backup links. The presence of preset backup links allows MPLS local protection to provide much faster recovery times than the protocols used in pure IP networks. The RSVP-TE signaling protocol is used to inform the switches along a given LSP about the links and labels to use for the LSP (path).

3.2.4.5.1 MPLS Management

The IETF MPLS Working Group is responsible for specifying the MIB modules covering base MPLS functionality. Thus far RFC 3812 (Traffic Engineering Link MIB) and RFC 3813 (MPLS LSR MIB) have been defined, and major vendors (especially Cisco) do support these MIBs. Because of the quickly evolving nature of existing implementations, it is also common for vendor-specific MIBs (e.g., Juniper) to be provided and supported.

3.2.4.6 Management Solutions

The present status of the technologies discussed can be summarized as follows.

- *Domination of proprietary solutions:* This means that the management protocols selected are controlled by the supplier of the equipment or facilities vendors. Working groups do not exist with the exception of the ATM Forum.
- *SNMP reengineering:* Many equipment vendors include SNMP agents in their devices to meet the requirements of customers. These SNMP agents provide information for performance management and reporting, but they usually do not change the real-time processing of status data within the devices. Most equipment is extremely intelligent. SNMP is simply too slow for certain decisions.
- *Heterogeneous management structures:* In most cases, domains are managed. Domain boundaries may be defined by the geography or by the family of managed objects. Most of the interfaces to element managers and managed devices are still proprietary. However, they will be defined clearly by TMN.
- *Low penetration of TMN:* Suppliers have recognized the need for a generic standard, but they are not willing to invest heavily in supporting it. Some providers support the Q3 interface. There are no implementation examples for the higher layers of TMN. The network element and network management layers are well understood.
- *Definition of MIBs:* Each technology requires the definition of public and private MIBs. It is merely a question of time until each of the emerging technologies uses MIB as the basis of management.
- *OSSs in transition:* The legacy-type OSSs will give way to new OSSs based on separating operations data from operating functionality, implementing very flexible software, and separating network management from service management.

3.2.5 Next-Generation Wireless Access Technologies

This section discusses next-generation wireless access technologies. After an introduction to generic cellular and wireless concepts, a few emerging next-generation technologies are discussed.

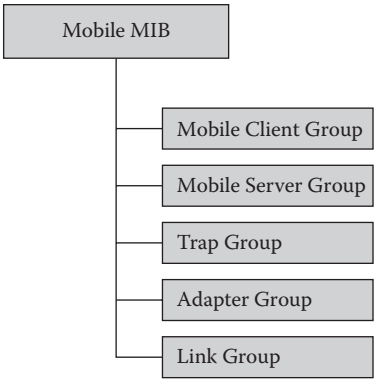
3.2.5.1 Mobile and Wireless Communication

The purpose of a mobile communications system is the provision of telecommunications services between mobile stations and fixed land locations or between two mobile stations. There are two forms of mobile communications: cellular and cordless.

The best approach is to examine the major attributes of these two alternatives and compare them to each other. First, a cellular system usually has a completely defined network that includes protocols for setting up and clearing calls and tracking mobile units through a wide geographical area. Thus, in a sense, it defines a UNI and an Network-to-Network Interface (NNI). With cordless personal communications, the focus is on access methods in comparison to a closely located transceiver, usually within a building. That is, the focus is on a rather geographically limited UNI.

Cellular systems operate with higher power than cordless personal communications systems. Therefore, cellular systems can communicate within large cells with a radius in the kilometer range. In contrast, cordless cellular communication cells are quite small, usually in the order of 100 meters.

Cellular will continue to be a preferred medium for the consumer and business market. Personal Communications Service (PCS) will emerge as a driver for wireless. Telecommunications providers will



FIGu RE 3.2.6 Mobile management information base.

use PCS to help reduce cellular churn, more closely joining customers with vendors by providing end-to-end service. Wireless data connectivity, driven by lighter and smaller equipment capable of being carried by humans as well as in vehicles, will drive wireless connectivity requirements deeper and further into the network infrastructure.

The technology is not new, but implementation has occurred over an extended period of time. Cordless systems are undergoing rapid technological changes. Different protocols and standards are in use.

3.2.5.1.1 MIB Availability

The mobile MIB can be summarized as follows:

- Network managers require more control over remote network workstations.
- To manage these remote machines, network managers require data on remote network access.
- Some of this data can be supplied via SNMP from the remote workstation.
- Other important information is available only on the local network at the point where the connection was created.
- This requires, therefore, two important SNMP agents with associated mobile MIB components, one at the remote workstation and the other at the local network connection point.

The following sources of information have been created (Figure 3.2.6):

- The *Mobile Client Group* contains information to be relayed from the remote workstation to a network management system on the attached network. This group stores information on mobile client name, description, location, phone number, power management configuration, connection hardware and software, client type (CPU, RAM, disk, video), system software, date, time, network adapter used, and configuration and statistics.
- The *Mobile Server Group* contains information to be relayed from the local network server to a network management system on the same network. This group stores information about the network server where remote connections can be originated, such as server name, remote network hardware, slot and port number, server uptime, connection speed, service type, and traffic statistics.
- The *Trap Group* describes SNMP trap types for remote workstations, such as mobile computer docked, undocked, suspended, resumed, PCMCIA inserted, and a trap table. This is a table of alerts that can be sent to the specified IP address via the specified protocol by setting the value of the mobileTrapTrigger object to the index of an entry in the table.
- The *Adapter Group* contains information about network adapters, including hardware information and type of connection.
- The *Link Group* provides data about mobile network links such as link status and link performance.

Management is more difficult than with wirelines because more information is processed in real time or near real time. In most cases, proprietary solutions dominate. Implementation of TMN and other standards is extremely slow.

3.2.5.2 Wi-Fi

Wi-Fi is the marketing name of a set of wireless access protocols based on the IEEE 802.11 standards. Although Wi-Fi and 802.11 standards mostly overlap, there are slight differences resulting from the fact that standardization lags behind industrial progress in this quickly advancing field of technology.

Wi-Fi is intended primarily for computer data communication in environments where cabling is not practical. The lower access speed offered by Wi-Fi is offset by wireless flexibility and range.

The 802.11 standards refer to a connection-oriented technology. Subscriber stations (SS, or endpoints) are allowed to communicate only after establishing a session with the Base Station (BS, or network infrastructure). Base Stations are also called access points (APs) or, in marketing terms, *hotspots*.

The standards define three aspects of the interface: the physical layer (PHY), the media access control (MAC) layer, and since the 802.16e amendment, the quality of service (QoS) classification aspect.

- *PHY:* Wi-Fi uses multiple channels and adaptive modulation and coding, and it supports features intended to provide good NLOS (non-line-of-sight) characteristics. Error correction is also provided by the physical layer.
- *MAC:* MAC 802.16 defines a number of convergence sublayers, which specify how to implement higher-level protocols such as Ethernet, ATM, and IP. MAC also defines protocols for secure communication relying on key exchange (Wired Equivalent Privacy [WEP], Wi-Fi Protected Access [WPA]) and encryption algorithms (e.g., Data Encryption Standard [DES]; Advanced Encryption Standard [AES]). Power-saving features such as sleep mode and idle mode are also defined in this layer.
- *QoS:* This protocol provides five QoS classes, which, in order of decreasing real-time guarantees, are UGS (Unsolicited Grant Service; real-time steam of fixed-size packets), (e)rtPS (Extended Real-Time Polling Service; two classes for real-time steams of variable-length packets), nrtPS (Non-Real-Time Polling Service; delay-tolerant streams of variable-length packets with guaranteed bandwidth), and BE (Best Effort; no guarantee provided on the network's behalf).

3.2.5.2.1 Advantages and Disadvantages of Wi-Fi

An advantage of Wi-Fi is that no cabling or extensive physical deployment is necessary. Also, minimal user training is required, and end users are typically capable of installing and activating their own devices (near-zero configuration). Wi-Fi is cheap, widely available, and secure.

In terms of disadvantages, performance is often unpredictable as a result of interference and signal path blocking issues. Also, network performance degrades rapidly with distance. Multiple, co-located Wi-Fi networks compete for the same frequency, which leads to interference problems. Older installations may use weaker encryption, which is a security hazard. Power consumption is high compared to other portable technologies (such as Bluetooth), resulting in battery life issues.

In some countries, there can be legal problems associated with operating Wi-Fi transmitters. Because of differences in regulations, different frequency bands are available in different regions of the world (United States, Europe, Japan).

3.2.5.2.2 Management

Enterprise Wi-Fi equipment vendors usually offer proprietary management solutions as well. A unified MIB, IEEE802dot11-MIB, is available, but only devices from major vendors (such as Cisco) support it. Other vendors may implement it partially, and others do not support it at all. The standard MIB is also slightly outdated, especially regarding newer Wi-Fi features such as improved authentication and encryption protocols.

3.2.5.3 WiMAX

WiMAX (Worldwide Interoperability for Microwave Access) is a cover term for a large and diverse family of emerging digital long distance wireless protocols. WiMAX is based on the IEEE 802.16 standard family, established to define a protocol stack for Wireless Metropolitan Area Networks (WirelessMAN).

WiMAX is comparable to Wi-Fi but offers a much wider range. It is also comparable to 3G Universal Mobile Telecommunication Service (UMTS), the packet switched radio service offered by modern cellular networks (WiMAX is even superior in most technical aspects). The main difference is that while, in practice, 3G is implemented exclusively by Groupe Speciale Mobile (GSM) mobile operators (who invested heavily in UMTS licensing), WiMAX is a more open market with lower entry costs. Although WiMAX is also licensed, licenses can be acquired at a much lower cost than for UMTS. Also, WiMAX services are expected to operate more like extended-range Wi-Fi hotspot services rather than existing 3G services, which are affordable only for SIM card holders for local providers. Also, WiMAX is expected to be a viable and affordable technology for fixed residential network access, replacing xDSL in some areas.

From the infrastructure point of view, WiMAX is similar to 3G. The systems are built up identically, that is, with base transceiver stations (BTSs), antenna systems, and affordable and easy-to-use network interface devices at the customer end. It is expected that WiMAX chipsets will soon be integrated into laptops as PDAs, as is the case with Wi-Fi today. Base Station Controllers (BSCs) are optional, as most BTSs readily connect to IP networks.

As of 2007, different vendors' WiMAX components were not interoperable, and management systems for WiMAX were proprietary systems developed by the vendors of the WiMAX device family. These are basically element management systems that allow for the controlling, monitoring, and provisioning of WiMAX services and for upgrading software on the BTSs operated by the customer.

3.2.5.3.1 MIB Availability

The IEEE 802.16 Network Management Task Group has defined MIBs for the management of 802.16 networks (IEEE 802.16 MIBs), and these MIBs are included in the upcoming version of the protocol definition proposal IEEE P802.16Rev2. Device vendor support for these MIBs is generally available. IEEE Draft 802.16f is an official amendment to the standard specifying MIBs for management of the 802.16 MAC and PHY layers.

3.2.6 Next-Generation Telecommunications Application Technologies

3.2.6.1 The Move to IP

During the last decade, new telecommunications services have proliferated in parallel with a gradual shift in customers' demands. Two fundamental technology factors have facilitated these trends:

- *Mobile networks:* The increasing convenience and decreasing costs of mobile communications attract more and more customers. In most countries of the developed world, the number of mobile phones has exceeded the number of fixed-line subscriptions. Accustomed to the convenience of accessibility anywhere, anytime, customers have begun to demand a new kind of converged telecommunications service portfolio.
- *Broadband Internet:* The wide availability and low cost of Internet connectivity also have reshaped customers' communication patterns. People who in the past used the telephone and sent letters to communicate now e-mail, Instant Message, chat, and migrate toward VoIP-based services. Broadband Internet connections also allow the delivery of new multimedia and entertainment services (IPTV, video on demand, music services, etc.). Completely new, community-based services (Internet multiplayer gaming and P2P networks) have appeared as well.

These two market forces squeeze legacy telecommunications providers from both sides. Their main area of business, fixed lines, is rapidly losing share to mobile networks and Internet-based services

(e-mail, messaging, VoIP). On the other hand, their existing infrastructure (subscriber loops), field expertise, and business organization have enabled them to recognize new revenue opportunities, and most have been quick enough to invest in the mobile and ISP areas.

Broadband Internet connections also have become the vehicles for new kinds of services. These new offerings (VoIP, IPTV, P2P networks, etc.) compete successfully with existing services by either being more cost efficient (VoIP) or delivering better-quality service in a more personalized manner (e.g., IPTV/VoD).

Consumer patterns are increasingly set by the "going mobile" phenomenon, with more and more customers "cutting the line." Such customers create a new demand: the demand for broadband services on mobile devices. This leads to an even stronger call for convergence in both technology and services.

The catalyst of this convergence is IP technology. IP enables the delivery of diverse content-rich services to consumers over a wide technology platform. Network convergence brings together three, formerly parallel networks (fixed/broadband, mobile, and enterprise) and unifies them into a single IP-based infrastructure.

In telecommunications terminology, this infrastructure is referred as NGN (Next-Generation Network), which is characterized by using packet-switched (IP/MPLS) technology, fiber optics, and software-based content delivery and control as opposed to circuit-switched technology, copper wires, and hardware-based switching.

MPLS technology enables high-performance, reliable, and scalable IP networks over a fiber-optic carrier infrastructure. Overcoming a key weakness of traditional IP networks, an MPLS-based infrastructure provides reliable QoS services. The Session Initiation Protocol (SIP) controls application-level signaling and enables multimedia communication. In the case of VoIP, the most important device is the softswitch, a VoIP controller and gateway between NGN and legacy (PSTN) networks. As the name implies, a softswitch is entirely programmable, and it executes as an ordinary computer program.

Multimedia content is converted between different networks by Media Gateway Controller (MGC) devices.

Part of the convergence process is Fixed-Mobile Convergence (FMC). Advances in mobile access technologies (3G, WiMAX, Wi-Fi) enable broadband access from mobile devices, and the IMS framework defines the protocols facilitating this kind of convergence.

In NGNs, convergence is a recurring theme:

- Anytime, anywhere access enabled through network and service convergence
- Convergence of services such as triple/quadruple play and FMC
- Convergence of IT and telecommunications networks into a single IP/MPLS backbone
- On the provider side, convergence of formerly parallel provider networks: enterprise (leased lines, trunks), residential (fixed lines), mobile (cellular networks)

These converged networks reduce operational costs and allow seamless, unified user access to content-rich services from any access technology.

An important aspect of NGN technologies is the pervasiveness of open standards and the prominence of scalable decentralized systems and solutions. Monolithic, proprietary hardware and software are replaced by open software-based systems and interfaces and commercial-off-the-shelf (COTS) components. Vertical software components (OS, middleware, applications) are also commoditized, and stronger competition results in cheaper, better solutions. Internet-based and other common standards such as XML, SSL, HTTP, and Simple Object Access Protocol (SOAP) are gaining in prominence. These open standards and protocols enable the implementation of transparent and trustworthy services, in that customers and third-party entities can reliably measure service performance. Thus, the technology shift enables service contracts with better service-level agreement (SLA) parameters, generally improving the quality of telecommunications services.

The NGN architecture consists of the following layers (planes in NGN terminology), all built on COTS components (hardware and software) and open standards:

- The *Access Network* consists of fixed and mobile IP-capable devices.
- The *Transport Plane* is responsible for routing and delivery of packets on the network.

- The *Control Plane* performs call session control and maintains subscriber data (e.g., location). It also configures "stove pipes" vertical, guaranteed-quality data paths between access devices and applications.
- The *Service Plane* hosts business services and applications (servers).

Horizontally coupled with each layer, the OSS/BSS layer performs service management, billing, and CRM.

It is important to note that NGN moves away from the vertical integration model of legacy networks. In legacy networks, classes of access devices are handled according to their specific characteristics vertically across service layers. This leads to parallel *service silos*; for example, each layer has to know that a particular call is made over a voice phone and handle the session accordingly. In NGN, multiple (if not all) services share common enablers (such as the Session Description Protocol [SDP]), thus reducing integration costs and increasing flexibility for present and future multimedia services.

3.2.6.2 Triple and Quad Play

Triple play refers to provision of three services (broadband Internet access, TV and video on demand, and telephone) over a single broadband connection. Triple play is offered over both CATV networks and DSL connections.

Television service is the CATV operators' main line of business, and using Internet access technologies over CATV, they can provision additional services such as home Internet access and VoIP. Telco operators provide telephone and DSL access service, and thus they provision TV service using IPTV technology over the DSL connection.

Quad play extends the offering by fixed-mobile convergence solutions, with advanced broadband technologies (such as the WiMAX family) enabling high-speed mobile access.

These modern services, fully based on IP, are the driving force behind the technological conversion of service infrastructure to next-generation IP-based systems.

Along with the new technologies, a new set of management systems has emerged. This has occurred in part because of the new requirements and opportunities offered by the new technologies and also because the management solutions and concepts of the old technologies have been replaced by concepts and software reflecting the evolution of management systems in the general IT field. The fact that operators need to be more efficient and cost-conscious than ever before has also propelled this change.

Consistent implementation of new management systems results in the emergence of *managed services*, in which all aspects of the service are managed by modern systems:

- Technical and service inventory, offering a cross-platform and cross-technology of logical connections of different layers and services provided over them
- Service provisioning, the process of decomposing customer-level service packages into elementary technical services and implementing them automatically or as a controlled and monitored workflow
- Service assurance, the process of monitoring the performance of services, providing consolidated and filtered alarms and multilayer, end-to-end diagnostic tools that enable early problem detection and accelerate troubleshooting
- Monitoring and reporting of service levels and utilization trends for internal use by the provider or directly to customers
- Integration to billing systems and monitoring of the consistency of used and billed services

Quite a few management system vendors are offering a complete set of solutions covering the above items. These companies are either traditional telco players (e.g., Telcordia, Agilent, Cramer, Metasolv) or organizations with impressive high-end enterprise traditions (e.g., IBM, HP, EMC, BMC, CA).

A good example of the new management standards is the DSL Forum's TR-069, which is a CWMP (CPE WAN Management Protocol). Legacy protocols such as SNMP and syslog are not always adequate for the management of such devices, and thus these new standards fill a critical need.

3.2.6.3 IMS and SIP

IMS (IP Multimedia Subsystem) is a protocol stack and architecture framework for delivering IP multimedia services to mobile users. Originally developed for GPRS (cellular GSM packet switched data service), its coverage has been extended to other mobile access technologies (Wi-Fi, CDMA2000, etc.) as well. In order to bridge the paradigm gap between IP and mobile communities, IMS builds on existing IETF protocols such as SIP, SDP, and DIAMETER (authentication) to provide a robust and complete multimedia platform.

Technically, IMS forms a middle, horizontal layer between access technologies and the application layer. Extension points and operational profiles are provided to support auxiliary business functionality (e.g., operator control, billing). Within this horizontal layer, IMS provides vertical interfaces between application services and access endpoints. For such channels, IMS coordinates and ensures QoS parameters, accounting correlation between participating services/transports, and provider control over the transmitted content. In addition, comprehensive security and user authentication (single sign-on) services are provided as well as advanced network services for application servers (e.g., presence, location).

IMS codification began in 1999, and most of the architecture decisions were made not long after. Thus, the recent technology and paradigm shift in the mobile world (e.g., the proliferation of cellular IP access methods) makes the value of a full IMS platform questionable. For example, many mobile phones already come with Wi-Fi support available, and thus they are able to connect to Wi-Fi IP networks independent of their cellular connections. This makes efforts to provide IP service over mobile technologies less rewarding.

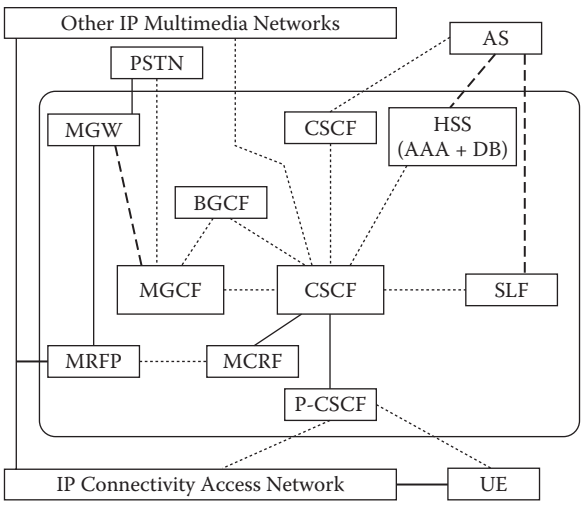
IMS services can be used through many access technologies as long as the technology supports standard IP. Native IMS devices (also called direct IMS terminals, such as mobile phones and PDAs) register directly on the IMS network and access services coordinated by their SIP protocol agents. Other, nonnative clients (e.g., legacy telephone systems, VoIP systems) connect through gateways.

3.2.6.3.1 IMS Functional Architecture

A simplified view of the IMS functional architecture is shown in Figure 3.2.7.

Descriptions of each component are as follows:

- *UE (User Equipment):* access device using IMS services.
- *AS (Application Server):* application providing multimedia content.



FIGu RE 3.2.7 Functional architecture of IMS.

- *HSS (Home Subscriber Server):* stores the master database of user profiles and performs authentication and authorization.
- *SLF (Subscriber Location Function):* maintains user addresses/locations when multiple HSSs are present.
- *CSCF (Call Session Control Function):* manages SIP sessions within the IMS network and is one of the core services of the IMS infrastructure. CSCFs may assume the following roles:
  - The S-(Serving)CSCF Session is the control point for UEs; sessions originate and terminate at S-CSCFs from the IMS network's point of view. S-CSCFs also ensure that session parameters are within the UE's capabilities. In SIP terminology, S-CSCF is called a *registrar*.
  - The I-(Interrogating)CSCF is the contact point for the UE's home network to other networks; it is called a *stateless proxy* in SIP terminology.
  - The P-(Proxy)CSCF is a technical contact point between the IMS network and the UE. All UEs connect to P-CSCFs first, which may co-locate other CSCF functionality as well.
- *MGW (Media Gateway Function):* implements the connection to non-IMS services such as PSTN (legacy phone network). MGWs perform the physical and logical conversions between different networks (e.g., packet-switched and circuit-switched).
- *MGCF (Media Gateway Control Function):* controls the behavior of MGWs and processes out-of-band signaling if present.
- *BGCF (Breakout Gateway Control Function):* selects the MGCF and MGW for a particular session and provides all functionality (e.g., security) not handled by them.
- *MRFP (Media Resource Function Processor):* provides and processes media streams internal to IMS operations (e.g., dial tones, system text messages in different languages, signal mixing and synchronization for multiple parties). It also provides media analysis and statistical functions.
- *MRFC (Media Resource Function Control):* controls the MRFP according to the information pertaining to the current session (e.g., controls audio and video synchronization, commands generation of dial tones).

Since IMS is a large-scale mobile technology for individual subscribers, identification is a key issue. IMS uses the following information to identify users or their devices:

- *International Mobile Subscriber Identity (IMSI):* unique ID within the worldwide cellular network
- *Temporary Mobile Subscriber Identity (TMSI):* ID generated on the basis of the IMSI and used in sessions to avoid disclosing the sensitive IMSI
- *International Mobile Equipment Identity (IMEI):* device identifier
- *Mobile Subscriber ISDN Number (MSISDN):* the telephone number associated with the user
- *IP Multimedia Private Identity (IMPI):* textual, URI ID specific to the subscriber
- *IP Multimedia Public Identity (IMPU):* textual, URI ID that is not specific to the subscriber (subscribers may have multiple IPMUs for each IPMI, e.g., for multiple cell phones)

These identifiers are maintained by HSS servers, occasionally assisted by SLF servers with locating subscribers. In this environment, session control is also a crucial issue. Session control is performed via the SIP protocol.

3.2.6.3.2 Session Control

The SIP protocol is an IETF (RFC3261) signaling protocol used for the management of multimedia communication sessions between two (unicast) or more (multicast) peers. SIP is a helper protocol, offloading the burden of signaling from communication applications. SIP assists with applications by:

- Determining the end user's location (network address) and availability
- Negotiating capabilities (media and parameters)
- Orchestrating session setup ("ringing" and establishment parameters)

- Performing session management (e.g., transfer, termination, modification of parameters, participant management)

SIP is concerned with session setup and maintenance only and is designed to work in concert with other protocols responsible for actual multimedia communication (e.g., RTP [Real-Time Transport Protocol], RTSP [Real-Time Streaming Protocol], and SDP [Session Description Protocol]).

SIP is designed to be open, extensible, and easy to implement. The protocol leverages existing standards and industry best practices; for example, SIP addresses are well-formed Internet Uniform Resource Identifiers (URIs) in which the user address part follows the structure of e-mail addresses (e.g., sip:bob@sipserver.company.com).

Additionally, SIP follows an HTTP-like textual request–response transaction model, and the message header syntax is very similar to that of an HTTP message; SIP even reuses most of the HTTP status codes (e.g., "404 Not Found"). This makes SIP user-friendly for developers and operators already familiar with Internet technologies. To summarize, SIP relies on the following Internet practices:

- *Text-based messages:* These are easy to debug and troubleshoot.
- *MIME-type support:* Handler applications can easily recognize content types.
- *Reuse of and leverage on mature Internet services and protocols:* These include DNS (Domain Name Service), RTP, RSVP (Resource Reservation Setup Protocol), and so forth.
- *Ease of extension with new message types/parameters:* Older, different SIP participants can simply ignore the fields they do not support.
- *Transport-layer independence:* SIP works over both UDP and TCP and involves no requirements on the protocol stack below the IP layer.
- *Multidevice and multiservice support:* Different capabilities of user agents are handled and negotiated separately; for example, clients with audio support only can still communicate with their audio-only peers.

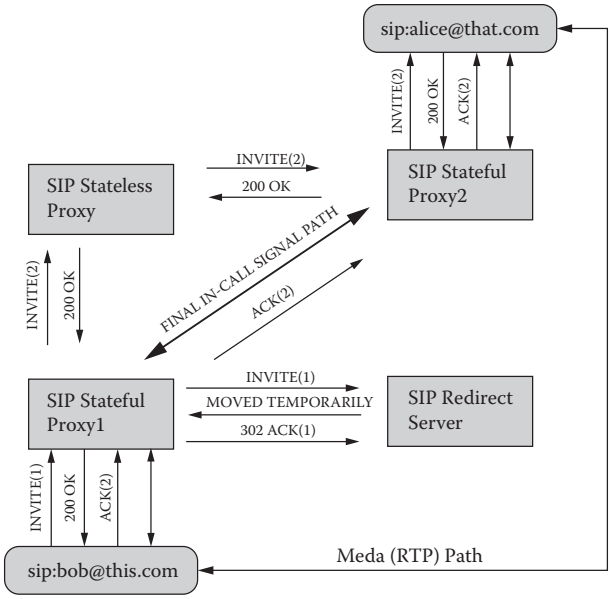
SIP relies on the presence of well-known SIP servers users can turn to. Each SIP user (called a user agent) is registered to a *home server* (called a *registrar* in SIP) and has an SIP URI uniquely identifying the user within the SIP network.

SIP servers may also act as proxy servers, accepting user requests and forwarding them to other SIP servers better qualified to serve the particular requests. Thus, users registered to different registrars may also establish sessions if the SIP network is able to route the establishment (INVITE) request to the addressee's registrar.

Figure 3.2.8 shows an example of an SIP session between sip:bob@this.com and sip:alice@that.com. In this example, we assume that Proxy1 also hosts Bob's home server and location service (which is the typical configuration), and the same is true for Alice and Proxy2.

The register calls, made previously by Bob and Alice's user agents toward their home server (registrar), are not shown.

1. Bob dials Alice in the SIP User Agent application.
   a. The UA contacts its configured proxy server (Proxy1) and sends an INVITE message.
2. Proxy1 retrieves Bob's contact information (IP address) from the SIP Registrar/Location Service using a QUERY call (not shown).
   a. The called party, Alice, is located in a different SIP domain Proxy1 does not explicitly know about. Thus, Proxy1 turns to a Redirect Server, which reveals that Alice can be reached through yet another proxy. This proxy is a Stateless Proxy (e.g., located at an external SIP provider/ISP) that is merely able to forward SIP messages between the two SIP domains.
   b. The INVITE message arrives to Proxy2, Alice's local proxy and home server. Proxy2 INVITEs Alice's User Agent, which notifies her (e.g., rings) and, when Alice answers the call, sends the OK message back.



FIGu RE 3.2.8 Visualization of a SIP session.

3. The INVITE message contains Bob's IP address and additional protocol information, which enables the parties to establish an RTP session independent of the SIP mechanism.
4. During the session, an SIP In-Call Signal Path is maintained between Proxy1 and Proxy2 for session management purposes.

The participation of the third-party stateless proxy is optional. Such proxies enable the construction of large-scale SIP networks of multiple domains. The session discussed above is an example of the most basic extra-domain SIP interaction. The protocol contains several additional features and message types (e.g., for holding, transferring, multiparty management), making it a full-scale signaling and session management vehicle.

3.2.6.3.3 SIP Management

The IETF SIP Working Group has released an MIB for the SNMP management of SIP devices, published in RFC 4780. The management information is structured as follows:

- Common MIB module defining objects common to all SIP devices
- User agent MIB module defining management objects for both user agent servers (UASs) and user agent clients (UACs)
- Server MIB module broken down into the following sections: proxy server, redirect server, and registrar server

There are a multitude of commercial and open-source SIP servers and implementations available. SNMP support among these servers and implementations is rare. Some implementations, such as BEA's WebLogic SIP server, support a proprietary MIB. As of early 2008, we were not aware of any implementations fully implementing RFC 4780.

3.2.6.4 IPTV and VoD

IPTV (IP television) is the generic (marketing) name of systems delivering digital television service over IP as a transmission medium. On the mass consumer (residential) market, it is often bundled with video on demand (VoD) services sharing the same technology for delivering multimedia content.

With respect to terminology, there is a marked difference between "Internet TV" and IPTV. In the first case, providers set up Internet TV servers, and individual subscribers connect to them using their broadband connection provided by a third-party ISP. Providers have no control over transmission quality as the commercially unrelated ISP makes best-effort delivery. Users view the digital video content on their PCs using client software.

In the second case (IPTV), the entire closed system is set up by the broadband service provider; thus, transmission quality is guaranteed, and only broadband customers of the ISP in question can access the service. Providers usually deploy specialized equipment (a set-top box, or STB) at the consumer's premises. STBs receive the IPTV transmission and convert it to analog signals suitable for consumer TV sets. Deploying an IPTV system requires considerable commitment on the part of the telecommunications service provider, as it necessitates changes to the entire existing IP service infrastructure.

3.2.6.4.1 Protocols and Technology

IPTV channels are encoded and compressed via MPEG-2 or MPEG-4 encoding and transmitted as MPEG transport streams through a connectionless UDP. In the case of premium channels, transport streams are also encrypted. IP multicast technology (IGMPv2) is used to broadcast the same channel to multiple subscribers; customers' STBs join the multicast group carrying the required channel. VoD content is transmitted via IP unicast, using RTSP (Real-Time Streaming Protocol) to control the transmitter.

Transport streaming protocols do not explicitly require guaranteed delivery and maximum latency from the underlying transfer network, as they were developed for best-effort networks and employ sophisticated error correction and predictive coding. In reality, user experience deteriorates quickly with network errors (packet losses) and nonuniform packet latency. Thus, network providers often have to restructure their core networks in order to provide IP service of higher quality than necessary for legacy IP. For this purpose, IPTV subnetworks are separated from "regular" Internet traffic to the extent possible. Network security (such as protection from Distributed Denial of Service attacks, for example) is also a concern, given that it is relatively easy for malicious attackers to disrupt multicast networks after gaining access to the core network.

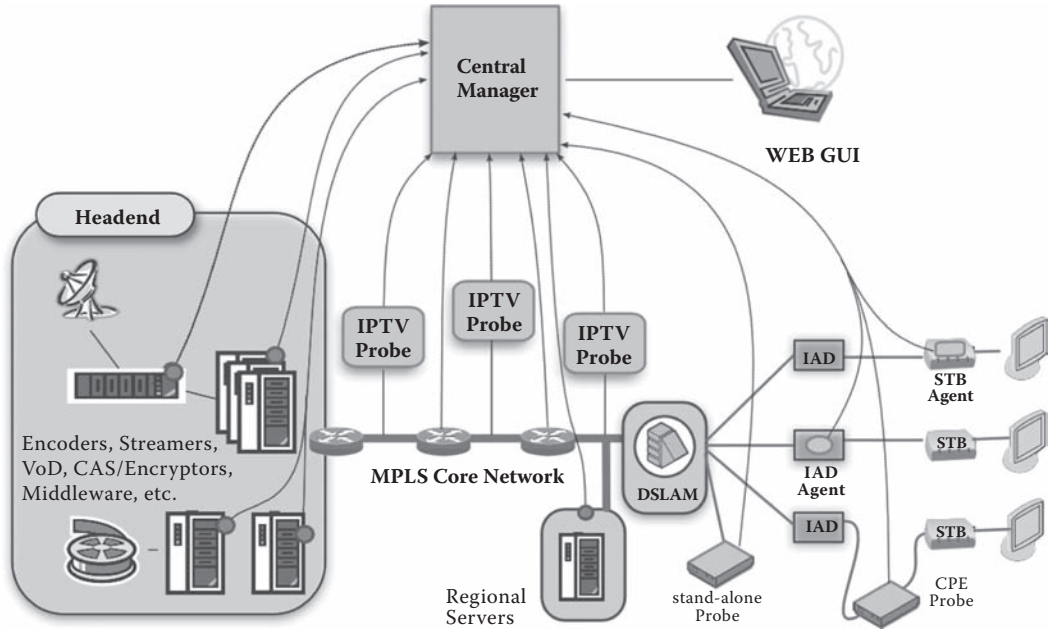
Zapping time, or channel switch delay, is also an important quality indicator of an IPTV infrastructure, as users do not typically tolerate delays longer than 1 or 2 seconds. This delay mostly depends on the quality of the IGMP infrastructure (routers and network edge devices).

In order to attract customers from legacy, cable, or satellite TV systems, IPTV providers strive to make their service more interactive. This includes features, such as pausing, resuming, rewinding, and recording the TV stream, which is not possible with legacy TV technology. These features require communication between the server infrastructure and individual STBs and temporary storage of recordings either on the STB or within the network.

3.2.6.4.2 IPTV Architecture

There are three primary components of an IPTV service infrastructure (Figure 3.2.9):

- *Headend:* where digital content (usually received from satellite) is encoded, optionally stored, and encrypted. Middleware servers managing subscriber data and channel package (e.g., billing) information are also part of the headend, as well as VoD servers.
- *Distribution network:* multicast-capable protected IP network responsible for distributing transport streams. The distribution network consists of core routers, transmission links, and multicast-aware edge devices (e.g., DSLAMs for DSL networks). Sometimes load-balancing and redundancy devices for headend servers (e.g., VoD and recorded content servers) are also deployed at key points of the distribution networks.
- *Customer premises:* end user locations with the IAD (Internet Access Device or home router and modem) and STB.



FIGu RE 3.2.9 Typical IPTV service infrastructure with end-to-end management and monitoring.

In the headend, receiver units accept channels from content providers and produce DVB-quality signals for encoder or transcoder units, which compress these channels according to configurable parameters and produce IP unicast or multicast transport streams. These are optionally stored (for pause and playback support) or, in the case of premium content, encrypted. After this stage, transport streams are routed into the distribution network and transmitted to subscribers' set-top boxes, which decode the TV content.

At the customer endpoint, the IAD takes care of separating IPTV traffic from legacy Internet traffic (and optionally from VoIP traffic as well). These two or three services are always bundled by the provider to make the service more attractive. Packets belonging to different services are transmitted in different VPNs/VLANs (or other, equivalent classification technology is used, such as MPLS) because distribution routers rely on this classification in enforcing QoS policies.

3.2.6.4.3 Management

The above three architectural components each pose unique challenges for IPTV management. In headend, a comprehensive solution consolidating many kinds of highly specialized devices from different vendors (e.g., receivers, transcoders, encryptors) must be deployed. Typically, all devices support SNMP with vendor-specific MIBs. As for the distribution network, some of the parameters important to monitor (e.g., latency variation, IGMP delays) are not included in standard MIB definitions and are not supported by standard Layer 2 and Layer 3 devices in the distribution network. Thus, operators have to resort to proprietary solutions yet again. These solutions typically utilize purpose-built network probes with vendor-specific interfaces.

The same is true for the highly vendor-specific IADs and STBs. SNMP is generally supported with standard MIBs, but the really important metrics (e.g., MDI [Media Delivery Index], an aggregate transmission quality indicator) are rarely available.

Operators of large-scale residential networks are wary of using SNMP for IAD/CPE management as a result of SNMP's security problems. Newer devices support the TR-069 management protocol, which is much better suited for this role; however, as of late 2007, TR-069 support for mainstream devices had yet to emerge.

3.2.7 Telco and Web 2.0 Applications

The recent infusion of Web 2.0 innovation is poised to change the telecom industry dramatically. Internet pioneers are working to blend their services with telco assets, bringing about a range of compelling applications, devices, services and revenue possibilities.

This is a fundamental shift, compared with the first generation of Web applications, which were essentially delivered "over the top" of best-effort IP connectivity, as if the telcos were a public road system. Instead, Web 2.0 applications, and so-called mesh-ups, are implemented "across the middle" of telco networks, directly controlled by network policies and supported by back-office functions to ensure that they work properly and deliver a quality user experience. The following examples are typical for such innovative applications (LUCA07):

- *iPhone*: Apple's latest is the ultimate mobile device: a phone, a powerful Web browser (with zoom and orientation features), and a media player, with a range of productivity and messaging applications. But, it is useless without a mobile network.
- *Software as a Service (SaaS)*: Today, CRM and sales force automation applications are leading this market. However, just about every software supplier has a SaaS strategy. Given that these are distributed, mission-critical applications, the SaaS does not work without fault-resistant connectivity and robust security.
- *Gaming*: Today's gaming experience is unparalleled. In fact, the technology in a PlayStation 3 or Xbox makes a standard PC look like a calculator. The growth in the market is driven by distributed, multiplayer games that integrate voice, video, and rich media. The user experience cannot be met without low-latency, low-error broadband networks.
- *Video*: Video is fast becoming telecom's next killer application. Hollywood is excited, as IPTV and mobility together have unlimited potential for personalized, on-demand, interactive content. User-generated content via YouTube has opened up unforeseen markets. AT&T's recent mobile push-to-video service allows users to capture, publish, and archive video clips directly from the handset. And there are other applications like security, remote monitoring, and training among others. But video is a very high bandwidth requester—it will not work without quality network distribution.
- *Mobile applications*: These are popping up all over. On the enterprise side, custom applications for real estate, transportation, medicine, workforce management, sales, and other activities are announced almost daily. Cisco, Microsoft, and some operators are working on integrated communications technologies that are making Fixed-Mobile Convergence (FMC) a reality. On the customer side, navigation, friend-finder, messaging, banking, social networking, and entertainment applications are gaining in momentum. None of these work well without integrated call handoff over 3/4G networks.

The emerging market reality is that new value is created only when telcos and Web innovators integrate their offerings. This point is perfectly illustrated by the recent announcement that Google has agreed to provide the portal, interface, and application suite for WiMAX network services from Sprint. This is an excellent opportunity telcos have been waiting for. Operators of both wireless and fixed networks are very well positioned to enable the Web 2.0 innovation. Their multibillion-dollar investment in fiber, spectrum, broadband, and access networks will bring 10 or 100 Mbps access everywhere, all the time. Investment in IMS and rich communication functionality —such as location, presence, bandwidth policy, identity/authentication, QoS, rich messaging, profiling, and other network services—is bringing powerful communications elements to bear. The investment in standardized, open APIs and service delivery platforms is helping to expose these assets to the development community.

By opening their networks to third-party developers and allowing them to create, integrate, and blend their products easily and rapidly with communications assets, the telcos no longer settle for a role as bit-pipes. Instead, they sit between the consumer and developer, and get to participate in the action. From a revenue perspective, it could not get any better for a telco. The question is: Will the operator's back

office be ready to monetize their network capacities and services? This is critical for the telcos, because the Web companies will not care. They are perfectly comfortable monetizing their applications via ads and click-throughs or, in Apple's case, via device sales. That is their business model, and experiene with content has shown that it is very difficult to reach a revenue-sharing agreement.

Therefore, operators will have to carve out and charge for Web 2.0 services on a discriminatory basis. They will have to have tight control over their network assets, carefully defining and executing network access and usage policies based on the service. They will have to support complex partnering agreements and meet delivery expectations.

3.2.8 Summary and Trends

Customers will show little interest in management solutions at lower layers. The primary target is to offer management capabilities at the service and business management layers. Significant extensions of existing management products are needed. This segment has shown the status of management capabilities at the element and network management layers only. In order to offer metrics for higher layers, MIB extensions and eventually new tools will be required.

New tools and new management applications are expected to help multiple management layers by providing consolidated and detailed results, respectively. Network infrastructure components, themselves, software and hardware sensors, B/OSS applications will provide a wealth of data. These data are the basis for control instruments, such as balanced scorecards, dashboards, and for analytics, such as business intelligence tools. Elaboration on these tools and applications will follow in subsequent sections.

Acronyms

1×RTT	1× Radio Transmission Technology
ADM	Add-Drop Multiplexer
ADSL	Asymethric Digital Subscriber Line
AP	Access Point
AS	Application Server
ATM	Asynchronous Transfer Mode
BSS	Business Support System
BGCF	Breakout Gateway Control Function
CAP	Carrierless Amplitude Phase
CATV	Community Antenna Television
CEM	Customer Experience Management
CES	Circuit Emulation Switching
CLEC	Competitive Local Exchange Carrier
CMTS	Cable Modem Termination System
CNM	Customer Network Management
COTS	Component Off-the-Shelf
CPE	Customer Premise Equipment
CRM	Customer Relationship Management
CSCF	Call Session Control Function
DHCP	Dynamic Host Configuration Protocol
DLC	Digital Loop Carrier
DSLAM	Digital Subscriber Line Access Multiplexer
DSL	Digital Subscriber Line
DTE	Data Terminal Equipment
EMS	Element Management System
FDM	Frequency Division Multiplexing

FMC	Fixed–Mobile Convergence
FMS	Fixed–Mobile Substitution
GIS	Geographical Information System
GPON	Gigabit Passive Optical Network
GSM	Group Speciale Mobile
HDSL	High-Bit-Rate Digital Subscriber Line
HDTV	High-Definition TV
HSPA	High Speed Packet Access
HSS	Home Subscriber Server
IAD	Internet Access Device
ICT	Information and Communications Technologies
IGMP	Internet Gateway Management Protocol
ILMI	Interim Local Management Interface
IMEI	International Mobile Equipment Identity
IMPI	IP Multimedia Private Identity
IMPU	IP Multimedia Public Identity
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
ISV	Independent Software Vendor
IEX	Inter Exchange Carrier
LAN	Local Area Network
LEC	Local Exchange Carrier
LER	Label Edge Router
LCN	Logical Channel Number
LSP	Label Switched Path
MAC	Media Access Control
MGCF	Media Gateway Control Function
MGW	Media Gateways
MRFC	Media Resource Function Control
MRFP	Media Resource Function Processor
MGW	Media Gateway Function
MDI	Media Delivery Index
MIB	Management Information Base
MGC	Media Gateway Controller
MPLS	Multiprotocol Label Switching
MSS	Marketing Support System
NGN	Next-Generation Network
NGOSS	Next-Generation OSS
OSS	Operations Support System
P2P	Peer To Peer
PBX	Private Branch Exchange
PDA	Personal Digital Assistant
PDU	Protocol Data Unit
POP	Point Of Presence
POTS	Plain Old Telephone System
PSTN	Public Switched Telecommunication Network
PVC	Permanent Virtual Circuit
QoS	Quality of Service
RADSL	Rate-Adaptive Digital Subscriber Line
RTP	Real-Time Transport Protocol

RTSP	Real-Time Streaming Protocol
RTUs	Remote Terminal Units
SDH	Synchronous Digital Hierarchy
SI	System Integrator
SIP	Session Initiation Protocol
SLF	Subscriber Location Function
SOAP	Simple Object Access Protocol
SP	Service Provider
SPVC	Semipermanent Virtual Circuit
STB	Set-Top Box
SVC	Switched Virtual Circuit
SNMP	Simple Network Management Protocol
STDM	Statistical Time Division Multiplexing
TMSI	Temporary Mobile Subscriber Identity
TDM	Time Division Multiplexing
TMN	Telecommunications Management Network
UAC	User Agent Client
UAS	User Agent Server
UDP	User Datagram Protocol
UE	User Equipment
UMTS	Universal Mobile Telecommunication Service
UNI	User Network Interface
VCC	Virtual Channel Connection
VCI	Virtual Circuit Identifier
VPC	Virtual Path Connection
VoD	Video On Demand
VoIP	Voice over IP
WLAN	Wireless Local Area Network
WiMax	Worldwide Interoperability for Microwave Access
XMP	Extended Markup Language

References

ABER97: Aber, R. 1997. xDSL supercharges copper, *Data Communications* (March 1997): 99–105.

BLAC94: Black, U. 1994. *Emerging Communications Technologies,* Prentice-Hall Series in Advanced Communication Technologies, Englewood Cliffs, NJ: Prentice Hall.

IETF *RFC 3261*, *RFC 4780*

LUCA07: Lucas, M., and Chamarthy, R. 2007. Charging for telco & Web 2.0 services, *Billing World & OSS Today* (September/October): 9–12. Telestrategies Publication.

ROGE95: Rogers, C. 1995. Telcos versus cable TV: The global view, *Data Communications* (September 1995): 75–80.

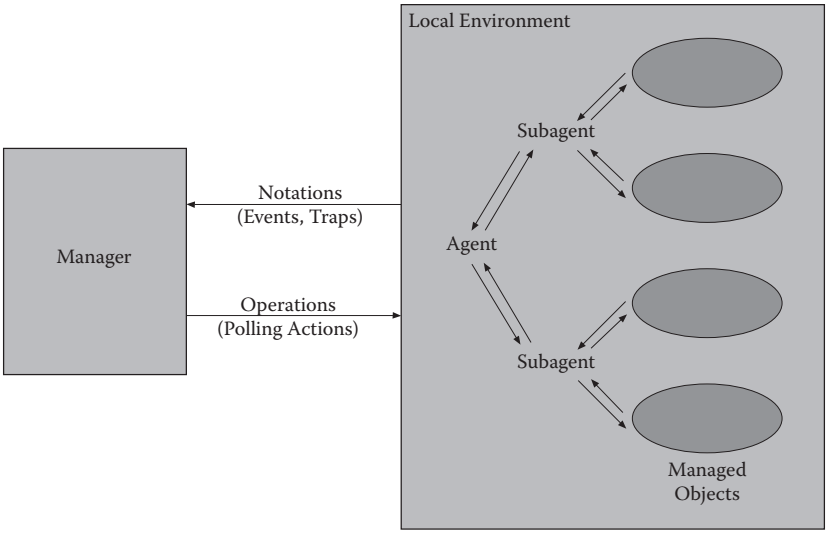
TERP98: Terplan, K. 1998. *Telecom Operations Management Solutions* (Boca Raton, FL: CRC Press).

Understanding SIP: Ubiquity White Paper available from www.sipcenter.com

3.3 Management-Related Standards

Tivadar Szemethy

Industry standards help suppliers, vendors, and customers to communicate with each other more efficiently, but finding the common denominator of all interests is not easy. Service creation, provisioning, assurance,



FIGu RE 3.3.1 Communication paths between manager, agents, subagents, and managed objects.

delivery, and service management cannot wait for fully completed standards. This chapter summarizes the most important management standards for multimedia communications. Besides Telecommunications Management Network (TMN), enhanced Telecom Operations Map (eTOM), Simple Network Management Protocol (SNMP), Remote Network Monitoring (RMON), Open Management Architecture (OMA), and Desktop Management Task Force (DMTF), OSSJ is also introduced, which may help to unify and simplify management frameworks and applications. The TeleManagement Forum may help a lot to improve the interoperability among multiple suppliers by providing practical guides, specifications, solution sets, and conformance documents. Web services, Web technologies, and Web protocols will change the management protocol landscape significantly. This segment will introduce the dominating trends led by basis technologies, description and location services, business process description, and databases.

3.3.1 Manager–Agent Relationship

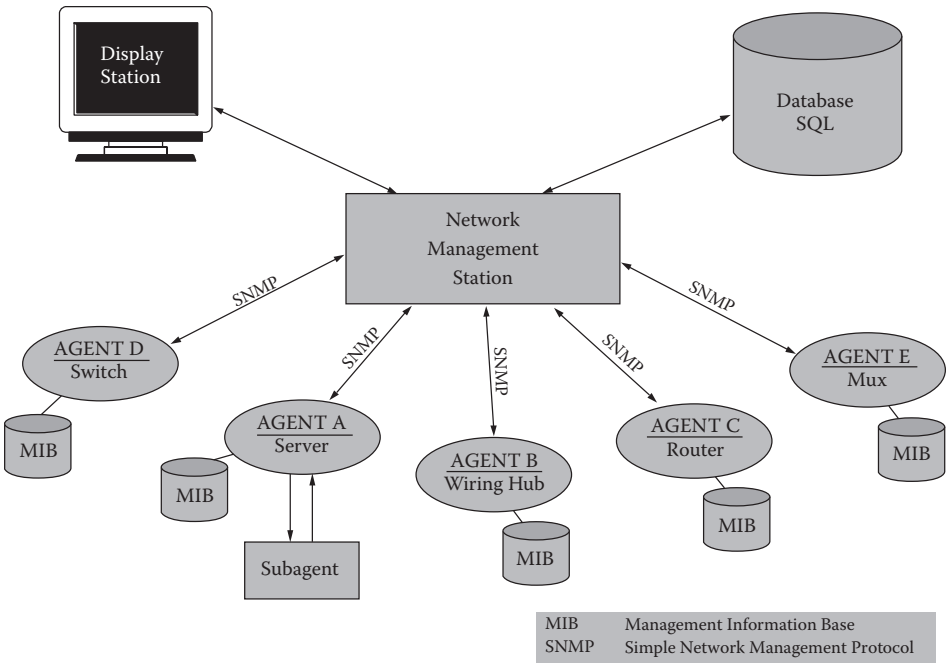
Protocols represent agreements between communicating parties. Management–agent relationships have been frequently implemented in the area of network management. Figure 3.3.1 shows an example of such a relationship. The agent side may be hierarchical in that subagents are implemented into specific devices. Depending on the nature of the management protocol, either the manager or the agents begin the dialogue. There are always exceptions for high-priority networking events.

In the case of Common Management Information Protocol (CMIP), event techniques are used. Assuming that agents are capable of capturing, interpreting, filtering, and processing events, they will notify the manager about alarming conditions. Of course, the manager can interrupt and send inquiry messages to the agents.

In the case of SNMP, the manager polls the agents periodically. The agents respond by sending information on device status and performance. Typically, agents wait for the poll unless unusual events occur in the network. Special traps can be defined and implemented for such events.

3.3.2 Simple Network Management Protocol (SNMPv1, SNMPv2, SNMPv3)

In the SNMP environment, the manager can obtain information (Figure 3.3.2) from the agent by periodically polling managed objects. Agents can transmit unsolicited event messages, called *traps*, to the manager. The management data exchanged between managers and agents is called the *management*



FIGu RE 3.3.2 Structure of SNMP-based management services.

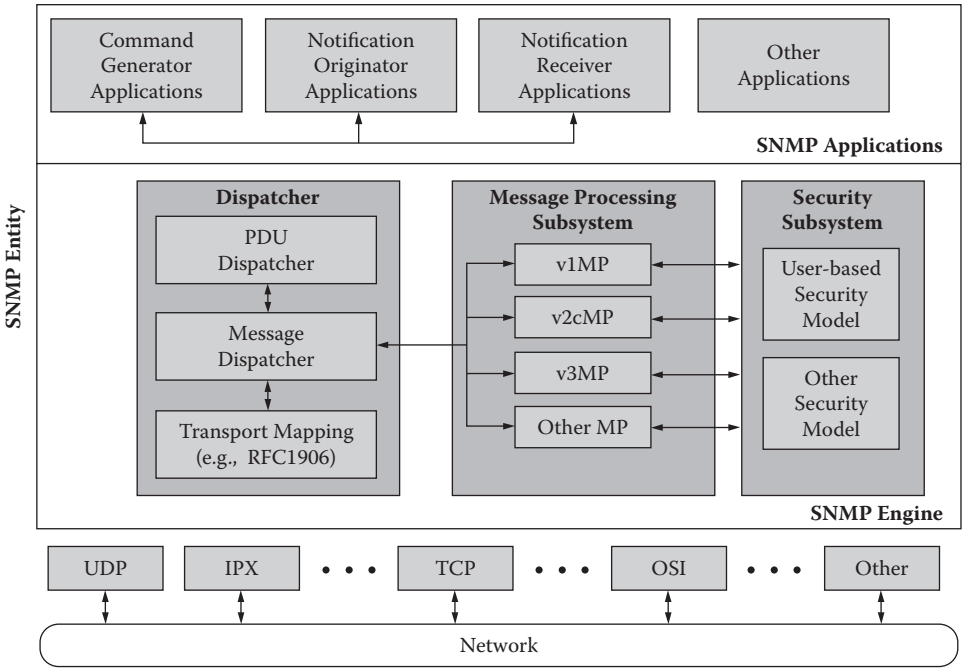
information base (MIB). The data definitions outlined in Structured Management Information (SMI) must be understood by both managers and agents.

The manager is a software program residing within the management station. The manager has the ability to query agents using various SNMP commands. The management station is also in charge of interpreting MIB data, constructing views of systems and networks, compressing data, and maintaining data in relational or object-oriented databases. A traditional SNMP manager is shown in Figure 3.3.3 (STAL99).

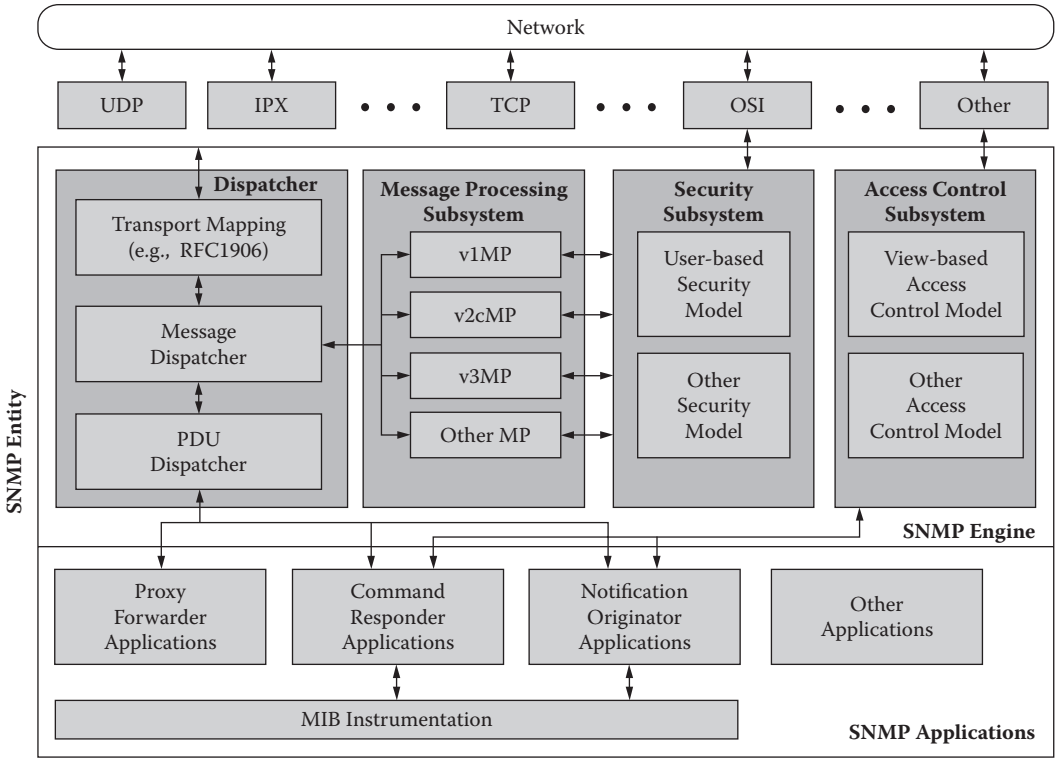
This figure shows typical functions that are not only valid for managers, but also for managed agents. Figure 3.3.4 shows the typical functional blocks of SNMP agents.

In a traditional manager, the SNMP engine contains a dispatcher, a message processing subsystem, and a security subsystem. The dispatcher is a simple traffic manager. In the case of outgoing protocol data units (PDUs), the dispatcher accepts these units from applications and performs the following functions. For each PDU, the dispatcher determines the type of message processing required—which may be different for SNMP versions 1, 2, and 3—and passes the PDU on to the appropriate message processing module in the message processing subsystem. Subsequently, the message processing subsystem returns a message containing that PDU and including the appropriate message headers. The dispatcher then maps this message onto a transport layer for transmission. In the case of incoming messages, the dispatcher accepts these messages from the transport layer and performs the following functions. The dispatcher routes each message to the appropriate message processing module. Subsequently, the message processing subsystem returns the PDU contained in the message. The dispatcher then passes this PDU to the appropriate application.

The message processing subsystem accepts outgoing PDUs from the dispatcher and prepares them for transmission by wrapping them in the appropriate message header and returning them to the dispatcher. It also accepts incoming messages from the dispatcher, processes each message header, and returns the enclosed PDU to the dispatcher. An implementation of the message processing subsystem may support a single message format corresponding to a single version of SNMP or may contain a number of modules, each supporting a different version of SNMP.



FIGu RE 3.3.3 Traditional SNMP manager.



FIGu RE 3.3.4 Traditional SNMP agent.

TABl E 3.3.1 Components of SNMP Entities

Dispatcher: Allows for concurrent support of multiple versions of SNMP messages in the SNMP engine; it is responsible for (1) accepting PDUs from applications for transmission over the network and delivering incoming PDUs to applications, (2) passing outgoing PDUs to the message processing subsystem to prepare as messages and passing incoming messages to the message processing subsystem to extract incoming PDUs, and (3) sending and receiving SNMP messages over the network
Message processing subsystem: Responsible for preparing messages for sending and for extracting data from received messages
Security subsystem: Provides security services such as authentication and privacy of messages; it potentially contains multiple security models
Access control subsystem: Provides a set of authorization services that an application can use for checking access rights; access control can be invoked for retrieval or modification request operations and for notification generation operations
Command generator: Initiates SNMP Get, GetNext, GetBulk, or set request PDUs and processes responses to requests it has generated
Command responder: Receives SNMP Get, GetNext, GetBulk, or set request PDUs destined for the local system, as indicated by the fact that the contextEngineID in the received request is equal to that of the local engine through which the request was received; the command responder application will perform the appropriate protocol operation, using access control, and will generate a response message to be sent to the request's originator
Notification originator: Monitors a system for particular events or conditions and generates Trap and/or Inform messages based on these events or conditions; a notification originator must have a mechanism for determining where to send messages and which SNMP version and security parameters to use when sending messages
Notification receiver: Listens for notification messages and generates response messages when a message containing an Inform PDU is received
Proxy forwarder: Forwards SNMP messages; implementation of a proxy forwarder application is optional

The security subsystem performs authentication and encryption functions. Each outgoing message is passed to the security subsystem from the message processing subsystem. Depending on the services required, the security subsystem may encrypt the enclosed PDU, and it may generate an authentication code and insert it into the message header. The processed message is then returned to the message processing subsystem. Similarly, each incoming message is passed to the security subsystem from the message processing subsystem. If required, the security subsystem checks the authentication code and performs decryption. An implementation of the security subsystem may support one or more distinct security models.

The SNMP engine for a traditional agent has all of the components found in the SNMP engine for a traditional manager along with an access control subsystem. This subsystem provides services to control access to MIBs for the reading and setting of management objects. These services are performed on the basis of the contents of PDUs. An implementation of the security subsystem may support one or more distinct access control models. Security-related functions are organized into two separate subsystems: security and access control. This is an excellent example of modular design in that the two subsystems perform quite distinct functions; thus, it makes sense to allow standardization of these two areas to proceed independently. The security subsystem is concerned with privacy and authentication and operates on SNMP messages. The access control subsystem is concerned with authorized access to management information and operates on SNMP PDUs.

For both SNMP managers and agents, management functionalities comprise a number of components. These components are listed in Table 3.3.1 (STAL99).

There are continuing changes with respect to MIBs. In addition to standard MIBs, such as MIB I and II (Table 3.3.2), the Internet Engineering Task Force (IETF) has defined a number of adjunct MIBs covering hosts, routers, bridges, hubs, repeaters, Fiber Distributed Data Interface (FDDI) networks, AppleTalk networks, frame relay networks, switches, asynchronous transfer mode (ATM) nodes, mobile components, and applications.

The following trends are expected with SNMP. SNMP agent-level support will be provided by an even greater number of vendors. SNMP manager-level support will be provided by only a few leading vendors

TABl E 3.3.2 MIB II Structure

11 Categories of Management (2) Subtree	Information in the Category
System (1)	Network device operating system
Interfaces (2)	Network interface specific
Address translation (3)	Address mappings
IP (4)	Internet protocol specific
ICMP (5)	Internet control message protocol specific
TCP (6)	Transmission protocol specific
UDP (7)	User datagram protocol specific
EGP (8)	Exterior gateway protocol specific
CMOT (9)	Common management information services on TCP specific
Transmission (10)	Transmission protocol specific
SNMP (11)	SNMP specific

in the form of several widely accepted platforms. Management platforms provide basic services, leaving customization to vendors and users.

Wider use of intelligent agents is also expected. Intelligent agents are capable of responding to a manager's request for information and performing certain manager-like functions, including testing for thresholds, filtering, and processing management data. Intelligent agents enable localized polling and filtering on servers, workstations, and hubs, for example. Thus, these agents reduce polling overhead and management data traffic, forwarding only the most critical alerts and processed data to the SNMP manager.

Remote monitoring (RMON) MIB will help to bridge the gap between the limited services provided by management platforms and the rich sets of data and statistics provided by traffic monitors and analyzers.

The strengths of SNMP include the following:

- Agents are widely implemented.
- Implementation is simple.
- Agent-level overhead is minimal.
- The polling approach is good for LAN-based managed objects.
- It is robust and extensible.
- It offers the best direct manager–agent interface.

Weaknesses include:

- It is too simple and does not scale well.
- There is no object-oriented data view.
- Unique semantics make integration with other approaches difficult.
- Communication overhead is high as a result of polling, particularly for WAN-based management objects.
- There are many implementation-specific (private MIB) extensions.
- No standard control definition is available.
- One agent per device may be inappropriate for systems management.

SNMP is continuously being improved and extended. SNMPv1 is the earliest specification.

As mentioned earlier, SNMPv1 is defined over IP/UDP, OSI CLNS, AppleTalk DDP, and Novell IPX. In SNMPv1, the only security method supported is authentication via a plain text password (community string).

SNMPv2 addresses many of the shortcomings of version 1. SNMPv2 can support either a highly centralized management strategy or a distributed one. In the latter case, some systems operate in the role of both manager and agent. In its agent role, such a system will accept commands from a superior

manager; these commands may deal with access of information stored locally at the intermediate manager or may require the intermediate manager to provide summary information about subagents. The principal enhancements to SNMPv1 provided by version 2 fall into the following categories (STAL99):

- The structure of management information is being expanded in several ways. The macro used to define object types has been expanded to include several new data types (such as 64-bit counters) and to enhance the documentation associated with an object. A noticeable change is that a new convention has been provided for creating and deleting conceptual rows in a table. This capability originates from RMON.
- Transport mappings help in allowing different protocol stacks to transport SNMP information, including user datagram protocol, OSI connectionless-mode protocol, Novell internetwork (IPX) protocol, and AppleTalk.
- The protocol operations with the most noticeable changes include two new PDUs. The GetBulkRequest PDU enables the manager to efficiently retrieve large blocks of data. In particular, it is powerful in retrieving multiple rows in a table. The InformRequest PDU enables one manager to send trap-type information to another manager.
- MIB extensions contain basic traffic information about the operation of the SNMPv2 protocol; this is identical to SNMP MIB II. The SNMPv2 MIB also contains other information related to the configuration of SNMPv2 manager to agent.
- Manager-to-manager capability is specified in a special MIB called M2M (manager to manager), which provides functionality similar to the RMON MIB. In this case, the M2M MIB can be used to allow an intermediate manager to function as a remote monitor of network media traffic. Also, reporting is supported. Finally, two major groups, Alarm and Event, are supported.
- SNMPv2 security includes a wrapper containing authentication and privacy information as a header to PDUs.

The SNMPv2 framework is derived from the SNMP framework. The evolution from SNMP to SNMPv2 is intended to be seamless. SNMPv2 is different from version 1 not only in the supported data types and protocol primitives but in PDU format as well. Thus, version 2 is not upwardly compatible with version 1. RFC 1908 outlines two interoperability strategies: (1) proxy agents and (2) bilingual network management systems (NMSs). Proxy agents are version 2 agents that relay and translate messages between version 2 NMSs and version 1 agents. Bilingual NMSs communicate with both version 1 and version 2 agents. The version supported by each agent is stored in a database, and the NMS uses the appropriate protocol version for each different agent. From an implementation point of view, it is easy to extend a version 2 implementation to support version 1 as well.

A new security model was also proposed and standardized in SNMPv2, but it proved overly complex and impractical. SNMPv2c (Community-Based Simple Network Management Protocol version 2) is a de facto standard (RFCs 1901, 1908) that includes all of the version 2 improvements with the exception of the security model, retaining the community-based authentication from version 1.

SNMPv2u (User-Based Simple Network Management Protocol version 2, RFCs 1909, 1910) proposes a different, simpler security mechanism as an alternative to SNMPv2. Although later evolving into the Simple Network Management Protocol version 3 (SNMPv3) security framework, SNMPv2u has not really caught on. SNMPv3, which is "SNMPv2 plus security plus administration," is the current IETF standardized version. The key new feature of SNMPv3 is better security. The design goals for version 3 can be summarized as follows (STAL99):

- Use existing work for which there is implementation experience and some consensus as to its value. Thus, the SNMP architecture and SNMPv3 security features rely heavily on SNMPv2u and SNMPv2*.
- Address the need for secure set request messages over real-world networks, rectifying the most important deficiency of SNMPv1 and SNMPv2.

- Design a modular architecture that will (1) allow implementation over a wide range of operational environments, some of which require minimal, inexpensive functionality and some of which may support additional features for managing large networks; (2) make it possible to move portions of the architecture forward in the standards track even if consensus has not been reached on all pieces; and (3) accommodate alternative security models.
- Keep SNMP as simple as possible, despite the many necessary and useful extensions.

On the basis of these design goals, developers have made the following design decisions (STAL99):

- *Architecture:* An architecture should be defined that identifies the conceptual boundaries between documents. Subsystems should be defined that describe the abstract services provided by specific portions of an SNMP framework. Abstract service interfaces, as described by service primitives, define the abstract boundaries between documents and the abstract services that are provided by the conceptual subsystems of an SNMP framework.
- *Self-contained documents:* Elements of procedure along with the MIB objects that are needed for processing a specific portion of an SNMP framework should be defined in the same document and, to the extent possible, should not be referenced in other documents. This allows pieces to be designed and documented as independent and self-contained parts that are consistent with the general SNMP MIB module approach. As portions of SNMP change over time, the documents describing other portions of SNMP are not directly affected. This modularity allows, for example, security models, authentication and privacy mechanisms, and message formats to be upgraded and supplemented as the need arises. The self-contained documents can move along the standards track on different time lines.
- *Remote configuration:* The security and access control subsystems add a whole new set of SNMP configuration parameters. The security subsystem also requires frequent changes of secrets at the various SNMP entities. To make this deployable in a large operational environment, these SNMP parameters must be able to be remotely configured.
- *Controlled complexity:* It is recognized that manufacturers of simple managed devices want to keep the resources used by SNMP to a minimum. At the same time, there is a need for more complex configurations that can devote more resources for SNMP and thus provide more functionality. In the design, there is an attempt to keep the competing requirements of these two environments in balance and allow the more complex environment to logically extend the simple environment.
- *Threats:* The security models in the security subsystem should protect against the principal threats, such as modification of information, masquerade, message stream modification, and disclosure. They do not need to protect against denial of service and traffic analysis.

SNMPv3 is secure against the following threats:

- *Modification of information:* Any entity can alter an in-transit message generated by an authorized entity in such a way as to effect unauthorized management operations, including the setting of object values. The essence of this threat is that an unauthorized entity can change any management parameter, including those related to configuration, operations, and accounting.
- *Masquerade:* An entity that does not have authorization for access to management operations may attempt to assume the identity of an authorized entity to engage in those operations.
- *Message stream modification:* SNMP is designed to operate over a connectionless transport protocol. There is a threat that SNMP messages could be reordered, delayed, duplicated, or replayed to effect unauthorized management operations. For example, a message to reboot a device could be copied and replayed later.
- *Disclosure:* An entity could observe exchanges between a manager and an agent and thereby learn the values of managed objects and learn of notifiable events. For example, observation of a set command that changes passwords would enable an attacker to learn the new passwords.

However, SNMPv3 is not intended to secure against denial of service (an attacker preventing exchanges between a manager and an agent) or traffic analysis (an attacker observing the general pattern of traffic between managers and agents).

In practice, few networks use SNMPv3. Those that rely on SNMPv2c (the most common variant) have already dealt with the lack of protocol support for security and administration and see little to be gained from an expensive upgrade process. Additionally, it is not uncommon for network device vendors to acknowledge that their version 2 agent implementation is much more "tried and trusted" than the newer and much more complex version 3 implementation, which sees little practical use.

3.3.3 Remote Monitoring (RMON1 and RMON2)

- Remote network monitoring is an IETF standard for monitoring and protocol analysis LANs (Ethernet and Token Ring). RMON1 focuses on Layer 1 and Layer 2, while RMON2 adds network- and application-layer concepts. A further extension, SMON, adds support for switched networks. RMON information is accessed via an SNMP manager through the RMON MIB.

The RMON MIB intends to bridge the gap between the limited services provided by management platforms and the rich sets of data and statistics provided by traffic monitors and analyzers. RMON defines the next generation of network monitoring with more comprehensive network fault diagnosis, planning, and performance tuning features than any current monitoring solution. The design goals for RMON are as follows (STAL99).

- *Off-line operation:* In order to reduce overhead associated with communication links, it may be necessary to limit or halt polling of a monitor by the manager. In general, the monitor should collect fault, performance, and configuration information continuously, even if it is not being polled by a manager. The monitor simply continues to accumulate statistics that may be retrieved by the manager at a later time. The monitor may also attempt to notify the manager if an exceptional event occurs.
- *Preemptive monitoring:* If the monitor has sufficient resources and the process is not disruptive, the monitor can continuously run diagnostics and log performance. In the event of a failure somewhere in the network, the monitor may be able to notify the manager and provide useful information for diagnosing the failure.
- *Problem detection and reporting:* Preemptive monitoring involves an active probing of the network and the consumption of network resources to check for error and exception conditions. Alternatively, the monitor can passively—without polling—recognize certain error conditions and other conditions, such as congestion and collisions, on the basis of the traffic that it observes. The monitor can be configured to continuously check for such conditions. When one of these conditions occurs, the monitor can log the condition and notify the manager.
- *Value-added data:* The network monitor can perform analyses specific to the data collected on its subnetworks, thus relieving the manager of this responsibility. The monitor can, for instance, observe which station generates the most traffic or errors in network segments. This type of information is not otherwise accessible to the manager not directly attached to the segment.
- *Multiple managers:* An internetworking configuration may have more than one manager in order to achieve reliability, perform different functions, and provide management capability to different units within an organization. The monitor can be configured to deal with more than one manager concurrently.

Table 3.3.3 summarizes the RMON MIB groups for Ethernet segments, and Table 3.3.4 defines the RMON MIB groups for Token Ring segments. At present, there are only a few monitors that can measure both types of segments using the same probe.

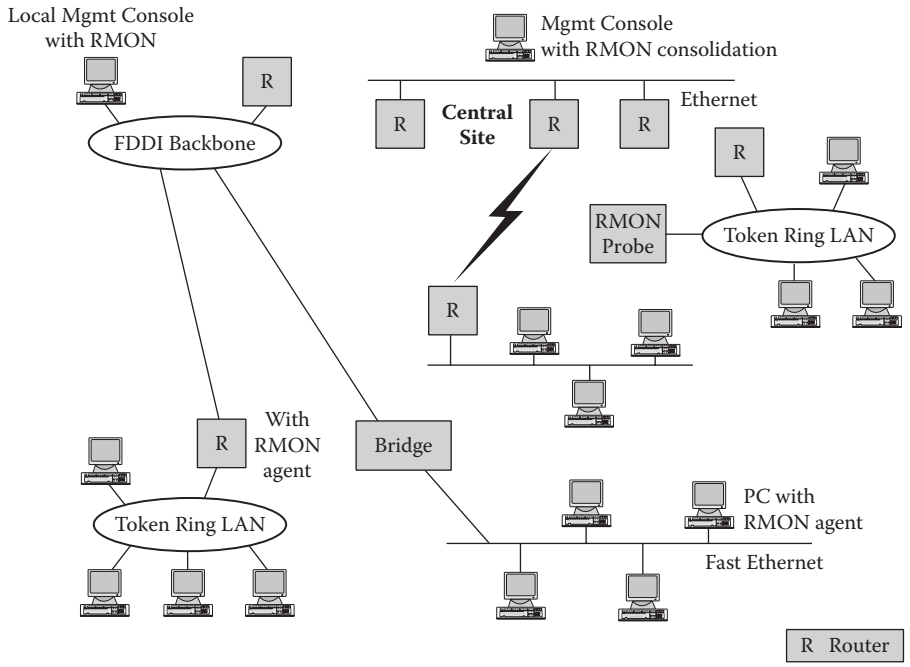
TABl E 3.3.3 RMON MIB Groups for Ethernet

Statistics group	Features a table that tracks approximately 20 different characteristics of traffic on the Ethernet LAN segment, including total octets and packets, oversized packets, and errors
History group	Allows a manager to establish the frequency and duration of traffic observation intervals, called buckets; the agent can then record the characteristics of traffic according to these bucket intervals
Alarm group	Permits the user to establish the criteria and thresholds that will prompt the agent to issue alarms
Host group	Organizes traffic statistics by each LAN node, based on time intervals set by the manager
HostTopN group	Allows the user to set up ordered lists and reports based on the highest statistics generated via the host group
Matrix group	Maintains two tables of traffic statistics based on pairs of communicating nodes; one is organized by sending node addresses, the other by receiving node addresses
Filter group	Allows a manager to define, by channel, particular characteristics of packets; a filter might instruct the agent, for example, to record packets with a value that indicates they contain DECnet messages
Packet capture group	Works with the filter group and allows the manager to specify the memory resources to be used for recording packets that meet the filter criteria
Event group	Allows the manager to specify a set of parameters or conditions to be observed by the agent; whenever these parameters or conditions occur, the agent will record an event into a log

RMON is rich in features, and there is a very real risk of overloading the monitor, the communication links, and the manager when all of the details are recorded, processed, and reported. The preferred solution is to do as much of the analysis as possible locally, at the monitor, and send just the aggregated data to the manager. This assumes powerful monitors. In other applications, monitors may be reprogrammed during operations by the managers. This is very useful when diagnosing problems. Even if the manager can define specific RMON requests, it is still necessary to be aware of the trade-offs involved. A complex

TABl E 3.3.4 RMON MIB Groups for Token Ring

Statistics group	Includes packets, octets, broadcasts, dropped packets, soft errors, and packet distribution statistics; statistics are at two levels: MAC for the protocol level and LLC statistics to measure traffic flow
History group	Contains long-term historical data for segment trend analysis; histories include both MAC and LLC statistics
Host group	Collects information on each host discovered on the segment
HostTopN group	Provides sorted statistics that allow reduction of network overhead by viewing only the most active hosts on each segment
Matrix group	Reports on traffic errors between any host pair for correlating conversations on the most active nodes
Ring station group	Collects general ring information and specific information for each station; general information includes ring state (normal, beacon, claim token, purge), active monitor, and number of active stations, and ring station information includes a variety of error counters, station status, insertion time, and last enter/exit time
Ring station order	Maps station MAC addresses to their order in the ring
Source routing statistics	Provides information on the number of frames and octets transmitted to and from the local ring in source-routing bridges; other information includes broadcasts per route and frame counters per hop
Alarm group	Reports changes in network characteristics based on thresholds for any or all MIBs; this allows RMON to be used as a proactive tool
Event group	Logs events on the basis of thresholds; events may be used to initiate functions such as data capture or instance counts to isolate specific segments of the network
Filter group	Defines packet matches for selective information capture; these include logical operations (AND, OR, NOT) so that network events can be specified for data capture, alarms, and statistics
Packet capture group	Stores packets that match filtering specifications



FIGu RE 3.3.5 RMON probes in LAN segments.

filter will allow the monitor to capture and report a limited amount of data, thus avoiding overhead on the network. However, complex filters consume processing power at the monitor; if too many filters are implemented, the monitor will become overloaded. This is particularly true if the network segments are busy, which is probably the time when measurement is most valuable.

Figure 3.3.5 shows the RMON probes in the segments. These probes can be implemented in three different ways: as stand-alone monitors; as modules of hubs, routers, and switches; and as software modules in Unix, NT operating systems, or PC workstations. Each of these alternatives involves benefits and disadvantages.

3.3.3.1 Probe as a Stand-Alone Monitor

The benefits of this alternative are:

- Performance is excellent.
- All functions are supported.
- Various options are available, such as stackable or rack-mountable.

The disadvantages are:

- Costs are high for an average LAN segment.
- Multiple probes are required when segmentation of LANs is deployed by switches without using probe ports.
- Most advanced LAN technologies might not be supported immediately.

3.3.3.2 Probe as a Module of Hubs, Routers, and Switches

The benefits of this alternative are:

- It is a convenient solution because the networking components have already been deployed.
- Costs are much lower than with stand-alone probes.

- Integration of probes into switches is less expensive than deploying an individual probe in each switched segment.

The disadvantages are:

- Networking components need upgrades or customization to support the probes.
- Upgrades are not always economical.
- Performance might become a problem.
- Conformance to standards may not be sufficient.
- Problems with the probe may affect the performance of the networking component.
- RMON functionality may be limited; not all RMON indicator groups are supported.
- Processing programs are very simple in comparison to stand-alone probes.
- Integration with management platforms is usually incomplete.
- RMON modules may be provided by different vendors, possibly leading to incompatibility problems.

3.3.3.3 Probe as a Software Module in Unix, NT Operating Systems, or PC Workstations

The benefits of this alternative are:

- Costs are much lower than with other alternatives.
- Performance is good when RISC or Pentium processors are used.
- Scalability and extendibility are excellent.
- Support of state-of-the-art technology is easier than with other alternatives.
- It is possible to use a combination of supervising Ethernet and Token Ring.
- Outband access to probes is possible with proper configuration.

The disadvantages are:

- Purchase of adapters and additional workstations may be required.
- The user is responsible for the installation.

The extension of this alternative might also be utilized for switched LANs. RMON can be installed into the adapters of end-user devices. Usually, only the filter and packet capture groups are supported at the end-user device level. The other groups are supported by the collector. Overhead is minimal, as are performance effects in the switch and in end-user devices.

RMON probes are extremely helpful in collecting data on Web site access frequency and activities. Independently, depending on how probes are implemented, vendors of probes are expected to work together. Standards are continuously being improved to offer even more functionality with respect to capturing and processing Web site–performance-related data.

The existing, widely used RMON version is basically a MAC standard. It does not give LAN managers visibility into conversations across the network or connectivity between various network segments. The extended standard is targeting the network layer and higher. It will provide visibility across the enterprise. With remote access and distributed workgroups, there is substantial intersegment traffic. The following functionalities are included.

- **Protocol distribution and protocol directory table:** The issue here was how to provide a mechanism that will support the large number of protocols running on any one network. Current implementations of RMON employ a protocol filter that analyzes only the essential protocols. RMON2, however, will employ a protocol directory system that allows RMON2 applications to define which protocols an agent will use. The protocol directory table will specify the various protocols an RMON2 probe can interpret.
- **Address mapping:** This feature matches each network address with a specific port to which the hosts are attached. It also identifies traffic-generating nodes/hosts by MAC, Token Ring, or Ethernet address, helping to identify specific patterns of network traffic. It is useful in node

discovery and network topology configurations as well. In addition, the address translation feature provides duplicate IP address detection, resolving a common trouble spot with network routers and virtual LANs.

- **Network-layer host table:** This table tracks packets, errors, and bytes for each host according to a network-layer protocol. It permits decoding of packets based on their network layer address, in essence allowing network managers to look beyond the router at each of the hosts configured on the network.
- **Network-layer matrix table:** This table tracks the number of packets sent between a pair of hosts by a network layer protocol. The network manager can identify network problems quicker by using this table, which shows the protocol-specific traffic between communicating pairs of systems.
- **Application-layer host table:** This table tracks packets, errors, and bytes by host on an application-specific basis (e.g., Lotus Notes, e-mail, Web). Both the application-layer host table and the matrix table trace packet activity of a particular application. This feature can be used by network managers to charge users on the basis of how much network bandwidth was used in their applications.
- **Application-layer matrix table:** This table tracks packet activity between pairs of hosts by application (e.g., pairs of hosts exchanging Internet information).
- **Probe configuration:** Currently, vendors offer a variety of proprietary means for configuring and controlling their respective probes. This complicates interoperability. The probe configuration specification, based on the Aspen MIB, defines standard parameters—such as network address, SNMP error trap destinations, modern communications with probes, serial line information, and downloading of data to probes—for remotely configuring probes. It provides enhanced interoperability between probes by specifying standard parameters for operations, providing one vendor's RMON application with the ability to remotely configure another vendor's RMON probe.
- **u ser history collection group:** The RMON2 history group polls, filters, and stores statistics based on user-defined variables, creating a log of the data for use as a historical tracking tool. This is in contrast to RMON1, where historical data is gathered on a predefined set of statistics.

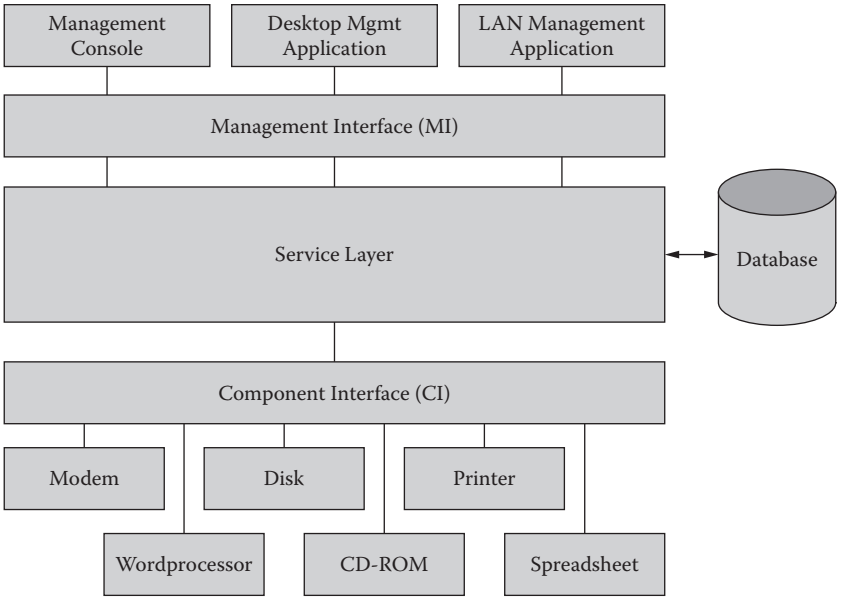
After implementation, steadily increasing amounts of complete information will be available for performance analysis and capacity planning.

3.3.4 Desktop Management Interface

Basically, SNMP is utilized to manage systems, assuming that system components accommodate SNMP agents. However, there are no MIBs as of yet that describe principal indicators for management purposes. An important emerging standard for desktop management is the Desktop Management Interface (DMI). According to the Desktop Management Task Force (DMTF), the DMI should accomplish the following goals:

- Enable and facilitate desktop, local, and network management
- Solve software overlap and storage problems
- Create a standard method for management of hardware and software components using MIFs (management information formats)
- Provide a common interface for managing desktop computers and their components
- Provide a simple method to describe, access, and manage desktop components

DMI is independent of the hardware, OS, and management protocol used to encapsulate it, and it can be used over a network or in stand-alone mode. Mappings are provided for easy adaptation into existing management protocols (e.g., SNMP). The scope of management under DMTF includes CPUs, I/Os, motherboards, video cards, network interface cards, faxes, modems, mass storage devices, printers, and applications. Figure 3.3.6 shows the structure of this standard. There is a clear separation between the



FIGu RE 3.3.6 Structure of the desktop management interface.

managed components (CI, or component interface) and the services offered for management (MI, or management interface).

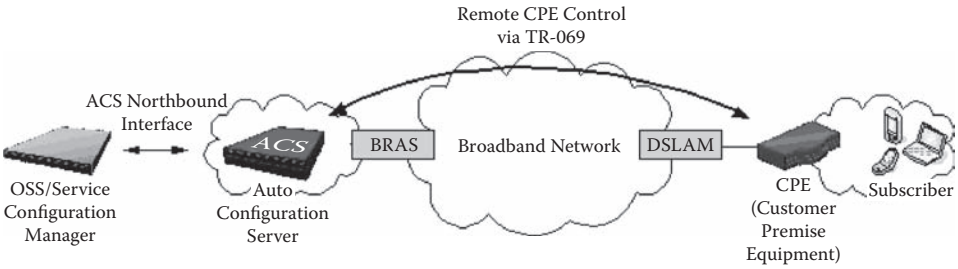
The DMI has four primary components:

- *Management information format (MIF)*: The MIF is a textual description of the hardware and software components of a particular PC, structured into groups of attributes with different values. Each MIF file must contain the special ID group, which includes the product name, version, serial number, and the time and date of the last installation. MIF files are collected in the central MIF database.
- *Component interface (CI):* The CI is an application programming interface (API) that can be used to manipulate the system's MIF file and send Events. Commands include the Get and Set commands to query and modify the MIF and the Event command to send notifications to the management software.
- *Management interface (MI):* The MI is also an API that allows administrators to issue Get, Set, and List commands to list and access all DMI-manageable devices.
- *Service layer:* The service layer is the *DMI agent*, a memory-resident code that accesses both the MI and CI and allows management and component software to access the MIF database. The service layer is provided as operating system software (driver) and is accessible by all programs executed by the OS.

3.3.5 TR-069 CPE WAN Management Protocol

3.3.5.1 Overview

The TR-069 protocol, defined and promoted by the DSL Forum, is a communication standard between customer premise equipment (CPE) deployed over a broadband TCP/IP connection and an autoconfiguration server (ACS) operated by the broadband service provider. The protocol defines methods for the automatic configuration and management of CPE devices such as DSL modems, home gateways, VoIP devices, and set-top boxes. The scope of TR-069 is not limited to DSL or related devices. The ACS



FIGu RE 3.3.7 The CPE WAN management protocol.

provides basic device management (e.g., firmware updates and device status) as well as dynamic service provisioning and configuration (e.g., setting service provider information or enabling functions such as parental control).

Figure 3.3.7 shows the components of the CPE WAN management protocol.

3.3.5.2 Functional Components

The protocol consists of the following primary functional components:

- *Autoconfiguration and dynamic service provisioning:* Initially, CPEs connect to an ACS and obtain their initial configuration. The URL of the ACS may be preconfigured or obtained via ACS discovery. During device operation, dynamic reprovisioning may be initiated by either side asynchronously. The protocol defines a number of standard elements (settings) and allows a wide range of vendor extensions. In the case of ACS-initiated provisioning, strong support for mass provisioning is provided in the form of versatile CPE criteria for selection.
- *Software/firmware image management:* TR-069 provides tools for downloading firmware images and software components to CPE devices using digitally signed transfer files. Firmware or software package version control is also included. Downloads may be initiated by either the ACS or CPE device. The ACS is notified about the success or failure of the download operation.
- *Status and performance monitoring:* A common set of status/performance parameters for CPE devices is defined and made accessible. Vendor-specific extension of this parameter set is also supported.
- *Diagnostics:* Similar to status and performance monitoring, an extensible common set is defined for device and service diagnostics.
- *Identity management for Web applications:* Web services accessed from within the CPE device's local network can obtain additional information from the TR-069-managed device (e.g., home access router associated with subscriber information).

3.3.5.3 Communication Protocol and Security

TR-069 uses RPC (remote procedure call) mechanisms for communication, implemented over Simple Object Access Protocol (SOAP)/HTTP. Use of Secure Sockets Layer/Transport Layer Security (SSL/TLS) for authentication and encryption is recommended by the protocol, as it is intended for WAN environments with no physical security guarantees.

Using certificate-based SSL authentication, retail-distributed CPE devices can authenticate automatically discovered ACSs and connect to trusted ones operated by the service provider. Alternatively, different security mechanisms (such as shared secret–based encryption) can be used. The protocol includes additional provisions for secure communication as well (such as the signed voucher and signed package format mechanisms discussed below). These mechanisms enable secure communication independent of the transport layer security provided by the ACS operator. Thus, secure third-party services (e.g., for providers of third-party services) can be created and maintained.

TABl E 3.3.5 The TR-069 Protocol Stack Encapsulation Scheme

Layer	Description
CPE/ACS management application	Local application (not specified within TR-069) using protocol services
RPC	TR-069-defined RPC methods and data structures
SOAP	SOAP 1.1; encodes remote procedure calls
HTTP	HTTP 1.1; transfer protocol for SOAP
SSL/TLS	SSL 3.0 or TLS 1.0 standard transport layer security protocols; their use is recommended by TR-069 but not mandatory
TCP/IP	Standard TCP/IP

3.3.5.4 Architecture

The TR-069 protocol stack uses the encapsulation scheme shown in Table 3.3.5. The protocol uses a number of architecture components in order to implement the functional components listed previously. These components are as follows:

3.3.5.4.1 Parameters

TR-069 defines a generic mechanism (defined by specific RPC methods) for getting/setting CPE device parameters by the ACS. These parameters can be used to configure the device or obtain monitoring, statistics, and status data. Parameters have textual names following a hierarchical naming convention, similar to a file name within a directory structure. The separator character in TR-069 is a dot. Parameters are typed, and TR-069 supports seven types: *object* (a container for other parameters), *string*, *int*, *unsignedint*, *boolean*, *dateTime*, and *base64* (binary). IP and MAC addresses are represented as strings. For example,

InternetGatewayDevice.LANDevice.2.LAN-EthernetInterfaceConfig.3.Stats

has the type "object" and contains statistical information for the second LAN device's third Ethernet interface. Members of this object are *unsignedint* variables such as

InternetGatewayDevice.LANDevice.2.LAN-EthernetInterfaceConfig.3.Stats.BytesSent.

Parameters can be defined as read-only or read-write. A discovery mechanism (*GetParameterNames* RPC) is also defined for the ACS to allow discovery of supported parameters by the managed device.

3.3.5.4.2 File Transfers

The protocol defines an RPC mechanism for downloading or uploading files between the device and the ACS. The standard supports designating HTTP, HTTPS, FTP, and TFTP as transport protocols (only HTTP support is mandatory).

3.3.5.4.3 CPE-Initiated Notifications

This protocol element allows the CPE device to notify the ACS about various conditions and to ensure that periodic CPE–ACS communication takes place. Initial "bootstrap" communication (the CPE "signing in" to the ACS for the first time) is also defined within this element. Each time a CPE device initiates a notification, it identifies itself by supplying manufacturer information and serial number.

3.3.5.4.4 Asynchronous ACS-Initiated Notifications

This protocol element enables the ACS to notify managed devices about configuration changes, allowing for near-real-time reconfiguration.

3.3.5.5 Procedures

In this section, a number of procedures for implementing the functionality and architecture components are discussed.

3.3.5.5.1 ACS Discovery

TR-069 defines three possible scenarios for a managed device to locate an ACS after reboot or initial deployment. The address of an ACS is always a valid HTTP(S) URL. If an HTTPS URL is specified, a non-HTTPS-capable client will attempt the same URL with an unencrypted HTTP. ACS discovery scenarios are as follows:

- Default URL stored with the device's internal configuration.
- Locally configured ACS URL (e.g., from the LAN side of the device).
- ACS discovery using DHCP: The CPE asking for a DHCP address uses DHCP Option 60 (Vendor Class Identifier) and includes the string "dslforum.org" within the ID string. The DHCP server includes Option 43 (Vendor Specific Information) in the response and encodes the *URL of the ACS* and (optionally) a *ProvisioningCode* string, which is used to select the primary service provider. These values can also be accessed (and set) later by the ACS as parameters InternetGatewayDevice.ManagementServer.URL and InternetGatewayDevice.DeviceInfo. ProvisioningCode. If this mechanism (ACS discovery) is used, the CPE may disable the local configuration of ACS and ProvisioningCode.

3.3.5.5.2 Connection Establishment

The CPE may initiate connection establishment with the ACS at any point. In the following cases, the CPE device is mandated to establish connection and issue the *Inform* RPC method:

- Initial installation/first-time activation on the access network.
- On power-up or reset.
- Periodically within each *PeriodicInformInterval* (configurable parameter) or as instructed by the ACS via the *ScheduleInform* method.
- On receiving a valid Connection Request from the ACS.
- Whenever the ACS URL changes.
- Whenever a parameter value is modified, which requires an *Inform* call to be issued. For example, such parameters are IP address (management address or broadband address), provisioning code, and software version.
- Parameters can be marked for "active notification." Upon the change of such parameter values, an Inform notification must be sent unless the change was caused by the ACS itself.

The ACS can also initiate a connection with the CPE device using the Connection Request notification. Responding to this, the CPE device *must* perform a connection initiation as discussed above. The Connection Request notification is an HTTP Get request to an URL serviced by the CPE device (HTTPS should *not* be used). The CPE device should limit the number of Connection Requests it accepts during a specified period of time in order to avoid denial of service attacks.

Not all CPE devices are accessible from the ACS (e.g., as a result of NAT or firewalls). In this case, ACS-initiated connection cannot be used, and the protocol relies on the CPE to connect within *PeriodicInformInterval*.

3.3.5.6 Protocol Encoding and Sessions

The protocol uses W3C SOAP 1.1 for RPC. SOAP is implemented over HTTP 1.1, where the CPE device is the client and the ACS is the server. The standard recommends using SSL or TLS (for which the required version is SSL 3.0 or TLS 1.0).

If no encryption (SSL or TLS) is used, the ACS must use HTTP digest authentication.

The role and number of SOAP envelopes during a session are as follows:

- A SOAP request from an ACS to a CPE device is sent over an HTTP response. The CPE answers this request within a subsequent HTTP post.
- Each HTTP message may contain more than one SOAP envelope. These envelopes are independent of each other.
- When there are multiple SOAP envelopes within an HTTP message, the *SOAPAction* header must be empty.

The maximum number of SOAP envelopes supported by a CPE device is encoded within the *Inform* message initially sent to the ACS. The maximum number of envelopes accepted by the ACS is encoded within the *Inform* response.

Use of the *ID* tag within a SOAP header to associate responses with requests is optional. SOAP request envelopes are processed and answered in sequence, even if there are multiple envelopes within an HTTP message.

3.3.5.6.1 Sessions and File Transfers

The CPE device should maintain a TCP connection throughout the duration transactions forming a single session. If this is not possible (i.e., across an HTTP proxy), the ACS should use HTTP session cookies. The CPE device cannot be expected to store the cookies beyond a single session's duration.

If the CPE is instructed to perform a download or upload operation, it can use the session's TCP connection to transfer the file only if the file's URL resolves to the IP address of the ACS. In any other case, a new session must be opened, either while maintaining the current session with the ACS or closing it and performing the transfer.

3.3.5.6.2 Authentication

CPEs must be authenticated, either via SSL/TLS or HTTP. If SSL/TLS is being used with no authentication, HTTP Basic or Digest authentication must be performed in order to identify the CPE device. With no SSL/TLS, only Digest authentication is acceptable. In any case, each CPE has to have a unique identifier (username) and password. The password is a shared secret; thus, it must be unique among CPEs. TR-069 defines a recommended format for the username and does not specify how to initially share the password between the ACS and CPEs. However, it does recommend a format for the username string, which is "OUI-SERIAL." OUI is a six-digit hexadecimal "IEEE Organizational ID," and SERIAL is the unit's serial number.

3.3.5.6.3 SOAP Encoding

TR-069 RPC messages are encoded using standard SOAP 1.1:

- The envelope namespace identifier is http://schemas.xmlsoap.org/soap/envelope/.
- The serialization namespace identifier is http://schemas.xmlsoap.org/soap/encoding/.
- The namespace for TR-069 elements and attributes is urn:dslforum-org:cwmp-1-0.
- TR-069 data types correspond directly to SOAP 1.1 data types.
- Faults are indicated by SOAP fault elements with the following conventions: *faultcode* indicates the source of the fault (in this context, Client is the request's originator, and Server is the responder); the *faultstring* sub-element must contain the string *CWMP fault*; and fault details are contained in the *Fault* structure, which is defined by TR-069 within the CWMP namespace.

TR-069 utilizes three SOAP header fields:

- *ID:* This can be used to associate requests and responses.
- *HoldRequests:* An ACS may regulate traffic from CPE devices using this boolean field. When set, the CPE must not send any further requests within the session.
- *NoMoreRequests:* This can be used by any of the peers to indicate that the sender will not send more requests during the remainder of the session.

3.3.5.6.4 Signed Vouchers and Signed Package Format

A *voucher* is a digitally signed data structure that is used to enable or disable a set of Options on the CPE. An Option is an optional capability or feature of the CPE whose enabling or disabling requires secure tracking (e.g., linked to a payment). Options can be Enabled, Disabled, or Enabled with Expiration. Vouchers can also be digitally signed, and the certificate used for this signature can be different from that used for the ACS-CPE communication.

The signed package format defines a file format that can be used to securely download files to a recipient device. One or more files can be encapsulated into a signed package, and apart from the digital signature authenticating the source, the package may also contain instructions for the extraction and installation of the contents. The package includes the following fields:

- *Header:* includes the preamble (magic), version number, and lengths of command and payload parts
- *Command list:* commands for extracting and installing the payload (in the form of type-length-value abstract commands)
- *Signatures:* a PKCS#7 digital signature block containing one or more signatures
- *Payload files:* files in user-specific formats

The protocol defines an extensive, device-independent command set to be used in the Command List, defining 25 various commands such as Extract, Add File, Move File, Format File System, and Required Minimum Storage. The command set also supports querying version information on the host system and extracting/installing files depending on the version numbers obtained.

3.3.6 Telecommunications Management Network

The Telecommunication Management Network (TMN) is a special network in its own right that is implemented to help manage service providers' telecommunication networks. As such, it interfaces to one or more individual networks at several points in order to exchange information. It is logically separate from the networks it manages, and it may be physically separate as well. However, TMN may use parts of the telecommunication networks for its own communications.

The TMN effort is chartered by the International Telecommunications Union Telecommunications Standardization Sector (ITU-TS). Development began in 1988 and has concentrated primarily on the overall architecture, using the Synchronous Digital Hierarchy (the international version of the North American Synchronous Optical Network, or SONET) technology as a target. However, the TMN techniques are applicable to a broad range of technologies and services.

TMN is an extension of the OSI standardization process. It attempts to standardize some of the functionality and many of the interfaces of managed networks. When fully implemented, the result will be a higher level of integration. TMN is usually described via three architectures:

- The functional architecture describes the appropriate distribution of functionality within TMN, appropriate in the sense of allowing for the creation of function blocks from which a TMN of any complexity can be implemented. The definition of function blocks and reference points between them leads to the requirements for the TMN-recommended interface specifications.
- The information architecture, based on an object-oriented approach, gives the rationale for the application of OSI systems management principles to the TMN principles. The OSI systems management principles are mapped onto the TMN principles and, when necessary, are expanded to fit the TMN environment.
- The physical architecture describes interfaces that can actually be implemented together with examples of physical components that make up TMN.

The management functions are grouped into the five functional areas identified as part of the OSI model. Examples are:

- Fault management (alarm surveillance, testing, trouble administration)
- Configuration management (provisioning, rating)
- Performance management (monitoring quality of service [QoS], traffic control)
- Security management (managing access and authentication)
- Accounting management (rating and billing)

The management requirements that helped shape the TMN specifications address planning, provisioning, installing, maintaining, operating, and administering communications networks and services.

The TMN specifications use standard CMIP application services wherever appropriate. However, one of the key concepts of the TMN specifications is their introduction of *technology-independent* management, which is based on an abstract view of the managed network elements. Diverse equipment can be managed through this abstract view and a single communications interface. Thus, TMN-managed networks can consist of both TMN-conforming and TMN-nonconforming devices.

The TMN specifications define an intended direction, with many details to be determined. The published TMN specifications address the overall architecture, the generic information model, management services, management functions, management and transmission protocols, and an alarm surveillance function. The TMN working groups will next focus on the service layer, traffic (i.e., congestion), and network-level management.

The relatively slow pace of TMN specification development has not prevented companies from recognizing the benefits of the TMN approach to management. The TeleManagement Forum is incorporating TMN into its specifications, and many companies are beginning to build, or specify, management systems and components that comply with TMN principles. Management systems that comply with these principles can reduce costs and improve services for the following reasons (Terplan 2001):

- Standard interfaces and objects make it possible to rapidly and economically deploy new services.
- Distributed management intelligence minimizes management reaction times to network events.
- Mediation makes it possible to handle similar devices in an identical manner, leading to more generic operations systems and vendor independence.
- Mediation makes it possible to manage and transparently upgrade the existing device inventory.
- Distributed management functions increase scalability.
- Distributed management functions isolate and contain network faults.
- Distributed management functions reduce network management traffic and the load on operations systems.

Many of the benefits that accrue from the TMN principles are due directly to TMN's distributed architecture and its mediation function.

The TMN architecture addresses communications networks and services as collections of cooperating systems. By managing individual systems, TMN has a coordinated effect on the network.

This coordination can be illustrated through a simple example. Within an enterprise, one operations system may deal with the network-element inventory, another may deal with traffic planning, and several element managers may deal with network elements of various types. When a customer requires a circuit of specific bandwidth and quality of service, all of these systems must be coordinated to meet the customer's needs. The TMN architecture not only facilitates this effort but allows it to be distributed among several systems. This distribution has the side effect of allowing a TMN-based system to scale to handle global networks by enabling the workload to be spread across multiple systems.

This ability to subdivide and distribute the total management effort requires clear definition of functions, interfaces, and the information model. These topics are defined in the TMN specifications and are outlined in the paragraphs to follow.

The TMN architecture identifies specific functions and their interfaces. These functions allow a TMN to perform its management activities. The TMN architecture provides flexibility in building a management system by allowing certain functions to be combined within a physical entity. The following

function blocks, along with their typical methods of physical realization, are defined within the TMN specifications (Terplan 2001):

3.3.6.1 Operations Systems Function (OSF)

The OSF monitors, coordinates, and controls the TMN entities. It is a TMN-compliant management system or set of management applications. The system has to enable general activities such as management of performance, faults, configuration, accounting, and security to be performed. In addition, specific capabilities for planning of operations, administration, maintenance, and provisioning of communications networks and systems should be available. These capabilities are realized in an operations system. Such a system can be implemented in many different ways. One possibility is a descending abstraction (e.g., business, service, and network) wherein the overall business needs of the enterprise are met by coordinating the underlying services. In turn, individual services are realized through coordinating network resources.

3.3.6.1 Workstation Function (WSF)

The WSF provides TMN information to the user. This typically involves such elements as access control, topological map displays, and graphical interfaces. These functions are realized in a workstation.

3.3.6.2 Mediation Function (MF)

This function acts on information passing between an OSF and a network element function or Q adapter function to ensure that the data emitted by the MF complies with the needs and capabilities of the receiver. MFs can store, adapt, filter, threshold, and condense information. They provide the abstract view necessary to treat dissimilar elements in a similar manner. MFs may also provide local management for their associated network element functions (in other words, MFs may include an element manager). The MF function is realized in a mediation device. Mediation can be implemented as a hierarchy of cascaded devices using standard interfaces. The cascading of mediation devices and the various interconnections to network elements provide a TMN with a great deal of flexibility. This also allows for future design of new equipment to support a greater level of processing within the network element, without the need to redesign an existing TMN.

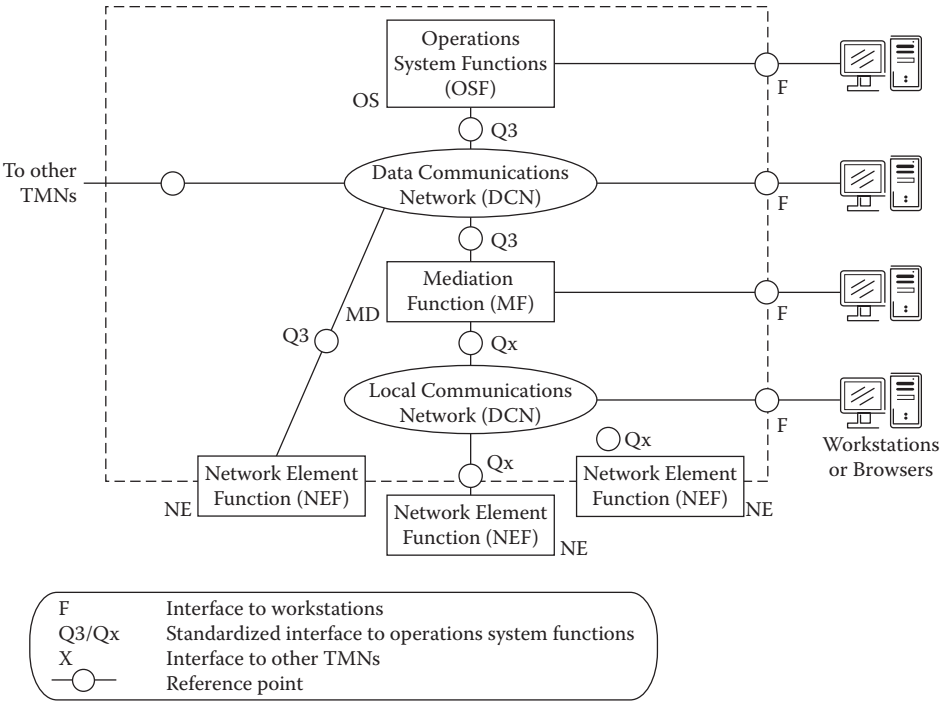
3.3.6.3 Q Adapter Function (QAF)

The QAF connects non-TMN-compliant network element functions to the TMN environment and is realized in a Q adapter. A Q adapter allows legacy devices (i.e., those that do not support the TMN management protocols, including SNMP devices) to be accommodated within a TMN. A Q adapter typically performs interface conversion functions (i.e., it acts as a proxy).

3.3.6.4 Network Element Function (NEF)

This function is realized in the network elements themselves, which can present a TMN-compliant or TMN-noncompliant interface. Examples include physical elements (switches), logical elements (virtual circuit connections), and services (operations systems software applications). Figure 3.3.8 illustrates the functions within a TMN environment. The portions that are outside the TMN environment are not subject to standardization. For example, the human-interface portion of the workstation function is not specified in the TMN standard.

Within the TMN specification are well-defined reference points to identify the characteristics of the interfaces between the function blocks. These reference points identify the information that passes between the function blocks. The function blocks exchange information using the data communications function (DCF). The DCF can perform routing, relaying, and internetworking, acting at Open Systems Interconnection (OSI) Layers 1 to 3 (i.e., physical, data link, and network layers) or their equivalents. These functions are performed in the data communications network. Figure 3.3.8 also shows the reference points that have been defined in the TMN specification. These reference points, characterized by the information shared between their endpoints, are as follows:



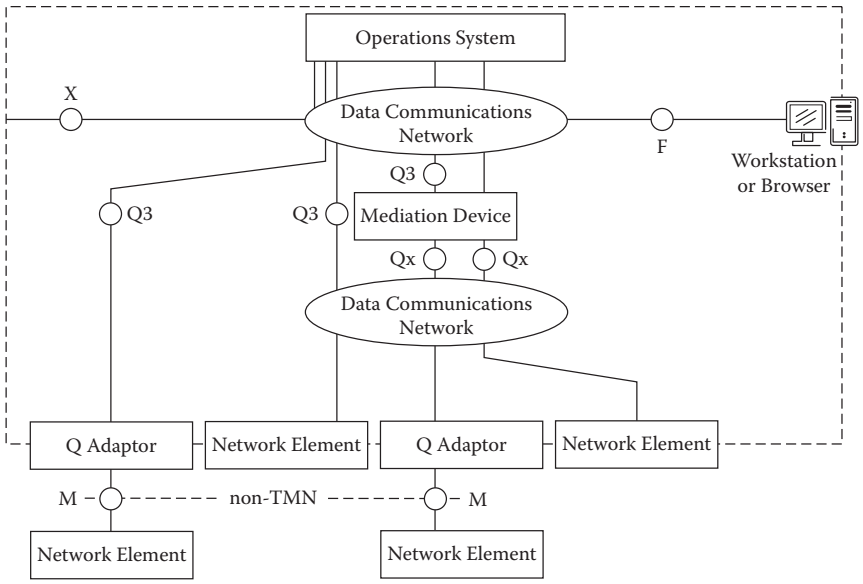
FIGu RE 3.3.8 Functions within the TMN architecture.

- F is the interface between a workstation, an operations system, and a mediation device.
- G is the interface between a workstation and a human user. The specification of this interface is outside the scope of TMN.
- M is the interface between a Q adapter and a non-TMN-compliant network element. This interface, which is not specified by TMN, is actually one of the most important, as today's networks primarily consist of devices that do not comply with the TMN standard.
- Qx is the interface between a Q adapter and a mediation device, between a TMN compliant element and a mediation device, and between two mediation devices.
- Q3 is the interface between a TMN-compliant element and an operations system, between a Q adapter and a mediation device, between a mediation device and an operations system, and between two operations systems.
- X is the interface between operations systems in different TMNs. The operations system outside the X interface may be part of either a TMN environment or a non-TMN environment. This interface may require increased security over the level required by the Q interfaces. Access limitations may also be imposed.

Currently, only the Q3 interface has been specified to any degree of detail. The definition includes its management protocol (CMIP), alarm surveillance capabilities, and operations on the generic model used to describe the network. Alarm surveillance refers to a set of functions that enable the monitoring and interrogation of the network concerning alarm-related events or conditions.

The information model presents an abstraction of the management aspects of network resources and the related support management activities. This model consists of the management-protocol object classes required to manage a TMN. Information about these objects is what is exchanged across the TMN-standard interfaces.

The TMN specifications provide a generic information model that is technology independent. This independence allows management of diverse equipment in a common manner, through an abstract view



FIGu RE 3.3.9 Sample TMN physical architecture and interfaces.

of the network elements. This concept is vital for TMN to achieve its goals. The generic information model also serves as a basis for defining technology-specific object classes. While enabling more precise management, the resulting specific object classes still support a technology-independent view. For example, a TMN definition of a switch could be used to perform common management activities such as provisioning and performance gathering. In addition, this generic switch definition could be extended to cover the peculiarities of a particular vendor's switch. The extended definition could be used for such specialized activities as controlling the execution of diagnostic routines. TMN generic modeling techniques can be used by a resource provider or management-system provider to define its own objects.

The TMN information model is common to managed communications networks. It can be used to generically define the resources, actions, and events that exist in a network.

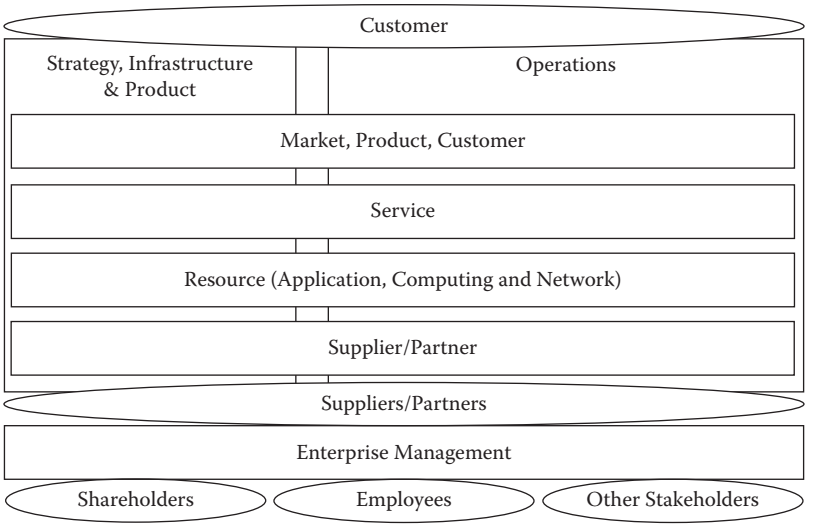
The TMN architecture is an excellent means of visualizing network and service management solutions. Figure 3.3.9 shows a simple solution (Terplan 2001). This figure represents a typical network management system (NMS)-layer solution integrating various element management systems (EMSs). Additional management applications help support Session Management Layer (SML) and Business Management Layer (BML).

The visualization technology may be used to support various regional network management systems. On top of these network management systems, service-oriented messages, events, and alarms can be extracted and displayed in various service centers.

3.3.7 TOM and eTOM

3.3.7.1 Enhanced Telecommunications Operations Map (eTOM)

The eTOM business process framework serves as the blueprint for process direction and the starting point for development and integration of Business Support Systems and Operations Support Systems; also, it helps drive TM Forum members' work in regard to developing NGOSS solutions. It offers service providers a neutral reference point as they consider internal process reengineering needs, partnerships, alliances, and general working agreements with other providers. For suppliers, the eTOM framework outlines potential boundaries of software components and the required functions, inputs, and outputs that must be supported by products.



FIGu RE 3.3.10 eTOM business process framework: Level 0 processes.

The eTOM framework provides the enterprise processes required for a service provider. However, it is not a service provider business model. It does not address the strategic issues or questions of who a service provider's target customers should be, what market segments should be served, what the provider's vision and mission should be, and so forth. A business process framework is one part of the strategic business model and plan for a service provider.

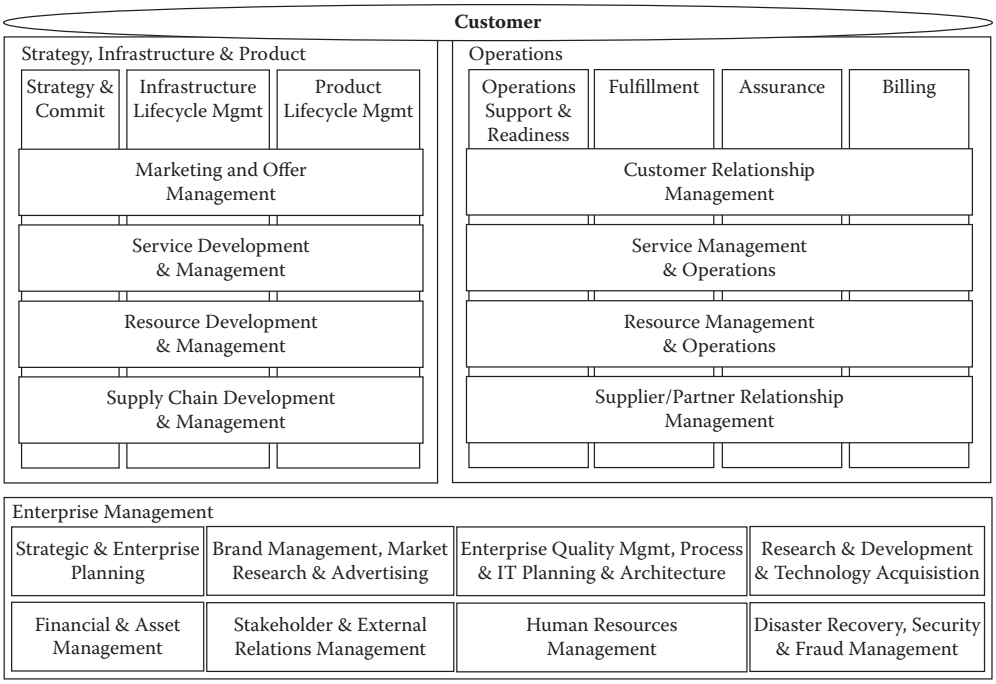
Figure 3.3.10 presents a detailed conceptual view of the eTOM business process framework. This view provides an overall context that differentiates strategy and life-cycle processes from the operations processes in two large groupings shown as two separate boxes. It also differentiates the key functional areas in five horizontal layers. In addition, the figure illustrates the internal and external entities that interact with the enterprise.

Figure 3.3.11 shows the view of Level 1 processes; this detail is needed to position and analyze business processes. The figure presents seven vertical process groups. These are the end-to-end processes required to support customers and manage a business. The focal point of eTOM is the core customer operations processes of fulfillment, assurance, and billing (FAB). Operations Support and Readiness is now differentiated from FAB real-time processes to increase the focus on enabling support and automation in FAB (i.e., online immediate support of customers). The Strategy and Commit vertical, as well as the two Lifecycle Management verticals, are also now differentiated because, unlike operations, they do not directly support the customer, are intrinsically different from the operations processes, and work on different business time cycles.

The horizontal process groupings in Figure 3.3.11 distinguish functional operations processes and other types of business functional processes (e.g., marketing versus selling, service development versus service configuration). The functional processes on the left (within the Strategy and Commit, Infrastructure Lifecycle Management, and Product Lifecycle Management vertical process groupings) enable, support, and direct the work in the operations verticals.

In summary, eTOM has led to a number of improvements:

- It has expanded scope to all enterprise processes.
- It distinctly identifies marketing processes owing to heightened importance in an e-business world.
- It distinctly identifies enterprise management processes so that all members of the enterprise are able to identify their critical processes, thereby enabling process framework acceptance across the enterprise.



FIGu RE 3.3.11 eTOM business process framework: Level 1 processes.

- It brings FAB to a high-level framework view emphasizing customer priority processes as the focus of the enterprise.
- It defines an Operations Support and Readiness vertical process grouping applicable to all functional layers other than enterprise management. To integrate e-business and make customer self-management a reality, the enterprise must understand the processes it needs to enable for direct—and increasing—online customer operations support and customer self-management.
- It recognizes three enterprise process groupings that are distinctly different from operations processes (i.e., Strategy and Commit, Infrastructure Lifecycle Management, and Product Lifecycle Management).
- It recognizes the different cycle times of strategy and lifecycle management processes and the need to separate these processes from the customer priority operations processes where automation is most critical. This is done by decoupling the Strategy and Commit and the two Lifecycle Management processes when automation is most critical (i.e., the day-to-day, minute-to-minute cycle times of customer operations processes).
- It moves from customer care or service orientation to a customer relationship management (CRM) orientation that emphasizes customer self-management and control, increasing the value customers offer to the enterprise and the use of information to provide customization and personalization for individual customers. It adds more elements to this customer operations functional layer to better represent selling processes and to integrate marketing fulfillment within CRM.
- It acknowledges the need to manage resources across technologies (i.e., application, computing, and network) by integrating the network and systems management functional process into resource management and operations. It also moves management of IT into this functional layer as opposed to having an outbound process grouping.

3.3.8 OSS-J

OSS-J is a TM Forum initiative intended to provide standardized NGOSS interfaces (APIs) and design guidelines for the development and integration of OSS/BSS systems based on component-off-the-shelf (COTS) components. OSS-J defines 12 APIs—Common, Trouble Ticket, Inventory, Service Activation, Billing Mediation, QoS, Fault Management, Order Management, Discovery, SQM, Pricing, and Performance Management—for multiple integration profiles: Java, XML, and Web services. Each integration profile contains specifications, a reference implementation, and a conformance test suite (TCK). APIs are developed as a community effort and are available at no charge.

OSS-J APIs are grouped into different *views*:

- *SOA Enablement:* contains the subset of APIs designed to facilitate the deployment of SOA-ready environments
- *Prosspero Approved APIs:* include APIs that have been approved as part of Prosspero packages
- *Prosspero Solution Center:* APIs that are being worked on for inclusion into Prosspero solutions

Prosspero is a TM Forum effort devoted to providing usable reference implementations and "ready-to-go" solution packages for the standards developed by various TM Forum workgroups.

In spite of the early standardization and significant effort in implementation, OSS-J, along with other NGOSS technology solutions, still awaits mass acceptance and adoption.

3.3.9 Transitioning from IPv4 to IPv6

IPv4, today's dominant internetworking protocol, was designed in the 1970s. Today, in the Internet era, the demands for such a protocol are quite different, and IPv4 is showing its age. The successor, IPv6, designed within the last decade, addresses many of IPv4's shortcomings. Most important among these shortcomings is depletion of address space. IPv6 aims to improve security, privacy, convergence, management, and service delivery as well.

The exhaustion of the 32-bit IPv4 address pool (some 4 million unique addresses) is a hotly debated issue: Some predict that all addresses will be assigned by 2010, while others—debating the prevalence of NAT and firewalling technologies—believe that the process will be slower. One thing is certain: It is bound to happen sooner or later.

IPv6 offers 128-bit addresses (potentially 3.4×10^{38} unique addresses), solving the address pool exhaustion problem. Additionally, the larger address space allows more efficient addressing for routing. The larger address space also simplifies LAN management, as Ethernet and other link-layer MAC addresses map trivially onto IPv6 addresses within an organization, resolving difficult-to-manage issues such as maintaining MAC↔IP address assignment with protocols like DHCP (Dynamic Host Configuration Protocol). In order to simplify network administration, IPv6 also contains autoconfiguration mechanisms (such as neighbor discovery) that allow an instantly usable IP network to be set up without configuration or special servers. IPv6 contains built-in IPSec security, enabling secure networking at the native IP level and alleviating administration of VPN gateways and routers.

QoS is also supported at the lowest level. Each packet has two fields for traffic categorization: an 8-bit traffic class or priority field (similar to that of IPv4) and a 20-bit flow label, specifically created with QoS in mind.

Probably the most powerful feature is the *extension header*, which is the generalization of IPv4's protocol field. This 8-bit field may hold protocol identifiers similar to those of IPv4 (although the assigned numbers for common encapsulated protocols are different), or it may signal the presence of a "next header" within the IPv6 packet. IPv6 uses this mechanism to define options and also enables the encapsulation of arbitrary higher-level protocols.

IPv6 was designed with easy transition from IPv4 in mind. In the context of IPv6 networks, transition mechanisms refer to coexistence schemes between version 4 and version 6 networks and hosts. The

main purpose of these mechanisms (apart from enabling gradual v4 to v6 transition) is to allow "isolated" version 6 hosts and networks to access each other over a version 4 infrastructure.

The following schemes have been defined:

- *Dual stack:* This approach, defined in RFC 4213, outlines how to have hosts implement both protocol versions. Since IPv6 is a logical, conservative extension of IPv4, it is possible to extend an IPv4 protocol stack code with IPv6 support by reusing a large part of it.
- *Tunneling:* This refers to the encapsulation of IPv6 packets into IPv4 packets, enabling transparent transport of IPv6 traffic over a legacy infrastructure. There is an IPv4 protocol number (41) allocated for this purpose. This method is also known as *6in4 tunneling*. Encapsulating IPv6 into UDP packets is also standardized in order to enable transfer over NAT or firewalls. There are two tunneling approaches. In *configured tunneling*, endpoints are configured manually (or by a *tunnel broker* service, based on rules).
- Configured tunneling uses direct IPv4 encapsulation (protocol 41).
- In *automatic tunneling*, endpoints are automatically deduced by the network infrastructure (routers) based on some kind of in-band signaling. The recommended method is to use 6in4 encapsulation, encode IPv4 information within the IPv6 addressing on the local side, and use an IPv4 anycast address on the remote side. *ISATAP* is an automatic tunneling mechanism with direct IPv4↔IPv6 address mappings. Finally, *Teredo* tunneling uses UDP encapsulation and is able to transfer IPv6 over multiple NAT translations. Teredo is enabled by default in Windows Vista and supported in Windows XP SP2.

Accessing IPv4-only services from IPv6 computers is typically accomplished through an application-level proxy installed on a dual-stack machine.

3.3.10 Web Service Technologies and SOA: SOAP, UDDI, BPEL

SOA (service-oriented architecture) is a new paradigm in application integration. SOA systems consist of distributed, loosely coupled cooperating entities offering services to each other. SOA is a modern, Internet-era, large-scale manifestation of the client–server paradigm.

SOA provides a language (BPEL, or Business Process Execution Language) for modeling and executing high-level business processes. BPEL is intended as a common language that both IT and business/management professionals can understand and use; IT as well as business concepts can be expressed in BPEL. By executing BPEL processes (which link and sequence lower-level services and primitives), SOA provides orchestration for services (i.e., it implements coordination of the services and interactions of a business environment).

The SOA paradigm is built over a number of lower-level Web service technologies that provide data exchange, remote procedure call (client–server transaction), and service locator facilities. SOA data (and service) modeling and representation are always done in XML, and the role of XML is pervasive throughout the paradigm.

It is important to note that although, in theory, SOA systems can be implemented over multiple technologies (particularly transport), in practice SOA is practically always bound to specific technologies. The following sections briefly introduce and discuss SOA-enabling technologies in a bottom–up fashion, with the lowest-level technologies first.

3.3.10.1 Fundamental Technologies: XML-RPC and SOAP

XML-RPC is a very simple RPC protocol that encodes its messages in XML and uses the HTTP protocol for transport. The most important aspect of XML-RPC is simplicity: The entire specification fits on two printed pages, and implementations are typically a few hundred lines long. XML-RPC supports seven basic data types and one compound (*struct*); the protocol does not support asynchronous RPC invocations.

SOAP (originally Simple Object Access Protocol, but the spelled-out version of the acronym has officially been dropped with the evolution of the protocol) is a generalization of the simple XML-RPC. SOAP also uses XML for encoding and data representation. HTTP is the prominent transport protocol, with the SOAP standard defining bindings for other transports as well.

The protocol's basic data unit is the *message*, which consists of three parts:

- The *header* contains transport-specific information such as an HTTP request/reply header. The header is transport specific, and the remainder of the message is XML encoded.
- The *envelope* identifies the SOAP version and the encoding (serialization) rules used.
- The *body* is the container for the protocol payload (request/reply data).

SOAP does not specify the content or format of the message body (apart from XML well-formedness). The only body type specified in the standard is the fault message, which contains SOAP error messages.

The SOAP protocol specification also consists of three parts:

- *Envelope*: XML framework describing message content and processing
- *Encoding rules*: data-type definitions and their XML encodings
- *Bindings*: transport protocol specifications

SOAP supports a number of data types, with the atomic types coming from the XML schema standard. Compound types are also supported, as well as polymorphic (run-time-dynamic) types.

3.3.10.1.1 Advantages and Disadvantages

SOAP is versatile, platform independent, simple, and extensible. With the most popular HTTP binding, using SOAP in different networking scenarios (e.g., firewalls, relays) is easy and straightforward. Because of the XML encoding, SOAP is easy to debug and troubleshoot.

On the other hand, data encoding in textual XML is terse and inefficient, and SOAP parsing is resource intensive; as a result, SOAP users must cope with significant overhead. With HTTP encoding, transactions are unidirectional, and thus in many situations a party waiting for an event has to resort to polling instead of a more efficient notification mechanism.

3.3.10.2 Description and Location Services: WSDL and UDDI

SOA technologies were envisioned for large distributed and frequently changing environments, such as corporate intranets. In such networks, describing and locating services and interfaces is an ever-present challenge. Service providers (servers) need to announce their presence and the services they implement. Similarly, clients require automated means to identify and locate the particular server providing the service they are requesting. Such challenges are solved by the two further building blocks of SOA: WSDL (Web Services Description Language) and UDDI (Universal Description Discovery and Integration).

3.3.10.2.1 Web Services Description Language

WSDL describes a Web service interface as well as the bindings (e.g., HTTP address) to be used to access it. WSDL defines data and message types and abstract service definitions accepting and producing messages. WSDL describes four service paradigms:

- *One-Way:* The endpoint accepts message(s).
- *Request-Response:* The endpoint accepts message(s) and sends (corresponding) response(s).
- *Solicit-Response:* The endpoint sends message(s) and awaits response(s).
- *Notification:* The endpoint sends message(s).

As indicated above, WSDL describes interfaces as collections of the ports and message formats they accept (and produce). WSDLs also contain definitions of bindings, which assign concrete encoding rules and protocol endpoints to services. Finally, WSDL service definitions contain logical groups of services.

By obtaining and parsing WSDLs, a Web services client can determine whether a certain functionality is offered by the server the WSDL describes and how to access that functionality (message encoding rules and concrete endpoint location). Functionality matching is performed by matching the input and output data types, and WSDL's rich typing facilities ensure that the functionality can be described in the necessary degree of detail.

3.3.10.2.2 Universal Description Discovery and Integration

UDDI is an OASIS standard for a platform-independent, XML-based registry of business services rendered over the Web service platform. The goal of UDDI is to enable businesses to publish and discover service listings and to define the technical means for the interaction necessary for services to be used. According to the standard's intent, UDDI catalogues WSDL service descriptions made available by service providers.

UDDI business registrations have three components:

- *White Pages:* address, contact, and known identifiers
- *Yellow Pages:* industrial categorizations based on standard taxonomies
- *Green Pages:* technical information about services exposed by the business

UDDI was created in August 2000. Originally, UDDI reflected the vision, dominant at the time, that ultimately automated and secure Web services would proliferate on the public Internet. Such a vision requires a global directory wherein users locate the services they need (e.g., credit card verification) along with vendor, technical, and billing information, and these services can be used in a completely automated manner. With respect to this vision, UDDI acts as a public brokerage service that pairs user searches with vendor offerings.

The vision of libraries of publicly available Web services has yet to materialize, and the initial enthusiasm is fading. The pioneers behind UDDI, IBM, Microsoft, and SAP closed their public UDDI nodes in January 2006. There is a clear need for the Green Pages functionality in the SOA-based business software environment to become prominent on corporate intranets. Thus, UDDI may live on in a greatly reduced role, as a broker dynamically binding client systems to implementations seeking a given service.

UDDI is served in the form of UDDI registries. Hosts providing UDDI service are termed *UDDI nodes*.

3.3.10.3 Business Process Execution Language

As the name implies, BPEL, or more precisely WS-BPEL (Web services BPEL), is an executable business process modeling language aimed at the formal definition of business processes and their interaction protocols. BPEL implements the *programming in the large* paradigm, focusing on the high-level state transitions of the process being modeled. BPEL is also called an *orchestration language*: The analogy is that of a conductor who synchronizes the performance of a group of musicians, each of them playing a sophisticated piece of music.

BPEL describes processes using state automata. Processes are either abstract (with fewer details) or executable (with full behavior specification). A process is constructed from a series of activities that are either flow constructs (e.g., *RepeatUntil*) or interaction elements (e.g., sending a message). As BPEL is a high-level language, in practice it needs a companion language, which is used to specify individual actions or activities. Java is often used in this role. BPEL native data-manipulation functions are not sufficient for meaningful data-processing tasks.

BPEL is an XML-based language; it leverages on XML encoding and data types. Messages used in BPEL interactions are described via WSDL.

BPEL does not define any visual representation for the language. Software vendors usually provide their own nonstandard mappings or translate BPEL processes into BPMN (Business Process Modeling Notation) models, which do involve visual notation. Although the two languages are related, one-to-one mapping cannot be established; thus, BPMN translation is not always possible.

3.3.10.4 Lightweight Directory Access Protocol (LDAP)

Directory services are fast becoming the key to the enterprise, allowing applications to locate the resources they need and enabling network managers to authenticate end users. Corporate networkers need to be aware about what LDAP is capable of, where it is headed, and what it was never intended to do.

LDAP was intended to offer a low-cost PC-based front end for accessing X.500 directories. Due to high overhead and acceptance delays of X.500, LDAP has emerged to fill the gap, somehow expanding its role. It rapidly became the solution of choice for all types of directory services applications on IP networks. LDAP applications can be loosely grouped into three categories: those that locate network users and resources, those that manage them, and those that authenticate and secure them. Network managers who want to put the protocol to work need to go into detail, coming to terms with standard components and features. This protocol can save companies time and money. It can help network managers keep pace with the rising demand for directory services. New applications appear almost every day. But there are limits to what a protocol can do for distributed computing. It cannot store all the types of information that network applications need. Knowing the difference between LDAP facts and fiction is the only way to avoid potential pitfalls.

3.3.10.4.1 Attributes of LDAP

The current specification comprises eight features and functions (HOWE99):

- Information model: Organized according to collections of attributes and values, known as entries, this model defines what kinds of data can be stored and how that data behaves. For example, a directory entry representing a person named Jim Fox might have an attribute called sn (surname) with a value "Fox." The information model, inherited almost unchanged from X.500, is extensible: almost any kind of new information can be added to a directory.
- LDAP schema: Defines the actual data elements that can be stored in a particular server and how they relate to real-world objects. Collections of values and attributes—representing such objects as countries, organizations, people, and groups of people—are defined in the standard, and individual servers can define new schema elements as well.
- Naming model: Specifies how information is organized and referenced. LDAP names are hierarchical; individual names are composed of attributes and values from the corresponding entry. The top entry typically represents a domain name, company, state, or organization. Entries for subdomain, branch offices, or departments come next, often followed by common name entries for individuals. Like the LDAP information model, the naming model derives directly from X.500. Unlike X.500, LDAP does not constrain the format of the namespace; it allows a variety of flexible schemes.
- Security model: Spells out how information is secured against unauthorized access. Extensible authentication allows clients and servers to prove their identity to one another. Confidentiality and integrity also can be implemented, safeguarding the privacy of information and protecting against active attacks such as connection hijacking.
- LDAP functional model: It determines how clients access and update information in a LDAP directory, as well as how data can be manipulated. LDAP offers nine basic functional operations: add, delete, modify, bind, unbind, search, compare, modify distinguished name, and abandon. Add, delete, and modify govern changes to directory entries. Bind and unbind enable and terminate the exchange of authentication information between LDAP clients and server, granting or denying end users access to specific directories. Search locates specific users or services in the directory tree. Compare allows client applications to test the accuracy of specific values or information using entries in the LDAP directory. Modify distinguished name makes it possible to change the name of an entry. Abandon allows a client application to tell the directory server to drop an operation in progress.
- LDAP protocol: Defines how all the preceding models and functions map onto TCP/IP. The protocol specifies the interaction between clients and servers and determines how LDAP requests and responses are formed. For example, the LDAP protocol stipulates that each request is carried in

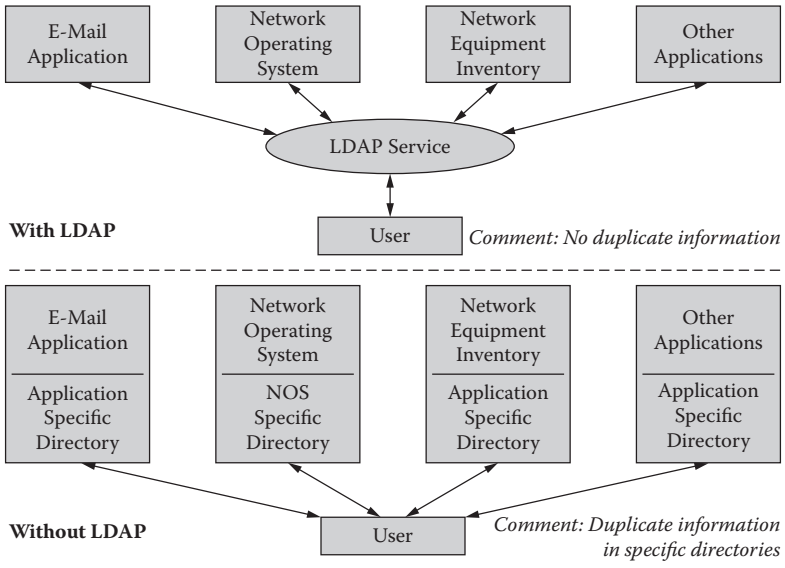
a common message format and that entries contained in response to a search request are transported in separate messages, thus allowing the streaming of large result sets.
- Application program interface (API): Details how software programs access the directory, supplying a standard set of function calls and definitions. This API is widely used on major development platfroms running C, C++, Java, Javascript, and Perl.
- LDAP data interchange format (LDIF): Provides a simple text format for representing entries and changes to those entries. The ability helps synchronize LDAP directories. LDIF and the LDAP API, along with scripting tools like Perl, make it easy to write automated tools that update directories.

LDAP directories and operating systems are melding to create intelligent environments that can locate network resources automatically. Examples include:

- Active Directory and Windows NT (Microsoft)
- HP-Unix and LDAP (Hewlett-Packard)
- Sun Solaris and LDAP (Sun Microsystems)
- Irix and LDAP (Silicon Graphics)
- Digital Unix and LDAP (Compaq)

In this new role as operating system add-on, LDAP furnishes a way to locate printers, file servers, and other network devices and services. LDAP makes these services standard, more accessible, and in many cases, more powerful and flexible. LDAP is also starting to play a critical role in network management, where it can be a great help to network administrators. Without LDAP, managers and administrators have to maintain duplicate user information in many specific and separate directories across the network. With LDAP, it is possible to centralize this information in a single directory accessed by all applications (Figure 3.3.12). Of course, replacing key legacy applications with LDAP-enabled ones takes time, but big changes are already underway.

LDAP also has an important role to play in tighter security, with the directory acting as gatekeeper and deciding who has access to what. In this capacity, LDAP performs two critical jobs. First, it serves as an authentication database. Second, once the identity of a user has been established, it controls access to resources, applications, and services using stored policies and other information. LDAP also permits corporate networkers to use their directories to implement PKI (public key infrastructure) security.



FIGu RE 3.3.12 Directory centralization with LDAP.

From the user's point of view, LDAP provides the directory in which certificates of other users are found, enabling secure communication. From the administrator's point of view, LDAP directories are the way in which certificates can be centrally deployed and managed.

3.3.10.4.2 Limitations of LDAP

LDAP has three major limitations. First, the protocol cannot and will not make relational databases redundant. It lacks the heavy update, transaction processing, and reporting capabilities of these products. Nor does it offer two-phase commits, true relational structure, or a relational query language like SQL. Using LDAP to implement an airline reservation system would be a serious mistake. Second, it is not reasonable to expect that LDAP serves as a file system. Its information model is based on the simple pairing of attributes and values. Thus, it is not well suited to binary large object data that is managed by typical file systems. It is also not optimized for write performance and is unable to furnish byte-range access to values, both critical features of a file system. Finally, it does not have the locking semantics needed to read- and write-protect files. Third, LDAP is not a stand-in for DNS, which may well be the world's largest distributed database. Although LDAP's abilities are more or less a superset of DNSs—whose biggest job is translating names like home.netscape.com into IP addresses—here is a very good argument for not penetrating tasks of DNS; DNS is working fine. Also, LDAP cannot contend with the connectionless transport that DNS usually runs over. Ultimately, LDAP may have a role in managing and augmenting the information found in DNS. For example, it could link contact information to host information, but it cannot take the place of the DNS database itself.

In summary, LDAP has its place among the successful network management tools.

3.3.11 Summary and Trends

There are multiple standards for network management. All of them have advantages and disadvantages, and of course, also different application and implementation areas. Telecommunications suppliers and customers will have to live with multiple standards. The question is how these standards can seamlessly interoperate. There are basically three alternatives:

- Management gateway: The interoperability is realized by a special system responsible for translating management information and management protocols. Looking at the practical realization of such a gateway, it is important to target the use of OMA for both OSI and Internet-based management. Many existing object specifications for management could be taken over by the OMA-based management.
- Platforms with multiple architecture: The interoperabiltity is realized by a multilingual platform, understanding multiple protocols. Protocol conversion is not necessary. Management information can be interpreted and transformed by the platform or by applications. Different architectures are supported simultaneously, but without deep integration.
- Agent with multiple architectures: The interoperability is realized at the agent level. In this case, the management agent understands multiple protocols and languages. It requires some intelligence for the agent. If selected, agent software must be implemented in many, practically in all, networking components. This number is considerably higher than in the case of management platforms.

There is a new group—the Joint X/Open TeleManagement Forum Inter-Domain Management Group—that addresses in particular the interoperability between OSI–Management, Internet–Management, and OMG–OMA. This type of work takes a lot of time. In the meantime, practical solutions are absolutely necessary. In most cases, gateways deliver the quickest solutions.

Standardization is absolutely necessary to ensure interoperability of various components of communication systems. This chapter has laid down the basics. Management frameworks and platforms may support some of the standards, but there is no product that supports all of them.

Open database connectivity (ODBC) is an application programming interface (API) allowing a programmer to abstract a program from a database. When writing code to database, the user usually has to

add code that talks to a database using a particular language. If the user wants his/her program to talk to an Access, Fox, and Oracle database, code the program with three different database languages. This can cause some problems.

When programming to interact with ODBC, the user only needs to talk the ODBC language (a combination of ODBC API function calls and the SQL language). The ODBC manager will outline how to contend with the type of database the user is targeting. Regardless of the database being used, all of the calls will be to the ODBC API. All that the users need to do is install an ODBC driver specific to the type of database selected.

Directory-enabled networking (DEN) is a specification to save information about network devices, applications, and users in central directories. DEN addresses the integration of application- and user-level directory information with network control and management, building on open Internet standards, such as LDAPv2 and Wbem/CIM. The CIM initiative is being extended to meet the needs of the DEN initiative. In the future, management applications will have access to authoritative information on the relationships among network elements, services, and individuals, so that preferences, privileges, and profiles can be enforced according to enterprise policies but with personal customization, and so that policies governing network applications can make use of directory-based information. In summary, information will be consistent in DEN directories and in CIM management systems.

IPv4 will migrate to IPv6, but slower than expected. Every managed object will get its IP address with the result of better accessibility and manageability. But management tools may run into a volume problem, having too much information in real time and near real time.

Acronyms

ACS	Autoconfiguration Server
API	Application Programming Interface
BPEL	Business Process Execution Language
BPMN	Business Process Modeling Notation
CIM	Common Information Model
CMIP	Common Management Information Protocol
COTS	Component Off The Shelf
CRM	Customer Relationship Management
CPE	Customer Premises Equipment
DMI	Desktop Management Interface
ETOM	Enhanced Telecommunications Operations Map
LAN	Local Area Network
DEN	Directory Enabled Networking
LDAP	Lightweight Directory Access Protocol
M2M	Manager To Manager
MF	Mediation Function
MIB	Management Information Base
MIF	Management Information Format
NMS	Network Management System
ODBC	Open Database Connectivity
OSF	Operations Systems Function
PDU	Protocol Data Unit
QAF	Q Adapter Function
RMON	Remote Monitoring
SIP	Session Initiation Protocol
SNMP	Simple Network Management Protocol
SOA	Service-Oriented Architecture

SOAP	Simple Object Access Protocol
TMN	Telecommunications Management Network
UDDI	Universal Description Discovery and Integration
UDP	User Datagram Protocol
VPN	Virtual Private Network
WSDL	Web Services Description Language
WSF	Workstation Function

References

[HawE99] Howes, T.: LDAP: Use as Directed, *Data Communications Magazine,* New York, February 1999, pp. 95–103.

[TR-069] DSL Forum TR-069 Technical Report http://www.dslforum.org/techwork/tr/TR-069.pdf

[WIKI] Wikipedia article about TR-069 (figure) http://en.wikipedia.org/wiki/TR-069

Terplan, K. OSS Essentials—Support System Solutions for Service Providers, John Wiley & Sons, Inc., New York, 2004.

3.4 Management Function

József Wiener

There are many ways to define management functions. The views are slightly different for telecommunications service providers and with enterprises. Standard organizations are busy to help. This section will introduce three different alternatives: first, using the recommended Telecommunications Management Network (TMN) layers; second, using the traditional FCAPS (fault, configuration, accounting, performance, and security), and, third, using the FAB (fulfillment, assurance, and billing) grouping of the TeleManagement Forum. A comparison of these three alternatives will be provided.

In addition to the standard description of well-understood functions, there are emerging areas that need particular attention. Three selected areas will be addressed: event correlation for supporting fault management automation, telecom expense management to control and reduce operating expenses in enterprises, and security disciplines to lower security risks and improve readiness for vulnerabilities.

3.4.1 Management Functions

Network and service management is critical in today's telecommunications business activities to ensure network availability, quality of service, secure communications, and ease of use of the network and related technologies. It is also important in effectively managing network devices and supporting users.

Various standards organizations and vendors have proposed many management architectures and models. Some are implemented in the real world, while others remain concepts for consideration. The key purpose of these models is to partition the task of managing a network into smaller functions, making the entire network management job much easier. Table 3.4.1 lists some of the models, along with key technologies.

3.4.1.1 TMN Functions

The Telecommunications Management Network (TMN) model was introduced by ITU-T in Recommendation M.3000 in 1985 as a reference model for operations support systems (OSSs) of telecommunications service providers. The TMN concept is an architectural framework for the interconnection of different types of OSS components and network elements (i.e., it is intended for use by service providers to manage their service delivery networks). It consists of management architectures at the functional, physical, informational, and logical abstraction levels. TMN also describes the standardized interfaces and protocols used for the exchange of information between OSS components and network elements, and the overall functionality needed for network management.

TABl E 3.4.1 Management Models and Technologies

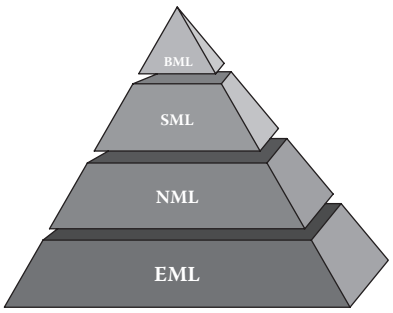
Model	Sponsor Organization(s)	Target Network(s)	Main Functions	Status
FCAPS (fault, configuration, accounting, performance, and security)	ISO, ITU-T	All	Fault, configuration, accounting, performance, security management	Conceptual framework for network management
TMN (Telecommunications Management Network)	ITU-T	Telecom network	Business management, service management, network management, element management	Conceptual framework for service provider network management systems
OAM&P (Operation, administration, maintenance, and provisioning)	Service providers	Telecom network	Operation, administration, maintenance, provisioning	Widely adopted by large service providers in their network management systems
eTOM (enhanced Telecom Operations Map)	TeleManagement Forum; approved by ITU-T	Service provider network	Network and systems management, service development and operations, customer care	Accepted standard, replaces OAM&P

The TMN model is composed of two main parts, a management layer functional definition and a large number of network interoperability standards for communicating between systems and network elements. The TMN interoperability standards are numerous and focus exclusively on technology.

Logical layers of management functions are often illustrated by a pyramid (Figure 3.4.1), in which the greatest amount of elementary data can be found at the bottom level, with the degree of complexity (during processing) increasing up through the layers.

The focus of the TMN pyramid is on managing networks via layers of responsibility divided among functional areas. TMN management layers and their functions are as follows.

- *Business management layer:* performs functions related to business aspects (e.g., analyzes trends and quality issues) or provides a basis for billing and other financial reports.
- *Service management layer:* performs functions related to handling of services in the network, such as definition and administration of and charges for services.
- *Network management layer:* performs functions for distribution of network resources, such as configuration, control, and supervision of the network.
- *Element management layer:* contains functions for the handling of individual network elements, including alarm management, handling of information, backup, logging, and maintenance of hardware.



FIGu RE 3.4.1 TMN layers.

The TMN layers provide a simple way of thinking about the different types of management information used in running a network services business and how that information is filtered, summarized, or decomposed as needed in the various management layers.

The main problem with the TMN standards process has been the industry's inability to demonstrate conformance to the standards established in the TMN model. The pyramid model still has much relevance and significance to the industry today, but the TMN interoperability standards work may have already moved beyond its prime in terms of usefulness. In addition, the TMN model has been partly incorporated into newer models (e.g., eTOM).

3.4.1.2 FCAPS Functions

FCAPS is a network management functional model defined by ITU-T (International Telecommunication Union) and ISO (International Standard Organization) in Recommendation M.3400. The FCAPS model splits general telecommunications management functionality into five key areas (Table 3.4.2):

- Fault management
- Configuration management
- Accounting management
- Performance management
- Security management

These categorizations are functional ones and do not describe the business-related role of a management system within the telecommunications network. The idea of FCAPS stems directly from ITU-T recommendations and describes the five different types of information handled by management systems.

Table 3.4.2 presents a subset of the FCAPS functions.

3.4.1.2.1 Fault Management

Fault (or maintenance) management (sometimes also understood as event and error management) refers to a set of functions enabling the detection, isolation, and correction of abnormal operations in the telecommunication network and its environment. Its purpose is to detect and record events that have occurred in different parts of the network and establish the cause of these events at the highest level of detail and accuracy possible. Errors are explored and repaired in the shortest time possible.

Fault management may be confined to simply recording alarm states originating from separate network elements and generating appropriate error messages to inform the operator. A more sophisticated and effective method is to correlate these individual events and evaluate their correlation by means of appropriate algorithms. Correlation algorithms may be needed to identify causes of malfunctions and locate their sources precisely in complex situations. That is, one elementary error may generate a number of error reports, or conversely, one error report may refer to several events or alarm states (e.g., equipment breakdown, cable rupture, traffic congestion). If necessitated by numbers or types of failures, automatic loopback tests or other special test routines can be initiated to locate faulty system elements. Analyzing interrelations of elementary alarm states and evaluating results of test routines may lead to a precise diagnosis even when individual methods have been unsuccessful.

Application of event and error management can result in many malfunctions being discovered and eliminated before they cause any noticeable problems for users. In some cases, the impact of the emerging failure can be eliminated by an automatic and dynamic reconfiguration of network routes. Automatic reconfiguration is generally effective in meshed networks and (if only a single cable rupture has occurred) in ring network configurations or cases of traffic congestion.

3.4.1.2.2 Configuration Management

The functions of configuration management are exercising, controlling, identifying, and collecting data from or providing data to network elements. In terms of managing up-to-date, broadband

TABl E 3.4.2 FCAPS Functions

Fault Management	Configuration Management	Accounting Management	Performance Management	Security Management
Fault detection rates	Resource initialization	Tracking of service/ resource usage	Utilization and error rates	Selective resource access
Fault correction	Network provisioning	Cost for services	Consistent performance level	Enabling of network element (NE) functions
Fault isolation	Autodiscovery	Accounting limits	Performance data collection	Access logs
Network recovery	Backup and restoration	Combination of costs for multiple resources	Performance report generation	Security alarm/event reporting
Alarm handling	Resource shutdown	Setting of quotas for usage	Performance data analysis	Data privacy
Alarm filtering	Change management	Audits	Problem reporting	Monitoring of user access rights
Alarm generation	Preprovisioning	Fraud reporting	Capacity planning	Handling of security breaches and attempts
Clear alarm correlation	Inventory/asset management	Support for different modes of accounting	Performance data and statistics collection	Security audit trail log
Diagnostic testing	Copy configuration	Tracking of service usage	Maintenance and examination of historical logs	Security-related information distributions
Trouble detection and collection	Remote configuration	Billing for services	Data analysis	Control of access to resources
Error logging	Job initiation, tracking, and execution	Discounting	Report generation	Enabling/disabling of functions
Error handling	Automated software distribution			
Error statistics	System turn-up			
Testing and acceptance	Network provisioning			
Root cause analysis	Database handling			

telecommunication networks, configuration management generally includes two essential, logically different functions: static and dynamic configuration management.

Static configuration management involves assigning and unassigning network elements to or from the network logically (*attaching* and *detaching*), as well as recording, indicating, displaying, and reporting network topology and lists of network equipment along with their system parameters (e.g., type, topological location, symbolic and physical addresses). In the case of small and simple networks (along with recording equipment parameters), static configuration management may also involve inventory control; however, with complex networks, this function should be handled separately.

Dynamic configuration management involves establishing the actual routes for the required interconnections via the network. It implies network reconfiguration by establishing a new possible route, if the actual route breaks down, or taking down the route if a request arrived requiring that the connection be canceled. In terms of dynamic configuration management, network topology displays must reflect the actual routes and connections in the network.

3.4.1.2.3 Accounting Management

Accounting management (sometimes also referred to as billing and accounting management) refers to a set of functions that enable network service use to be measured and the costs of such use to be determined. Accounting management should provide facilities to collect accounting records and set billing parameters for use of the service.

In the accounting management process, time and other characteristics of users' network access are measured, and charges are calculated according to several parameters (e.g., price lists, subscriber contracts, time of use, services used). Billing and accounting information is collected, classified, and recorded. On the basis of charges, data bills can be prepared and sent to customers, income can be calculated and recorded, and so forth.

3.4.1.2.4 Performance Management

The functions of performance management (sometimes also referred to as traffic and performance management) are to evaluate and report on the functioning of telecommunication equipment and the effectiveness of the network and/or network elements. Performance management may involve measuring the intensity of data flow along the different routes of the network; collecting, evaluating, and displaying data measured in this way; determining efficiency indices; and calculating trend analyses. Information gathered and evaluated in this process can be used similarly to and in connection with data gained in the scope of fault management. These data can be used to establish traffic load levels and determine whether a given network complies with the necessary performance requirements. (If any congestion occurs, overloaded network routes can be relieved through system reconfiguration or alteration of the actual routing strategy. The network management system can automatically intervene in network operations. If a permanent lack of network capacity has been observed, the decision should be made to initiate new investments and increase network capacity.)

3.4.1.2.5 Security Management

Security management involves establishing classes of authentication, checking users' authorization to access the network, controlling passwords, and taking other possible measures to prevent the network from any unauthorized access. There may be a need to protect management terminals from unauthorized interventions according to the given security requirements. Depending on the special requirements outlined in accordance with the purpose of a network, functions contained within security management may differ from application to application.

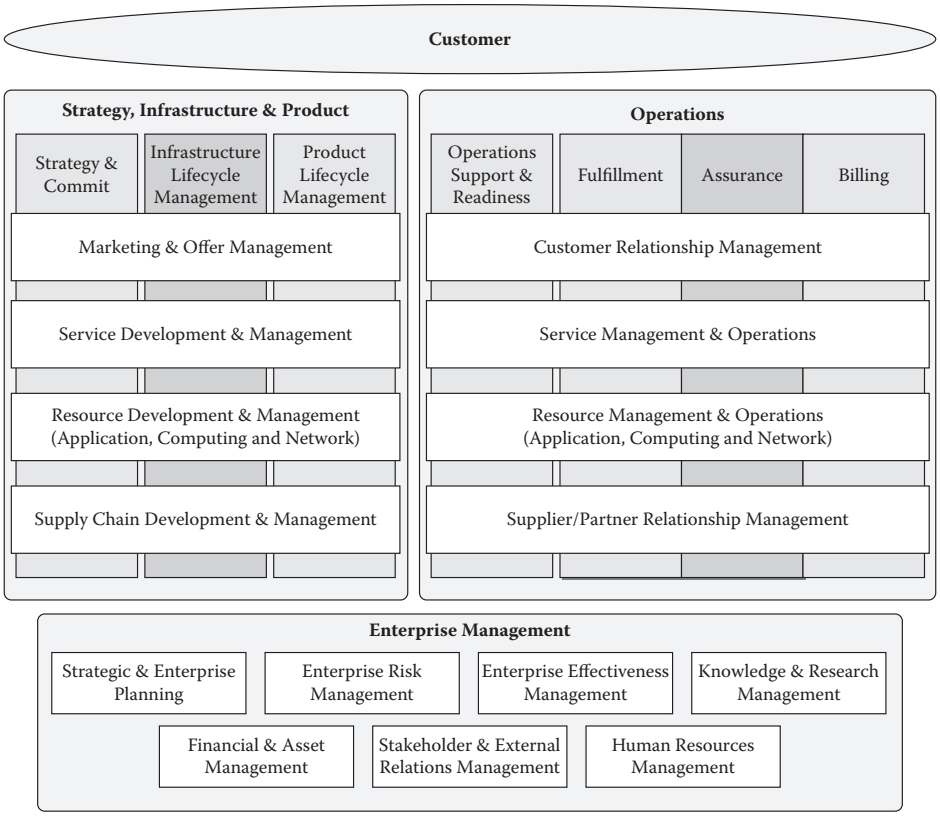
Security refers to authentication, access control, data confidentiality, data integrity, and nonrepudiation. Security management includes maintenance and distribution of authentication and authorization information such as passwords and encryption keys. In other words, security management allows administrators to control who has access to which resources.

Security management has expanded greatly in the past several years. Today it also includes configuration and control of intrusion detection systems, firewalls, and traffic filtering by URL, address, or protocol with the goals of monitoring activity and detecting or deflecting malicious actions.

3.4.1.3 Fulfillment, Assurance, and Billing processes (eTOM)

The Enhanced Telecom Operations Map (eTOM) was developed by the TeleManagement Forum and later adopted by ITU-T in its Recommendation M.3050. eTOM is a business process framework for use by service providers and their suppliers and partners within the telecommunications industry. It describes all of the enterprise processes required by service providers and analyzes them at different levels of detail according to their business significance and priority. eTOM represents an improvement to the original TMN model and incorporates both that model and FCAPS functionality.

Essentially, eTOM provides a common view of the major operational functions of a communications service provider using terms that are understood by business operations personnel while maintaining a communications channel to systems developers. It takes the perspective of service providers, looking



FIGu RE 3.4.2 eTOM framework.

internally at business processes, specifically the processes supporting customer care, service development, and operations management, with links to traditional network operations functions including fault, configuration, and performance management.

eTOM defines a complete enterprise management framework and addresses the impact of e-business environments and business drivers. It can be a starting point for standardizing business processes as well as OSSs and business support systems (BSSs). Another area of improvement is process-modeling methodology, which provides the linkage necessary for Next-Generation Operations Support Software (NGOSS).

The eTOM model has multiple levels of abstraction covering three main areas. The highest, conceptual view of the eTOM framework involves the following major process areas (Level 1 processes (Figure 3.4.2).

- *Operations:* This area includes customer relationship management, service management, resource management, and supplier/partner relationship management.
- *Strategy, infrastructure, and product:* This area includes processes that support the creation of strategies for marketing, development of new services, resource development and management, and supply chain development and management.
- *Enterprise management:* This area includes the basic business processes required to run and manage an enterprise, such as human resource management, financial asset management, and disaster recovery management.

Figure 3.4.2 shows the major process areas and their decomposition into Level 1 processes. This provides an overall view of the eTOM framework. (In practice, it is the next level [Level 2, decomposition] at which users tend to work, as this degree of detail is needed in analyzing their businesses.)

The eTOM framework's aim is to categorize process elements and business activities so they can subsequently be combined in numerous different ways to implement end-to-end business processes (e.g., fulfillment, assurance, and billing [FAB]) that deliver value for the customer and the service provider.

The figure also shows seven vertical process groupings representing the end-to-end processes required to support customers and manage businesses. The focal point of the eTOM framework is the core customer operations processes of fulfillment, assurance, and billing. Operations Support and Readiness (OSR) is differentiated from FAB real-time processes to highlight the focus on enabling support and automation in FAB.

In the Strategy, Infrastructure, and Product (SIP) process area, the Strategy and Commit vertical, as well as the two Lifecycle Management verticals, are differentiated. They do not directly support the customer, they are intrinsically different from the operations processes, and they work on different business time cycles.

The Logical Layered Architecture, which includes business, service, and network layers, was originally used to help organize core business processes, as this facilitated mapping of the management functions defined in TMN to processes. Given that the TMN layering approach is still relevant, this loose coupling has been maintained in the evolution of the eTOM framework.

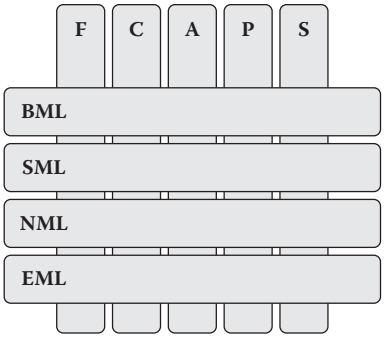
In the OSR process area of the eTOM framework, there are four functional process groupings that support operations processes and management of operations to support customer, service, resource, and supplier/partner interactions. The following horizontal functional groups can be defined within the operations (management) area of eTOM:

- *Customer relationship management (CRM):* This functional process grouping considers the fundamental knowledge of customers' needs and includes all functionalities necessary for the acquisition, enhancement, and retention of relationships with customers. It focuses on customer service and support, whether storefront, telephone, Web, or field service. CRM applies to both conventional retail customer interactions and wholesale interactions, such as when an enterprise is selling to another enterprise that is acting as the *retailer*.
- *Service management and operations:* This grouping focuses on knowledge of services (e.g., access, connectivity, content) and includes all functionalities necessary for the management and operations of communications and information services required by or proposed to customers. The emphasis is on service delivery and management as opposed to management of the underlying network and information technology.
- *Resource management and operations:* This grouping maintains knowledge of resources (application, computing, and network infrastructures) and is responsible for managing the resources (e.g., networks, IT systems, servers, routers) used to deliver and support services required by or proposed to customers. It also includes the functionalities responsible for the direct management of all such resources (e.g., network elements, computers, servers) utilized within the enterprise. These processes are responsible for ensuring that the network and information technology infrastructure supports the end-to-end delivery of required services. Resource management and operations processes manage the complete service provider network, subnetwork, and information technology infrastructures.
- *Supplier/partner relationship management:* This grouping supports core operational processes, including fulfillment, assurance, and billing and functional operations processes. Supplier/partner relationship management processes align closely with a supplier's or partner's CRM processes.

The eTOM model is intended primarily for the telecommunications industry, including both service providers and their suppliers. With the convergence of the telecommunication and IT industries, their own models—namely eTOM, TMN, and FCAPS for telecom and the ITIL approach for IT—are also converging. More details on eTOM and the eTOM-ITIL convergence are provided in Section 3.6 (Support Processes).

3.4.1.4 Comparison of Management Models

The different models and architectures are not independent of each other. They are the result of competition between international bodies, they have followed the evolution of the technology associated



FIGu RE 3.4.3 Mapping FCAPS into TMN layers.

with operational and management practices, and they are a consequence of the natural evolution of standards.

In the previous decades, the industry focused on network management, and the FCAPS and TMN models were widely accepted. Only a few individuals realized that formulating customer service management solutions was a far greater challenge than addressing the performance and capacity management needs of a communications network, even though all of the standards activities at the time were devoted to network management.

3.4.1.4.1 Management Service/Function Approach

In the first TMN recommendation regarding the management service/function approach, a set of TMN application functions were identified, building on the FCAPS management categories described in ITU-T Recommendation M.3400. Management function sets and management functions are categorized according to their FCAPS application and specified together with generic end-to-end flow scenarios that relate them to management services and managed areas (according to ITU-T Recommendation M.3200) and TMN logical layers (according to ITU-T Recommendation M.3010). The resulting scheme consists of horizontal groupings into layers and vertical groupings into functional flow-through areas, as illustrated in Figure 3.4.3.

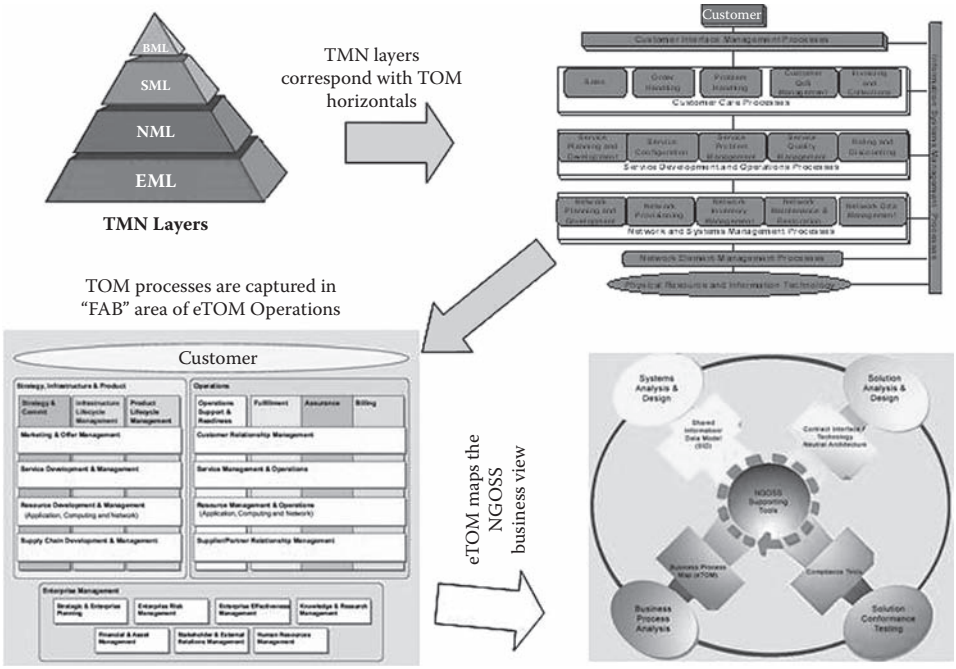
3.4.1.4.2 Business Process Approach

The business process approach has built on the concepts of management services and functions in order to develop a reference framework for categorizing all of a service provider's business activities. This is done through a business-oriented definition of each area of business activity, in the form of a process view that describes the service provider's enterprise in a top–down, structured way with progressive decomposition to expose increasing details. The identified individual process elements can then be positioned within a model allowing analysis of organizational, functional, and other relationships and combined within process flows that trace activity paths through the business.

This process view of the service provider is expressed in the eTOM business process framework. The eTOM framework can serve as a blueprint for standardizing and categorizing business activities (i.e., process elements), which will help set a direction and starting point for development and integration of BSS and OSS. In the context of TMN, the eTOM framework provides a business-oriented view of service provider requirements that management services and functions need to support, and the mapping from individual eTOM processes to management functions, and vice versa, is documented to assist and support the application of these processes and functions within management solutions.

3.4.1.4.3 Relationship between the Management Service/Function and Business Process Approaches

The process-oriented perspective can be related to the functional view provided elsewhere in TMN to allow the relevant management and resource/network capabilities to be linked with the business needs they support.



FIGu RE 3.4.4 Evolution of TMN to eTOM.

The TMN management services/functions can be related to eTOM business processes, but they provide a different perspective on the management environment. The main difference between the two approaches is that the TMN approach has been built on the requirement to manage network equipment and networks (bottom up), while eTOM has been built on the need to support processes of the entire service provider enterprise (top–down).

Both approaches can be used in identifying generic and specialized management function sets to support management activities and, together with other recommendations, can capture a technology- and resource-oriented view of the management domain, which is often valuable and relevant when considering the structure and organization of a management solution. The eTOM framework provides an additional business-oriented viewpoint that is important in considering the business requirements of the service provider as the user of a management solution and in ensuring that the arrangement of management functions is meaningful and useful for the way the service provider does business. Figure 3.4.4 shows the evolution of TMN to eTOM via the Telecommunications Operations Map (TOM).

3.4.1.5 Expectations and Trends

Network and service management are critical in terms of ensuring network complex services, quality of service, as well as effectively managing network devices and supporting users. Service providers must manage a complex set of products, networks, and services in a dynamic, competitive market. The focus is on automating operations, management, and business processes. To achieve this, standard requirements need to be identified through a market-centric approach involving key industry players.

The network of different operators cannot be connected without commonly accepted and implemented standards. This area is of common interest to all players in telecommunications.

Some of the earlier attempts to create standards for the network operation and management were not really successful. The first successful standard was the TMN standard, developed by ITU-T. The TMN is a series of very detailed standards that provide a framework for carriers to achieve interconnectivity

and communication across heterogeneous networks and systems. TMN has had its greatest impact on OSS applications residing in the element management layer and network management layer.

The TMN model is composed of two main parts, a management layer functional definition and a large number of network interoperability standards for communicating between systems and network elements. The interoperability standards are numerous and focus mainly on technology, and lost their relevance as a result of their complexity and the changes in the technology over the last 15 years. The layered architectural model has often been used as an organizational model and is still relevant.

Another widely accepted standard is the FCAPS model. This model works fine in describing carrier operations, but in truth is primarily focused on the concept of technology management. It provides valid ways of understanding what goes on in carrier environments. Neither TMN nor FCAPS deals with business processes, operational practices, or the business value chain.

The first complete framework document that was consistent with the TMN architecture and resolved some of these issues was TOM. This document provided an overall process framework spanning the breadth of the operational processes that exist within a typical wireline-based telco, structured to be consistent with the four architectural layers described within the TMN architecture model.

The TOM was expanded in the eTOM document. This document overcame the limitations of the previous models. A single enterprisewide business process framework was produced that captured the operational processes within a telecommunications enterprise.

Essentially, eTOM provides a common view of the major operational functions of a communications service provider using terms that are understood by business operations personnel while maintaining a communications channel to systems developers. eTOM is a reference framework that categorizes the business processes that a service provider will use. It defines a complete enterprise management framework and addresses the impact of e-business environments and business drivers. eTOM can be considered a blueprint for standardizing business processes as well as OSS/BSS. Another area of improvement is process-modeling methodology, which provides the linkage necessary for NGOSS.

With the convergence of IT and the telecommunications industry, the dependence on telecommunications in IT must be reflected in standards and practices. The main driver for IT standards is ITIL. ITIL is organized into sets of texts that are defined by related functions: service support, service delivery, managerial, software support, computer operations, security management, and environmental. After becoming the worldwide de facto standard in service management, ITIL is currently one of the fastest growing business optimization initiatives; its processes are being adopted by organizations both big and small because of its ability to improve business processes. ITIL focuses on best practices, and as such it can be adapted and adopted in different ways according to individual organizational needs.

Both the eTOM and ITIL frameworks are concerned with management of services and their delivery to the customer. ITIL and eTOM are not in conflict; they are, in fact, complementary, compatible, and mutually supportive. Each framework has it own strengths that can be used to complement and support the other.

3.4.2 In-Depth Considerations of Selected Management Functions

In addition to the standard description of well-understood functions, there are emerging areas that need particular attention. In order to achieve the highest level of automation in the area of fault management, event correlation is key. This is the basis of expert systems using various forms of reasoning technologies. Enterprises spend a considerable amount of money for telecommunications services. Telecom expense management streamlines isolated responsibilities with the result of substantial cost savings. Security and risk considerations are a never-ending responsibility. This segment will address assessment, prevention, detection, response, and vigilance.

3.4.2.1 Event Correlation

Fault management may be confined to simply recording alarm states originating from the separate network elements and generating appropriate error messages in order to inform the operator. However, in a big network a single fault can cause a burst of failure events, and operators are not able to decide what is the real (root) problem. Event correlation simplifies and speeds monitoring of network events by consolidating alerts and error logs into a short, easy-to-understand package. Network administrators can deal with, say, 25 events based on cross-referencing intrusion alerts against firewall entries and host/asset databases much more efficiently than when they must scan 10,000 mostly normal log entries. In one real-life instance, more than 200,000 alarms were recorded per day on an Synchronous Digital Hierarchy (SDH)/Plesionchronous Digital Hierarchy (PDH) network, and more than 80% of these alarms were generated by two dozen pieces of equipment (this equipment generated so-called *prelling alarms*).

Event correlation approaches are applied to handle a large number of events and to find the root cause of a fault. A root cause is the most basic reason for an undesirable condition or problem. If the real cause of the problem is not identified, one is merely addressing the symptoms and the problem will continue to exist. Thus, identifying and eliminating root causes of problems is of utmost importance. Root cause analysis is the process of identifying causal factors using a structured approach with techniques designed to focus on identifying and resolving problems. Tools that assist groups and individuals in this process are known as *root cause analysis tools*.

Some tools provide the means for administrators to visually "correlate" among data from different points in the infrastructure. In this case, the tools themselves have no intelligence. Often a great deal of expertise is needed to analyze the data and isolate problem conditions. Moreover, the manual correlation effort is time consuming and laborious. A variety of technologies and relatively simple operations are associated with event correlation and root cause analysis:

- Deduplication identifies incident events that are duplicates of a previously received event. Duplicate events may occur because of a continuing problem within a system. Deduplication can either drop the duplicate event or create a "child" event and attach it to the original event. The original event is known as the "parent" event.
- Compression takes multiple occurrences of the same event, examines them for duplicate information, removes redundancies, and reports them as a single event. As a result, 1,000 "route-failed" alerts become a single alert stating "route failed 1,000 times."
- Counting reports a specified number of similar events as one event. This differs from compression in that it does not just tally the same event, and there is a threshold to trigger a report.
- Suppression removes or drops incident alerts that are generated or created by systems that are downstream of a failed system. For example, servers that are downstream of a failed router will fail. Alert suppression will prevent incident/problem trouble tickets from being generated in the support management system.
- Filtering removes or drops informational incident alerts or incident alerts derived from systems or functions that are not part of the support model implemented in the incident/problem support management system. An example of this is the filtering of informational data alerts from a firewall when the support model is related to hardware availability and faults only.
- Association identifies incident alerts that are a result of or are associated with problems in other systems or functions. This level of correlation creates "child" trouble tickets that are attached to the originally generated "parent" trouble ticket. For example, this can function to attach server availability fault alerts to an active change order (request for change, change management) during the time of a planned maintenance outage.
- Generalization associates events with higher-level events. This can be useful for correlating events involving multiple ports on the same switch or router in the event that it fails. One does not need to see each specific failure if one can determine that the entire unit has problems.

- Time-based correlation can be helpful in establishing causality, for instance, tracing a connectivity problem to a failed piece of hardware. Often more information can be gained by correlating events that have specific time-based relationships. Some problems can be determined only through such temporal correlation.

The main idea of correlation is to condense and structure events to retrieve meaningful information. These approaches primarily address the correlation of events as reported from management tools or devices. Therefore, they can be labeled device-oriented correlation approaches 0.

In today's telecom and IT environments, offering services with an agreed-upon quality becomes more and more important. These changes also affect event correlation. It has become a necessity for providers to offer such guarantees to differentiate themselves from other providers. To avoid violations of service-level agreements (SLAs), it is especially important for service providers to identify the root cause of a fault in a very short time or even act proactively. The latter refers to the case of recognizing the influence of a device breakdown on the services offered. As knowledge about services and SLAs is used in this scenario, it is referred to as *service-oriented correlation* 0.

The paradigm shift from device-oriented to service-oriented management also has implications for the area of event correlation. Today's event correlation addresses mainly correlation of events as reported from management tools. However, correlation of trouble reports from users needs to be addressed as well, because different reports could have the same cause. In such a case, the reports could be linked together, and processing would need to be performed only once. Therefore, the response time for trouble reports could be improved, and service-level guarantees could be maintained with less effort. Such correlation is labeled *service-oriented correlation* [1]. Taking into consideration the device-oriented (or network-oriented) and service-oriented approaches, correlation can be addressed from two perspectives, as follows.

- *Effort reduction (top–down perspective):* Instead of presenting every event/alarm to operators, the goal is to identify a resource as being the problem's root cause. For example, if a 155-Mbps SDH link fails, thousands of alarms are generated in the network, but only one is to be presented to the operator. All other alarms must be suppressed for the maintenance team. This is a form of root cause analysis.
- *Impact analysis (bottom–up perspective):* In the case of a fault in a resource, its influence on the associated services and affected customers can be determined. This analysis can be performed with respect to either short-term (a current resource failure) or long-term (e.g., network optimization) considerations. Impact analyses are often used with SLAs.

True correlation, root cause analysis, and impact analysis go hand in hand; correlation involves analyzing the state of resources, services, and subcomponents to identify where the cause of a problem resides and its impact on services. By differentiating between the cause and effects of a problem, a monitoring solution can allow administrators to focus on the former rather than being distracted by the latter 0.

Given that almost every monitoring solution claims to perform correlation and root cause diagnosis, what are the differences between these tools?

3.4.2.1.1 Model-Based Reasoning

In model-based reasoning (MBR), each component of an infrastructure is modeled with respect to its attributes, behavior, and relation to other models. A model can represent either a physical entity or a logical entity (e.g., LAN, WAN, domain, service, business process). The behavior of the entire infrastructure is a result of the interaction of the component models, where each of these models can be either a representation of a physical entity or a logical entity. Event correlation is a result of the collaboration of models.

In service-oriented event correlation, this approach is useful if it is possible to model each service as a logical entity. The behavior of a service may be difficult to describe in some scenarios since it depends on actual customer behavior.

As all components of a network are represented in the model, it is possible to perform simulations to predict how the overall network will behave. However, a large MBR system is not easy to maintain in all cases. It can be difficult to appropriately model the behavior of all components and their interactions correctly and completely (an example MBR system is NetExpert from OSI, which is a hybrid MBR/rule-based reasoning system).

3.4.2.1.2 Rule-Based Reasoning

In rule-based reasoning (RBR), a set of rules is used to actually perform the correlation. The rules have to be in a "conclusion if condition" form. The condition contains received events together with information about the state of the system, while the conclusion may consist of actions, which lead to changes in the system and can be input to other rules.

The rules in an RBR system are more or less readable, so their effect is intended to be intuitive. In practice, the rule sets may become quite large, which may lead to unintended rule interactions and difficulty in maintaining the system. In addition, the system will fail if an unknown situation occurs that is not covered in the rules.

Because of the correlation performance of rule-based reasoning algorithms, this approach is useful in service-oriented event correlation. However, the drawbacks of the approach must be minimized (i.e., the system must be maintained at an effective level, and there must be a plan for dealing with unknown situations).

3.4.2.1.3 Codebook Approach

As is the case with RBR, the codebook approach proposes a correlation algorithm. This approach uses experience from graphs and coding. A dependency graph is used consisting of events and root causes as nodes and directed edges to represent dependencies. The dependency graph contains two kinds of modeling nodes: faults (sometimes denoted as problems) that must be detected and observable events (symptoms) caused by faults or other circumstances. After a final input graph is selected, the graph is transformed into a correlation matrix in which the columns contain faults and the rows contain events.

It is possible to choose weights for the edges (e.g., a weight for the probability that fault/event A causes event B). Another possible weighting could indicate time dependencies. There are several alternatives with respect to reducing the initial graph. If, for example, the dependency of events is cyclical and there are no probabilities for the cycles' edges, all events can be treated as one event in isolation. Techniques derived from coding theory can be applied for optimization. For example, some event rows can be deleted if the events do not lead to discrimination in root causes.

This approach provides an advantage over RBR in that it can—in some situations—deal with unknown combinations of events. These combinations can be mapped onto known combinations using the Hamming distance. Another advantage of the codebook approach is that, as mentioned, it uses long-term experience with graphs and coding; this experience is used to minimize the dependency graph and to select an optimal group of events with respect to processing time and robustness against noise. In contrast, the RBR approach may allow for greater flexibility than the encoding schema.

The codebook technique is also useful in service-oriented event correlation, as an efficient correlation algorithm is provided. A disadvantage of the approach could be that, similar to the case with RBR, frequent changes in the environment make it necessary to regularly edit the input graph. SMARTS InCharge is an example of such a correlation system.

3.4.2.1.4 Case-Based Reasoning

In contrast to the preceding two techniques, the case-based reasoning (CBR) approach does not necessitate any prior knowledge about the infrastructure. It contains a database of cases that have occurred before together with the identified root causes. While the first root causes must be identified manually, automated matching to prior cases is performed at later stages. The ability of this approach to learn from prior cases is useful in service-oriented event correlation, even though case-based reasoning algorithms are less efficient than the algorithms for the two preceding techniques.

In contrast to other approaches, CBR systems do not use any knowledge about the system structure. The knowledge base saves cases with their values for system parameters and successful recovery actions for these cases. The recovery actions are performed not by the CBR system, but in most cases by a human operator.

If a new case appears, the CBR system compares the current system parameters with the system parameters in prior cases and attempts to find a similar one. To identify such a match, parameters for which cases can differ or must be the same have to be defined. If a match is found, a learned action can be performed automatically or the operator can be informed with a recovery proposal.

An advantage of this approach is that an integral part of it is the ability to learn, which is important in rapid changing environments. However, there are also difficulties in applying the approach. The fields used to find similar cases must be defined appropriately. If there is a match with a similar case, an adaptation of the previous solution to the current one has to be found.

An example CBR system is SpectroRx from Cabletron Systems (the part of Cabletron that developed SpectroRx became an independent software company) [2].

3.4.2.1.5 Hybrid Approach for Dynamic Situations

A hybrid approach that combines RBR and CBR has been proposed to deal with highly dynamic situations (e.g., battlefield scenarios). In the proposed architecture, an RBR and a CBR system run in parallel. The RBR engine uses temporal and spatial dependencies to correlate reported events, while the CBR engine makes use of prior situation templates. Because the approach has not been implemented as of yet, details about the collaboration of the approaches involved are not available. According to the authors, this work represents the first attempt to combine RBR and CBR techniques in the network and systems management domain.

3.4.2.2 Telecom Expense Management

Telecom expense management (TEM) tools are not new, but more businesses than ever are finding them necessary. Reasons include the fact that service providers continue to bill customers in error, as well as ongoing concerns over industry consolidation and, in some instances, outright scams. There are essentially three ways to bring TEM tools into the enterprise: buying, hosting, and managing software in-house; buying software-as-a-service (SaaS) in which the software is hosted on servers at an external data center; or outsourcing the entire operation.

TEM is not only about cost savings. It is also about improving bill payment and credit processes and getting a better handle on a company's network and device inventory. At the beginning, customers need to know how many circuits and routers they have and how many wireless devices are assigned to employees. In addition, there are a handful of facts telecom managers need to know. They include total expense per service provider/vendor, number of invoices processed, average time needed to process an invoice from receipt to payment, cost of late-payment penalties, and average time needed to receive a credit stemming from billing errors.

Asset and inventory management help customers identify areas they really want to fix or address before sitting down with a vendor. Executive sponsorship should occur on two levels: sign-off for economic benefits and sign-off for business process improvements.

Policies need to be drafted and enforced. When new policies come from the CXO level, they are usually more widely accepted and followed. TEM touches many parts of the enterprise, not only telecom and IT. Telecom managers are expected to gather information from the finance, human resources, and accounts receivable departments before making policy changes. Having these departments on board can help in drafting realistic new policies and can help ease employees into a new TEM world. Once management quantifies how much the enterprise spends and will spend on telecom services, facilities, and equipment, the expected TEM return on investment can be quantified.

Industry consultants recommend centralizing telecom management organizations in terms of invoices, payments, and ordering. Centralizing, which refers to putting all processing in one place, will

TABl E 3.4.3 TEM functionality checklist (*Source:* Aberdeen Group)

Function	Present Level of Use (%)	Intended Level of Use Within 12 Months (%)	Intended Level of Use After 12 Months (%)
Invoice presentation and analysis	67	17	7
Invoice contract term reconciliation	63	15	7
Service usage auditing and accounting	64	23	5
Service usage allocation and chargeback	55	14	5
Inventory and asset management	50	26	13
Spend analysis	40	26	15
Service rate database	38	23	10
Electronic contract management	33	22	13

give businesses a head start in controlling their telecom expenses. Many organizations are centralizing similar telecom-related tasks.

Typical TEM functions are summarized in Table 3.4.3. Also included are usage and intended usage percentages based on a 2007 survey of 90 enterprises.

Prior to deploying TEM services and tools, it is useful to consider the following factors.

- *Executive buy-in:* The appropriate CXO should be on the support board before initiation of RFIs and RFPs.
- *Cooperation:* There should be an investigation into how TEM needs gel with compliance regulations and internal policies.
- *Asset status:* Inventories of network elements and facilities should be up to date, along with knowledge of the time necessary to process invoices and receive credits.
- *Success monitoring policies:* KPIs should be set prior to initiating use of TEM services or tools.
- *Clear expectations:* Goals and expectations should be well documented in negotiations with vendors.
- *Centralization:* Many organizational units may be involved in TEM-related activities; centralization and governance are absolutely necessary.
- *Return on investment:* Numbers regarding expected return on investment (ROI) should be on the table before initiation of negotiations with vendors and service providers.

3.4.2.3 Examples of Products

Some product examples are summarized below.

Rivermine is a leading provider of automated solutions (both software and managed services) that enable organizations to gain visibility into, and control over, their telecom spending. These market-leading TEM solutions automate the entire telecom lifecycle, including contract sourcing, ordering, inventory management, invoice processing/auditing, and reporting/analytics.

Rivermine's TEM software suite delivers everything necessary for telecom contract negotiation, procurement, provisioning, inventory management, invoice processing, auditing, and business intelligence. This lifecycle solution automates and streamlines many of the most challenging telecom management tasks with four powerful TEM applications:

- *Rivermine Inventory Engine:* builds and maintains a current repository of wired, wireless, and data networking assets.
- *Rivermine Service Order Manager:* creates, validates, and tracks telecom orders from request through approval and provisioning.
- *Rivermine Finance Manager:* enables automated invoice processing, bill validation, cost allocation, and auditing.
- *Rivermine Clarity:* delivers unparalleled visibility into and control over telecom spending through packaged dashboards, reports, and drill-down analytics.

Rivermine's core strength resides in its inventory management and continuing development to bring business intelligence associated with telecom expenditures to market.

Rivermine recently secured a potentially influential industry partner to bring managed services and TEM Business Process Outsourcer (BPO) to market. IBM GTS agreed to use Rivermine's application and services as the core of its TEM offerings. In addition, IBM plans to offer its TEM solution as a standalone service as well as bundled within broader outsourcing agreements. IBM's presence in more than 170 countries could make this a compelling offering globally. Another provider, Invoice Insight, has shown success in the federal and civilian government sectors as well as in large commercial opportunities, which, along with its partnerships with Accenture and other Enterprise Service Providers (ESPs), has allowed the company to grow faster than much of its competition. Notably, 93% of its revenue is recurrent in nature. Invoice Insight remains the industry market driver in terms of SaaS solutions, growing its BPO services nicely from its meat-and-potatoes invoice auditing managed service. Key relationships with Accenture, Huron, and other outsourcers will continue to serve the company well.

A third provider, Tangoe, offers its CommCare suite for complete communications lifecycle management. CommCare goes beyond the savings results of TEM, allowing companies to experience the operational benefits of communications lifecycle management in transforming all facets of fixed and mobile communications. The CommCare suite of managed services brings control, visibility, and understanding to every critical process within the communications environment.

As a legacy application publisher, Tangoe has done an extremely effective job of transitioning its business model to managed services and TEM BPO in terms of revenue growth and focused execution. Tangoe offers licensed applications in addition to managed services and TEM BPO. Tangoe's core strength resides in its invoice and contract management capabilities.

CommCare's communications lifecycle management services are built on patented technologies and best-of-breed functionality that optimize all essential voice, data, and mobile communications from beginning to end. That means unparalleled understanding of the communications infrastructure, future needs, and financial investments.

More complete analyses of several TEM vendors can be found in Goodness and Redman 0.

3.4.2.4 Security and Risk Considerations

Five areas require particular attention by security officers: assessment, prevention, detection, response, and vigilance.

3.4.2.4.1 Assessment

The overall impact of Sarbanes Oxley Act (SOX) and other similar compliance measures is that inattention to security now has a clearly marked price tag for senior managers; if IT security is lax, they can wind up facing fines or even jail time. Either way, their company's stock options will take a beating.

Infrastructure risk management is no longer an unpopular exercise in looking for leaks even in the absence of obvious security threats. It is now understood as an essential element of due diligence, part of the practice of being a going concern that wants to stay that way. The resources available and the respect and consideration for people doing the work appear to have markedly changed for the better in the last couple of years.

There are many ways to enact risk assessment in form while failing to deliver it in substance. In the past, IT security stakeholders were primarily internal—operators needing to ensure facility uptime and in-house users needing access to applications and confidence in the quality of data protection. The environment now demands far more attention to external stakeholders. Technology-centric professionals are still likely to think in terms of IT asset protection—ensuring that servers are not taken down and applications are not taken out of service—while failing to appreciate the data-centric and process-centric viewpoints.

The National Security Agency (NSA) has come into the public spotlight as a center of excellence for security techniques, including the Information Assurance Methodology (IAM) and InFosec Assessment

Methodology (IFAM). IAM involves a notion of impact attributes that enterprise professionals should embrace and understand. Core impact attributes include confidentiality, integrity, and availability, all of which are essential for a secure infrastructure. Other guidelines focus on relevant attributes such as the following.

- *Accountability:* Who introduced what information, made what changes, or took what actions?
- *Nonrepudiation:* Who made what commitments, and how can others prove it?
- *Authorization:* Who has permission to take what actions? Who granted that permission? What combinations of permission should not be allowed?
- *Audit:* What actions were taken on the basis of access to what information? By whom? When? From where?
- *Access control:* Which entities have which specific modes of access to which resources, subject to what policies?
- *Information needs:* Stakeholders should be asking questions about the types of information that are important to an organization; they should annotate the resulting list with the characteristics that define their own mission-specific, type-specific standards of what it means for that organization's information needs to be securely met.

3.4.2.4.2 Prevention

In the past, security prevention advice focused on the need to harden outward-facing systems, particularly against external attacks. Protection against external threats is still a priority, but administrators now must also look inward to reduce the risks posed by company-internal users.

The threat landscape has changed significantly since the beginning of this decade. Largely gone are high-profile worms crafted to make noise and cause outages that make way for stealthier attacks. They have been replaced by increasingly sophisticated and targeted attacks designed to steal data for financial gain. Attackers have found deceiving users to be a highly effective way to establish a foothold on a network, either as a way to initiate further attacks or for data theft. Many organizations have stepped up their user security training, but it is still relatively easy to make a bad choice (e.g., opening attachments, installing innocent-appearing software, or even simply clicking on the wrong link).

The real key to security prevention is minimizing the amount of damage that unwitting users can do if and when they make a bad decision. For instance, administrative control to local resources may be denied. While there are certainly tactics that attackers can use to escalate privilege on a compromised host, limiting a user's rights at least would make the process more difficult. Companies should now be considered negligent if they unnecessarily allow users local administration rights.

The most difficult part of adopting the least-privileged user account (LUA) philosophy is attempting to get poorly written applications to work properly with limited rights. However, tools help identify where applications will run afoul of the LUA. Other tools help administrators adjust credentials in a targeted fashion, increasing privilege levels only when absolutely necessary.

Of course, an attacker with user privileges is still a problem as a result of the risk of additional attacks. Companies must therefore continue to maintain, and even streamline, an effective patch strategy that encompasses not only the operating system but also third-party applications and drivers. It is mandatory as well for administrators to track and monitor lists of known, unpatched vulnerabilities. They must evaluate the potential impact of these vulnerabilities on the network and weigh the costs and benefits of deploying temporary workarounds.

Loss of mobile data also remains a major concern, as can be seen from the numerous accounts of lost and stolen laptops containing personal records. The use of encryption products that secure files, folders, or entire volumes remains an obvious way to deal with this threat. However, encryption is still a workaround that ignores larger, systemic questions. For example, do the potential productivity gains derived from granting employees anywhere, anytime access to sensitive data outweigh the risks of this access? And, if so, is there a better way to manage the outflow of this information?

According to industry consultants, a preferred alternate strategy would be to invest in data access through tightly secured Web services with control over who can access what from where. Network connectivity is not at the point at which it is ubiquitous enough to make this scenario a reality. Nevertheless, the goal should be to have only a few core entry points to critical data.

3.4.2.4.3 Detection

Multiple reported data losses represent a huge challenge for IT managers responsible for security. How can an accurate account be maintained of what malicious activity has been prevented through prudent action? This is where the art of detection comes into play. To demonstrate that detection is effective, reports must be created in such a way that non-IT managers can use them as well. In some cases, trial use of programs that control user access (e.g., single sign-on) can show failed attempts to access data. Reports generated by trial implementations of leak prevention tools are even better at showing averted nefarious activity. It is recommended that IT managers place authorized use at the center of the security architecture and use risk assessment to determine how to protect valuable data and systems. IT managers must know and detect all activity that falls outside the boundaries of authorized and acceptable use of data and systems. However, detecting out-of-bounds data and system use is not sufficient to keep security systems in the good graces of upper management. Inflexible systems that cannot accommodate traffic spikes are barriers to productivity. To this end, products are available that allow IT technicians associated with business units to create policies as an essential part of implementing detection tools. Leak-detection tools should allow authorized users and administrators to make changes to monitoring functions. IT managers evaluating these types of tools should situate this functionality atop their list of essential features.

When it comes to detection, there is no substitute for an intimate knowledge of what traffic should be traveling through the network and what data and transactions are needed to carry on business. Although vendors of detection tools emphasize the simplicity of installation and integration of their products into the network, the fact is that unless a human being evaluates the traffic and usage patterns revealed by these tools, malicious activity can remain undetected. Finally, change management procedures can go a long way toward reducing the false-positive readings often associated with detection tools.

3.4.2.4.4 Response

How should IT managers respond when they find that a rootkit has turned a company's systems into its personal playground? Unfortunately, the best advice that can be given is to take down the system on which the rootkit has been implanted and rebuild it from scratch. However, this usually is not an option when the rootkit has had access to a number of resources. Everything that touched the infected system in any way, shape, or form must be considered suspect. And businesses will need to watch carefully for a significant period of time to make sure there are no hidden or remaining effects from the rootkit invasion.

With most standard infections, the first step once a problem has been detected is to literally pull the plug. However, while this works fine when one system is involved, should it be done for an entire network? If the network is an internal corporate segment, management should pull the plug on the entire segment. While this will cause a great deal of inconvenience for users, it is vital to disconnect the affected system from the Internet. In the case of resources that cannot be shut down, such as network segments that include externally facing Web, database, and application servers, it may be necessary to intentionally poison Domain Name System tables. This will mislead rootkit controllers about the location of affected systems.

Once all potentially infected systems are isolated, the rootkit itself should be located and removed. Standard applications, such as antivirus tools, can and will help. At this point, it is vital to trace all activity related to the rootkit infection. Everything that could have been touched or seen by the rootkit-affected system needs to be checked, and all activity on the infected system should be studied, starting from the time of the infection.

The shady characters who now have detailed information on all vital passwords and access mechanisms can do a great deal of harm. In order to avoid such a situation, everything—all passports, user

accounts, authentication systems, anything that could have been scanned or accessed by the infected system—must be changed. It may even be the right time to upgrade the network, the server infrastructure, and the security solutions.

The final step is to attempt to stop a rootkit infection from ever occurring again. A rootkit infection, and all of the turmoil it causes, is a good opportunity to reiterate the importance of good security practices. Of course, there also may be a need to educate IT staff. There is no rest for the security weary. The only effective response is continuous vigilance.

3.4.2.4.5 Vigilance

During the last couple of years, the standard of what constitutes due care for maintaining an enterprise security posture has risen almost beyond recognition. It can be difficult to obtain good estimates of associated costs, since organizations are understandably loath to discuss in detail their security efforts or spending.

Even a well-conceived security strategy can be executed to excess. That said, most organizations correctly suspect they have yet to reach the level of "good enough," let alone any fears of going too far. Any long-term progress in elevating enterprise security will have to involve a cultural fight against evolving technology. The IT world's processing, connectivity, and storage trends pave the way for intentional as well as merely careless leakage or abuse. Only organizational buy-in to the relevance of security awareness and to the appropriateness and necessity of broad participation in the security process can overcome adverse technology trends. Proactive design of useful and necessary business processes, identification of the data and privileges needed to carry them out, and instrumentation of systems to detect any violations of those boundaries are the techniques that will succeed.

3.4.3 Summary and Trends

Most of the management functions discussed here have changed from their original meaning and now involve less or more sophisticated implementations, mostly as a result of changes in technology and the environment. Some of the improvements in these functions have been discussed in this chapter.

Fault management may be confined to simply recording alarm states and generating meaningful error messages for the operator. In a big network however, a single fault can generate a burst of events, and operators need support decide what is the real (root) problem.

Event correlation approaches are applied to handle a large number of events and to find the root cause of a fault. If the real cause of the problem is not identified the problem will continue to exist. Thus, identifying root causes of faults is important. Root cause analysis is the process of identifying factors causing the problem. Tools that assist operators in the process of finding the root cause are known as root cause analysis tools.

As in a big network, manual correlation does not provide the required speed and accuracy, automated IT-based solutions are needed. This is an area with a good future, especially in a mobile, converged world.

TEM offers a report-based snapshot of wireless expenditures and attempts to control costs through monitoring and adjusting usage and service plans. While TEM documents historical costs and usage, this static, one-size-fits-all solution is a throwback to old-fashioned expense management and reporting. Today's enterprises need to understand, proactively manage, and reduce the costs and complexities of mobility services, not merely know (and control) where dollars are being spent.

Security management in the Internet world is an emerging issue. The overall impact of SOX and other similar compliance measures is that inattention to security now has a clearly marked price tag for senior managers; if IT security is lax, they can wind up facing fines or even jail time. Either way, their company's stock options will take a beating.

In the past, security prevention advice focused on the need to harden external attacks. Protection against external threats is still important, but looking inward to reduce the risks posed by internal users is also needed.

Multiple reported data losses represent a huge challenge for operators responsible for security. Reports must be created in such a way that non-IT experts can use them as well. It is recommended to use risk assessment to determine how to protect valuable data and systems. Products are available that allow to create policies as an essential part of implementing detection tools.

Acronyms

BPO	Business Process Outsourcer
BSS	Business Support System
CBR	Case-Based Reasoning
CRM	Customer Relationship Management
EMS	Element Management System
ESP	Enterprise Service Provider
eTOM	Enhanced Telecom Operational Map
FAB	Fulfillment, Assurance, and Billing
FCAPS	Fault, Configuration, Accounting, Performance, Billing
IAM	Information Assurance Methodology
IFAM	InFosec Assessment Methodology
ISO	International Standards Organization
ITU-T	International Telecommunications Union Telecommunications Sector
ITIL	IT Infrastructure Library
LUA	Least-Privileged User Account
MBR	Model-Based Reasoning
NE	Network Element
NSA	National Security Agency
NGOSS	Next-Generation OSS
OSR	Operations, Support, and Readiness
OSS	Operations Support System
PDH	Plesianchronous Digital Hierarchy
RBR	Rules-Based Reasoning
RFI	Request for Information
RFP	Request for Proposal
SDH	Synchronous Digital Hierarchy
SOA	Service-Oriented Architecture
SOX	Sarbanes Oxley Act
TEM	Telecom Expense Management
TMN	Telecommunications Management Network

References

1. Hanemann, A., Sailer, M., and Schmitz, D. *Assured Service Quality by Improved Fault Management—Service-Oriented Event Correlation*. Munich Network Management Team (Leibniz Supercomputing Center & University of Munich). Proceedings of International Conference on Service Oriented Computing, 2004.
2. Hanemann, A., and Schmitz, D. *Service-Oriented Event Correlation—the MNM Service Model Applied to E-Mail Services*. Munich Network Management Team (Leibniz Supercomputing Center). 11th International Workshop of the HP OpenView University Association (HPOVUA 2004). Paris, France, June 2004.
3. eG Innovations. *Choosing a Monitoring System for Your IT Infrastructure? What Should Your Key Considerations Be?* White Paper of eG Innovations, 2005. Available at www.eginnovations.com.

4. Goodness, E. and Redman, P. *MarketScope for Telecom Expense Management, Worldwide, 2H07*. Gartner, February 2008.
5. Hanemann, A. and Sailer, M. *A Framework for Service Quality Assurance Using Event Correlation Techniques. Munich Network Management Team* (Leibniz Supercomputing Center & University of Munich). Proceedings of Advanced Industrial Conference on Telecommunications, Lisbon, 2005.
6. Broadband Network Management 2004-2009. The Insight Research Corporation, Feb 2004.

3.5 Support Systems for Service Providers

József Wiener

The telecommunications industry shows both evolutionary and revolutionary signs. Evolution is seen with incumbent carriers; revolutionary attributes are visible with new entrants. The technology itself shows a mixture of wireline and wireless services (converged services and networks) supporting all major telecommunication forms, such as voice, data, and video on the same network (network convergence) and on a small number of customer devices (converged devices). Many such services and networks are rapidly evolving in the Internet Protocol (IP) era with a consequent requirement for radical improvements and streamlining in the design of support systems. These systems are the underlying resource for the management and delivery of all of the communications services that enable the handling of direct customer interactions. Support, administration, and management from day-to-day operations to traffic trending, capacity planning, forecasting, customer care, billing, provisioning, order processing, and network operational management are all functions implemented via support systems.

Operations Support Systems (also called Operational Support Systems, or OSSs) are computer systems used by telecommunications service providers. The term OSS most frequently describes network systems dealing with the telecom network itself, supporting processes such as maintaining network inventory, provisioning services, configuring network components, and managing faults.

The complementary term Business Support Systems (BSSs) is a newer label and typically refers to business systems dealing with customers, supporting processes such as taking orders, processing bills, and collecting payments. The two systems together are often abbreviated BSS/OSS or simply B/OSS.

BSSs are the components that a telephone operator or telco uses to run its business operations. Typical activities that count as part of BSSs are taking a customer's order, managing customer data, managing order data, billing, rating, and offering business-to-business (B2B) and business-to-consumer (B2C) services. BSSs are linked to OSSs in the enhanced Telecom Operations Map (eTOM) that maps processes into the functional areas of fulfillment, assurance, and billing, where assurance is typically covered by the OSS platform. BSS and OSS platforms are linked in the need to support various end-to-end services. Each area has its own data and service responsibilities.

The terms OSS and BSS are no longer limited to telephone operators offering mobile, fixed, and cable services but also can apply to service providers in all sectors, such as IT services and utility providers.

A Market Support System (MSS) or Marketing Decision Support System (MKDSS) is an information system that helps with decision making in the formation of a marketing plan. The reason for using an MKDSS is that it helps to support a software vendor's planning strategy for marketing products; it can assist in identifying advantageous levels of pricing, advertising spending, and advertising copy for the firm's products.

With the advent of multiservice converging/converged networks and BSS/OSSs supporting voice, data, and video from fixed and mobile networks, aligning processes utilizing separate systems for voice and data, fixed networks, and mobile networks during the migration to new systems has become a significant challenge. There is a strong need to provide interoperability among support systems while enabling ease of customization across a wide range of functionality. New systems are tailored to provide a competitive advantage while also working with legacy systems. This approach enables all players in

the OSS/BSS supply chain to use the elements appropriate for their business with the confidence that they all fit together.

This new approach—called NGOSS (Next-Generation Operations Support Software)—is a comprehensive, integrated framework for developing, procuring, and deploying operational, business, and marketing support systems and software. Through an integrated system of business and technical elements, NGOSS allows support systems to become interoperable like never before. Now there is no clear characterization of or differentiator among the OSS, BSS, and MSS acronyms, and the three types of support systems (OSS, BSS, and MSS) together are referred to as OSS/BSS here, since there is virtually no difference between them.

3.5.1 Status, Definitions, and Markets for Support Systems

3.5.1.1 State of the OSS/BSS/NGOSS Market

Over the past few years, there has been a greater concentration in some OSS/BSS markets, and the market has seen players emerging and consolidating through mergers and acquisitions (M&As) worldwide to address evolving requirements, technologies, and standards. Some vendors have achieved significant market power with product suites that provide complete solutions. As a result, these vendors maintain thought leadership positions in the critical network and OSS/BSS/NGOSS focus areas. Among the key differentiating factors are integration capabilities, innovation, and client relationships.

However, the market is still diversified, with many global and local players competing for market share in the various regional markets. Vendors are seeking to expand and complement their core offerings. The current market consists of large global vendors that offer end-to-end OSS/BSS solutions and a number of smaller vendors that offer highly specialized point solutions and products. Vendors differentiate themselves mainly by the scope of their product suites, the standards and advanced features they support, the scalability of their offerings, their ability to provide professional services, and the modularity and interoperability of their solutions.

The acquisition of new entrants and established OSS/BSS single-product and product suite suppliers will continue, however, consolidating the market further. This may change products and solutions, depending on the firm acquired and the strategies and market focus the acquiring company may have. Vendors' sales and market strategies continue to depend on global and local systems integrators (SIs) to implement and support product suites. This is understandable considering the competition and the demands for global reach. Systems integrators and network equipment providers (NEPs) offer great opportunities for expansion into new accounts with diverse solutions.

3.5.1.2 Industry Issues of Support Systems

There are several issues that the industry must take into account when developing and deploying OSS/BSS systems.

3.5.1.2.1 Upgrading of Cycles in Support Systems

As a result of global deregulation, carrier competition is driving the demand for new, more efficient back-office solutions. In addition to reducing operating expenses, advanced OSS/BSSs should improve time to market and facilitate fast and more economic introduction of new, revenue-producing solutions.

3.5.1.2.2 Product-Based, Vendor-Driven Solutions

Carriers in general increasingly demand solutions rather than raw technology and development kits for custom-developed OSS/BSS solutions. The advent of technology standards encourages the use of best-of-breed, standards-based vendor solutions instead of homemade solutions. However, some large carriers want to be able to do a portion of the work (e.g., customization) themselves. What is common

in most instances is that they want to have a role in the support system's functionality. In a majority of cases, they want the support services of the vendor or the vendor's local partner.

3.5.1.2.3 Emergence of Complex, Multiplatform Environments

The need for reliability and scalability of large centralized systems remains. Service providers very often incorporate a multiplatform strategy augmenting existing investments in legacy solutions with newer technologies targeted at profitable customer market sectors. Interoperability of new systems and existing (even legacy) systems is a high priority.

3.5.1.2.4 Emphasis on System Integration

Complex multiplatform, multivendor telecom networks require substantial system integration for interoperability. With multiple client-server and legacy OSS/BSSs in place, integration capabilities of vendors are in high demand. Carriers need support in their efforts to consolidate legacy OSS/BSS infrastructures and provide comprehensive integration capabilities beyond their own technologies.

3.5.1.2.5 Service Providers' Investment Strategies

Large telecom providers are looking to "future-proof" investments in OSS/BSS suites, focusing on application components and the ability to buy what they need instead of complete product suites while addressing cross-domain service and functional demands. Large telecom providers will continue to select best-of-breed solutions from a limited number of OSS/BSS vendors. Small to midsize providers may prefer a complete product suite for an end-to-end solution from a single vendor.

3.5.1.2.6 Importance of Quick ROI

OSS/BSS vendors that can provide solutions to specific OSS/BSS problems that quickly generate a return on investment (ROI) to the service provider will do better in this continued investment-adverse climate. Many service providers are demanding "instant ROI" in order to justify their OSS/BSS systems and projects.

3.5.1.2.7 OSS/BSS Growth

OSS/BSS growth is tied to share shifts among telecom end markets and carriers. The strongest near-term growth has been achieved by vendors targeting the fast-growing telecom end markets, emerging local carriers, and wireless (mobile) carriers.

3.5.1.2.8 Developing Markets

OSS/BSS growth is dominated by new carrier adoptions, acquisitions, and incumbent upgrades. Developing markets, such as data solutions, local number portability, broadband, triple play, carrier interconnection, and converged and bundled services, are likely to justify the next wave of OSS/BSS spending.

3.5.1.2.9 Convergence and Telecom Consolidation

Convergence and telecom consolidation may accelerate the use of advanced OSS/BSSs. Consolidation of carriers across multiple end markets creates advantages for OSS/BSSs targeting multiple end markets. It increases the complexity of telecom networks and the complexity and required functionality of support systems and demands for OSS/BSS integration, probably on the basis of the NGOSS approach.

3.5.1.2.10 Outsourcing

Ongoing structural changes in the telecom industry have resulted in new requirements for support systems. In order to concentrate on customer management, some back-office functions may be outsourced to service bureaus. These service bureaus may use support systems from the same vendors, but they will use them in a shared fashion among multiple service providers.

3.5.1.2.11 Independent Software Vendors (ISVs)

ISVs are heavily expanding their professional services capabilities. However, they are competing with traditional SIs that implement and deliver next-generation projects and integrate numerous complex OSS/BSS products. ISVs will focus on channel strategies in addition to direct sales to enhance their business potential and long-term growth opportunities.

3.5.1.2.12 Professional Services

Carrier requirements extend beyond network class software to increased consulting and professional services aimed at customized integration and customer support. Most OSS/BSS software providers set up multiple relationships with global and local SIs and NEPs to implement, integrate, and support their products while expanding the breadth of their solutions.

3.5.1.3 Carrier Support System Issues

Not only the industry but also carriers are facing several issues never before experienced.

3.5.1.3.1 The Legacy Dilemma

Most service providers today carry an operational burden in the form of their legacy architectures. This aspect not only applies to incumbents; even recently established service providers have managed to build a tangled web of support systems in quite a short time frame.

The root of the problem is the now-classic stovepipe syndrome: Service providers have built separate sets of systems architectures where each independently supports only one type of transport service (e.g., plain old telephone system [POTS], asynchronous transfer mode [ATM], F/R, SDH, mobile). Predominantly, they consist of a patchwork of niche systems developed in-house in the late 1970s and early 1980s. These systems are often huge applications serving several processes in several layers with a large number of peer-to-peer system interfaces creating rigidity that cannot possibly match existing requirements in terms of flexibility and time to market (or money). The inherent source of the problem, the legacy architecture, is crucial for day-to-day operations, and the data on which operations rely is stored in outdated databases; thus, a catch-22 situation is created. The consequence is that service providers still have to create reliable transformation processes to enable migration to new OSS/BSS environments, which is a prerequisite for survival in the Inter Connection Technology (ICT) industry.

3.5.1.3.2 The Migration Dilemma

Spilling over from the legacy dilemma, an even greater difficulty for service providers is being able to execute a manageable migration to a new, NGOSS-based OSS/BSS architecture that supports novel ways of conducting business and dealing with customers, products, and resources. Migrating to a new environment requires a substantial time period, often several years, since a "big-bang" approach is too dangerous, with the risk of fatally disrupting critical business operations. The long migration period in turn creates a number of costly problems, since the legacy still needs to coexist along with the new OSS/BSS to facilitate exchange of information between the two architectures. These high costs will create confidence problems in the organization, since the key motivator for deploying a new architecture is cost savings. Very often the parallel costs during migration (i.e., operating the legacy and new support systems) are not fully visible in the decision-making process. In the case of most service providers, senior management personnel have yet to come to terms with the real costs involved in implementing a new OSS/BSS architecture.

Another important issue to keep in mind is the close kinship between support system strategies, business plans, and service portfolio strategies. Moreover, if unforeseen costs are to be avoided, introduction of new services, possible retirement of existing ones, and deployment of new technologies must go hand in hand with OSS/BSS strategies when organizations are planning ahead for the scope of migration.

Table 3.5.1 presents a comparison of different approaches to migration and development of OSS/BSS systems.

TABl E 3.5.1 Comparison of Different Approaches to OSS/BSS System Migration and Development

Approach	Advantages	Disadvantages
Ad hoc tools	Fast to implement, flexible to change	High operational costs, high error rate, poor customer support; does not scale
Custom-specific application	Specific to current operation needs	Expensive and time-consuming to change for new services
Function-specific integrated application	Complete function-specific application; prebuilt functions; may support several different services through one platform	Expensive to change; inhibits customer-centric support for multiple services
Best-in-breed OM application	Incorporates commonly used OM functions; may support several different services through one platform	Requires integration with other OSS/BSS systems
EAI/BPM workflow-based application	Easy to add new capabilities; end-to-end process control across systems, work centers, and partners	Requires development of specific management (e.g., order management) functionality (unless prebuilt components are deployed)

3.5.1.3.3 Evaluation of Suppliers

Carriers should evaluate suppliers on their ability to customize and integrate products, as well as on their ability to provide business consulting services and financial conditions that help carriers develop new services with optimal chances of success. SIs are well qualified to take a leading role in managing the delivery and implementation of complex next-generation OSS/BSS projects. SIs have (in theory) crucial methodologies and resources and are certified for most vendor technologies. Suppliers should be evaluated according to a variety of parameters, including:

1. Ability to meet operating expenditure (OPEX) and capital expenditure (CAPEX) requirements
2. Implementation of service-level agreements (SLAs), from service creation to quality and performance of services
3. Adherence to evolving standards and technologies, providing interfaces to applications and databases that integrate with their products
4. Capability to provide consulting services beyond their own products
5. Financial conditions aligned with slow or even hypothetical revenue growth for new services

3.5.1.3.4 Convergent Billing

The customer may expect to receive one bill for all services, such as voice, data, video, and Internet. The minimal requirement is to receive multiple bills with electronic staples.

3.5.1.3.5 Rapid Fulfillment (Provisioning) of New Services

Based on additional OSS/BSSs, provisioning can be expedited by better interfaces and more accurate data.

3.5.1.3.6 Service Differentiation

New services can be created and deployed with existing infrastructures. If the value-added nature of these services is carefully defined, customers may consider them differentiators.

3.5.1.3.7 New Service Offerings

Incumbent service providers are expected to react rapidly to new communication needs, including offering Internet access at a reasonable cost and deploying xDSL (digital subscriber line), virtual private networks (VPNs), and voice over IP (VoIP).
In each of these cases, either the deployment of new OSS/BSSs or the customization of existing systems is required. In both cases, additional market opportunities are available for support system vendors.

3.5.1.3.8 Outsourcing

The deregulation of the telecommunication industry has led to outsourcing of operational support development activities, for example, to spin-off companies and other software vendors. Several carriers have evaluated the benefits of outsourcing their back-office services partly or entirely. Outsourcing will eliminate the need for carriers to invest scarce research and development dollars in support systems. Essentially, it allows carriers to focus on their core business. Customers may not be aware of where OSS/BSS services come from. Today's outsourced solutions are service bureaus. All or some of the carrier's support systems are outsourced. In the latter case, the vendor relies on remote access to the carrier's existing solution to deliver incremental functionality. For most emerging carriers, the benefits of outsourcing outweigh the negatives.

3.5.1.3.9 Professional Services

The trend with respect to professional services is unmistakable. Some geographic regions expect to have somewhat higher growth rates as a result of service providers' greater dependence on vendors for subject matter expertise, project management, and system integration.

3.5.1.4 Summary and Conclusions

Over the past few years, there has been greater concentration in some OSS/BSS markets, and the market has seen players emerging and consolidating worldwide through mergers and acquisitions (M&As).

However, the market is still diversified, with many global and local players competing for market share in the various regional markets. Vendors are seeking to expand and complement their core offerings. The current market consists of large global vendors that offer end-to-end OSS/BSS solutions and a number of smaller vendors that offer highly specialized point solutions and products.

The key differentiating factors are supporting convergent services and next-generation networks, integration capabilities, scalability, innovation, and client relationships.

The acquisition of new entrants and established OSS/BSS single-product and product suite suppliers will continue, however, consolidating the market further. This may change products and solutions, depending on the firm acquired and the strategies and market focus the acquiring company may have. At the same time, new players are entering the market, and the market still will be fragmented.

Vendors should support service providers in reducing operating expenses, improving time to market, and facilitating the fast and more economic introduction of new, revenue-producing solutions. Typically, best-of-breed, standards-based vendor solutions are preferred to homemade solutions, but some large carriers want to able to do a portion of the work themselves. Consolidation of carriers across multiple end markets creates challenges for OSS/BOSS vendors targeting multiple end markets. As mergers increase, the complexity of telecom networks, the required functionality of support systems, and demands for OSS/BSS integration increase. As a result of the ongoing structural changes in telecoms, some back-office functions may be outsourced to service bureaus.

If SPs are to survive and ride the convergence wave successfully, it is critical that they streamline their support systems and business processes. A strong OSS/BSS suite will help them cut costs, stay ahead of competitors, better retain customers, increase wallet share, and minimize losses from declining fixed-line voice revenues.

New-generation support systems pose OSS/BSS challenges for service providers with respect to legacy and migration. Most service providers today carry an operational burden in the form of their legacy architecture. These systems are costly and time consuming to upgrade to a convergent/NGN (next-generation all-IP network) environment. The legacy architecture is crucial for day-to-day operations, and the data on which operations rely is stored in outdated databases. Service providers still have to create reliable transformation processes to enable a migration from legacy systems to new OSS/BSS environments, a prerequisite for survival in the ICT industry. An even greater difficulty for service providers is being able to execute a manageable migration to a new NGOSS-based OSS/BSS architecture

that supports novel ways of conducting business and dealing with customers, products, and resources. Migrating to a new environment requires a substantial period of time, often several years, since a "big-bang" approach is too dangerous, with the risk of fatally disrupting critical business operations.

Carriers should evaluate suppliers on their ability to customize and integrate products as well as on their ability to provide business consulting services, financial conditions that help carriers develop new services with optimal chances of success. Other user-oriented points are fast fulfillment time and provision of differentiated services for customers. From an operational cost point of view, correct use of outsourcing and professional services offered by venders and independent companies is critical to analyze.

3.5.2 OSS/BSS Market Drivers

The market is changing very rapidly. OSS/BSSs should be positioned well and should meet telco expectations in a timely fashion.

OSS/BSS drivers are responses to the forces of change in the global telecom industry. These forces include technology improvements; deregulation; continued privatization and consolidation within the telecom sector; convergence of technologies, standards, and open systems; and—most of all—competition. OSS/BSS systems should support organizations in meeting three primary business goals:

1. *Revenue protection:* by facilitating a better customer experience, self-service capability, and converged services
2. *Revenue generation and growth:* by enabling new revenue-generating services to be launched more quickly
3. *Management of costs:* by providing the capabilities necessary to manage and leverage the NGN

These forces and the changes they are causing require carriers to adopt specific strategic goals in order to succeed. These strategic goals then create operating goals, which are met through new and improved information systems (OSS/BSSs).

Today, with revenue growth prospects rooted in reality, profitability is the new rallying cry. From the OSS/BSS perspective, the focus is on ROI. Automated online billing, network asset management, configuration management, automated provisioning, and customer self-service are experiencing renewed interest because of their ability to directly reduce the capital expenditures (CAPEX) and operational expenses (OPEX) associated with delivering services and maintaining the network.

3.5.2.1 Deregulation and Privatization

Deregulation is designed to encourage competition through the proliferation of new entrants. Looking to gain share, carriers are entering each other's markets, blurring traditional lines between services, geographic coverage, and communication platforms. Aggressive new carriers have moved rapidly to establish nationwide service networks, consolidating local, long distance, Internet, wireless, and cable services under one umbrella. Incumbent carriers are trailing this way of convergence.

As a result of deregulation and privatization, competition is everywhere: in long distance, local exchange, ISP, cable, wireless, mobile, multimedia, broadband, content, presence, and so forth. In many cases, OSS/BSSs are the differentiators. The best OSS/BSS opportunities are seen with competitive local exchange carriers (CLECs). OSS/BSS requirements vary substantially from carrier to carrier. As a result, CLEC-OSS/BSS strategies range from internal development to outsourcing, systems integrators, and third-party software/service providers. CLECs can be small or midsized, and they may or may not own facilities. In all cases, they must interoperate with incumbent local exchange carriers (ILECs) by opening OSS/BSSs to gain access in various phases of provisioning and order processing, as well as in service activation.

Telecommunications service competition began in the 1980s in the United States, led by MCI, with OSS/BSSs playing a key role. The AT&T divesture in 1984 marked a major breakthrough. The second significant milestone was the Telecom Act of 1996. The Telecom Act in the United States and "liberal-

ization" of markets across Europe and Asia created a boom market by lowering barriers to entry, which created waves of new "service providers."

Deregulation requires operators to open up their networks and provide a number of wholesale products to others in the Information and Communications Technologies (ICT) value chain, including their own new business units in these adjacent industries (M3050.2). In the initial stages, the former monopoly providers, regional Bell operating companies (RBOCs) in the United States and Post, Telegraph, and Telephone (PTTs) everywhere else, suddenly found themselves in a competitive marketplace, and they had to focus on customer care and roll out new services to avoid being rolled over by the upstarts.

It should be noted that the trend toward convergence of telecommunications, media, and Internet services has had different effects on traditional telecommunications regulations in the United States and the European Union (EU). In the United States, liberalization of local telecommunications services was first accelerated by intensive unbundling regulations. When infrastructure-platform competition increased, these regulations were subsequently curtailed dramatically. Today incumbents are obliged to offer unbundled access to the local loop only for narrowband voice telecommunications services. For broadband access, U.S. regulations currently rely on competition from other infrastructure platforms.

In Europe, full liberalization of telecommunications service provisions was introduced in 1998. Convergence was accounted for in the new EU regulatory framework for electronic communications of March 2002, which equally applies to different electronic communications platforms. In Europe the advent of network convergence has not had the effect of reducing regulation of wholesale telecommunications inputs. Rather, market power regulation has been extended to sectors previously not included.

It is assumed that other regions—Asia/Pacific, South America, and Africa—will follow these deregulation and privatization trends. All of this has led to unprecedented growth in the communications software market, the small number of OSS vendors suddenly being joined by hundreds of software competitors seemingly overnight.

3.5.2.2 Growth of the Global Telecommunications Market

Explosive telecom expansion (in addition to competition) is forcing telecommunications service providers to manage the lifecycle of their support systems, either assessing the productivity of their current support systems or replacing them with new, future-proof systems. Growth not only means increases in the number of subscribers for existing services, but new services are provisioned on existing infrastructures, and completely new services on new infrastructures are deployed or acquired. As a result, increasing the capacity of existing OSS/BSSs is not sufficient; rather, the capability to manage new infrastructures and new services should be enhanced as well.

3.5.2.3 Increasing Network Complexity

As a result of customer expectations and aggressive competition, the time to market of new services is extremely short. Telecommunications service providers do not have the time to build new infrastructures; instead, they must combine existing and new infrastructures, such as copper, fiber, and wireless. Use of existing infrastructures is also forced by the need to reduce costs.

Growth in the number of network elements has been accompanied by an increase in the complexity of items to be managed. Although carriers talk about "the network" as if their infrastructures were one monolithic system, in reality most carriers run their business on a hybrid mix of legacy (TDM, Sonet/SDH) and next-generation (IP/MPLS, Ethernet, WDM, WiMAX) technologies and equipment from a wide variety of NEPs. To make matters even more complicated, 2G, 3G, and 4G all require different and overlapping infrastructures. These infrastructures are extremely complex, with a high degree of interdependence among network elements. Whenever carriers add something new to their networks, there are hidden hassles and costs.

Carriers often deploy emerging services on the basis of a mixture of infrastructures as an overlay. Emerging services use both emerged technologies (voice networks, ISDN, circuit switching, packet switching, message switching, Frame Relay, Fast Ethernet, Fast Token Ring, Fiber Distributed Data Interface

[FDDI]/CDD, ATM, mobile, wireless, Sonet/SDH, xDSL) and emerging technologies (3G mobile and Wi-Fi/WiMAX, WDM, high-speed IP routers, optical switches, IP multimedia subsystem [IMS]).

This in turn makes service deployment, activation, assurance, and fault isolation a challenge, especially as the number of service providers increases, along with the number of interworking "foreign" interfaces. As networks shift from lower-speed, dedicated-rate, inflexible services to mobile, fully configurable, bandwidth-on-demand and high-speed services, OSS/BSSs must adapt to this new situation. When services are offered in combination, OSS/BSSs should be modified, reengineered, and connected to each other. The tough part today is not the choice of technology but how to fit all of these neat technology pieces together—and do so in a reliable, scalable, and cost-effective way. This opens new business opportunities for OSS/BSS vendors.

The introduction of standards for support systems is accelerating the demand for third-party OSS/BSSs. Legacy systems are primarily proprietary systems that are not integrated across functional areas. Service providers depend upon custom development by internal development staff and outside integrators to connect various support systems. The answer for carriers will be pluggable EMS/NMS/OSS/BSS frameworks, which are expected to solve most multivendor, multidomain issues.

3.5.2.4 Convergence

Traditionally, the term fixed-mobile convergence (FMC) has been used by the telecom industry when discussing the integration of wireline and wireless technologies. However, convergence not only involves FMC; it also involves convergence between the media, datacom, and telecommunication industries. Convergence is occurring at multiple levels—the network level, the service level, the operator level, and the vendor level—but it is occurring fastest at the device level.

A decade ago, construction of an international communications report was a reasonably straightforward affair. Nearly all of the main communications metrics were easily measurable and reflected accurately the way in which consumers used services. One could look to primary Public Switched Telephone Network (PSTN) fixed-line telephone network statistics (almost always the incumbent in each country) to examine the voice call patterns of consumers within their homes. Analyses of mobile operators could provide an accurate picture of the ways in which consumers used voice services when they were away from home. The Internet was still in its infancy, and virtually no residential broadband service was available; thus, one could capture use of the Internet by monitoring dial-up calls to Internet service providers (again, the vast majority of this traffic would be carried by the incumbent fixed-line operator). And pay TV services were provided over a limited number of platforms with a small number of price points.

However, the act of measuring and analyzing consumer behavior in the communications market has become far more complex. New sectors, products, and services have sprung up; new operators have become established in all major sectors; and both traditional and new services are being delivered to consumers over multiple platforms. This combination of new products, operators, and delivery channels has combined to fuel the process of convergence.

While convergence is redefining the way business works, consumers are the primary beneficiaries. Convergence provides the ability to mix and match services to cater to market demands for flexibility, quality, and value. Changing the communications landscape rapidly, convergence has affected devices, services, and networks. For example, FMC handsets now provide consumers with the ability to switch between their landlines and mobiles. This offers consumers the cheapest network available at a given location. While this reduces the cost of communications to the consumer, it exerts dynamic pressures on SP revenues. SPs have coupled these external drivers of convergence with internal drivers to create flexible and innovative models that will better meet the needs of consumers.

Convergence is fundamentally transforming the way SPs operate. Falling PSTN revenues coupled with triple/quad play are driving SPs to formulate convergence-led strategies for future growth. The critical question that SPs need to ask themselves is whether they have the process and IT flexibility to cut across silos of products and services.

3.5.2.4.1 Industry and Communications Convergence

Industry convergence occurs when firms and industries that were once independent become competitive, complementary (mutually dependent), or both. Often convergence is associated with industrial reorganization (e.g., mergers and divestitures) as firms adapt to changing realities.

Advanced technology, coupled with deregulation and aggressive competition, is driving communications convergence. Customers prefer to receive all types of services, including long distance and local voice, data/Internet, cable/video, and wireless access, from the same service provider. Voice is expected to support both local and long distance, requiring it to play an LEC and Inter Exchange Carrier (IEX) role at the same time. Data is gaining importance for both local and long distance and usually includes Internet access. Data is supposed to reach voice volumes within 5 years, necessitating complete rebuilding of circuit switching technology. Cable is expected to accommodate voice and data in addition to video. Wireless includes all kinds of mobile services and satellites supporting voice, video, and data.

SPs have traditionally focused on managing fault and performance problems more with regard to their networks and less with regard to their customers. Enabled by major technological advances, competitive forces are now bringing more choices, reduced pricing, and increased flexibility to all customer segments. Incumbent SPs are realizing what competitors already know: The network—wireline/wireless, circuit-switched/IP—is not a source of competitive differentiation anymore, and it is not the most important focus for long-term corporate success. The services riding these networks, often supplied through a combination of sources, are the real differentiators and revenue generators. Through improved capabilities and partnerships, SPs are improving in terms of delivering what customers want, including Web access, media and entertainment content, business applications, low-cost voice, and yet-to-be-named quality-of-life improvements.

3.5.2.4.2 Service Convergence

Convergence of services means that multiple services are offered for the customer on the same (common) device and the same (common) network. Services that formerly were offered separately are now being offered together; a multitude of services (person to person, person to content, and content to person) can be provided to the same user over different access networks and to different devices.

Service convergence generates new revenue streams, including Short Messaging Services (SMSs) sent between mobile and Instant Messaging users and video calls made between fixed and mobile devices. Other examples of revenue sources are presence services, surveillance services, and blogging.

Multimedia and multimodal services take convergence to new levels. VoIP, presence, push-to-talk, messaging, interactive applications, data or video sharing plus streaming, browsing, and downloading are being delivered over fixed and mobile packet networks. To launch new services and applications quickly, operators can use an IMS to eliminate the complexity of different network service platforms. A standards-based service delivery framework (SDF) provides comprehensive lifecycle management, making the launch of new services and applications quicker and easier to integrate and operate, delivering solutions more rapidly to market, and reducing the total cost of ownership. In effect, the operator can provision—and the end user can quickly and conveniently self-provision—new services.

However, the fact remains that voice is still the optimal way of communicating in most cases. Making voice services more convenient to use is just as important as adding new functionality. However, equal to end users' perceptions of greater simplicity is the complexity of delivery for operators against the background of a fiercely competitive market.

VoIP and Instant Messaging are two developments that helped kick-start service convergence. VoIP has had a seismic impact on telephony within enterprises, and as the penetration of broadband access increases, so does the availability of this transport mechanism within the home. Users also benefit from personalized VoIP, including same number, same contacts, and the same supplementary services (e.g., call barring, call waiting, ring back tones, one voice mail) through any access network. IP Digital

Subscriber Line Access Multiplexers (DSLAMs) allow operators to offer both DSL access and traditional two-wire POTS connections using an SIP client in the DSLAM. This development and others such as fixed VoIP phones, Analog Telephony Adapters (ATAs), and fixed soft switches place fixed-line operators in an excellent position. They can offer multimedia services via DSL and attractive tariffs for analog POTS connected to an IP network, thereby maintaining existing services where required and evolving the core network to an IP-based solution. Smart phones, on the other hand, have Wireless Local Area Network (WLAN) interfaces so that they can access fixed broadband networks. This allows the mobile phone to be used as an IP phone and users to continue employing their personalized services at home, or via WLANs, connected to DSL in hot spots or offices. Convergence in this case enables a practical combination of cellular and fixed broadband access. The user experience does not change; the same voice and multimedia services are used in the same way.

Fixed-to-mobile substitution and fixed VoIP are gradually replacing PSTN voice telephony. Multimedia services are being delivered over fixed and mobile packet networks. Operators must now decide on the kinds of services they wish to provide themselves or through partners, to whom, and in which regions. And what they can offer is no longer limited to traditional telecom services; for example, they can perhaps gain entry into new businesses such as surveillance solutions.

3.5.2.4.3 Fixed-Mobile Convergence and Substitution

One of the most common types of convergence is FMC, often confused with fixed-mobile substitution (FMS). FMC is the removal of distinctions between fixed-line and wireless telecommunication networks; a combination of fixed broadband and local access wireless technologies is used to create a seamless service. One of the most eagerly anticipated changes in communication and service provision for end users, FMC poses unique challenges for operators and service providers.

FMC and FMS are two warring forces. On one side, fixed-line incumbents are looking to FMC to allow them to better leverage their in-house fixed and mobile assets and offset dwindling fixed-line revenues. On the other, mobile operators are stepping up their substitution strategies to reroute more fixed traffic over their own mobile networks. FMC is not completely in force in Europe, but mobile operators have been quick to see the writing on the wall and spent much of 2005 building an arsenal of FMS services to offset its arrival. Driving substitution has become mobile operators' number one strategy against FMC, and the rate of innovation in this area over the past few years has been extraordinary.

Substitution services have been grouped into six categories: home zone services, fixed alternative services, flat rates, family tariffs, bundled, and in the business sector, mobile VPNs. All six are proving to be successful to differing degrees. Results in the form of higher mobile shares of total market outgoing minutes are likely to start trickling in by mid-2008.

In conclusion, FMC and FMS will continue to play off each other in the consumer space, with pros and cons on either side.

3.5.2.4.4 Network Convergence

Network convergence implies consolidation of the network to provide different user services, with telecom-grade quality, over several access types with an emphasis on operator cost efficiency and support with respect to user service convergence. Network convergence simplifies the end-user experience and dissolves the barriers and complexities that separate today's network islands. The same services are available across all networks and, in an ideal world, appear and perform in exactly the same way, making usage easy, transparent, and intuitive.

From an operator's perspective, the goal of network convergence is to migrate today's separate PSTN, Private Line Network Management (PLNM), backbone, and IP networks to a fully converged network that supports any access technology. The full evolution includes a cost-effective migration to an all-IP network using IMS as the unifying platform, allowing all new services to be accessed in a standard and consistent manner. Advancing in this evolution will be the key to an operator's ability to reduce OPEX and CAPEX and increase competitiveness and profitability.

Many locations (e.g., homes, enterprises) already have access networks available (xDSL, WLAN, cable, etc.). When operators launch new services such as video streaming or hosted e-mail, they can take advantage of these existing networks, extending service access to more potential subscribers. In turn, this will mean launching services to new market segments for new revenue opportunities. With multiple access networks, operators can attract existing and new customers with an enhanced convergence service portfolio using unified billing.

A converged core network is the key enabler for converged networks. Multiple access to a common, converged core network enables cost optimization for both mobile and hybrid operators. Reuse of existing access network infrastructure and integration with the service infrastructure results in both OPEX and CAPEX savings. In addition, multiple access enables operators to introduce end-to-end quadruple-play services to new customers.

Although wireless and wireline convergence is capturing headline attention, new IP-based technologies are changing the competitive landscape. For example, the voice calling market was dramatically altered when IP-enabled network capabilities (VoIP) were installed by competitive entrants and incumbents alike. This forever changed the way services such as long distance were packaged and offered. As the industrywide "IP over fiber" network infrastructure initiatives continue to take shape, several new service capabilities are emerging, many of which were never thought practical in the past.

Platform convergence is the most advanced—and fastest growing—form of convergence. It enables consumers to access multiple products and services over a single platform, and often over one device, with a single operator relationship. Examples include VoIP telephony—both fixed and mobile—and TV over mobile devices.

3.5.2.4.5 Device Convergence

Most of the devices used for communication are constrained in their functionality or limited by the network they access. Typically, a device is used only for a single purpose, with support for its other functions limited. PSTN phones, low-end mobile phones, and set-top boxes are good examples. Consumers use these devices for a single purpose. When they change tasks, they change devices and access networks. This results in service islands, which lead to mismatched user experiences from different public and private networks. What is needed are unifying devices that can access services in a similar and easy way. Convergence of end-user devices is necessary to use converged services.

Device convergence refers to common devices supporting several access types, such as CDMA2000, WCDMA, Group Speciale Mobile (GSM), fixed broadband, and WLAN. Device convergence allows multiple applications to be run reusing the same functions for identification and authentication. Furthermore, mobile devices are supporting more and more functions in addition to telephony (e.g., cameras, TV/video, and e-mail).

Device convergence brings diverse functionality together in a single device; for example, a phone with a camera, FM radio, TV, Internet browser, and MP3 player. Device convergence then works together with network convergence to provide connectivity to services using the access technology most suitable at a particular location or moment in time. Service convergence enables seamless and transparent delivery of services to users over any network. Concurrent delivery of all major media types—voice, data, and video/TV—over fixed lines has been around for some time as triple play. But FMC adds mobility to the mix (quadruple play) and allows the same services to be used with different devices and through different access networks while users are on the move. End-user demand for mobile access to services—traditionally available only in the fixed domain—has become one of the main business drivers.

The most well-known type of device convergence is that associated with mobile handsets. The mobile handset and laptop PC markets are driving service integration between carriers and traditional network technologies. For example, FMC is gaining momentum as new customer "interface devices" enable seamless roaming across both IP and cellular networks. This often results in improved customer perceptions of mobile service quality in instances in which cellular network coverage is less than opti-

mal. These devices are also ushering in new capabilities for the delivery of services and content, which generally involve complex interactions between network capacity and content delivery.

Device convergence allows consumers to access many different services from the same device, even if they are delivered over different platforms (an example might be a mobile phone with an integrated FM radio). The technological differences between PCs, mobile devices, e-books, TVs, and cellular phones will be eradicated.

The three worlds of communications, computing, and consumer electronics are coming together as new multimedia functions and multiple technologies are being built into handheld wireless devices. The new multifunction devices open up new opportunities for mobile operators to extend their service offerings, drive up average revenue per user (ARPU), and reduce churn. Yet the parallel trend of integrating noncellular technologies, such as Wi-Fi and Bluetooth, to create multimode devices may benefit fixed operators rather than mobile operators.

These trends are also bringing the devices and their suppliers—handset vendors and operators—into competition with traditional and new media devices and services such as digital cameras, music players, PCs, and even the ubiquitous TV set. The result could be collaboration previously unheard of between entertainment/media companies and telecom service providers, as well as between consumer electronics and telecom device vendors.

3.5.2.4.6 Digital Convergence

Digital convergence also presents challenges to the industry. The boundaries between consumer electronics, media, entertainment, software, and telecom firms will continue to erode. Consumer electronics firms may find themselves competing with unexpected rivals. The firms that survive and prosper through this difficult transition will be good at spotting opportunities and moving quickly to market. Speed is of the essence in the convergence game, as is collaboration. In addition, companies must be open-minded about potential new revenue streams, even if this means collaborating with erstwhile competitors.

3.5.2.4.7 Telecom-Media Convergence

No longer are telecom and media companies confined to their own markets. Fixed, mobile, and IP service providers can offer content and media services, and equipment providers can offer services directly to end users. Content providers are consistently looking for new distribution channels. Convergence refers to the combination of all of these different media into one operating platform—the merging of telecom, data processing, and imaging technologies. This convergence is ushering in a new multimedia epoch in which voice, data, and images are combined to render services to the user. The key result of convergence at a macro-business level is the merging of the telecommunications and media/entertainment industries.

3.5.2.5 Summary and Conclusions

Deregulation and privatization have forced aggressive competition in all areas: long distance, local exchange, ISP, cable, wireless, mobile, multimedia, broadband, content, presence, and so forth. Today, with revenue growth prospects rooted in reality, profitability is the new rallying cry. From the OSS/BSS perspective, the focus is on ROI. Thus, the basic drivers are financial forces (cost savings, revenue growth), and OSS/BSS drivers are responses to the forces of change in the global telecom industry.

Explosive telecom expansions (in addition to competition) are forcing telecommunications service providers to manage the lifecycle of their support systems, either to assess the productivity of their current systems or to replace them with new, future-proof systems.

As a result of customer expectations and aggressive competition, the time to market of new services is extremely short. Telecommunications service providers do not have the time to build new infrastructures; rather, they must combine existing and new infrastructures, such as copper, fiber, and wireless. This combination of old and new results in increased network complexity. Although carriers talk about "the network" as if their infrastructures were one monolithic system, in reality most carriers run their

business on a hybrid mix of legacy (TDM, Sonet/SDH) and next-generation (IP/ MPLS, Ethernet, WDM, WiMAX) technologies and equipment from a wide variety of network equipment providers.

Traditionally, the term fixed-mobile convergence (FMC) has been used by the telecom industry when discussing the integration of wireline and wireless technologies. However, convergence not only involves this particular kind of convergence; it also involves convergence between the media, datacom, and telecommunication industries. Convergence is occurring at multiple levels—the network level, the service level, the operator level, and the vendor level—but it is occurring fastest at the device level. Several types of convergence are having an impact on the market: industry convergence (mergers and acquisitions), service convergence (multiple services together can be offered on the same [common] device and the same [common] network, including FMC and FMS), device convergence (common devices supporting several access types and several services), digital convergence (boundaries between consumer electronics, media, entertainment, software, and telecom firms will continue to erode), and telecom–media convergence (fixed, mobile, and IP service providers can offer content and media services, and equipment providers can offer services directly to end users).

As a result, automated online billing, network asset management, configuration management, automated provisioning, and customer self-service are experiencing renewed interest because of their ability to directly reduce the capital expenditures (CAPEX) and operational expenses (OPEX) associated with delivering services and maintaining the network.

3.5.3 Future of OSS/BSS

The telecommunications industry, in support of the underlying developments in society as a whole, is undergoing fundamental change as service providers move toward the implementation of next-generation all-IP networks (NGNs). This movement toward NGNs is motivated by the need to support new market demands, as well as the need to exploit the synergies that new technologies enable. Service providers need to consider four elements:

1. New IP networks and their associated new IP multimedia services
2. Changes in the area of customer care
3. Reconsideration of ordering and billing processes
4. Transformation of OSS/BSSs

Many SPs suffer from outdated processes and systems associated with the management of customer and network processes. Incumbent legacy support systems account for a very large part of the operational expenditures of SPs—approximately 12 to 15%. However, these systems too suffer from outdated and fragmented processes and labor-intensive maintenance activities. As SPs move toward a system where product development lifecycles are shortening and there is a need to implement and retire services on demand, OSSs must be able to support this new approach. Flexibility and speed of change will be the keys to success. Table 3.5.2 compares the traditional and the new telecommunication models.

Table 3.5.3 presents differences from an OSS/BSS perspective.

3.5.3.1 OSS Transformation

Traditionally, service providers have managed their networks in a technology-focused way. The approach has been very network centric in focus, which has resulted in "stovepipes" that specialize in managing different segments and technologies of the infrastructure: customer premises equipment (CPE), edge, core, call path, non–call path, IT/applications, and third party. Mergers and acquisitions have increased the number of such silos in any given organization, and the introduction of new services (frequently on separate, dedicated platforms) has only increased the problem.

The result is that, in managing multiple networks across different lines of business and geographies, operational systems have become highly complex. This complexity manifests itself generally as a large number of self-built and COTS (Commercially Off the Shelf) applications, many point-to-point integrations

TABl E 3.5.2 Comparison of Traditional and New Telecommunications Models

Area	Traditional	Current/Future
Service control	One SP	Many third-party providers
Understanding of network	Network capacity well understood	Different quality of service required by each application
Service mode	Single-mode services	Multimode services
Focus of operation	Network	Customer
Access network	One access network transmitting one service	Many services over many access networks
Problem isolation	Single-domain problem isolation	Multidomain performance and root cause analysis
Role of software solutions	Serving operations and engineering	Serving operations, engineering, customer care, quality, and product management
Service providers' position	Monopolies	Competition
Financial	Guaranteed rate of return	No financial guarantees
Control	Government bureaucracies	Market; fight for market share
Efficiency	Inefficiency rewarded	Efficiency rewarded

TABl E 3.5.3 Differences from OSS/BSS Perspective

Area	Traditional	Current/Future
OSS/BSS	Silo-based, single-domain OSS/BSSs	Multidomain systems
Database	Isolated databases	Integrated with ERP systems
Type of asset handling	Asset management	Asset optimization
Workflow	Workflow management	Enhanced customer experience
Function	Supports services	Enables and supports services
Position of OSS/BSS	Cost center	Strategic investment

between applications, highly distributed data, and poor end-to-end process integration. Furthermore, new services are generally of a "bundle" nature, therefore implying various technologies in a single provisioning order. This obviously increases the destructive impact of having multiple silos for the same process.

The outcome is low efficiency driven by low data accuracy, high maintenance costs, manual process management, and poor collaboration with suppliers. The financial impact is low asset utilization, reduced cash flows resulting from increased intervals of time to complete orders, reduced revenue from higher order dropout rates and lower service quality, and higher costs of both planned and unplanned maintenance. Deploying new networks (including the transition to all-IP) and services and adding new businesses through acquisition place increasing pressure on operational support systems by increasing this complexity. Thus, in order to support the drive to new revenue, next-generation OSS must be much more flexible than current solutions.

SPs need to consider the effects of convergence, the new-generation IP networks (NGNs), and their associated IP multimedia services and make the necessary changes to their OSS/BSSs. They should undertake an urgent review of their OSSs and look to quickly initiate a transformation, thus generating immediate cost savings while positioning themselves for NGNs and the advent of IP multimedia system (IMS) services. OSS/BSS and process transformation are critical in lowering the costs of delivering existing and new services and improving the flexibility of SP businesses. The benefits of an OSS transformation are twofold:

1. It delivers cost savings and operational effectiveness improvements in running the existing business.
2. It delivers the necessary systems, infrastructure, and processes to enable a smooth, cost-effective transition to an NGN.

TABl E 3.5.4 Changes from Product-Centric Approaches to Customer-Centric Approaches

FROM	⇨	TO
Product/network centric		Customer centric
Organization stovepipes		One team
Vertical organization		Horizontal organization
Technology led		Business partners
Product stovepipes		Holistic solutions
Old portfolio		New wave services
Telco		IT and network (telco) services

3.5.3.2 Customer Orientation: Customer Care and Management

Competition is driving telco service providers to change their network-centric approach to a customer-centric approach. The main changes are summarized in Table 3.5.4.

Carriers will likely focus on improving the overall value of their services—quality, support, and price—as a means of retaining customers. Many of these improvements will come from advanced OSS/BSSs. In addition to improving the customer interface (e.g., offering Web access), granular data available with new OSS/BSSs can be utilized to retain key customers and reduce customer churn. Over a longer range, further differentiation is expected. Similar to other industries, high-margin customers may receive special treatment, while average customers receive an average level of service.

3.5.3.2.1 Customer Network Management

Effective end-to-end management of today's globalized, heterogeneous corporate customer networks, based on a mixture of several technologies, is essential for customers. Current management systems and OSS/BSSs generally involve different management policies and management information models. However, large-end customers increasingly want to buy network capacity and complete the final stages of service provisioning via a service offering called customer network management (CNM) on the basis of their own management policy. CNM can help facilitate the efficient allocation and management of network resources.

CNM incorporates a class of OSS/BSSs that enables end users to securely view, troubleshoot, and reconfigure their "own" network and generate reports on their subscribed telecommunication services. CNM provides strategic links to the customer and allows service providers to further differentiate their offerings. OSS/BSS vendors are expected to offer the following:

1. *Performance:* extraction of information from the network without slowing overall network operations
2. *Customization:* packaging information so that customers can receive an appropriate level of detail in a way they can understand
3. *Security:* delivery of information to the customer in a cost-effective and secure manner so that customers see only relevant information about their portion of the network

CNM capability provides various network information specific to the customer's service. Information regarding traffic patterns, operational status, and trouble spots can be obtained by using a standard Web browser application. A limited set of control capabilities should also be available through CNM. It is expected that CNM will eventually migrate to the customer contact point for subscription, activation, monitoring, and control by the customer. CNM systems are most often implemented on the corporate virtual private network.

3.5.3.2.2 Customer Experience Management (CEM)

CEM is a term borrowed from other industries and partially from the customer-facing side of the communications business. While specific uses of CEM vary by department, the basis for CEM and a focus on CEM principles is the bridging together of processes and data to efficiently manage a customer's use of the products and/or services offered by all involved suppliers (e.g., SPs and their content partners). In its broadest sense, CEM is the practice of collecting customer usage data from all practical sources (network devices, content servers, and management databases) to be used by both business and technical departments in accomplishing two specific goals:

1. *Establishing an internal view of the customer service experience:* This allows proactive improvement of customer service based on how the resources defining a service are used. It includes, for example, improving the trouble reporting process by coupling customer transaction data with network reported events. It also includes identifying which mobile handsets may not be compatible with certain downloaded content or how many times a user fails to achieve successful launch of a data service such as e-mail or Web access. CEM involves monitoring customers' use of their purchased services and analyzing this data to note trends, preferences, and usage problems.
2. *Establishing an external view of the customer experience:* This supplies customers with a means for understanding how use of their service subscription measures up to the definition prescribed by a business contract. Though providing an external view of service usage data offers customers an effective understanding of their service experience, most SPs, as a result of the limited use of such data for CEM-related business activities to date, first develop internally oriented CEM business practices and then expand a limited subset of such capabilities to their key customers.

With the ever-present threat of subscriber churn, enhancing customer service is critical for business success. Single bills, real-time self-service care, and instant responsiveness to requests are key to maintaining high levels of customer satisfaction. The proliferation of single-use and on-demand features such as location-based services, video on demand, and competitive entries means that subscriber loads have never been higher. And an important part of the experience is how many mouse clicks or keystrokes the customer must make, or must wait for the customer service representative (CSR) to make, to enable a service request.

The transformation of SPs to a customer-centric business strategy requires a strong understanding of the customer experience. With this understanding comes a need for change to existing business processes, systems, and even work groups. The focus is first on the customer and then on the components defining the service. Network-related issues that do not have an immediate customer impact can be categorized with other lower-priority problems and addressed after all customer-affecting issues have been resolved.

Self-service is the most important function from a customer experience perspective—hence the need for accurate product bundling and pricing information. Web technology will primarily be used to deliver this service. CNM represents a modest source of incremental growth for OSS/BSS suppliers.

3.5.3.3 Billing Convergence

Billing convergence—unlike other types of convergence—is not a driver, but rather a result of changes, especially service convergence. The customer may expect to receive one bill for all services, such as voice, data, video, and Internet. Billing convergence means that consumers can receive many different services on the same bill and possibly deal with one customer support center; this type of convergence also allows operators to offer bundled pricing of products as a means of attracting and retaining multiservice customers (e.g., a combined cable TV/fixed-line bundle). The minimal requirement is to receive multiple bills with electronic staples.

Billing processes will continue to undergo a transformation. Billing opportunities exist for vendors that can deliver Internet-based billing payment solutions and unified billing solutions across multiple service offerings.

In addition to supporting pure-play prepaid, pure-play postpaid, or hybrids of both, new billing systems should support postpaid subscribers with real-time credit control, support prepaid subscribers with order management and customer care, and support any combination of prepaid, postpaid, and hybrid accounts. The ability to use different billing models (e.g., fault rate, usage-based billing) is a requirement for convergent billing systems. Any deployment mode should be modified with either option over time to meet phased business transformations or changing business needs, because all modules are based on the same convergence architecture.

At the same time, billing convergence allows consolidation of infrastructure, marketing, client base, and customer care. It allows companies to stop treating their prepaid and postpaid consumers as separate markets and allows them to run more promotions targeted to specific user segments depending on usage or user characteristics (e.g., age groups, usage patterns). Consequently, offering users the ability to combine prepaid and postpaid elements will increase usage and produce additional revenue. This will require companies to offer their subscribers a greater range of payment options to facilitate uptake. A new billing approach is needed to support these new services, one that incorporates data sharing between different applications and real-time billing. Additional billing needs are as follows:

1. *Provision of incentives and promotions to subscribers:* For example, subscribers may receive an airtime bonus after watching a certain number of advertisements on their IPTV or mobile TV. Therefore, this type of approach allows roll out of the "new model" of sponsored calls.
2. *New business models:* New models such as revenue sharing, sponsorship, and advertising should be supported as well.
3. *More services on offer:* There is a need for more combinations between land, mobile, and Internet platforms and more varied content combining voice, images, and video.
4. *More business segments and more players in the value chain:* A simulation, pricing, and billing solution is needed that is capable of integrating numerous value-added criteria and is extremely versatile, interchangeable, and naturally compatible with IMS architecture.

3.5.3.4 Management Convergence

Similar to billing convergence, management convergence is not a driver but a result of changes. Finding ways to deliver this without increasing complexity or without making network management more complicated and still benefit from CAPEX and OPEX savings is a major challenge for the operator community. Operators are taking advantage of end-to-end solutions that ensure cost-efficient operations. Converged core networks, operations support systems, business support systems, and service machinery enable savings in service development, deployment, and network operations and maintenance. With FMC, operators can utilize the converged core network, share transport across IP networks, and employ the same solutions for all access networks using common service creation based on IMS for both fixed and mobile environments. Network operators must deliver what the market wants if they are going to enjoy long-term profitability and launch new services to meet end-user demands, attract new customers, and retain their existing customer base.

3.5.3.5 Fulfillment

Service fulfillment systems process the service orders and support the installation and confirguation of these services on the network. SPs use a wide variety of methods to implement service fulfillment. Even the most sophisticated SPs use many manual steps, including paper in-boxes, to manage low-volume, complex services. On the other hand, all modern SPs implement automated fulfillment systems for higher-volume, simpler services such as residential voice, mobile voice, or DSL.

Most SPs have a legacy of department-specific approaches to service fulfillment. In some ways, fulfillment has been the last bastion of manual processes and ad hoc engineering tools. They have resisted moves to automate aspects of the business. The complexity of underlying networks, entrenched operational groups, and the comparative lack of competition contribute to this resistance to the changes that

inevitably accompany new systems. This then leads to an approach that depends on manual activation using engineering tools. Small SPs with low service volumes use similar, ad hoc approaches; others buy COTS products with or without vendor support.

SPs tend to buy service fulfillment systems that solve immediate problems with the least risk. Most service fulfillment systems meet the needs of particular regional service operations. While there is a general move to more system consolidation to meet many of the needs described above, that move takes time. The general market is showing more interest in convergent fulfillment systems that can support immediate service needs while providing a basis for supporting future services. There is little interest in converting or consolidating old fulfillment systems.

Service fulfillment spending is devoted to a large number of small ISVs. This segment is consolidating as SPs pursue larger projects across multiple technologies. In this environment, some of the larger ISVs have been acquired by much larger companies that have decided it makes sense to get into the service fulfillment business.

A service fulfillment system involves a significant amount of software and a more significant amount of services. Typically, the services value will be five times the software value when customized features, integration with other OSSs, and data migration are taken into account. Often the SP IT staff will handle many of the services, but local IT integration firms frequently play a big role as well.

In the fulfillment environment, the role of standards is changing. SPs and vendors engage in a great deal of discussion about standards but rarely implement them. Now SPs are beginning to take standards more seriously. Many SPs, particularly in Europe, are making use of TeleManagement Forum standards in their own development and specifying their requirements in RFPs.

3.5.3.5.1 Activation Outlook

Activation systems are part of service fulfillment. Service fulfillment systems support the installation and configuration of networks and the processing of orders for service on those networks. The primary focus of service fulfillment is enabling the efficient use of a communications network to provide service to a large number of customers that share use of the network.

SPs use a wide variety of methods to implement service fulfillment and its associated activation elements. SPs take four basic approaches to fulfillment-related activation:

1. Have engineers use engineering tools in performing manual activation
2. Develop their own activation system
3. Use activation that supports a range of services and fulfillment systems
4. Use activation that is tightly linked to an integrated fulfillment system

Activation systems automate explicit commands to initiate a new service. Typically, activation systems map connection requests that come through other systems (e.g., order management or NRM) to the specific commands that control the network. Activation usually relates to high-volume connection services such as residential voice, mobile voice, or DSL.

Activation systems generally interact directly with network equipment or IT systems providing services. They are often closely tied to specific types and models of hardware. Other systems that touch the network include discovery and, most commonly, fault and performance monitoring.

SPs sometimes use Network Element Management (NEM) element management systems (EMSs) to handle direct network interfaces. This seems sensible in that NEMs have a grasp of the network equipment specifics and are best able to track needed changes. Consequently, it is rare for SPs to support activation through an EMS.

An activation system supports a number of actual functions, primarily activating new service. Other functions include deactivation and configuration changes. In a configuration change, the service itself will stay in place but a certain feature (e.g., adding call waiting for a voice line) may be added or removed.

The activation market is growing rapidly as SPs push to be more responsive to customers and add new services. Until recently, most commercial activation supported residential voice and broadband services.

Increasing competition and service complexity has caused SPs to invest in activation systems for mobile and business services as well. Continued high growth and new market opportunities are the result.

With convergent service networks, SPs are offering many services over the same connection. After the business, residential, or mobile connection is in place, customers can sign up for new services without the need for additional facilities or equipment. This enables an environment of quick activation of new services, leading to a significant payoff for SPs that implement automated service activation and a competitive disadvantage for those that do not.

Over the past two years, SPs have been aggressively increasing network investments, growing CAPEX faster than revenues. This aggressive network investment contributes to strong near-term activation growth. However, this CAPEX bulge will drop off by 2010, pulling down activation growth rates. Activation will increase at a 13% compound annual growth rate (CAGR) until 2012 (*OSS Observer*).

3.5.3.5.2 Order Management

Customer demand and increasing competition are driving SPs to introduce new services much more rapidly than in the past. New services need an overarching order management approach to provide a wide range of support, as described below. The traditional approach of supporting new services with manual processes as automated support in "catch-up" mode will not scale with this new service growth.

The unique impact of IP-based services forces SPs to shift from the one-service–one-network model to a model involving one customer interface and many service platforms. IP networks enable SPs to offer new services with little change in the underlying network. SPs are taking advantage of this capability to build service revenue, retain customers, and maximize the return on network investment. The advent of VoIP accelerates the shift in service revenue to the IP network.

Unlike many other IP services, voice is a communications service that everyone uses. SPs can take advantage of VoIP in three primary ways: (1) as a way to enter new consumer or business voice markets, (2) as a means to offer value-added services such as push-to-talk, or (3) as a way to drive cost out of traditional voice service. The push to VoIP will accelerate the conversion of order management from single-service silo approaches to multiservice order management platforms.

3.5.3.6 Service Assurance

Two key factors are driving telecom service providers to investigate IP service assurance: service scaling and the addition of more IP applications to their networks. The emergence of VoIP as a mainstream business and consumer service is combining with wide-scale commercial rollouts of IPTV services to stress carrier IP networks as never before, and that burden will only continue to grow as network operators add more services to their IP portfolios.

In order to survive in the competitive triple- and quad-play marketplace, companies will need to focus on quality. Customers, whether enterprise or residential, do not really care about the network across which their services are being delivered; rather, their only concern is whether their service is reliable. And IP service providers are beginning to accept that if they cannot consistently deliver quality services, they will lose customers to other providers that promise better quality and—most importantly—deliver it.

Although IP service assurance covers the full range of IP applications, the addition of IPTV appears to be the primary catalyst for network operator decisions to deploy IP service assurance. The number of service providers that have installed IP service assurance solutions largely follows the deployment of IPTV deployments globally, with countries in the Asia/Pacific region leading the way.

Entering into the next phase of build out on IP networks, the development of a good service assurance program is imperative. Many companies are recognizing the need to focus on subscribers to ensure that whatever service assurance platform they choose provides value to the customer. Doing so ultimately requires service providers to ask several questions: What brings value to the customer in a service assurance product? Is it the ability to fix a service quickly if it goes down, or the ability to prioritize uptime for voice specifically? What will customers pay for service assurance?

3.5.3.6.1 Probe System Market Outlook

Probe-based systems are effective in ensuring network availability and service quality because of their ability to handle detailed session trace and pinpoint problems in complex multivendor networks. Spending will follow network expansion and mass-market service deployments. The most significant factor influencing growth in the market is increasing competition.

Convergence of services, the focus on the customer experience, bundled services in the residential broadband market, the introduction of IPTV, the adoption of carrier Ethernet, and the introduction of new access technologies are leading to increased investments in probe-based systems.

SPs will find value in probe systems with respect to supporting new network technologies and protocols. Support of troubleshooting tasks in operations centers and the effective dispatch of the field workforce are core capabilities. SPs will want to short-list suppliers that have evolved their solutions to support service quality metrics in a fixed mobile convergent infrastructure.

The probe system market is the most mature segment within the service assurance market, but it continues to rapidly grow in supporting new services as SPs migrate from circuit-switched to IP technologies. Forecasts show (*OSS Observer*) that the probe system market will grow approximately from $870 million in 2007 to $1.15 billion in 2012 at a compound annual growth rate (CAGR) of 6%.

3.5.3.7 Inventory Management

Seemingly the most successful type of approach to OSS transformation is an inventory-centric one. Most service providers have different inventory systems for different technologies, and accuracy levels of these systems range between 40 and 80%.

OSS/BSS processes can be improved by consolidating inventory systems into a single system across geographies, business units, and networks, usually in two applications—one network inventory system that links logical and physical inventory (inside plant) and one geographical information system (GIS) (outside plant). These improvements can occur in three phases:

1. First, data accuracy can be increased, which can remove 10 to 30% of the downstream process exceptions. This data accuracy is obtained in a three-step approach that first includes the initial data load with its challenges of data cleansing and transformation, followed by the need to include a synchronization process with the network itself to prevent divergence of inventory from real data and, finally, system integration between the inventory system and the rest of the OSS to effectively modify inventory information.
2. Second, operational and capital costs can be lowered as asset utilization improves and processes are linked, automated, and centralized.
3. Finally, new services can be introduced quickly and at a lower cost by improving cash flow and competitiveness.

Estimates indicate that the inventory market will continue very strong growth through 2012 at a CAGR of 14% (*OSS Observer*). The market is driven primarily by rapid fulfillment of new service orders in a more competitive environment. SPs are investing much more in inventory management that supports such rapid fulfillment. The top six suppliers account for 66% of the total market. The areas that previously received significant inventory investments, business services and residential broadband, are growing less than the average, while mobile is expected to grow much faster than other areas.

The value of service fulfillment systems resides in helping SPs transform new orders into active subscriber support rapidly and efficiently. This involves tracking the steps of order fulfillment, allocation of resources, and configuration of services associated with networks and applications. It also involves installing and configuring network equipment and applications. The inventory portion of service fulfillment is now becoming a significant factor in nearly all service fulfillment deployments.

A decade ago, the emphasis shifted from PSTN to DSL and cable broadband. In the past two years the market has shifted from these simple, high-volume services to a much broader base of residential

broadband, mobile, and business services. Large SPs that have long invested in inventory management are spreading its application to many services. Smaller SPs, especially those in emerging markets, are investing in commercial inventory management for the first time.

SPs should be using commercial inventory systems to improve their ability to roll out new services, better serve their customers, and control the costs related to providing services. The wide range of successful deployments in the market confirms the applicability of commercial systems and dramatically reduces the risks from a few years ago. Inventory systems play a critical role in service fulfillment and trouble management, and it is essential to reduce costs through better network asset utilization.

3.5.3.8 Promising Market

Future support system trends need to be estimated in order to position OSS/BSSs and their vendors. Service providers' financial results are currently improving as a result of debt reduction and good sales performances from broadband and mobile. However, this situation will not be easily sustainable in today's complex and extremely competitive market. A number of factors support this supposition:

1. Broadband access is driving fundamental societal change by changing end users' expectations and capabilities. Companies such as Google and Yahoo! have changed the rules for both businesses and consumers, and we now live in a search-based economy. The bandwidth available today will be insufficient to meet user demand for new bandwidth-hungry services such as IPTV, and SPs and Internet Service Providers (ISPs) must invest so that they can deliver significantly more bandwidth and/or improve their OSS/BSSs to gain better bandwidth utilization.
2. Capital expenditures are required to migrate to NGNs as companies move toward a core IP network. There is also a need to deliver broadband fiber access technologies. In the mobile world, investment is needed to expand 3G coverage and implement HSDPA (High-Speed Download Packet Access). However, adoption of 3G is well below operator expectations and plans, and the mobile Internet has not yet matured sufficiently to deliver meaningful revenues.
3. VoIP is now, or will very shortly become, a fundamental element of the broadband proposition. This will have an impact on traditional fixed-voice and wireline revenues.
4. Regulatory changes (e.g., the European Union's drive to reduce mobile roaming charges by up to 50%) will place significant pressure on mobile SPs.
5. The advent of alternative network suppliers and the entry of companies such as Google, Skype, and Yahoo! into the arena are creating a highly competitive marketplace where only the fit and agile will survive.
5. Lines are blurring between traditional SPs and media companies, and as triple- and quadruple-play propositions begin to mature and broadband speeds and capacity increase, that blurring will increase and a content and service provider ecosystem will emerge.
6. Although still in its early days, service convergence (e.g., FMC and FMS) is becoming a reality.

These factors will have a number of obvious impacts on the market. Mergers and acquisitions are expected to increase as companies scale to face competition, redistribute coverage more efficiently, and transform to meet future business needs.

Despite consolidation among larger players, the overall OSS market remains fragmented owing to new entrants.

1. Rapid changes (technological advances, increased competition, globalization, and the drive toward network convergence) continue to shift market dynamics in the communications industry, creating new business opportunities and challenges.
2. Service provider organizations are consolidating into larger, full-service operators that offer multiple bundles of services for voice, video, data, and more—in triple- and quad-play configurations. Nontraditional service providers are also entering these markets, offering their own unique service bundles to entice new customers.

3. SPs will need to optimize their current business processes and operations, and there may be a significant need for optimization (i.e., "transformation") in the use of human resources.
4. Companies will offer bundled fixed and mobile services with VoIP as part of a basic flat-rate package.

Because of the need for transformation, the outlook for the support systems market is promising. Products have matured to the extent that vendors can offer comprehensive, integrated, and modular packaged OSS/BSS solutions. OSS/BSS products are evolving into integrated and modular product suites that offer solutions for telecom and service providers worldwide. Apart from the products' performance and scalability, the primary differentiators are the ability to integrate with new and legacy systems, enhanced functionality for cross-domain networks and services, and reduced time to market. During the next few years, OSS product suites will become more robust, extending into multiple network and service management. Many suppliers have at least a decade of experience, which gives them an in-depth understanding of their customers' requirements. Equally important, however, will be anticipation of future customer needs.

OSS suppliers will continue to provide new and expanding modules in their suites through mergers and acquisitions (M&As). This will lead to a consolidation of vendors with less differentiation in their product offerings and, consequently, less choice for customers.

Some vendors pursue a go-alone sales strategy at the expense of channel relationships, which may limit their growth potential with new accounts.

In order to match the rich service offerings of new entrants, service providers have implemented or will implement multiple upgrade strategies, including modifications by internal staff members, custom development by external system integrators, and integration of third-party products. Most likely, they will not completely replace their existing OSS/BSSs. Several incumbent carriers are incorporating best-of-breed solutions with their legacy systems. This will provide great opportunities for point OSS/BSS integration and for professional services.

In addition, when service providers upgrade their operational environment, the main responsibility is usually assigned to a system integrator, while OSS/BSS vendors provide software and expertise for certain processes and functions. This will also increase the market for SIs. For reasons related to revenue and market potential, several OSS/BSS software vendors are entering the system integration market, and new entrants are expected as well. ISVs will compete with traditional SIs. The result will be that end users will have more difficult choices to make.

The world's leading telecom providers want product suites and solutions from vendors that allow them to selectively invest in the components that meet their business and strategic requirements. Almost all telecom providers want to reduce the number of OSS/BSS vendors and the number of OSS/BSS products in their environments. The top requirements are evolving standards and technology, robust functionality, integration with legacy and other systems, and a future-proof solution. Business consulting beyond the product base is also an increasing demand. Generally speaking, carrier requirements extend beyond network-class software to increased consulting and professional services aimed at customized integration and customer support. Most OSS/BSS software providers have set up multiple relationships with global and local SIs and NEPs to implement, integrate, and support their products while expanding the breadth of their solutions.

New entrants stimulate significant demand in the OSS/BSS market. Most start from scratch and invite all types of OSS/BSSs vendors with point and integrated products. Larger new SPs with custom-designed, in-house solutions are enhancing these offerings to accommodate new services and technologies; they show some similarities with incumbent providers.

Less well-known and smaller new SPs either purchase or license point products from third parties or take advantage of service bureaus. Few are interested in in-house development and maintenance. OSS/BSS vendors can sell to these SPs directly or to service bureaus that may share their products among multiple SPs.

Some SPs may want to outsource their OSS/BSS services partly or entirely, and they must evaluate the benefits of doing so. Outsourcing will eliminate the need for carriers to invest scarce research and

development dollars in OSS/BSSs and allow them to focus on their core business. On the other side, this generates new opportunities for vendors in the area of outsourcing and professional services.

The multiplayer telecom service market and carrier interconnections provide excellent opportunities for OSS/BSS vendors. Resellers of local exchange services must provide electronic links to incumbent or worldwide carriers for ordering, service activation, troubleshooting, and billing. Today's OSS/BSSs do not have all of these features. There is a significant opportunity for incremental OSS/BSS sales by emerged as well as new vendors. Vendors in certain specialized service areas will play an important role during the next 10 years. Best-of-breed solutions are expected to offer provider portability, location portability, and service portability.

The growth of the market is most often expressed in terms of CAGR value. Unlike the average annual growth rate (AAGR) method, CAGR takes into account changes from year to year, not only in revenues but also in revenue growth rate. CAGR is the rate at which the amount in the final year represents the future value of the amount in the first year after a specific interval.

Estimates indicate a 10 to 15% CAGR over the next few years, until approximately 2011 or 2012. Some *OSS Observer* estimates differ: 10% CAGR for the global telecom software market but 25% for service delivery platforms, 16% for middleware, and 27% for service assurance.

3.5.3.8.1 Professional Services Revenue

The line between what constitutes OSS sales revenue and what constitutes professional services revenue continues to be blurred. In pricing their solutions, some vendors include consulting services as part of the final OSS delivery price. Other vendors price consulting services separately. Smaller OSS vendors, especially those who have taken their companies public, often include as many of their customization services in their product pricing as possible, since financial markets will reward them with a higher price-to-earnings ratio. Thus, any projections of professional services revenues associated with OSSs are highly definition dependent. Business drivers for procuring professional services include:

1. Provision of specialized services such as network planning and application integration
2. Gathering of operational/process information for the purpose of configuring the OSS
3. Project management
4. Support of user acceptance testing in the service provider's application lab
5. Support of application deployment (from the data center perspective) and operations sites
6. Training of operations staff on using the application
7. Creation of software requirement specifications for the customizations required
8. Development of software customizations

3.5.3.8.2 Outsourcing

Ongoing structural changes in the telecom industry place new requirements on support systems. Some back-office functions may be outsourced to allow a focus on customer management. The reasons for outsourcing are as follows.

1. *Lack of internal resources:* Most service providers are operating with lean workforces as a result of downsizing over recent years. It makes sense to outsource the specialized tasks associated with integrating OSSs into a service provider's operational environment.
2. *Managing costs:* While the supply of available and competent consultants increases as traditional telecom equipment and software manufacturers reconstruct their businesses, service providers can leverage competitive rates.
3. *Managing complexity:* Service providers need help in planning and executing their OSS integration projects, whether they are motivated by a merger, the need to reduce operational costs through flow-through automation of operational processes, or the need to put in place new systems or enhancements to existing systems in support of new services.

3.5.3.9 Summary

The telecommunications industry is undergoing fundamental change as service providers move toward the implementation of NGNs. Traditionally, service providers have managed their networks in a technology-focused way, resulting in "stovepipes" that specialize in managing different infrastructure segments and technologies. Mergers and acquisitions have increased the number of such silos in any given organization, and the introduction of new services (frequently on separate, dedicated platforms) only increases the problem.

SPs need to consider the impact of convergence, the new IP networks (NGNs), and the associated new IP multimedia services and make the necessary changes to their OSS/BSSs. OSS/BSS and process transformation are critical in lowering the costs of delivering existing and new services and improving the flexibility of SP businesses. OSS transformation delivers cost savings and operational effectiveness improvements in running the existing business and delivers the necessary systems, infrastructure, and processes to enable a smooth, cost-effective transition to an NGN.

Competition is driving telco service providers to change their network-centric approach to a customer-centric approach.

CNM incorporates a class of OSS/BSSs that enables end users to securely view, troubleshoot, and reconfigure their "own" network and generate reports on their subscribed telecommunication services. It provides various network information specific to the customer's service. The customer can obtain information regarding traffic patterns, operational status, and trouble spots by using a standard Web browser application. A limited set of control capabilities should also be available through CNM. It is expected that CNM will eventually migrate to the customer contact point for subscription, activation, monitoring, and control by the customer.

CEM, a term borrowed from other industries and partially from the customer-facing side of the communications business, is the practice of collecting data on customer usage from all practical sources (network devices, content servers, and management databases) to be used by both business and technical departments in accomplishing two specific goals: (1) establishing an internal view of the customer service experience, which allows proactive improvement of customer service based on how the resources defining a service are used, and (2) establishing an external view of the customer experience, which provides customers with a means of understanding how use of their service subscription measures up to the definition prescribed by a business contract.

Billing and management convergence—unlike other forms of convergence—is not a driver, but rather a result of change, especially service convergence. Billing convergence means that consumers can receive many different services on the same bill and possibly deal with one customer support center. This convergence also allows operators to offer bundled pricing of products as a means of attracting and retaining multiservice customers (e.g., a combined cable TV/fixed-line bundle). The minimal requirement is to receive multiple bills with electronic staples.

Billing will continue to undergo a transformation. Billing opportunities exist for vendors that can deliver Internet-based billing payment solutions and unified billing solutions across multiple service offerings.

Traditionally, each technology has had its own management system. With convergent management and OSS/BSS, carriers are taking advantage of end-to-end solutions that ensure cost-efficient operations. Converged core networks, operations support systems, business support systems, and service machinery enable savings in service development, deployment and network operations and maintenance.

Most SPs have a legacy of department-specific approaches to service fulfillment. In some ways, fulfillment has been the last bastion of manual processes and ad hoc engineering tools. The complexity of underlying networks, entrenched operational groups, and the comparative lack of competition contribute to resistance to the changes that inevitably accompany new systems. The market is showing interest in convergent fulfillment systems that can support immediate service needs with respect to different technologies while providing a basis for supporting future services.

Activation systems are part of service fulfillment. They automate explicit commands to initiate a new service. Typically, activation systems map connection requests that come through other systems (e.g.,

order management or NRM) to the specific commands that control the network. Activation usually relates to high-volume connection services such as residential voice, mobile voice, or DSL. Activation systems generally interact directly with network equipment or IT systems providing services. They are often closely tied to specific types and models of hardware. Other systems that touch the network include discovery and, most commonly, fault and performance monitoring.

Order management is part of the fulfillment process. The impact of IP-based services forces SPs to shift from the one-service–one-network model to a model involving one customer interface and many service platforms. IP networks enable SPs to offer new services with little change in the underlying network. SPs are taking advantage of this capability to build service revenue, retain customers, and maximize the return on network investment. The advent of VoIP accelerates the shift in service revenue to the IP network.

Entering into the next phase of build out on IP networks, the development of a good service assurance program is imperative. Many carriers are recognizing the need to focus on subscribers to ensure that whatever service assurance platform they choose provides value to the customer. With a converged assurance system, service providers are able to manage problems more rapidly, automate different tasks, and use sophisticated technologies (e.g., root cause analysis, impact analysis)—and thus shorten fault clearing times. These improvements result in better user satisfaction and less churn.

The value of service fulfillment systems resides in helping SPs transform new orders into active subscriber support rapidly and efficiently. This involves tracking the steps of order fulfillment, allocation of resources, and configuration of service associated with networks and applications. It also involves installing and configuring network equipment and applications. The inventory portion of service fulfillment is now becoming a significant factor in nearly all service fulfillment deployments. Inventory systems are essential in reducing costs through better network asset utilization.

Because of the need for transformation, the outlook for the support systems market is promising. Products have matured to the extent that vendors can offer comprehensive, integrated, and modular packaged OSS/BSS solutions. The multiplayer telecom service market and carrier interconnections provide excellent opportunities for OSS/BSS vendors. Estimates indicate a 10 to 15% CAGR over the next few years, until approximately 2011 or 2012. Some *OSS Observer* estimates differ: 10% CAGR for the global telecom software market but 25% for service delivery platforms, 16% for middleware, and 27% for service assurance.

3.5.4 Summary and Trends

The number and complexity of services have exploded in just a few years, and competition from aggressive entrants has shaken the very foundations of the incumbents, forcing them to act in order to survive. In turn, these organizations have restructured themselves as customer-centered, business-driven enterprises in which focused business roles are assumed by specified entities that increasingly operate as autonomous companies but leverage incumbents' latent resources. The differentiator among competitors has also shifted away from the customary network view, a view based on similar or even identical transport technologies from a small number of large-cap multinational equipment vendors toward a way in which service providers manage their customers, services, and resources. As a result, OSS/BSS systems have gradually grown in importance. There has been a clear revision of attitudes toward OSS/BSS as merely a support system; now OSS/BSS is seen as the competitive advantage for service providers. This is due to the realization that OSS/BSS is part of the solution, not part of the problem.

Despite growing awareness of the importance of the OSS/BSS environment, coupled with the fact that OSS/BSS is not a technical issue but a financial one, many service providers today are struggling with their old legacy architecture. Although important (as day-to-day operations completely depend on it), the legacy architecture should be considered a bottleneck when it comes to the rapid deployment and introduction of new services or the ability to respond to competitive offers in the marketplace.

Once deployed, advanced OSS/BSSs offer the following strategic benefits:

1. *Improved operating efficiencies in data, inventory, and network management:* It is expected that management of various objects (e.g., equipment, applications, databases) will be more integrated, requiring fewer human resources.
2. *Reduced support and maintenance costs associated with legacy systems:* Support and maintenance expenses are decreasing as a result of more automation and interconnection.
3. *Shorter product development (time to market) cycles:* Products and services can be created, tested, and deployed faster because of the advanced technologies used in OSS/BSSs.
4. *Speedier deployment of new services and pricing schemes:* Processes are connected to each other. Rapid service provisioning in combination with pricing guarantees rapid deployment.
5. *Bundled services:* The possibility is open for bundled services to be offered to customers.
6. *Flexibility to modifying pricing and marketing schemes:* As a result of interconnected processes, changes can be deployed very quickly. Even modeling and simulating resource utilization scenarios are easy to implement.
7. *Better customer satisfaction:* Customer satisfaction can be improved through shorter service deployments and faster fault clearing.
8. *Strategic marketing to target and acquire profitable business customers:* The availability of rich information on customers and their traffic generation patterns allows marketing strategies to be customized.
9. *Superior customer management to establish customer loyalty:* The significant improvements in customer care will help to avoid customer churn and to sell value-added communication services to loyal customers.

Acronyms

AAGR	Average Annual Growth Rate
ARPU	Average Revenue Per User
BSS	Business Support System
CAGR	Compound Annual Growth Rate
CDDI	Copper Distributed Data Interface
CEM	Customer Experience Management
CLEC	Competitive Local Exchange Carrier
CNM	Customer Network Management
COTS	Commercially off the Shelf
CPE	Customer Premise Equipment
DSLAM	Digital Subscriber Line Access Multiplexer
DSL	Digital Subscriber Line
EMS	Element Management System
F/R	Frame Relay
FDDI	Fiber Distributed Data
FMC	Fixed Mobile Convergence
FMS	Fixed Mobile Substitution
GIS	Geographical Information System
GSM	Group Speciale Mobile
HSDPA	High-Speed Download Packet Access
ICT	Inter Connection Technology
IMS	IP Multimedia Subsystem
IP/MPLS	Internet Protocol/Multi-Protocol Label Switching
ISV	Independent Software Vendor
IEX	Inter Exchange Carrier
LEC	Local Exchange Carrier

MKDSS	Marketing Decision Support System
NEM	Network Element Management
NEP	Network Equipment Provider
NGOSS	Next-Generation OSS
NRM	Network Resource Management
OSS	Operations Support System
PLNM	Private Line Network Management
POTS	Plain Old Telephone System
PSTN	Public Switched Telecommunication Network
PTT	Public Telephone & Telegraph
ROI	Return on Investment
SDH	Synchronous Digital Hierarchy
SI	System Integrator
SMS	Short Message Service
SONET	Synchronous Optical Hetwork Technologies
SP	Service Provider
TDM	Time Division Multiplex
SCDMA	Wideband Code Division Multiple Access
WDM	Wave Division Multiplexing
WLAN	Wireless Local Area Network
wiMax	Worldwide Interoperability for Microwave Access
xDSL	Digital Subscriber Line

References

ACTEXP: Goldman, L. *Activation Outlook Expert*. OSS Observer, February 2008.

CAI06: Mohr-McClune, E. Senior Analyst Enterprise Mobility Europe. *FMC: Did We Miss It?* Current Analysis Inc, May 2006.

CCC21: Messerschmitt, D. Department of Electrical Engineering and Computer Sciences University of California at Berkeley: The Prospects for Computing-Communications Convergence. Conference on "VISION 21: Perspectives for the Information and Communication Technology", Munich, Germany, Nov. 25, 1999.

CEO60: Goldman, L. *Activation Outlook*. OSS Observer CEO Digest, Oct 2007.

DI0307: Kurth, M. *OSS Market Overview and Strategic Scorecard for Vendors, 2006*. Gartner Datequest Insight, March 22, 2007.

DIOSS06: Kurth, M. *Dataquest Insight: OSS Market Overview and Strategic Scorecard for Vendors, 2006.* Gartner, 22 March 2007.

EMA071: Cotrupe, J. *OSS/BSS Shopping List for 2007 Part 1: Network Planning—An EMA Advisory Note*. Enterprise Management Associates, February 2007.

Emerging Best Practices for Carrier IP Transformation. Alcatel-Lucent White Paper, 2007.

ERICS07: Ruiz, L. *IMS and Convergence*. Ericsson Presentation, 2007.

F&S07: *Business Transformation: It's About the Customer, Not the Network*. IBM White Paper Stratecast (Frost & Sullivan) White Paper, 2007.

GB929: *Telecom Applications Map—The BSS/OSS Systems Landscape.* Telecommunication Managment Forum GB929 Guidebook, November 2006.

IBM07: *Stuart McIntosh Strategy Partner IBM Communications Practice: The Great Convergence of Content, Telecommunications and Business.* McIntosh, FutureNet, May 2007.

IBMTR: Bray, C. *Transforming Operational Support Systems—The reasons for OSS transformation and the approach for successful implementations.* A White Paper of IBM Communications Sector, 2006.

IDC0706: Winther, M. *Fixed-Mobile Convergence: Lowering Costs and Complexity of Business Communications*. An IDC White Paper, June 2007.

INVO: Goldman, L. *Inventory outlook*. OSS Observer CEO Digest, January 2008.

IROS07: The Insight Research Corporation, *Operations Support Systems 2007–2012,* The Insight Research Corporation, December 2007.

ISY07: Sreedharan, R., Jagirdar, R., and Ananthanarayanan, R. *Winning in the Age of Convergence: Product Framework for SPs*. Infosys Technologies Limited, 2007.

M.3050.2: ITU-T Series M: *TMN and Network Maintenance: International Transmissions Systems, Telephone Circuits, Telegraphy, Facsimile and Leased Circuits. Telecommunications management network. Enhanced Telecom Operations Map (eTOM)—Process decompositions and descriptions.* ITU-T Recommandation M.3050.2, June 2004.

M.3050.4: ITU-T Series M: *TMN and Network Maintenance: International Transmissions Systems, Telephone Circuits, Telegraphy, Facsimile and Leased Circuits. Telecommunications management network. Enhanced Telecom Operations Map (eTOM)—B2B integration: Using B2B inter-enterprise integration with the eTOM*. ITU-T Recommandation M.3050.4, June 2004.

NSFMC: *Fixed-Mobile Convergence*. White Paper, Nokia-Siemens Networks, 2007.

NTG: *Key Trends in the Carrier Service Provider Industry and Their Impact on OSS Systems*. NTS Operations Support System Executive Briefing, NTG Clarity Networks Inc, Oct. 2006.

OFC06: *The International Communications Market 2006*. Ofcom Research Publication, November 2006.

PIP0408: Greene, W. and Hayes, T. "Service Delivery Frameworks: The Service Provider's Mashup." *Pipeline* 8(4), 2006.

PROBO: Kelly, P. *Probe Systems Outlook*. OSS Observer CEO Digest, January 2008.

SQNGN: Kelly, P. et al. *Manage service quality: Create, execute and manage for quality of next-generation communications services.* IBM GRM Webcast, Sept. 27, 2007.

Terplan: Terplan, K. "Support Systems for Telecommunication Providers." (The CRC Handbook of Modern Telecommunications), CRC Press, 2001.

Tieto: Tenevall, T. OSS/BSS—The challenges ahead. TietoEnator White Paper, Oct 2002.

WikiSS: Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Operations_Support_Systems.

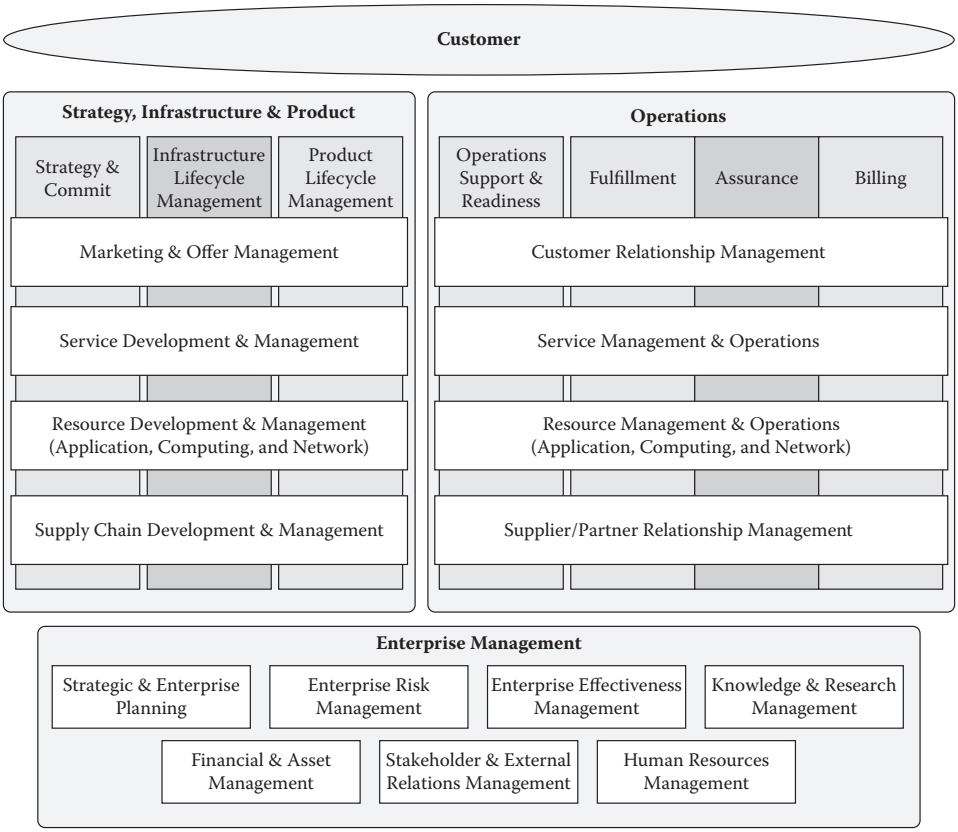
3.6 Support Processes for Service Providers

Árpád Bakay and József Wiener

Telecommunications service providers can efficiently and effectively conduct business by managing its essential business and support processes. These processes can be aggregated to deliver the major requirements common to any service-oriented business.

There are many different reference models, technologies, systems and tools to cover the various functions and processes of a telecommunications service provider. In terms of the reference models, the most well-known models include the International Organization for Standardization (ISO) FCAPS (Fault, Configuration, Accounting, Performance and Security). The International Telecommunications Union (ITU-T) proposed the model called the Telecom Management Network (TMN). The newer one proposed by the TeleManagement Forum and intended for use in telecommunications is called eTOM (enhanced Telecommunications Operations Map).

The TeleManagement Forum's Business Process Framework (also known as eTOM, enhanced Telecom Operations Map) is a business process model or framework for use by telecommunications service providers and their suppliers and partners within the telecommunications industry. It describes all the enterprise processes required by a service provider and analyzes them to different levels of detail according to their significance and priority for the business. eTOM was developed by the TeleManagement Forum, and adopted by ITU-T in its Recommendation M.3050.



FIGu RE 3.6.1 Decomposition of the Level 0 process into Level 1 process groupings.

Essentially, eTOM provides a common view of the major operational functions of a communications service provider, using terms that are understood by business operations personnel while maintaining a communications channel to the systems developers. It takes the service provider perspective, looking internally at business processes, specifically those processes supporting customer care, service development, and operations management, with links to traditional network operations functions including fault, configuration, and performance management.

eTOM is a reference framework that categorizes the business processes that a service provider will use. It defines a complete enterprise management framework and addresses the impact of e-business environments and business drivers. eTOM can be considered a blueprint for standardizing business processes as well as operations support systems (OSSs) and business support systems (BSSs). Another area of improvement is process-modeling methodology, which provides the linkage necessary for Next-Generation Operations Support Software (NGOSS).

The eTOM is also a model that provides the enterprise processes required for a service provider, but it is not a service provider business model. It does not address the strategic issues or questions of who a service provider's target customers should be, what market segments should the service provider serve, what are a service provider's vision, mission, etc. A business process framework is one part of the strategic business model and plan for a service provider.

For suppliers, the eTOM framework outlines potential boundaries of software components, and the required functions, inputs, and outputs that must be supported by products.

The eTOM model has multiple levels of abstraction, covering three main areas. The highest, conceptual view of the eTOM framework provides the following major process areas (Level 0 processes, Figure 3.6.1):

Operations: This area includes customer relationship management, service management, resource management, and supplier/partner relationship management.

Strategy, Infrastructure & Product: This area includes processes that support the creation of strategies for marketing, development of new services, resource development and management, and supply chain development and management.

Enterprise Management: This area includes those basic business processes that are required to run and manage an enterprise. Enterprise management includes process areas such as human resource management, financial asset management, and disaster recovery management, etc.

Figure 3.6.1 shows how the major (Level 0) process areas are decomposed into their Level 1 process groupings. This view thus provides the Level 1 decomposition of the Level 0 processes and gives an overall view of the eTOM framework. However, in practice it is the next—the Level 2 decomposition of the Level 1 processes—at which users tend to work, as this degree of detail is needed in analyzing their businesses.

3.6.1 High-Level Breakdown of Support Processes

As the eTOM framework is better regarded as a business process framework, rather than a business process model, its aim is to categorize the process elements and business activities so that these can then be combined in many different ways, to implement end-to-end business processes (e.g., fulfillment, assurance, billing, or FAB) that deliver value for the customer and the service provider.

3.6.2 Enterprise Management

Enterprise management covers corporate or business support management.

This grouping includes those processes that have a knowledge of enterprise-level actions and needs, or have application within the enterprise as a whole. This grouping encompasses all business management processes necessary to support the rest of the enterprise, including processes for financial management, legal management, regulatory management, process cost and quality management, etc.

Enterprise management processes are, in part, responsible for setting corporate strategies and directions and providing guidelines and targets for the rest of the business. This includes strategy development and planning for areas such as information systems that are integral to the direction and development of the business.

Enterprise management also contains those processes that are responsible for providing support services required throughout the rest of the enterprise. These services are centralized within enterprise management to avoid unnecessary duplication, and to provide a clearer focus for the relevant process responsibilities.

Note that functionality associated with a process grouping that is not required throughout the enterprise will not normally be located within enterprise management (for example, human resource management issues specific to call centers are likely to be associated with the processes in operations directly involved in this area).

The specific process groupings included within enterprise management are:

- Strategic and enterprise planning
- Brand management, market research, and advertising
- Financial and asset management
- Human resources management
- Stakeholder and external relations management
- Research and development, technology acquisition
- Enterprise; quality management, process and IT planning and architecture
- Disaster recovery, security, and fraud management
- Enterprise security management

Since enterprise management processes are aimed at general support within the enterprise, they may interface as needed with almost every other process in the enterprise, be they operational, strategy, infrastructure, or product processes.

3.6.2.1 End-to-End Vertical Process Groupings

Figure 3.6.1 shows the seven end-to-end vertical process groupings that are the end-to-end processes, which are required to support customers and to manage the business. Among these end-to-end vertical process groupings, the focal point of the eTOM framework is on the core customer operations processes of fulfillment, assurance, and billing (often referred to as FAB).

3.6.2.2 Strategy, Infrastructure, and Product Process (SIP)

The Strategy, Infrastructure, and Product Process (SIP) area includes processes that develop strategy, commit to the firm, build infrastructure, develop and manage products, and that develop and manage the supply chain. In the eTOM, infrastructure refers to more than just the IT and resource infrastructure that supports products and services. It includes the infrastructure required to support functional processes, e.g., customer relationship management, or CRM. These processes direct and enable operations. In the SIP process area, the Strategy and Commit vertical, as well as the two Lifecycle Management verticals (Infrastructure and Project Lifecycle Management), are differentiated.

3.6.2.3 Strategy and Commit Processes

The Strategy and Commit process grouping is responsible for the generation of strategies in support of the Infrastructure and Product Lifecycle processes. It is also responsible for establishing business commitment within the enterprise to support these strategies. This embraces all levels of operation from market, customer, and products, through the services and the resources on which these depend, to the involvement of suppliers and partners in meeting these needs. Strategy and Commit processes are heavily focused on analysis and commitment management. These processes provide the focus within the enterprise for generating specific business strategy and gaining buy-in within the business to implement this strategy. Strategy and Commit processes also track the success and effectiveness of the strategies and make adjustments as required.

3.6.2.4 Infrastructure Lifecycle Management Processes

The Infrastructure Lifecycle Management process grouping is responsible for the definition, planning, and implementation of all necessary infrastructures (application, computing, and network), as well as all other support infrastructures and business capabilities (operations centers, architectures, etc.). These processes identify new requirements, new capabilities, and design and develop new or enhanced infrastructure to support products.

3.6.2.5 Product Lifecycle Management Processes

The Product Lifecycle Management process grouping is responsible for the definition, planning, design, and implementation of all products in the enterprise's portfolio. The Product Lifecycle Management processes manage products to required profit and loss margins, customer satisfaction and quality commitments, as well as delivering new products to the market. These lifecycle processes understand the market across all key functional areas, the business environment, customer requirements, and competitive offerings in order to design and manage products that succeed in their specific markets. Product Management processes and the Product Development processes are two distinct process types.

3.6.2.6 Fulfillment, Assurance, Billing (FAB) Processes—Operations Area

The focal point of the eTOM framework is on the core customer operations processes, representing three major customer service processes of Order Fulfillment, Service Assurance, and Billing are located in the Operations Processes Area. These FAB process groupings are sometimes referred to as Customer Operations Processes.

FAB is an attempt to show the process groups that are usually needed to support the three major customer service functions. In truth, both FAB and FCAPS provide valid ways of understanding what goes on in carrier environments. Yet no conceptual model had yet tied together day-to-day operational processes with fast-emerging service provider requirements around their business processes and objectives. The Next-Generation OSS (NGOSS) of TMF (the TeleManagement Forum), and its Enhanced Telecom Operations Map (eTOM), is a sophisticated conceptual framework in terms of mapping the totality of a service provider's business and operations processes and understanding the synergies and linkages between the two.

3.6.2.7 Fulfillment (Order/Service Fulfillment)

The fulfillment: this vertical process grouping is responsible for providing customers with their requested products in a timely and correct manner. It translates the customer's business or personal need into a solution, which can be delivered using the specific products in the enterprise's portfolio. This process informs the customers of the status of their purchase order, ensures completion on time, as well as a delighted customer.

Service fulfillment refers to the complete process from receiving the customer order (e.g., for broadband or business VPN [virtual private network], Internet, VoIP, etc.) to activating and testing the service in the network. The fulfillment process is initiated predominantly by the customer. Thus, fulfillment systems are those systems that are used to receive and handle orders, provision service, and create/update information for customer billing.

The value of service fulfillment systems is in helping Communications Service Providers (CSPs) turn new orders into active subscriber support rapidly and efficiently. This involves tracking the steps of order fulfillment, allocating resources, and configuring the service serving networks and applications. It also involves installing and configuring the network equipment and applications. The inventory portion of service fulfillment is becoming a significant factor in nearly all service fulfillment deployments.

Service Fulfillment processes start with order entry (located on the CRM level), either through Sales Force Automation (SFA) tools, or through Web-based customer order entry portals. Either way, orders must be entered, validated, and inserted into order management workflows.

The Design and Assign function is the part of the provisioning process when circuits and services are designed and assigned to available physical/logical inventory. These systems typically interact with network inventory systems.

The primary engineering system that most carriers employ is a network inventory system. These systems support inventory entry and tracking of both physical and logical inventory. An ideal inventory system also integrates tightly with Design and Assign systems, and produces capacity threshold events and detailed inventory tracking reports.

Other engineering-related systems include traffic analysis and planning systems. These systems provide Network Engineering with software tools for performing sophisticated trend analysis and forecasting of capacity requirements.

Workforce Management Systems (WMS) are used in the service fulfillment process for scheduling and routing field technicians to perform physical provisioning activities, such as cross-connecting subnetworks at peering points, installing Customer Premises Equipment (CPE), deploying probes for surveillance, etc. The primary data these systems work with are work orders, skills of technicians, truck inventory, and schedules. It is important to note that WMS systems have a role in service assurance, as well; technicians are also dispatched to repair network problems that cannot be solved from the NOC (Network Operating Center). Gateway systems are used for two primary purposes: to order capacity on a trading partner's network, and to provision nonnetwork resources from third-party providers.

3.6.2.8 Assurance (Service Assurance)

Once a service is operational, the telecommunications service provider's primary responsibilities are to ensure that the service stays operational, to provide data to the customer on the performance of their service, and to provide customer support in the use of the service.

The Assurance processes are responsible for the execution of proactive and reactive maintenance activities to ensure that services provided to customers are continuously available and to Service-Level Agreement (SLA) or Quality of Service (QoS) performance levels. It performs continuous resource status and performance monitoring to proactively detect possible failures. It collects performance data and analyzes it to identify potential problems and resolve them without impact to the customer. This process manages the SLAs and reports service performance to the customer. It receives trouble reports from the customer, informs the customer of the trouble status, and ensures restoration and repair, as well as a delighted customer.

The key system in this area of provider business processes is CRM. This system is the source of record for all customer and service data.

There is often a separate trouble ticketing system in use for both customer-originated and NOC-originated trouble tickets. Both network and customer trouble ticketing plays a key role in service assurance processes. Automation of these processes invariably involves the automated creation, tracking, and closure of trouble tickets, along with appropriate customer and field personnel notifications throughout the ticketing workflow.

3.6.2.9 Billing and Revenue Assurance

This process grouping is responsible for the production of accurate bills in time, or providing pre-bill use information and billing to customers, processing their payments, and performing payment collections. In addition, it handles customer inquiries about bills, provides billing inquiry status, and is responsible for resolving billing problems to the customer's satisfaction in a timely manner. This process grouping also supports prepayment for services.

Billing systems can be simple enough to handle flat-rate monthly billing, or complex enough to rate and present bills for usage-based or value-based services. They are usually the last workflow destination in customer-facing processes. At the point an order is fully provisioned, tested, and sent to billing, it is considered a revenue-generating service.

In order to successfully automate various business processes, it is critical that engineering and traffic analysis/planning systems are tightly coupled with logical and physical inventory events and queries across Element Management System (EMS) interface boundaries.

The billing process offers the following services:

- Rapid creation of new service plans and rates; flexible rating and discount options ("pricing")
- Real-time event processing
- Convergent billing
- Prepaid and postpaid billing
- Customer self-management of services via the Web
- Analysis of customers, rates, plans, usage, etc.

3.6.2.10 Operations Support and Readiness (OSR-Operations Area)

Operations Support and Readiness (OSR) in the Operations Process Group is differentiated from FAB real-time processes to highlight the focus on enabling support and automation in FAB, i.e., online and immediate support of customers. It ensures that the operational environment is in place to let the FAB processes do their job. This process is responsible for support to the FAB processes, and for ensuring operational readiness in the fulfillment, assurance, and billing areas. In general, the processes are concerned with activities that are less "real-time" than those in FAB, and which are typically concerned less with individual customers and services and more with groups of these. They reflect a need in some organizations to divide their processes between the immediate customer-facing and real-time operations of FAB and other Operations processes, which act as a "second line" in carrying out the operational tasks.

3.6.3 Customer Relationship Management (CRM)

The key system in the customer service area of provider business processes is CRM. This system is the source of record for all customer and service data and often plays a dominant role in service delivery processes, such as order management.

Customer Relationship Management processes consider the fundamental knowledge of customers' needs and include all functionalities necessary for the acquisition, enhancement, and retention of a relationship with a customer. It is about customer service and support, whether storefront, telephone, Web, or field service. It is also about retention management, cross-selling, up-selling, and direct marketing for the purpose of selling to customers. CRM also includes the collection of customer information and its application to personalize, customize, and integrate delivery of service to a customer, as well as to identify opportunities for increasing the value of the customer to the enterprise. CRM applies to both conventional retail customer interactions, as well as to wholesale interactions, such as when an enterprise is selling to another enterprise that is acting as the retailer.

This system often includes workflow data for customer support processes, such as trouble ticketing and change requests. CRM systems are most commonly used by Customer Support Representatives (CSR) in call centers, but may also have Web-based front-ends for customer self-service.

3.6.3.1 Customer Contact Management, Retention, and Loyalty

This process is part of the Operations Support and Readiness (vertical) grouping and the CRM (horizontal) process group as well.

Customer contact management, retention, and loyalty processes deal with all functions related to the retention of acquired customers, and the use of loyalty schemes in the potential acquisition of customers, functions that are generally sold as part of a Customer Relationship Management (CRM) suite of applications.

They establish a complete understanding of the needs of the customer, a determination of the value of the customer to the enterprise, determination of opportunities and risks for specific customers, etc. These processes collect and analyze data from all enterprise and customer contacts. These functions allow an operator to create, update, and view the customer's information, and record and view all customer interactions, so that whoever is speaking to a customer can see the history of issues that have concerned that customer. More sophisticated systems allow capabilities to highlight customers as being at risk of switching to an alternative carrier (churn indicator) and provide comparisons with other operator's service packages to allow customer care agents to try to persuade a customer that their current operator can provide the best value for the money. These indicators can be provided via integration with business intelligence platforms.

In general, these processes provide the following functions:

- Verify customer relationship
- Interaction management
- Build customer insight
- Analyze and manage customer risk
- Personalize customer profile for retention and loyalty
- Validate customer satisfaction

3.6.3.2 Customer Information Management

This process is part of the Operations Support and Readiness (vertical) grouping and the CRM (horizontal) process group as well.

Customer Information Management is a core piece of any CRM solution. This is used by all users across all channels, to allow creation, update, lookup/search, and view of customer information. Customer information includes, but is not limited to, the customer identification, profile, accounts, service profiles, etc.

This process includes the following information:

- Customer name, address, contact information
- Customer organizational hierarchy (relevant for business and corporate customers)
- The customer's existing products/services
- The customer's billing accounts (the CRM application should be capable of linking multiple billing accounts that may be managed by different billing systems)
- The customer's current and past orders (linked to order management)
- The customer's current and past trouble tickets (linked to service/account problem resolution)
- The customer's interactions with the service provider

3.6.3.3 Customer Interface Management

This process may be distinct or may be performed as part of the individual Customer Relationship Management Processes on an individual service or cross-service basis.

These processes directly interact with customers and translate customer requests and inquiries into appropriate events, such as the creation of an order or trouble ticket or the adjustment of a bill. The process logs customer contacts, directs inquiries to the appropriate party, and tracks the status to completion. In those cases where customers are given direct access to service management systems, this process assures consistency of images across systems, and security to prevent a customer from harming their network or those of other customers. The aim is to provide meaningful and timely customer contact experiences as frequently as the customer requires.

Principal functions are:

- *Manage Contact:* Manage all contacts/requests between potential or existing customers and the enterprise. It deals with the identification of the contact, its development, enhancement, and update.
- *Manage Request (Including Self-Service):* Manage all requests (inbound and outbound) made by potential and existing customers. It receives the request and either enables its originator to automatically fulfill it or identifies and activates the opportune process to accomplish the request.
- *Analysis:* Perform all necessary analysis on closed (completed or unfulfilled) requests and on customer contacts. It generates related reports to be utilized for process improvement activities, proactive problems prevention, up-sell opportunities definition, etc.

3.6.3.4 Selling/Sales Process

This process is part of the Fulfillment end-to-end (vertical) grouping and the CRM (horizontal) process group as well.

This process encompasses learning about the needs of each customer, and educating the customer about the communications services that are available to meet those needs. It includes working to create a match between the customer's expectations and the service provider's ability to deliver. Depending on the service provider process it can be pure selling or can include various levels of support. The sales process may include preorder work and interfaces. The aim is to sell the correct service to suit the customer's need and to set appropriate expectations with the customer. SLA negotiation, Request for Proposal (RFP) management, and negotiation are led from this process.

Principal functions include:

- Learn about customer needs
- Educate customer on services
- Match expectations with respect to offerings and products
- Arrange for appropriate options
- Forecast service demand
- Manage SLA and RFP negotiations

3.6.3.5 Service Order Management, Order Handling

Order management is part of an overall service fulfillment end-to-end process, and is part of the Customer Relationship Process Group. Service providers use order management as they shift from accepting the order to actually processing the order to provide service. Order management orchestrates the activities to complete the customer order.

Service Order Management Processes manage the end-to-end lifecycle of a customer request for services. This includes capturing the order, configuring the products and services within the order, decomposing the order for provisioning activities, and orchestrating the activation and fulfillment and billing notification processes. Order management typically serves all the customer touch points and channels, including call center, retail, self-service, dealers, affiliates, etc. The order may be initiated by any channel and visible to the other channels if needed. It also may include creating the customer's billing profile, tracking order status, assigning individual orders to specific employees, and allowing task management within a predefined ordering team.

Telecom service providers and operators are currently investing in next-generation IP networks in order to provide their customers a richer set of services (VoIP, IPTV, etc.) and achieve profitable growth based on this. A central business challenge for service providers is delivering and managing the services for their customers efficiently and in a timely manner. To achieve this, the operators need integrated IT systems to manage the process and information flows when activating subscribers and services in the network (provisioning) and storing the required business and technical information.

Coordinating this activity is particularly complex for (currently fixed-line) broadband operators. They require a central system (Order Management) to manage the process in an efficient way. With the introduction of all-IP networks, the service providers will be offering convergent broadband services through both fixed and mobile channels. This is a major growth opportunity, but requires investment in a convergent OSS (Operations Support System), where fulfillment is a major component.

The ordering process includes all the functions of accepting a customer's order for service, tracking the progress of the order, and notifying the customer when the order is complete. Orders can include new, change, and disconnect orders for all or part of a customer's service, as well as cancellations and modifications to orders. Preorder activity that can be tracked is included in this process. The development of an order plan may be necessary when service installation is to be phased in, and the need for preliminary feasibility requests and/or pricing estimates may be part of this process when certain services are ordered. The aim is to order the service the customer requested, support changes when necessary, and to keep the customer informed with meaningful progress of the order, including its successful completion.

Principal functions of this process include:

- Accept orders
- Determine preorder feasibility
- Prepare price estimates and SLA terms
- Develop order plan
- Perform credit check
- Request customer deposit
- Initiate service installation
- Reserve resources
- Issue orders, and track status
- Complete orders, notify customers
- Initiate billing process

Service orders to other providers can be generated by the ordering process or by the service configuration process, depending on the nature of the service ordered by the customer. They can also be generated from network and systems management processes when part of a network infrastructure.

An order management system implements two types of functions. The first type is the management of end-to-end processes, which includes process management composed of several subprocesses. The other one is support of order management, service configuration, and service activation. Customer interface management, resource configuration and allocation, and S/P relationship management are fulfilled by other systems.

3.6.3.6 Customer Problem Handling

This area is responsible for receiving trouble reports from customers, resolving them to the customer's satisfaction, and providing meaningful status information on repair and/or restoration activity to the customer.

Principal functions are:

- *Isolate Problem and Initiate Resolution:* Receive and isolate the problem and initiate resolution actions. The process registers and analyses received trouble reports from customers; registers received information about customers impacted by service-affecting problems; and reports problem information to isolate the source/origin of the problem in order to determine what actions have to be taken, and to initiate the resolution of the problem.
- *Report Problem:* Generate and manage all reports that will be issued to the customer or to other processes concerning the problem.
- *Track and Manage Problem:* Track and manage the evolution of the problem during its lifecycle. The process can proactively or passively obtain information about a problem state, get its attributes, or obtain its archived form after its closure. Moreover, the process is responsible for the management of the escalation of the problem.
- *Close Problem:* Ensure that a problem affecting the customer is solved, that the customer is contacted if necessary to inquire about the customer's satisfaction with resolution of the problem, and agree to correct reporting on SLA/QoS violations.

3.6.3.7 Customer Quality of Service and Service-Level Agreement

This process is part of the Assurance end-to end (vertical) processes and the CRM (horizontal) process group as well. The process is responsible for the CRM part of resolving a problem, and must interwork with other related Service and Resource Management processes.

Customer Quality of Service (QoS) and Service-Level Agreement (SLA) Management (Contract and SLA Management) is a set of functions that assists operators in ensuring that their customers get the level of service for which they are paying. This process encompasses monitoring, managing, and reporting of quality of service (QoS) as defined in service descriptions, service-level agreements, and other service-related documents. It includes network performance, but also performance across all service parameters, e.g., orders completed on time. Outputs of this process are standard (predefined) and exception reports including, but not limited to, dashboards, performance of a service against an SLA, reports of any developing capacity problems, reports of customer usage patterns, etc. In addition, this process responds to performance inquiries from the customer. For SLA violations, the process supports notifying the staff responsible for problem handling, and for QoS violations, notifying management. The aim is to provide effective monitoring. Monitoring and reporting must provide SP management and customers with meaningful and timely performance information across the parameters of the services provided. The aim is also to manage service levels that meet specific SLA and standard service commitments.

This is not to be confused with the related set of functions and applications that exist at the service management and resource management layers to help operational managers understand the performance of services and network resources respectively.

Customer QoS functions aim to measure the customer's perceived quality of service. An example of this is the approach taken to measure voice or video quality as perceived by a human being. QoS measurements may be applied either on a per-customer basis, against a group of customers service quality

(e.g., a corporate account), or across an entire service. It may be applied where one operator is retailing another operator's wholesale service (e.g., Mobile Virtual Network Operators).

If QoS applications measure the actual level of service being offered, SLA management applications (Contract Management) provide the ability to compare actual QoS with the level of service promised. This is particularly important where service-level guarantees (SLGs) have been contractually offered—the primary purpose of the SLA management application(s) is to ensure that the operator knows at any time which services to which customers may potentially or actually be in breach of a service-level guarantee.

Principal functions of this application area are:

- *Measure Perceived QoS:* Measure or estimate the actual quality of service being received by the customer against preset thresholds.
- *Manage QoS/SLA Violation:* Ensure that the customer and the relevant internal processes are informed of service quality degradations and violations and that action is undertaken to resolve the degradation or violation.
- *Manage Reporting:* Report on the customer's QoS performance, manage the production and presentation of reports, and prepare reports for internal processes and respond to specific inquiries on the performance of the customer service.

3.6.3.8 Customer Self-Service, Customer Self-Management

This process is part of the FAB end-to end (vertical) process and the CRM (horizontal) process group as well.

Customer self-management provides an Internet technology–driven interface to the customer that allows the customer to undertake a variety of business functions directly for themselves. It interacts with the customer to provide fully automated service or assisted service over various customer touch points. Although customer self-management primarily triggers functionality defined in the rest of the CRM, Service Management and Resource Management applications, they should also contain functionality specific to customer self-empowerment.

Customer self-management generally provides a comprehensive collection of self-service functionality supporting all stages of the customer lifecycle, registration and fulfillment, assurance and billing management activities. As the various features provided for customer lifecycle management are often portlet type applications that are integrated to a CSP's overall customer self-management portal, the self-management functions can be broken down into three major subfunctions:

- Customer self-empowered fulfillment
- Customer self-empowered assurance
- Customer self-empowered billing

New business realities require self-service systems to support the following criteria:

- One-and-done fulfillment across the service portfolio
- Multidisciplinary customer service (customer service/account problem resolution)
- Syncronized multichannel interoperability
- Total convergent self-directed billing (view/pay/dispute all)
- Reconciliation interoperability
- Personalization and usability
- Visualization of SLAs across subscribed services (cf. Customer Service/Account Problem resolution)
- Portfolio–driven guided selling (Product Catalogue, Product Lifecycle Management)
- Leveraging the 360-degree customer view (Customer Information Management)
- Customer self-management applications enable service providers to increase profitability across the organization. These operations expect to gain more customer loyalty, service stickiness, and average revenue per user (ARPU) for the service provider.

These results are expected through:

- Rapid order-to-activation mechanism across the service portfolio
- Commodity-like enablement for telecom services (rapid introduction, easy amendment, cross-bundling)
- Universal platform supporting multiple users (consumers, business, dealers) and multiple lines of business (LOBs_ (wireline, wireless, IPTV) through a single point of contact
- Reducing costs through operating efficiencies

3.6.4 Customer Billing and Collections Management

This process is part of the Billing end-to-end process group, and the Customer Relationship (horizontal) processes. The process can also be called Invoicing and Collection.

Billing and Collections Management processes encompass creating and maintaining a customer's billing account, sending bills to customers, processing their payments, performing payment collections, monitoring the status of the account balance, and the handling of customer-generated or systems-reported billing and payment exceptions. These processes are accountable for assuring that enterprise revenue is billed and collected.

The aim is to provide a correct bill and, if there is a billing problem, resolve it quickly with appropriate status to the customer. An additional aim is to collect monies due the service provider in a professional and customer-supportive manner.

Some providers allow invoicing and collections functions for other providers as a service.

The principal functions include:

- Create and distribute invoices
- Collect payments
- Handle customer billing inquiries
- Support Retention and Loyalty
- Support Marketing Fulfillment and Selling
- Support Customer QoS/SLA
- Manage debt
- Bill on behalf of other providers

3.6.5 Service Management and Operations (SM&O)

Service Management and Operations processes focus on the knowledge of services (access, connectivity, content, etc.) and includes all functionalities necessary for the management and operations of communications and information services required by or proposed to customers. The focus is on service delivery and management as opposed to the management of the underlying network and information technology. Some of the functions involve short-term service capacity planning for a service instance, the application of a service design to specific customers, or managing service improvement initiatives. These functions are closely connected with the day-to-day customer experience.

The processes in this horizontal functional process grouping are accountable to meet, at a minimum, targets set for service quality, including process performance and customer satisfaction at a service level, as well as service cost.

3.6.5.1 Service Planning and Development Process

Service Development and Management is part of the SIP (vertical) group and Service Management and Operations (SM&O).

Service Development and Management is a horizontal functional process grouping that focuses on planning, developing, and delivering services to the Operations domain. It includes processes necessary

for defining the strategies for service creation and design, managing existing services, and ensuring that capabilities are in place to meet future service demand.

This process encompasses the following functional areas and functions:

- Designing technical capability to meet specified market need at a desired cost
- Ensuring that the service (product) can be properly installed, monitored, controlled, and billed
- Initiating appropriate process and methods modifications, as well as initiating changes to levels of operations personnel and training required
- Initiating any modifications to the underlying network or information systems to support the requirements
- Performing preservice testing to confirm that the technical capability works and that the operational support process and systems function properly
- Ensuring that sufficient capacity is available to meet forecasted sales
- Developing and implementing technical solutions
- Developing and implementing procedures
- Defining and implementing systems changes
- Developing and implementing training
- Developing customer documentation
- Planning rollout, testing, start service, and project management
- Set product/service pricing

3.6.5.2 Service Configuration and Activation Process

The Service Configuration and Activation process is part of the Fulfillment end-to-end process group, and Service Management and Operations processes as well.

This process encompasses the installation and/or configuration of services for specific customers, including the installation/configuration of customer premises equipment (CPE). It also supports the reconfiguration of service (either due to customer demand or problem resolution) after the initial service installation. The aim is to correctly provide service configuration within the time frame required to meet ever-decreasing intervals. They also support the reconfiguration of the service (either due to customer demand or problem resolution) after the initial service installation. This can include modifying capacity and reconfiguring in response to requests from other providers.

This includes the following principal functions:

- *Designing:* Provide a system architecture that complies with a particular customer requirement.
- *Activate Service:* These processes trigger the end-to-end activation/deactivation of all installed elements. These processes also include additional functions such as reporting configuration completion to client processes as well as updating and maintaining the customer's network records and service infrastructure records.
- *Track and Manage Work Orders:* The purpose of the these processes is to launch all the operational tasks needed to fix each solution requirement. The information flow associated with these processes includes the communication with the supplier/partner in order to accept requests for service configuration, service configuration changes, and/or for additional resource capacity.
- *Allocate/Assign Resources to Services:* The purpose of this process is to issue identifiers for new services and to manage identifier pools (blocks, ranges, and subnetting) for services (e.g., phone numbers, IP address, voice mailbox number, etc.). Deactivation of existing services is also handled by this process.
- *Implement and Configure Service:* This process delivers a final configuration that is optimal for customer service requirements. These processes install, configure, and reconfigure services for specific customers, including customer premises equipment.
- *Test Service End-to-End:* This process ensures that all components are operational, and that the service is working to agreed levels before its activation for the customer.

- *Handle Configuration Requests:* This process initiates/accepts configuration requests to/from other providers.
- *Upgrading and Reporting:* This process upgrades customer records and reports the completion of the configuration/activation tasks.

3.6.5.3 Service Design and Assign

This process is part of the fulfillment process on those services that require specific custom design or topology changes. These are typically complex business services that can include service to multiple users and locations, and possibly multiple service elements. As a simple example, imagine a typical business requirement for VPN services between two or more locations. Historically the design of these paths to enable the service was planned by engineering planning departments and the service was assigned or enabled manually.

Due to the multitechnology, multinetwork, and multiplayer nature of most services and the fact that networks typically contain thousands of network elements and nodes, the manual design and creation of services is a time-consuming and error-prone process.

To simplify service design, intelligent path analysis functions and other tools are needed, usually built into the inventory to automate the process. These features enable network engineers to quickly identify—after selecting just a few key criteria—an optimal service design based on rules such as least-cost routing, diversity, and preferred carrier. The tools should also return multiple service design options. Once an optimal service design is identified and selected, the inventory OSS/BSS assigns network elements to the end-to-end service to ensure effective tracking of service details and capacity requirements.

3.6.5.4 Service Problem Management

The Service Problem Management process responds immediately to customer-affecting service problems or failures in order to minimize their effects on customers, to invoke the restoration of the service, or to provide an alternate service as soon as possible. They encompass the reporting of problems, making a temporary fix or work-around, isolating the root cause, and finally recovering the complete functionality of the service and providing information for future enhancements.

This process encompasses isolating the root cause of service-affecting and non-service-affecting failures and acting to resolve them. Typically, failures reported to this process impact multiple customers. Actions may include immediate reconfiguration or other corrective actions. Longer-term modifications to the service design or to the network components associated with the service may also be required. The aim is to understand the causes impacting service performance and to implement immediate fixes or initiate quality improvement efforts.

Principal functions are:

- *Evaluate and Qualify Problem:* Determine the nature of a problem that has been reported by a customer and whether the customer is using the service properly.
- *Reporting to Customers in the Event of a Disruption:* This is done whether the disruption was reported by the customer or not.
- *Diagnose Problem:* Isolate the root cause of the problem. To achieve this, this process performs the appropriate tests and/or initiates queries for information. These processes perform a problem escalation to report the severity and, if necessary, to solve the incident.
- *Plan and Assign Resolution:* Identify the necessary steps in order to activate the different units that will be involved in fixing the problem.
- *Resolve the Problem:* Resolve the problem to the customer's satisfaction.
- *Track and Manage Work Orders:* Launch all the operational tasks needed to accomplish each solution requirement. The information flow associated with these processes includes the communication with the supplier/partner in order to accept requests for service configuration, service configuration changes, and/or for additional resource capacity. It coordinates the different

internal and external activities, and notifies the other processes that are responsible for the implementation and configuration.
- *Certify the Recovery of Normal Service Performance:* These processes will perform the necessary testing to achieve this purpose and make the necessary reports about the problem that occurred, the root cause, and the activities carried out for restoration. It also will issue the trouble clearance report to inform the CRM layer.

When trouble is reported by the customer, a trouble report may be sent to service problem resolution for correction. When a trouble is identified by service problem resolution, then problem handling is notified in order to inform the customer of the problem.

The processes deal with reactive versus proactive problem management.

Reactive problem management actions are needed where a problem has already affected the services being provided to customers. However, it is desirable to detect and resolve problems prior to users being affected by them, which is addressed by Proactive Problem Management.

Proactive Problem Management also includes planned maintenance outages. The aim is to have the largest percentage of problems proactively identified and communicated to the customer, to provide meaningful status, and to resolve them in the shortest time frame.

3.6.5.5 Service Quality Management Process

Service Quality Management (SQM) and impact analysis applications are designed to allow operators to determine what levels of service they are delivering to their customers. Ideally these take a customer-centric view, that is, the quality of service perceived by customers, but may measure additional service metrics to allow the operator to be aware of approaching problems or degradations to service. Impact analysis applications extend this capability to predict the likely impact of service degradations or network problems on specific customers.

The Service Quality Management processes encompass monitoring, analyzing, and controlling the performance of the service perceived by customers. These processes are responsible for restoring the service performance for customers to a level specified in the SLA or other service key quality indicator (KQI) descriptions as soon as possible. Once a problem has been detected and the root cause diagnosed, this information is presented to the operations team using a variety of integrated mechanisms. These mechanisms provide real-time notification and assignment of problems based on a variety of alarm metrics, forward or escalate the problem details (problem cause, impact assessment, and recommended corrective actions) to operational personnel or the systems responsible for service restoration.

These processes monitor the service quality using events from Resource Management, using the quality to forecast whether SLA promises will be met, and to improve the service quality.

This process supports monitoring service or product quality on a service-class basis in order to determine whether:

- Service levels are being met consistently
- There are any general problems with the service or product
- The sale and use of the service is tracking to forecasts

Principal processes are:

- *Monitor Service Quality:* Extract the necessary information to feed the different quality analysis processes. These processes collect and store all quality indicators related to the service, such as congestion events and resource alarm events. The processes also perform automated service testing using simulated standard user calling behavior, and collect data related to service usage, which may supply information to other processes.
- *Receive Alarm Information:* Receives alarm information from network and network element systems, and tests and diagnoses information from test systems.

- *Analyze Service Quality:* Assess the effectiveness of the service by identifying the current quality level against forecast or specified quality levels. Using the raw data from Service Quality Monitoring, these processes will correlate events in order to filter repetitive alarms and failure events that do not affect the quality delivered, and they will calculate key service quality indicators (such as mean time between failures and other chronic problems).
- *KPIs and KQIs:* Collates the key performance indicators (KPIs) and converts them to Key Quality Indicators (KQIs) against which the service quality can be measured.
- *Improve Service:* Assess and recommend improvements using the information from Service Quality Analysis in order to improve and/or correct deviations from the forecast KQIs. This recommendation will be passed to the customer layer, the resource layer, or the Service Configuration and Activation Processes.
- *Identify and Report Service Constraints:* Identify constraints that can affect service quality standards. These constraints may include resource failures, capacity shortages due to unexpected demand peaks, etc. These processes send this information to the CRM layer in order to keep customers informed.

In sophisticated systems, root cause analysis and service impact analysis are also incorporated into Service Quality Management Systems.

3.6.5.6 Service-Level Agreement Management

SLA Management use the output of the SQM applications to provide a comprehensive view of the level of service provided to customers compared to pre-agreed, often contractually binding agreements. Typically SLA agreements will be negotiated between operator and customer to measure a variety of service-oriented issues and impacts. These may be stated either in terms of service characteristics or in terms of the business impacts on the customer. Service-oriented characteristics could include:

- Availability
- Security
- Latency
- Transmission speed
- Time to respond to the initial fault report
- The escalation process
- The time to repair
- Spares holding
- The algorithm for calculating rebates
- Contact details
- Time to deliver from order confirmation

3.6.5.7 Rating and Discounting

Rating processes are the traditional heart of any billing system. They should ensure that the customer receives an invoice that is reflective of all the billable events delivered by the service provider dictated by their business relationship. In addition, it ensures that the appropriate taxes, rebates (i.e., missed customer commitments) and credits are applied to the customer's invoice(s).

Rating and discounting applications manage the customer's account and customer-specific pricing, charges, discounting, credits, and taxation for services delivered. They accept events that have been collected, translated, correlated, assembled, guided, and service rated. It takes these events and determines the account or customer-specific pricing, charges, discounts, and taxation that should be reflected in the invoice(s) for the customer.

In traditional circuit-switched voice networks, rating systems take the output of circuit switches as a call detail record and, based on preloaded algorithms, produce an appropriate charge for the call based on factors such as time, distance, customer tariff plans, etc. Historically, rating applications worked on a batch basis as calls were traditionally billed on a post-paid, after-the-event basis.

Principal functions are:

- *Mediate Usage Records:* Validate, normalize, convert, and correlate usage records collected from the resource layer. These processes also group usage records that relate to a specific service usage.
- *Rate Usage Records:* Identify and apply tariffs and charging algorithms to specific parameters encapsulated in usage records, in order to produce a charge that is then inserted in the usage record.
- *Apply Rating Rules:* Apply the correct rating rules to usage data on a customer-by-customer basis, as required.
- *Apply Discounts:* Apply any discounts agreed to as part of the ordering process, and any promotional discounts and charges.
- *Analyze Usage Records:* Generate reports on usage records based on requests from other processes. These processes produce reports that may identify abnormalities, which may be caused by fraudulent activity or related to customer complaints.

3.6.6 Resource Management and Operations (RM&O)

Resource Management and Operations functional processes maintain knowledge of resources (application, computing, and network infrastructures) and is responsible for managing all these resources (e.g., networks, IT systems, servers, routers, etc.) utilized to deliver and support services required by or proposed to customers. It also includes all functionalities responsible for the direct management of all such resources (network elements, computers, servers, etc.) utilized within the enterprise. These processes are responsible for ensuring that the network and information technologies infrastructure supports the end-to-end delivery of the required services. These processes ensure that infrastructure runs smoothly, is accessible to services and employees, is maintained, and is responsive to the needs, whether directly or indirectly, of services, customers, and employees. RM&O also has the basic function to assemble information about the resources (e.g., from network elements and/or element management systems), and then integrate, correlate, and in many cases, summarize that data to pass on the relevant information to Service Management systems, or to take action in the appropriate resource.

The Resource Management and Operations processes manage both service provider networks/subnetworks and information technology architectures.

Resource Management is often referred to as Network and System Management.

3.6.6.1 Inventory Management Process

This process works on the Resource Management and Operations level, and supports mainly Assurance, but also Fulfillment and Billing.

This process encompasses anything to do with physical equipment and the administration of this equipment. The process is involved in the installation and acceptance of equipment, with the physical configuration of the network, but also with handling of spare parts and the repair process. Software upgrades are also a responsibility of this process.

In recent years, the need for accurate inventory has become even more critical, as wireless operators roll out a variety of new services, wireline operators seek to realize cost savings through back-office consolidation, multiple services are offered over IP transport, and triple- and quad-play offerings drive significantly increased network and service complexity.

Despite the importance of inventory to a service provider's business, maintaining inventory data that accurately reflects real assets in place remains a challenge to many service providers. Depending on the line of business, inventory accuracy can be as low as 40%, and many service providers report accuracies around the 80% mark. For this reason, many are focusing on inventory management as a strategic imperative, realizing that it is time to "finally get it right."

Inventory management is in a period of transition and growth. As with many OSSs/BSSs, inventory systems are moving from solutions addressing a single "point of pain" to integrated, open solutions

that interoperate more effectively with a carrier's legacy environment, and can be configured to handle multiple lines of business. The bottom line: significantly reduced management costs, and inventory data that more accurately reflects inventory reality, tracking inventory changes in close to real time. As the technology has shifted, so has the vendor community, with significant ongoing M&A activity, revenue growth (for some), and shifting customer demographics.

Inventory systems must support a multitechnology and multilayer approach that captures the complete end-to-end environment. By capturing and integrating all technology domains—from DSL to IP—in a common inventory system, you can finally optimize operations by applying consistent network management principles by identifying where and how different technologies interact in the network.

With this perspective, you can quickly navigate between the different open systems interconnection (OSI) layers to identify dependencies and perform root cause analysis in the event of a network or service problem. This functionality also allows you to establish and enforce capacity utilization rules across and between layers, as well as understand complex relationships among different technologies and resources from two perspectives: the service view and the network resource view.

Principal functions are:

- Install and administer the physical network
- Supports Assurance Processes
- Supports Fulfillment and Billing Processes
- Perform work in the network
- Manage the repair activities
- Align inventory with network
- Manage spare parts
- Manage faulty parts
- Network Discovery—keeps the inventory up-to-date, with reduced manual work

3.6.6.2 Workforce Management

Workforce Management applications manage field forces to make optimum use of manpower and other resources such as vehicles. They are used to schedule resources, provide a map of field skill sets, and provide forecasting and load-balancing capabilities.

Workforce Management can be used to manage both internal and external (customer) resources in both service assurance and provisioning areas.

Principal functions are:

- *Scheduling:* Applications are usually designed to build schedules for groups and individuals, taking into account shift patterns, daily duties, multiple skill sets, resource availability, schedule preferences, and fluctuating nature of the workload.
- *Forecasting:* Application usually calculates optimal staffing requirements with input of historic statistics, service level goals, call center costs, change parameters, and expected workload. They may include resources that are required by date, time, queue, resource pool, etc.
- *Dynamic Management:* Application provides for immediate and unexpected changes in resource status, such as sick leave, or when unforeseen changes in the workload dictate that conditions be constantly monitored and spontaneous adjustments made.
- *Operational Support:* This function typically tracks and reports Work Force Management data such as actuals to forecasts and gathering individual and group statistics.

3.6.6.3 Resource, Network Planning, and Development Process

This process encompasses development and acceptance of strategy, description of standard network configurations for operational use, and definition of rules for network planning, installation, and maintenance.

This process also deals with designing the network capability to meet a specified service need at the desired cost and for ensuring that the network can be properly installed, monitored, controlled, and billed. The process is also responsible for ensuring that enough network capacity will be available to meet the forecasted demand and support cases of unforecasted demand. Based on the required network capacity, orders are issued to suppliers or other network operators and site preparation and installation orders are issued to the network inventory management or a third-party network constructor. A design of the logical network configuration is provided to network provisioning.

3.6.6.4 Resource Provisioning

This process is part of Fulfillment, and the Resource Management and Operations as well.

Its purpose is to allocate and configure resources to individual customer service instances in order to meet service requirements.

The process encompasses the configuration of the network, to ensure that network capacity is ready for provisioning of services. It carries out network provisioning as required, to fulfill specific service requests and configuration changes to address network problems. The process must assign and administer identifiers for provisioned resources and make them available to other processes. Note that the routine provisioning of specific instances of a customer service—in particular, simple services such as plain old telephone service (POTS)—may not normally involve network provisioning, but may be handled directly by service provisioning from a preconfigured set.

The processes encompass allocation and configuration of resources to individual customer service instances in order to meet the service requirements, and includes activation as well as testing to ensure the expected performance of the service.

The principal functions are:

- *Allocate and Deliver Resources:* Identify resources required to support a specific service instance. Such allocation request can be placed as part of a preorder feasibility check, to see whether there are adequate resources available to fulfill the request. In addition, the Allocate and Deliver Resource processes will possibly ensure that the appropriate resources are delivered to the appropriate location for installation and configuration.
- *Configure and Activate Resources:* Configure and activate the resources reserved for supporting a specific service instance. The Configure and Activate Resource processes can receive configuration requests for adding, changing, and complementing services, as well as for fixing resource troubles and adding resource capacity to cope with performance problems.
- *Test Resources:* The responsibility of this process is to test resources supporting a specific service instance. The objective is to verify whether the resources work correctly and meet the appropriate performance levels. If these tests succeed, the resources will be marked as in-service, which means the resources are available for use.
- *Update and Report on Resources:* These processes ensure the Resource Inventory Database reflects resources are being used for a specific customer.

Requirements for successful provisioning:

- Allocated resources must exist
- Resources must not be allocated to other services
- Physical location of resources must be known
- Physical and logical connections between resources must be known
- Requires that network elements are modeled into inventory system
- Requires extensive Network Element Interface (NEI) library from service fulfillment platform

It should be noted that NEI implementations vary a lot because there are no commonly accepted standards. Many network elements are managed though Command Line Interface (CLI), which is easy for humans, but problematic for machine-to-machine communication.

3.6.6.5 Resource Trouble Management Process

Resource Trouble Management is an essential part of the Assessment end-to-end process group.

Resource Trouble Management processes are responsible for the management of troubles with allocated resources. The objectives of these processes are to report resource failures, isolate the root causes, and act to resolve them.

Principal processes and functions are:

- *Survey and Analyze Resource Trouble:* This process takes care of the monitoring of resources in real time by resource failure event analysis, alarm correlation, and filtering as well as failure event detection and reporting. Responsibilities of these processes include, but are not limited to, Resource failure event analysis, alarm correlation and filtering, and failure event detection and reporting. The alarm correlation in particular aims at the matching of redundant, transient, or implied events to a specific root cause event.
- *Localize Resource Trouble*: This process is responsible for finding the root cause of resource trouble, which can be done using the following methods: verification of resource configuration for the targeted service features, performing resource diagnostics, running resource tests, and scheduling resource tests.
- *Correct and Recover Resource Trouble:* Failed resources are either restored or replaced by this process. These processes are also responsible for isolating a unit with a fault and managing the redundant resource units (e.g., hot standby). They will also report successful restoration or an unsuccessful attempt at restoration to cooperating processes.
- *Track and Manage Resource Trouble:* This process monitors the progress of the repair activities in the previous process.
- *Report Resource Trouble:* This process reports changes in resource troubles to other interested processes (e.g., Service Trouble Management).
- *Close Trouble Report:* To close a trouble report, this process verifies the successful elimination of the problem. This includes ensuring appropriate measures will be taken in order to prevent similar resource troubles from occurring in the future. It also includes interaction with the Resource Trouble processes for reporting purposes.

Event and alarm correlation are core components, as well as integrated impact analysis.

3.6.6.6 Resource Performance Management Process

Resource Performance Management processes encompass monitoring, analyzing, controlling, and reporting on the performance of resources. They work with basic information received from the Resource Data Collection and Processing processes.

Principal processes are:

- *Monitor Resource Performance:* These processes monitors received resource performance information. These processes monitor normal performance, and will also apply appropriate thresholds against resource performance. If a performance threshold is crossed, a resource trouble report will be instituted and handled through the Resource Trouble Management processes.
- *Analyze Resource Performance:* These processes analyze and evaluate received resource performance information and report the findings of this analysis. This also includes further analysis of the performance anomalies identified by the Monitor Resource Performance processes. Resource fault conditions will be detected by the Resource Trouble Management processes. However these processes, while still impacting service quality, may not detect intermittent fault conditions in multiple resources. The Analyze Resource Performance processes determine the overall quality, using specific measurements in order to detect service quality degradations. They may also detect degradation trends before the resource performance has dropped below an acceptable level.

- *Traffic Status and Performance Analysis:* If the performance analysis identifies a resource quality problem, this will be reported by the Report Resource Performance processes. Once reported to Resource Trouble Management or Service Quality Management, the Analyze Resource Performance processes will continuously monitor the resource performance and possibly escalate the problem, ensuring the resource performance is restored to an acceptable level in a timely manner.
- *Control Resource Performance:* The objective of the Control Resource Performance processes is to apply controls to resources in order to optimize the resource performance. The need for installing performance controls can be identified by the Analyze Resource Performance processes or requested by the Service Quality Management processes. This also includes the establishment of preplans, which have been defined in order to cope with abnormal situations caused by, for example, natural disasters or mass calling events. In the case of, for example, circuit switching and SS7 networks, this encompasses the application of traffic controls that affect call processing and traffic routing in order to optimize the call completion rate. In the case of packet networks, this encompasses various types of traffic conditioning, such as policing and shaping.
- *Report Resource Performance:* This process reports the findings of the Analyze Resource Performance processes. If the performance analysis identifies a resource quality problem, this will be reported to the Resource Trouble Management and/or the Service Quality Management processes. The latter processes are responsible for deciding on and carrying out the appropriate action/response.

3.6.6.7 Resource Data Collection and Processing

This process supports the Assurance and Billing end-to-end processes on the Resource Management level.

Resource Data Collection and Processing collects usage, network, and information technology events and performance information for distribution to other processes within the enterprise.

Principal processes are:

- *Collect Resource Data:* This process collects usage, network, and information technology events and performance information. Before this information is distributed to other processes within the enterprise, it will be passed to Process Resource Data for processing.
- *Process Resource Data:* The processes are responsible for processing the raw data collected from the resources. This includes the filtering of Resource Data based on well-defined criteria, as well as providing summaries of the Resource Data through aggregation. These processes are also responsible for formatting the Resource Data before distributing it to other processes within the enterprise.
- *Report Resource Data:* The Report Resource Data processes are responsible for distributing processed Resource Data to other processes within the enterprise for further analysis and/or reporting.
- *Audit Resource Usage Data:* The Audit Resource Usage Data processes are responsible for auditing the resource data collection stream in order to identify possible anomalies such as loss of usage data in the different collection and processing steps.

3.6.7 Supplier/Partner Relationship Management (S/PRM)

Supplier/Partner Relationship Management functional processes support the core operational processes, both the customer instance processes of Fulfillment, Assurance, and Billing and the functional operations processes. Supplier/Partner Relationship Management (S/PRM) processes align closely with a supplier's or partner's Customer Relationship Management processes. The inclusion of distinct Supplier/Partner Relationship Management processes in the eTOM framework enables a direct interface with the appropriate lifecycle, end-to-end customer operations or functional processes with suppliers and/or partners. The processes include issuing RFPs as part of the buy process, issuing purchase orders and tracking them through to delivery, mediation of purchase orders as required conforming to external processes, handling problems, validating billing and authorizing payment, as well as quality management of suppliers and partners.

When the enterprise sells its products to a partner or supplier, this is done through the enterprise CRM processes, which act on behalf of the supplier or the enterprise in such cases. Supplier/Partner (S/P) processes only cover the buying of services by the enterprise.

Principal processes are:

- *S/PRM Support and Readiness:* These processes are responsible for ensuring that all necessary facilities related to the interaction with suppliers and partners are ready and functioning. Moreover, these processes are responsible for the resolution of problems related to these facilities.
- *S/P Requisition Management:* S/P Requisition processes manage requisitions with partners and suppliers to ensure on-time and correct delivery of the product or service requested by the enterprise. According to the appropriate policy and practices of the enterprise, supply chain processes in SIP may be involved as well as, or instead of, S/P Requisition Management to achieve this. This process interfaces with the supplier's CRM process for order handling.
- *S/P Problem Reporting and Management:* S/P Problem Reporting and Management processes manage problems associated with supplier/partner interactions, whether identified within the enterprise or notified by the supplier/partner.
- *S/P Performance Management:* These processes track, measure, and report the performance of services or products from suppliers and partners, as well as how this relates to the performance of the supplier/partner against any contracts or business agreements. They interface with the supplier's CRM processes of Customer QoS/SLA Management.
- *S/P Settlements and Billing Management:* Manage all settlements and billing for the enterprise, including bill validation and verification and payment authorization.
- *S/P Interface Management:* Manage the contacts between the enterprise and its current or future suppliers/partners for products or services.

3.6.8 Support Processes Taxonomy

The grouping of support processes, functions, and products is not yet unique. Standard organizations, industry associations, market research companies, powerful service providers, and suppliers give their input and recommendations for process, function, and product taxonomies. This segment will introduce a few actionable examples.

3.6.8.1 TMN-OSS Model

Besides the eTOM model, the Telecommunications Management Network (TMN) model is still often used in the iindustry. Mapping of several Support Processes to the TMN layers is shown in Table 3.6.1. For all typical TMN layers, typical responsibilities are listed. It is obvious that there is no clear separation among processes, functions, and responsibilities. This table is useful as an entry point for building a blueprint of support process hierarchies.

TABl E 3.6.1 Mapping of Several Support Processes to the TMN Layers

Business Management	Management Strategic Direction, Planning, Budgeting Services, Service Level Agreements, Sales, Staffing, Payroll
Services Management	Service Order and Activation, Provisioning, Customer Care, Customer Access, Metrics Reporting, QoS, SLA Monitoring, Assets/Inventory, Performance, Capacity, Accounting, Billing
Network and Systems Management	Network Monitoring, Change Control, Configuration Management, Access Control, Statistical Data Collection/Analysis, Status Monitoring
Network Element Management	Agent/SNMP Manager, Alarms, Notifications, Fault Monitoring, Correlation, Diagnostics, Element FCAPS

3.6.8.2 Taxonomy by Insight Research

Companies performing market analysis on telecommunications usually work with their own model (market segmentation) in order to distinguish themselves from others.

Insight Research Corporation defines some segments and subsegments as shown on Table 3.6.2. This is a very individual grouping of processes and functions. Basically, the core support processes and functions are well represented without claiming general applicability. By including workforce management, middleware solutions, and professional services, this taxonomy goes beyond the core functionality and recommend service orientation.

3.6.8.3 Taxonomy by Amdocs

The following table (Table 3.6.3) illustrates the *OSS Observer*'s telecom software market taxonomy as seen by Amdocs. Amdocs has reached the size and power by now to be able to redefine existing taxonomy standards. The list of support processes and functions is complete and state-of-the-art. As expected from a large service company in the area of OSS/BSS, the structure is simpler than eTOM and easier to implement. Amdocs will most likely follow the time-to-market demand of its customers.

3.6.8.4 Taxonomy by the Gartner Group

Gartner defines the OSS market as follows (Gartner, March 2007):

- Inventory—tracks and manages network assets. In this ongoing process, installed and on-hand network assets are tracked for efficient inventory, procurement, repair, and reuse.
- Provisioning and Activation—includes systems and steps related to the process of implementing orders for customers.
- Network Management—includes configuration, traffic, fault, security, element, and performance management.
- Planning and Engineering—includes the steps from network planning to construction (for example, budgeting, procurement, and line and service testing).
- Workforce Management—encompasses activities surrounding work assignment, coordination, and tracking. The process involves ensuring that personnel with the appropriate qualifications are given the correct equipment at the right time and place. Examples of IT applications and systems supporting workforce management are dispatch, workflow management, and project tracking.

3.6.8.5 eTOM Taxonomy

Table 3.6.4 gives a summary of several processes used in the practice when the eTOM recommendations are followed.

3.6.9 eBusiness

In general eBusiness is understood as the interaction among business partners with the help of information technologies. It refers not only to buying and selling over the Internet (or other computer network), but also to servicing customers and collaborating with business partners.

The term eBusiness has often been interchanged with the term eCommerce. However, it is becoming increasingly accepted that the use of eCommerce should be restricted to referring to just those Web transactions (mainly business-to-consumer) that are used while buying and selling services and goods over the Internet.

An eBusiness enterprise is, then, an enterprise that utilizes Internet and related technologies to compete effectively in its business space.

TABl E 3.6.2 Insight Research Corporation's Segmentation

Segment	Subsegment	Definition of the Subsegment
Billing	Mediation	Gathering of call detail records (CDRs) and other data from network elements and preparation of this data for operation on by the bill rendering systems
	Rendering	Rating of the customer usage data and generation of the billing statement at the appropriate time
Customer Care		Front-end system supporting customer service agents in response to service requests, trouble reporting, and billing inquiries
Engineering and Planning		Systems related to network design, central office engineering, outside plant design, equipment engineering, capacity planning, etc.
Provisioning	Provisioning	Systems that assign network resources to a specific service
	Inventory Management	Tracking and reporting on equipment, tools, and materials with regard to quantity, location, and identity
Trouble Repair	Order Management	The scheduling and tracking of work related to installing a new service or changing an existing service
	Trouble Management	The scheduling and tracking of work related to trouble resolution
	Security Management	Systems that prevent unauthorized access to systems and data
	Testing	Measurement of network characteristics by taking network resources out of service or reducing the amount of resources available for service
	Surveillance	The in-service collection of network resource data to report on network events and the analysis of these events to determine alarm conditions
	Data Collection	Collects customer network usage data for the purpose of billing
Business Management		Payroll, marketing and sales, accounting, and HR systems
Workforce Management		Work assignment and the tracking of technicians handling trouble repair and service requests in the central office and at customer sites
Element Management		Systems that manage telecommunications equipment, including setting options, collecting equipment alarms, health, and performance data.
Middleware		The software that supports interactions between client and server software
Professional Services	IT Consulting	Consulting related to data networking, system performance, capacity sizing, database design, and other topics relating to computer-based architecture.
	Systems Integration	Integration and testing of OSS modules from different suppliers
	Business Process Re-engineering	Evaluation and analysis of business processes to align them with strategic business objectives
	Project Management	Involves overall coordination and management of project
	Customized Software	The adaptation of existing software, or the design and programming of new software, to meet project and customer-specific needs.

Source: Insight Research Corporation.

TABl E 3.6.3 *OSS Observer*'s Telecom Software Market Taxonomy

Functional Area/Segment	Associated Functions/Systems
Service Delivery Platforms	Charging
	Content Management
	Telecom Applications Server
	Device Management
Billing	Rating & Pricing
	Partner & Interconnect
	Fraud & RA
	Mediation
Customer Care	Customer Interactions Management
	CRM
	Subscriber Management
	Workforce Automation
Service Fulfillment	Order Management
	NRM/Inventory
	Activation
	Engineering Tools
Service Delivery	Service Management
	Fault & Event Management
	Performance Management
	Probe Systems
Network Management Systems	Mobile
	Residential Broadband
	Business Data Services
	PSTN
Middleware	Middleware

Source: OSS Observer.

3.6.9.1 Implications of eBusiness for Service Providers

As new technologies and markets emerge, enterprises have to adapt or die. Technologies affect customer needs while customer needs influence business designs. As business designs emerge, they affect processes and processes influence both customer expectation and the next generation of technology.

In response to this new paradigm, it is imperative that enterprises integrate business, technology, and processes. They must redefine the way in which they operate by using new technology-based business designs, creating new interenterprise processes, and integrating operations to support changing customer requirements.

The three principal reasons that service providers must integrate eBusiness and traditional business processes are:

- Customer expectations and the need to move to an approach that focuses on the management of customer relationships and the importance of improving customer retention and increasing the value customers contribute to the enterprise.
- Productivity gains and the need to ensure that these can continue to be obtained.
- Provision of a broader range of products and services to customers. For the information and communications services industry (more than almost any other industry), this requires a focus on better collaboration between and integration of processes.

The processes required in an eBusiness environment are fundamentally different from those in a traditional business environment.

TABl E 3.6.4 Summary of Principal Processes Used in the Practice of eTOM

Process or Function Name	Definition
Customer Relationship Management (CRM)	These processes use the knowledge of customers' needs and include all functions necessary for the acquisition, enhancement, and retention of a customer by supporting storefront, telephone, Web, or field services. It also involves churn analysis, cross-selling/up-selling, and direct marketing activities targeted to increasing sales. CRM applies to both conventional retail customer interactions and wholesale interactions.
Assurance	Once a service is operational, the telecommunications service provider's primary responsibilities are to ensure that the service stays operational, to provide data to the customer on the performance of their service, and to provide customer support in the use of the service. The Assurance processes are responsible for the execution of proactive and reactive maintenance activities to ensure that services provided to customers are continuously available and to SLA or QoS performance levels. It performs continuous resource status and performance monitoring to proactively detect possible failures. It collects performance data and analyzes it to identify potential problems and resolve them without impact to the customer. This process manages the SLAs and reports service performance to the customer. It receives trouble reports from the customer, informs the customer of the trouble status, and ensures restoration and repair, as well as a delighted customer.
Fulfillment	The fulfillment end-to-end process is responsible for providing customers with their requested products in a timely and correct manner. It translates the customer's business or personal need into a solution, which can be delivered using the specific products in the enterprise's portfolio. This process informs the customers of the status of their purchase order, ensures completion on time, as well as a delighted customer. Service fulfillment refers to the complete process from receiving the customer order (e.g., for broadband or business VPN, Internet, VoIP, etc.) to activating and testing the service in network. The fulfillment process is initiated predominantly by the customer. Thus, fulfillment systems are those systems that are used to receive and handle orders, provision service, and create/update information for customer billing.
Customer Contact, Retention, & Loyalty	Customer contact management, retention, and loyalty are a varied group of functions that are generally part of a Customer Relationship Management (CRM) suite. It allows an operator to create, update, and view the customer's information, record and view all customer interactions across different communication channels and departments, so that whoever is speaking to a customer can see the history of issues that have concerned that customer. More sophisticated systems allow capabilities to highlight customers as risks of switching to an alternative carrier (churn indicator).
Customer Information Management	Allows the creation, update, lookup/search, and view of customer information.
Customer Interface Management	These processes directly interact with customers and translate customer requests and inquiries into appropriate events, such as the creation of an order or trouble ticket or the adjustment of a bill. The process logs customer contacts, directs inquiries to the appropriate party, and tracks the status to completion.
Selling/Sales	This process encompasses learning about the needs of each customer, and educating the customer about the communications services that are available to meet those needs.

Continued

TABl E 3.6.4 Summary of Principal Processes Used in the Practice of eTOM (*Continued*)

Process or Function Name	Definition
Order Management, Order Handling	Order management functions manage the end-to-end lifecycle of a customer request for services. This includes capturing the order, configuring the products and services within the order, decomposing the order for provisioning activities, and orchestrating the activation and fulfillment and billing notification processes. Order management typically serves all the customer touch points/channels, including call center, retail, self-service, dealers, affiliates, etc. The order may be initiated by any channel and visible to the other channels if needed. It also may include creating the customer's billing profile, tracks the order status, assigns individual orders to specific employees, and allows task management within a predefined ordering team.
Service Order Processing	Based upon customer requests submitted to customer service representatives, creation and activation of services to customers.
Handling Service Change Requests	Based upon customer service change requests submitted to customer service representatives, changing and reactivation of services to customers.
Customer Problem Handling	This process is responsible for receiving trouble reports from customers, resolving them to the customer's satisfaction, and providing meaningful status on repair and/or restoration activity to the customer.
Customer Quality of Service & Service-Level Agreement	The process is responsible for the CRM part of resolving a problem, and must work with other related Service and Resource management processes. The process contains set of functions that assist operators in ensuring that their customers get the level of service for which they are paying. This process encompasses monitoring, managing, and reporting of quality of service (QoS), service-level agreements, and other service-related documents. Outputs of this process are standard (predefined) and exception reports including, but not limited to, dashboards, performance of a service against a SLA, reports of any developing capacity problems, reports of customer usage patterns, etc.
Customer Self-Service, Self-Management	Provides a comprehensive collection of self-service functionality supporting all stages of the customer lifecycle. These processes may include the following subprocesses: • Customer Self-Empowered Fulfillment • Customer Self-Empowered Assurance • Customer Self-Empowered Billing
Customer Care	Evaluates historical customer requirements, traffic patterns, expectations; reports and solves technical and billing problems.
Customer Analysis and Acquisition	Billing platforms tend to maintain the most complex picture of telecom customers in terms of resource usage, habits, and traffic patterns. Using these data intelligently, customer churn can be avoided and new services can be sold to customers.
Customer Billing & Collections Management	Billing & Collections Management processes encompass creating and maintaining a customer's billing account, sending bills to customers, processing their payments, performing payment collections, monitoring the status of the account balance, and the handling of customer-generated or systems-reported billing and payment exceptions. These processes are accountable for assuring that enterprise revenue is billed and collected.
Service Management and Operations	Service Management and Operations processes focus on the knowledge of services (access, connectivity, content, etc.) and includes all functionalities necessary for the management and operations of communications and information services required by or proposed to customers. The focus is on service delivery and management as opposed to the management of the underlying network and information technology.

TABl E 3.6.4 Summary of Principal Processes Used in the Practice of eTOM (*Continued*)

Process or Function Name	Definition
Service Planning and Development Process	This process grouping focuses on planning, developing, and delivering services to the Operations domain. It includes processes necessary for defining the strategies for service creation and design, managing existing services, and ensuring that capabilities are in place to meet future service demand.
Service Creation	Process of creating and testing new or advanced services on the basis of the existing infrastructure of the service providers
Service Activation, Provisioning, and Assignment	Process of allocating equipment, assigning numbers, and activating circuits or ports at switches and activating customer services.
Service Configuration and Activation	This process encompasses the installation and/or configuration of services for specific customers, including the installation/configuration of customer premises equipment (CPE). It also supports the reconfiguration of service (either due to customer demand or problem resolution) after the initial service installation. The aim is to correctly provide service configuration within the time frame required to meet ever-decreasing intervals.
Service Activation	This process is part of the fulfillment process on those services that require specific custom design or topology changes. These are typically complex business services that can include service to multiple users and locations, and possibly multiple service elements. As a simple example, imagine a typical business requirement for VPN services between two or more locations. Historically the design of these paths to enable the service was planned by engineering planning departments and the service assigned/enabled manually.
Service Problem Management	The Service Problem Management process is to respond immediately to customer-affecting service problems or failures in order to minimize their effects on customers, and to invoke the restoration of the service, or provide an alternate service as soon as possible. They encompass the reporting of problems, making a temporary fix or work-around, isolating the root cause and finally recovering the complete functionality of the service and providing information for future enhancements.
Service Quality Management	Service Quality Monitoring (SQM) and impact analysis applications are designed to allow operators to determine what levels of service they are delivering to their customers. Ideally these take a customer-centric view, i.e. the quality of service perceived by customers but may measure additional service metrics to allow the operator to be aware of approaching problems or degradations to service. Impact analysis applications extend this capability to predict the likely impact of service degradations or network problems on specific customers.
Service-Level Agreement Management	SLA Management uses the output of the SQM applications to provide a comprehensive view of the level of service provided to customers compared to pre-agreed, often contractually binding agreements. Typically SLA agreements will be negotiated between operator and customer to measure a variety of service-oriented issues and impacts. These may be either stated in terms of service characteristics or in terms of the business impacts on the customer.
Service Assurance	Allows continuous supervision of service-level agreements on service indicators, such as availability, throughput, call congestion, packet losses, CDR losses, and others. For violations, the billing module is informed to initiate discounts or reimbursements.
Service Operations	This term is often used instead of Service Assurance.
Resource Management & Operations (RM&O)	Resource Management & Operations functional processes maintain knowledge of resources (application, computing, and network infrastructures) and is responsible for managing all these resources (e.g., networks, IT systems, servers, routers, etc.) utilized to deliver and support services required by or proposed to customers.
Network and Systems Management	Very often used as a synonym for Resource Management & Operations (RM&O).

Continued

TABl E 3.6.4 Summary of Principal Processes Used in the Practice of eTOM (*Continued*)

Process or Function Name	Definition
Inventory Management	Allows maintaining, first of all, technical inventory data about equipment and circuits for the geographical reach of the service providers. Both CAD/CAM and GIS solutions may be implemented. Connections to the data warehouse are obvious. Sophisticated Inventory Systems have built-in autodiscovery functions, which updates the inventory with minimal manual effort.
Network Systems Administration	May be considered as part of the maintenance process, limited, however, to version control, backup, archiving, and distribution of software to equipment.
Capacity Management	Process of periodic surveillance of capacity in equipment and circuits. If capacity thresholds are exceeded, capacity extensions are initiated automatically.
Traffic Management	Process of observing typical traffic patterns by customers, customer groups, geographical areas, equipment, and facilities types. As a result, parameters and controls can be changed in equipment and facilities.
Workforce Management	Allows the central, policy-based, dispatch of workforce to monitor, test, maintain, inspect, and install equipment and facilities.
Resource Planning and Development	This process encompasses development and acceptance of strategy, description of standard network configurations for operational use, and definition of rules for network planning, installation, and maintenance. This process also deals with designing the network capability to meet a specified service need at the desired cost and for ensuring that the network can be properly installed, monitored, controlled, and billed. The process is also responsible for ensuring that enough network capacity will be available to meet the forecasted demand and support cases of unforecasted demand.
Network Planning and Development	Term often used instead of Resource Planning and Development.
Design and Planning	Allows, as a result of capacity bottlenecks, initiation of design processes that may include the deployment of new technology to equipment and facilities.
Resource Provisioning	The process encompasses the configuration of the network to ensure that network capacity is ready for provisioning of services. It carries out network provisioning as required to fulfill specific service requests and configuration changes to address network problems. The process must assign and administer identifiers for provisioned resources and make them available to other processes. Note that the routine provisioning of specific instances of a customer service—in particular, simple services such as plain old telephone service (POTS)—may not normally involve network provisioning, but may be handled directly by service provisioning from a preconfigured set.
Installation and Inspection	Allows, as part of the provisioning process, the physical deployment of equipment and facilities on the basis of provisioning and service order change requests of customers.
Testing	Process of testing equipment and facilities prior to deployment or as a part of the error repair process.
Resource Trouble Management	Resource Trouble Management processes are responsible for the management of troubles with allocated resources. The objectives of these processes are to report resource failures, isolate the root causes, and act to resolve them. Principal processes and functions are: • Survey & Analyze Resource Trouble • Localize Resource Trouble • Correct & Recover Resource Trouble • Track & Manage Resource Trouble • Report Resource Trouble
Reactive Fault Management	Process of determining, diagnosing, and resolving faults detected and reported by customers or by fault monitoring devices.

TABl E 3.6.4 Summary of Principal Processes Used in the Practice of eTOM (*Continued*)

Process or Function Name	Definition
Proactive Fault Management	In order to detect problems early, allows the continuous supervision of fault indicators, the identification of causes for chronic troubles, and the evaluation of vendor performance.
Preventive Fault Management	Allows evaluation of usage statistics, the causes of performance threshold violations, and the impact of additional payload on equipment and circuits.
Error Repair and Maintenance	Allows repair of chronic faults and deployment of preventive maintenance techniques to equipment and to facilities.
Resource Performance Management	Resource Performance Management processes encompass monitoring, analyzing, controlling, and reporting on the performance of resources. They work with basic information received from the Resource Data Collection and Processing processes. • Monitor Resource Performance • Analyze Resource Performance • Traffic status and performance analysis • Control Resource Performance • Report Resource Performance
Performance Monitoring	In order to further support preventive fault management, equipment and facilities (circuits) are monitored continuously. In addition, performance metrics are maintained in a repository, which can be part of the data warehouse. Allows repair of chronic faults and deployment of preventive maintenance techniques to equipment and to facilities.
Resource Data Collection & Processing	Resource Data Collection & Processing processes collect usage, network, and information technology events and performance information for distribution to other processes within the enterprise. Principal processes are: • Collect Resource Data • Process Resource Data • Report Resource Data • Audit Resource Usage Data
Call Data Collection	Collects call detail records (CDRs) from switches and transmits them to a billing database or mediation device. State-of-the-art solutions use complete automation.
Supplier/Partner Relationship Management (S/PRM)	Supplier/Partner Relationship Management functional processes support the core operational processes, both the customer instance processes of Fulfillment, Assurance, and Billing and the functional operations processes.
Billing	The billing process is an end-to-end process responsible for the production of accurate bills in time, processing, and collecting payments and providing prebilling and billing information to customers. In addition, it involves handling customer inquiries about bills, providing inquiry status and resolving billing problems to ensure customer satisfaction in a timely manner. This process also supports prepaid services. The billing process offers the following services: • Rapid creation of new service plans and rates • Flexible rating and discount options (pricing) • Real-time event processing • Convergent billing • Prepaid and post-paid billing • Customer self-management of services via the Web • Analysis of customers, rates, plans, usage, etc.
Rating and Discounting	Prices call data according to current plan; it does include threshold plans currently popular among wireless carriers. Also, discounts are considered with this function.

Continued

TABl E 3.6.4 Summary of Principal Processes Used in the Practice of eTOM (*Continued*)

Process or Function Name	Definition
Billing Management	In the Front its function it performs: • Setup and maintenance of the customer's billing profile (account, payment method, bill cycle, billing address, etc.). • One-and-done billing inquiries and dispute resolution (and possible escalation). In the back office, it provides aggregation of all rated billable events and charges for products and services delivered to the customer by the service provider and respective trading partners and the production of a timely and accurate invoice.
Bill Compilation, Processing, and Formatting	This process aggregates the rated service/call detail records and adds data for multiple services; also handles advanced charges and payments. Bill processing/formatting processes predefined numeric, text, and image content into print-ready and Web-ready streams that can be reproduced on a variety of media. For instance, telecommunications companies can process data from a billing system into standard industry print streams to produce paper bills. In addition, it aggregates the rated call detail records and adds data for multiple services, and handles advanced charges and payments.
Bill Presentment	Customizes bill formats on a customer or service provider basis, may consolidate multiple statements; delivers bills via mail, online, e-mail, tape, or Internet.
Data Analysis and Mining	Process of analyzing call data details collected from switches and transmits to other billing processes.
Mediation	An intermediate step for preprocessing and analyzing CDRs; fraudulent calls can be removed, data input from different switches in multiple formats can be converted into a format appropriate for bill processing. Also, pricing schemes can be inserted here, rather than by the call rating module. Call data can be selected and then transmitted to individual billing platforms, such as for voice, data, wireless, Internet, etc. Mediation is increasingly used for convergent and real-time billing.
Revenue Assurance	Factoring, finding of receivables, credit checks, remittance processing, and customer deposit management.
Collection and Credit Analysis	Collecting outstanding debt, usually with the help of third-party collectors.
Data Warehousing	Call detail records and additional data sources can be transmitted into warehouses. Data mining and other applications help to determine customers and end-product profitability.
Local Number Portability	Allows customers to retain their telephone numbers with multiple service providers. Also, access to value-added services can be retained.
Security Management	Process of identifying security risks in equipment and facilities, deploying security procedures and tools, creating and evaluating security logs, and protecting operations, business, and marketing support systems.

Considerations that should influence process design include:

- Exceptions should be handled excellently. In other words, process problems are identified in real time and actions to support the customer are taken in real time.
- Business rules should be easily configured and applied automatically.
- The ability to treat a process as an asset that can be assessed, replaced, or outsourced as appropriate to improve the operation of the business.

3.6.9.2 Service Provider Migration toward eBusiness

There are several alternative approaches to implementing eBusiness. Some companies are treating eBusiness (and eCommerce) as separate units. Some are overlaying eBusiness on traditional business operation. Other businesses are approaching eBusiness as a replacement of traditional business channels. The most successful eBusiness enterprises integrate eBusiness and traditional business channels where cost, quality, and profit can be best rationalized. This is much more than just throwing together a set of Web

pages to front an organization, although integrating storefront and Web operations is clearly a key part of the model for some businesses.

The integration of eBusiness and traditional business channels is the model that is most applicable to information and communications service providers. Undertaking such an integration is typically a substantial exercise.

The use of systematic business process frameworks (e.g., eTOM) also makes it easier to evaluate and improve the processes themselves. Employing business process modeling techniques contributes to the goals and profitability of service providers. Using consistent modeling techniques for business development and information systems development brings noticeable efficiency improvements and removes barriers within those enterprises and across cooperative, intercorporation projects.

Service providers that use systematic business process modeling to manage and improve their businesses have a much greater chance of migrating their existing organizational structure to encompass new challenges, the current of which is fully embracing the eBusiness paradigm.

3.6.9.3 An eBusiness Reference Model

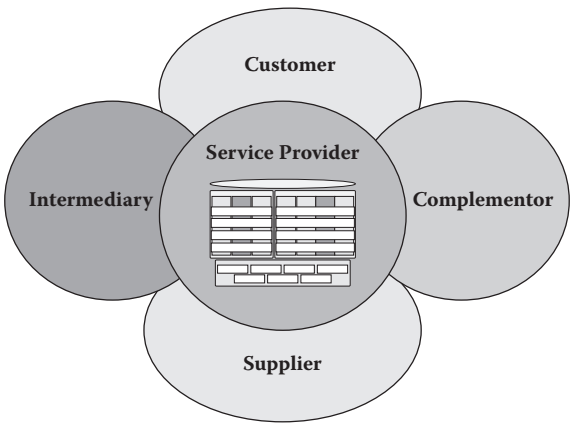
eBusinesses can be characterized as communities of complementary organizations linked together to create unique business entities that are easy to reconfigure in response to evolving customer needs. The central theme of eBusiness becomes the delivery of value by creating and utilizing end-to-end value streams that are based on an integrated and customer-centric technological foundation.

eBusiness involves increasingly complex networks of relationships to operate. Figure 3.6.2 depicts the sets of relationship groupings involved in a value network in the Interconnection Technology (ICT) industry.

The value network must operate with the efficiency of a self-contained enterprise, which requires managing the network on a process rather than an organizational basis. The model explicitly shows the use of the eTOM Business Process Framework by the service provider at its core. It is only shown here to simplify the figure, and its presence is not intended to imply that its use by the service provider is prescribed, just that the service provider would probably benefit from its use. Likewise, it is not intended to preclude the use of eTOM by the other entities shown within the value network. These entities may or may not make use of the eTOM Business Process Framework. The roles of the entities in the Value Network are described below.

The **customer** is responsible for ordering, using, and (usually) paying for service products. The customer may represent an end customer, where the product provided by the value network is consumed, or a wholesale customer that resells the product provided, generally with some added value.

The **service provider** presents an integrated view of service products to the customer. It is responsible for the contractual interface with the customer to sell products to the customer, provide the customer



FIGu RE 3.6.2 Relationships in eBusiness.

with contact and support, and bill the customer for the products supplied. The service provider can deliver some or all of a service product to the customer itself, or it might subcontract out provision of parts, or even all, of the product to other service providers while maintaining the customer-facing role of the one-stop shop. The service provider is responsible for acting on behalf of the value network it represents in relationships with intermediaries as well as with the customer.

The **complementary provider** extends the product provided by the service provider and offers additional capability that the service provider is not offering to the customer; that is, it complements the product being provided by the service provider and adds value to it, but is not essential for provision of the product itself. The complementary provider is in a partnership with the service provider and can enhance the service provider's product with its own products, thus making interactions with the service provider more attractive and convenient for the customer.

The **intermediary** supplies a service for a fee. For example, a localized selling function in a market where the service provider has a limited presence or understanding, is a typical service provided by an intermediary. The service provided could be an information service enabling customers to locate service providers most appropriate to their specific needs, or the provision of an environment in which providers can make their products known to customers in an electronic marketplace or trading exchange (infomediary).

The **supplier** interacts with the service provider in providing hardware, software, solutions, and services, which are assembled by the service provider in order to deliver its solutions or services to the customer. The service provider is bounded by its suppliers' ability to deliver.

3.6.10 IT Management Frameworks and the Information Technology Infrastructure Library (ITIL)

Information technology is not only a critical element of business for the telecommunications industry, but also for many other industries as well. The convergence of telecommunication to IT-originated technologies (characterized by the widespread use of the Ethernet and IP protocols, traditional computer architectures serving traffic-switching functions) has made issues faced by telecom technology management similar to that of enterprise networks.

The last two decades has brought the formation and evolution of numerous IT service management frameworks that all target the comprehensive and systematic categorization and description of concepts, methodologies, and best practices to be followed by modern and responsible IT management. From among initiatives like BS15000 (ISO/IEC 20000), AS8018, COBIT, or MOF (Microsoft Operations Framework), another related British government initiative for establishing an Information Technology Infrastructure Library (ITIL) has emerged as the most well-known and widely referenced standard.

ITIL as a trademark is owned by the UK Office of Government Commerce, but the IT Service Management Forum (itSMF), an international nonprofit organization for IT service management professionals, contributed to its content and had a key role in making it accepted worldwide. (Unfortunately there have recently been frictions between these two "parents.") IT management materials published originally by IBM during the early 1980s are also significantly incorporated in the library.

As of today, ITIL is the de facto reference point for IT service management and governance. This is especially true in Europe, while America seems to be less enthusiastic, and probably following less thoroughly documented but more pragmatic forms of IT management and operations guidance in practice.

ITIL has seen three versions so far (v1: 1987, v2: 2001, v3: 2007), each differing significantly both in content and in structure.

ITIL version 2 is structured in eight books called *sets*. Two of these sets, "Service Delivery" and "Service Support," form the core of ITIL by describing the key requirements for processes to be followed by IT departments. The rest of the documents link IT service management to related areas like technology ("ICT Infrastructure Management" and "Security Management"), business ("The Business Perspective" and "Application Management," "Software Asset Management") and provide guidance for implementing ITIL at an organization ("Planning to Implement Service Management").

As for the core sets of ITIL version 2, "Service Delivery" is focused on establishing quality IT services. The "Service Delivery" set is discussed in terms of the following processes:

- Service-Level Management
- Financial Management for IT Services
- Capacity Management
- IT Service Continuity Management
- Availability Management

The "Service Support" set discusses the requirements, rules, and best practices of ongoing operations and maintenance:

- Service Desk
- Incident Management
- Problem Management
- Configuration Management
- Change Management
- Release Management

ITIL version 3 takes a significantly different approach by structuring the above management topics into the following:

- Service Strategy—containing the recommendations of building a long-term IT service evolution
- Service Design—the procedures of preparing to implement a concrete new service
- Service Transition—the delivery procedures of new services, including their incorporation into the environment of existing production IT services
- Service Operation—discusses the efficient operation of existing services in terms of maximizing customer experience and lowering service operational costs
- Continual Service Improvement—the procedures to follow technological changes and the continual realignment of the IT services to changing business needs

Somewhat independent from the structure of the documents, it may be more important to recognize that there exists a simplified perception of the professional community, which mostly identifies ITIL with a few easy-to-remember focus areas like:

- A conscious approach of having definite goals with respect to knowing what to do about, taking control of, and continually optimizing IT services.
- Customer/user relationship is recognized as a key to successful and efficient services.
- Definition of Service-Level Agreements (SLAs, between customer and the provider) as part of the Service Delivery process that describes the ways to monitor the quality of services provided. SLAs also play a key role in the Services Support phase as a quantitative target for service quality (Service-Level Management).
- Operating a credible and authoritative Configuration Management Database with Change Management procedures in place for consistent updates.
- For running services, feedback is applied for continual service quality improvement.
- Operating a responsive and efficient service desk to respond issues (Change Requests and Incidents) and gather feedback from users and customers.

3.6.10.1 Judging the Value of ITIL

On the plus side, ITIL has helped in forming the basic requirements or IT service management, and making concepts like the Service Desk, Change and Configuration Management, Service-Level Assurance well known, understood, and accepted by everyone in this field. This is a very significant achievement, which should not be underestimated.

On the other hand, ITIL cannot really cope with the versatility of IT services and the consequential variety of viable approaches to IT service management. This is further aggravated by the rapid change in the number and types of IT services. While in the 1980s operation of an IT service was a serious and formal effort affordable only to huge and well-funded organizations like government agencies or telecommunication companies, recently we have seen innovative IT services started after one night of "hacking" on a PC in a college dormitory and evolving into billion-dollar businesses that practically drive the Internet of today. Clearly, IT service management principles that apply to mainframes are not applicable in today's typical environments.

ITIL does not seem to be successful in responding these challenges; indeed, as many critics say, it became rather indoctrinated (especially with release version 3) by merging into details and producing lengthy documents that are less and less applicable to modern, high-quality IT services.

3.6.10.2 Matching ITIL to eTOM

As seen, eTOM and ITIL are frameworks with many similarities, and the confluence of IT and telecommunication technologies obviously raises the question of a possible coexistence of the two frameworks at a telecommunication provider.

This topic has been researched and discussed actively, and many, often sharply different views have emerged. Still there is an overwhelming consensus that as IT (including networking) serves as the technological basis of modern telecommunication services, the most useful and practically implementable aspects of ITIL should be applied as the foundation of telecommunication operations specified by eTOM. In short, eTOM is a business framework, while ITIL is a technology management framework.

The TeleManagement Forum itself has also started a working group to investigate ITIL—eTOM compliance. This group created the document named *Application Note V.: An Interim View of an Interpreter's Guide for eTOM and ITIL Practitioners* (GB 921 V). This document provides and overview of mapping eTOM and ITIL processes in both directions (Figure 3.6.3).

The essence of these mappings is described in Figure 3.6.3. The position of the ITIL bubbles reflects the level of correlation: for example, the Service Configuration and Activation eTOM Level 2 process is strongly related to Change Management and Release Management, while it is partially related to Configuration Management.

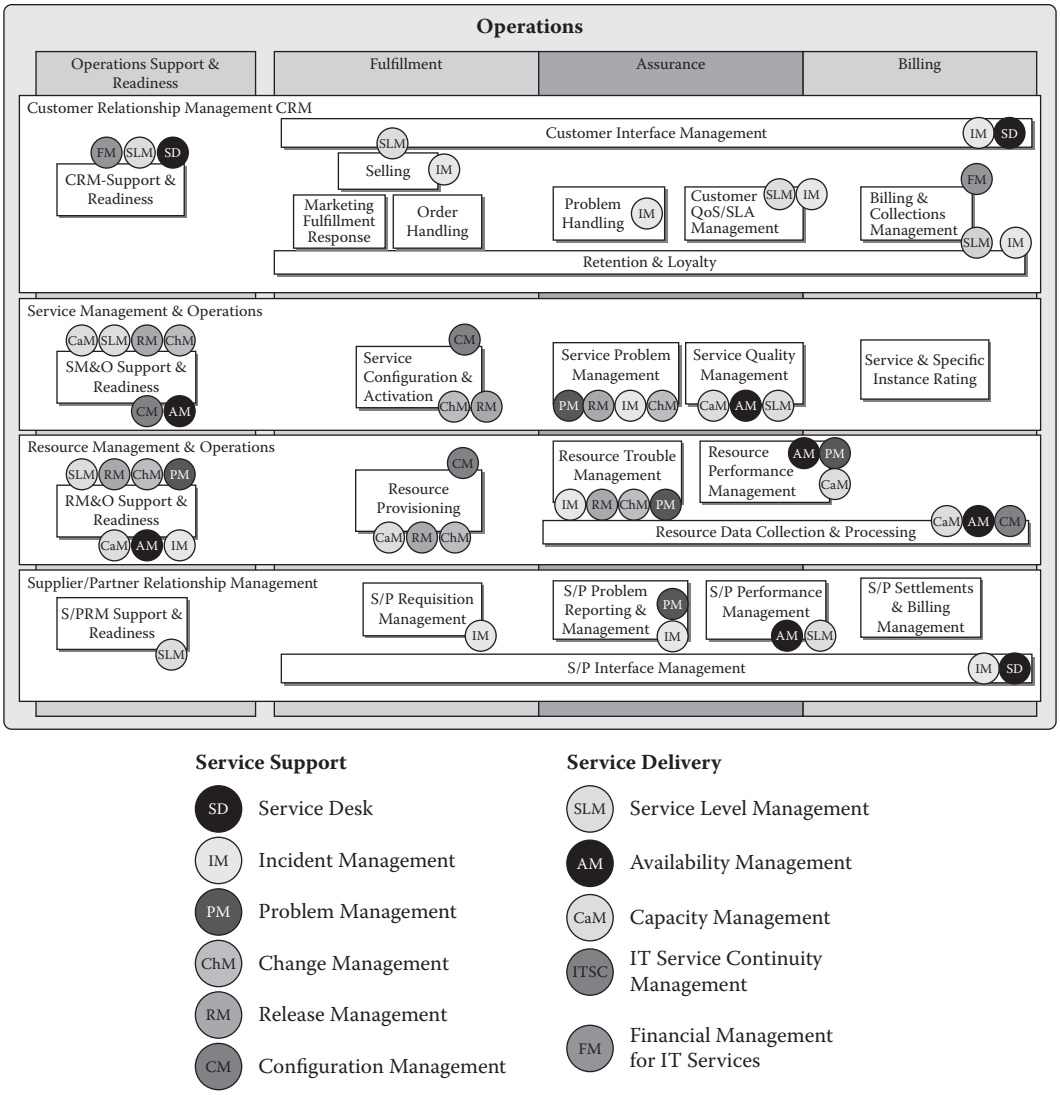
The document also investigates mapping from another aspect and at a more detailed level: how certain ITIL processes can be possibly implemented by eTOM level 3 process elements. Such an investigation is available for three of the most typical ITIL processes: Change Management (for a software release), Incident Management (infrastructure failure), and Customer Service Request for preapproved changes.

Although following two process structures simultaneously causes a number of potential conflicts (e.g., naming, documentation patterns, process definitions), as the above-referenced examples demonstrate, at least in some cases the two frameworks can coexist, delivering the added (albeit partially overlapping) advantages of both approaches.

3.6.11 Summary and Trends

The telecommunication industry has embraced the TMN model as a way to think logically about how the business of a service provider is managed. The model itself is simple, although its implementation is complex. The sheer number of standards now available that address the various interfaces among management systems sometimes makes it difficult to see and appreciate the big picture. These ITU standards are mainly concentrated in the Element Management and Network Management Layers. They have been developed from the bottom up, making it difficult to apply the standards as part of a business case. It is also difficult to have a customer-centric focus.

eTOM is a holistic, business-driven approach to implementing support processes for service providers. It provides investment direction as well as development specifications needed to produce management systems. It starts with a layered model, but goes much further, to address concrete business



FIGu RE 3.6.3 Mapping ITIL processes on eTOM Operations (from TMF Document GB 921 V).

problems in a pragmatic way. Specifically, it brings three key elements to help design and implement support processes:

A business process-driven approach: It starts from the premise that service providers and network operators need to automate their business processes, which means information needs to flow from end to end across many different systems. Only when a process is understood and the linkages are clear is it possible to apply standards in a way that delivers business value. Unless that is known, a great deal of money can be spent implementing standards that simply don't contribute to the overall business objective.

Technology-independent agreements: Business agreements about what information will flow between processes must be kept independent of the protocols used to implement those agreements. Technology will continue to change, becoming cheaper and easier to use, and delivering more power. eTOM applies the right technology for the job instead of forcing a single technology to serve every need. Further, the eTOM approach documents all agreements in technology-neutral form, so that the same agreement can be implemented in multiple technologies as they continue to evolve.

Products, not just paper: A main premise of eTOM is that paper standards are not sufficient to solve business problems. Products, not paper, are the end goal, provided documentation is produced to support the replications of industry agreements across multiple vendors' products.

This contribution has put emphasis on the enhanced telecom operations map (eTOM), which is a common model for telecommunications operations processes and a guide for reengineering such processes. In addition, ITIL—an enterprise IT standard guideline—is included to help service providers to streamline their IT operations. When business processes and IT processes are focusing on the business goal of the service provider, then the alignment criteria are fully met. In order to deepen this alignment, an eTOM/ITIL mapping was also addressed in the final part of this contribution.

Acronyms

ARPU	Average Revenue Per User
BSS	Business Support System
CDR	Call Detail Record
CLI	Command Line Interface
CPE	Customer Premises Equipment
CRM	Customer Relationship Management
CSP	Communications Service Provider
CSR	Customer Service Representative
DSL	Digital Subscriber Line
ETOM	Enhanced Telecommunications Operations Map
FAB	Fulfillment, Assurance, Billing
FCAPS	Fault, Configuration, Accounting, Performance, Security
GIS	Geographical Information System
ICT	Inter Connection Technology
ISV	Independent Software Vendor
ITIL	Information Technology Infrastructure Library
KQI	Key Quality Indicator
LEC	Local Exchange Carrier
LOB	Line of Business
M&A	Management and Administration, Mergers and Acquisitions
NGOSS	Next-Generation OSS
NOC	Network Operating Center
NEI	Network Element Interface
OSR	Operations Support and Readiness
OSS	Operations Support System
POTS	Plain Old Telephone Service
PSTN	Public Switched Telecommunication Network
QoS	Quality of Service
SFA	Sales Force Automation
SI	System Integrator
SIP	Strategy, Infrastructure, and Product Process
SLA	Service-Level Agreement
SLG	Service-Level Guarantee
SQM	Service Quality Monitoring
SP	Service Provider
TMF	TeleManagement Forum
VoIP	Voice over IP
VPN	Virtual Private Network

WFM	Workforce Management
WLAN	Wireless Local Area Network
WMS	Workforce Management System

References

DIOSS06: Kurth, M. *Dataquest Insight: OSS Market Overview and Strategic Scorecard for Vendors, 2006.* Gartner, 22 March 2007.

GB921-B: *Enhanced Telecom Operations Map® (eTOM)—The Business Process Framework. Addendum B: eTOM—B2B Integration: Using B2B inter-enterprise integration with the eTOM*. Telemanagement Forum, March 2004.

IIM-EDU: Yones, M. *Telecom Management—Preparing for Next Generation E-Telco, Operations Support Systems (OSS).* International Institute of Management Executive Tutorial Series, 2006.

IRR-WOS: *Wireless Operations Systems 2005–2010*. The Insight Resarch Report, Nov. 2005.

M.3050.0: ITU-T Series M: *TMN and Network Maintenance: International Transmissions Systems, Telephone Circuits, Telegraphy, Facsimile and Leased Circuits. Telecommunications management network. Enhanced Telecom Operations Map (eTOM)—Introduction.* ITU-T Recommandation M.3050.0, July 2004.

M.3050.1: ITU-T Series M: *TMN and Network Maintenance: International Transmissions Systems, Telephone Circuits, Telegraphy, Facsimile and Leased Circuits. Telecommunications management network. Enhanced Telecom Operations Map (eTOM)—The business process framework.* ITU-T Recommandation M.3050.1, June 2004.

M.3050.2: ITU-T Series M: *TMN and Network Maintenance: International Transmissions Systems, Telephone Circuits, Telegraphy, Facsimile and Leased Circuits. Telecommunications management network. Enhanced Telecom Operations Map (eTOM)—Process decompositions and descriptions.* ITU-T Recommandation M.3050.2, June 2004.

M.3050.2: ITU-T Series M: *TMN and Network Maintenance: International Transmissions Systems, Telephone Circuits, Telegraphy, Facsimile and Leased Circuits. Telecommunications management network. Enhanced Telecom Operations Map (eTOM)—B2B integration: Using B2B inter-enterprise integration with the eTOM.* ITU-T Recommandation M.3050.4, June 2004.

M.3050S2: ITU-T Series M: *TMN and Network Maintenance: International Transmissions Systems, Telephone Circuits, Telegraphy, Facsimile and Leased Circuits. Telecommunications management network. Enhanced Telecom Operations Map (eTOM)—Supplement 2: Public B2B Business Operations Map (BOM).* ITU-T Recommandation M3050 Supplement 2, June 2004

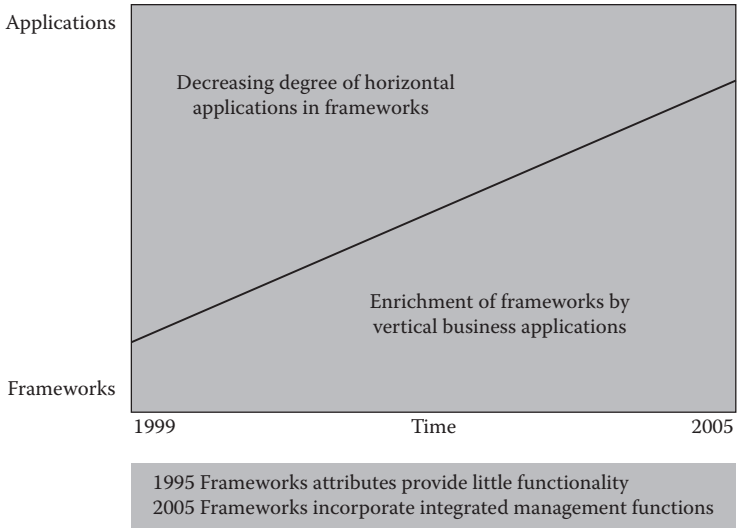
Obs0207: *Amdocs Snapsot*. OSS Observer, Feb. 2007.

TerpCRC: Terplan, K. *Support Systems for Telecommunication Providers.* The CRC Handbook of Modern Telecommunications, CRC Press, 2001.

3.7 Management Frameworks and Applications

Árpád Bakay, Tivadar Szemethy, and József Wiener

Management frameworks consist of an application platform and management applications. The application platform comprises management services, computing hardware, and operating systems. Management frameworks exhibit unique features that differentiate them from management systems, particularly proprietary management solutions. This section will introduce these differentiating features first. Management applications can be categorized as device dependent—provided by vendors of networking equipment, and device independent—usually provided by independent software vendors (ISVs). For the integration between application platforms and applications, application programming



FIGu RE 3.7.1 Evolving management frameworks.

interfaces (APIs) are provided by the framework suppliers, with the assumption that these APIs will be supported and used by application providers.

Typical framework solutions from suppliers such as Telcordia, Amdocs, IBM, Hewlett-Packard, and Oracle are illustrated. In addition to portfolios from these companies, operations support system/business support system (OSS/BSS) players will be categorized according to whether, from the perspective of mergers and acquisitions, they are traditional suppliers or potential or new entrants. Finally, near-term and long-term consolidation trends will be discussed.

3.7.1 Evolving Management Frameworks

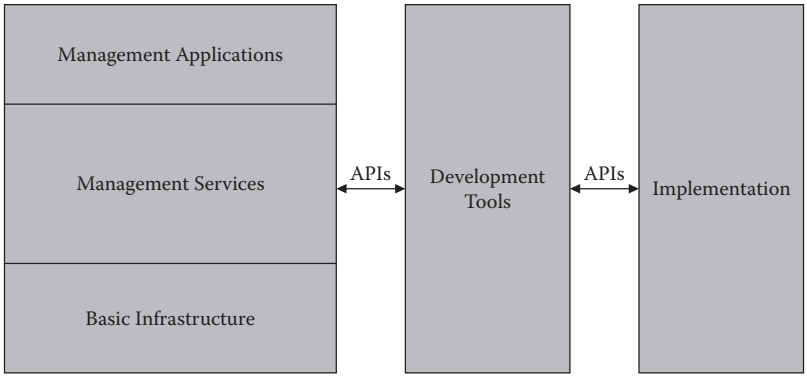
Enrichment of application platforms through generic and specific management applications makes them more powerful. Figure 3.7.1 illustrates this evolutionary process. When application platforms are introduced to the market, they usually provide support for a few management applications in addition to the core services that are part of the framework. If the framework is well accepted by users, the number of vertical management applications grows. Step by step, they are partially or fully integrated into the framework.

3.7.2 Features and Attributes of Management Frameworks

There is no actual scientific definition of a management framework. In order to determine whether certain products qualify as a framework, an elaborated list of attributes must be addressed first. When products are able to support basic framework attributes, they are said to meet the qualifications of a framework. Advanced attributes may then serve as differentiators between frameworks. Figure 3.7.2 shows the basic architecture of management frameworks, which consists of a runtime environment, development runtime tools, and APIs. The runtime environment is subdivided into management applications, management services, and the basic infrastructure (MORR00). Management services can be further subdivided into basic and advanced services. APIs interconnect the runtime environment, development tools, and the implementation pieces with each other.

3.7.2.1 Basic Infrastructure

The basic infrastructure concentrates on the hardware and software features of the management frameworks. The most important attributes are listed below.



FIGu RE 3.7.2 Architecture of management frameworks.

3.7.2.1.1 Hardware Platform

The hardware platform is considered a commodity. Both operating systems and management applications are expected to run on a wide variety of processors.

3.7.2.1.2 Operating Systems

There should be a healthy combination of proprietary and open-source software for the operating systems supporting service providers' principal business processes.

3.7.2.1.3 Directory Services

Management frameworks deal with a great variety of entities and a large number of resources. Allocation of human-readable names to each managed resource (object) is the goal of directory services. Directory services are based on commonly agreed upon standards that model the naming paradigm, provide naming notations, allocate identifiable names to managed resources, translate names into physical addresses of resources, and ultimately provide location transparency for the resource in the system. All of these considerations are valid for network and systems management frameworks. Framework management services and management applications use naming information from directory services to perform their functions in relationship to managed resources and other management frameworks. The principal directory service requirements are as follows (MORR00):

- Global information directory service and universal access to directory information
- Separation between the names of managed objects and the underlying physical networks
- Translation capabilities between various directory systems
- Translation between logical addresses and network addresses or routing addresses
- Storage of directory information and access to directory information, including metadata
- APIs in order to easily incorporate directory services into applications

In terms of directory service capabilities, the following components are examples (MORR00).

- *Directory service users:* people, management applications, electronic messaging, routers/servers, other management framework services
- *Resources requiring naming:* people, organizations, computers, processes, files, mail boxes, network devices, printers, object class abstractions, object instances, management applications, management services, management agents
- *Directory system types:* centralized, distributed, standard, proprietary, interpersonal communications directory (human), intersystem communications directory (computers and software systems)
- *Directory service generic operations:* query (read, list, search, compare), modification (add/remove entry, modify entry, modify naming space, quit), binding/unbinding (security authentication)

3.7.2.1.4 Time Services

In distributed systems and network environments, where processes and applications are running on different machines, time differences may occur between systems. Such time differences become critical when correct time stamps determine sequencing of events, job scheduling, measurement timing, and reporting intervals. Consistent use of time services is imperative when dealing with global networks, which span multiple time zones. Time service general requirements can be summarized as follows (MORR00):

- Use of an absolute, universal, coordinated time reference source
- Consistent synchronization services across hardware and software components
- Translation of universal time to local time for networks spanning multiple time zones
- Automatic resynchronization of manager and agent platforms after service interruptions
- Ability to operate in a heterogeneous computer and network environment
- Ability to keep the service running in case of major network instabilities
- Ability to provide both clock corrections and time source synchronization

In most cases, the Internet Network Time Protocol (NTP) is used. Its primary reference time source is the absolute Universal Time Clock (UTC) or sources directly derived from the UTC.

3.7.2.1.5 Software Distribution

In complex environments, management systems are usually distributed. These systems consist of servers, clients, and the communication paths between them. In order to ensure that they remain in synch, software versions or releases running on servers and clients must be compatible with each other. Manual software distribution is too slow and not sufficiently reliable. Electronic software distribution offers two popular alternatives: push and pull. Distributing software by "push" allows easier scheduling and better automation and does not require the physical presence of administrators. However, receiver servers and clients should be prepared for the distribution. Distributing by "pull" offers better control by administrators and changes during distribution at the price of low automation. Scheduling is flexible and depends solely on human decisions. At present, pull is the usual choice in Web environments.

3.7.2.1.6 Security Services

Open distributed network environments consist of an increasing number of interconnected computing resources, networks, and users. Networks are no longer closed networks but mixtures of private and public networks. These networks include heterogeneous components, which has a bearing on security services as well. The security of a network depends on the security of adjacent networks and other trusted partners. Frequent changes, such as adding new resources and new users, lead to additional concerns regarding security. Security can be seen as the security management functional feature built into certain management applications, namely security management applications. Since management frameworks control resources, security becomes an issue, as other framework services must operate securely if the system as a whole is to be secure. Security is often embedded in framework services such as communications, database management, and object manipulation services, which perform management operations.

Basic security requirements include (MORR00):

- Support for basic security features such as authentication, access control, and data integrity
- Ability to protect the system against potential intrusions
- Security features in the entire software development life cycle
- Distinctions in user security access profiles according to their role in the network
- Ability to group resources and users and apply common security policies to them
- Need to test security features and services against possible violations
- Protection of passwords and encryption keys by storing them in protected, encrypted files
- Mechanism to provide automatic clearing of disabled user accounts, user IDs, and passwords
- Capacity to communicate security data in a secured fashion

In typical management environments, management frameworks are protected by using a combination of the following security services.

- *Identification of users:* Unique representation of a user, computer, application, or remote system designed to provide accountability and record the actions of the identified entity.
- *Access control:* Allows the requesting party to actually access the system and networking resources if the party is authorized to do so. It is supported by log-in functions in which passwords are built in.
- *Authentication:* Verification of the entity prior to accessing the system and networking resources. In this case, the entity should prove its identity by using various techniques such as personal attributes, digital signatures, and others.
- *Data privacy:* An encryption mechanism using trusted third-party secret keys. Through a combination of private and public keys, the encrypted information can be verified for integrity and accessed for processing.
- *Data integrity:* A security feature that allows verification of cryptographic data checksums. Correctness of this verification represents proof that the data was not tampered with or corrupted through network transmission.
- *Security auditing:* Allows the generation of audit logs. These logs should be encrypted and protected against unauthorized access attempts.

3.7.2.2 Management Services

Management services address more specific items associated with management applications. The most important features are listed below.

3.7.2.2.1 Communication Services

3.7.2.2.1.1 Network Architectures There are significant differences in the various types of targeted networks to be managed. Many products are expected to manage legacy networks and more IP-based open networks at the same time.

3.7.2.2.1.2 Network Management Protocols Products are expected to provide at least Simple Network Management Protocol (SNMP) support, and it is an additional advantage when they can do more. SNMP support may include the capability of working with proxy agents capable of converting non-SNMP into SNMP. Protocols to be supported include CMIP, CMOT, LNMP, NMVT, ReMONitoring (RMON), SNMPv1, SNMPv2, and eventually Desktop Management Interface (DMI) to manage desktops.

The management platform provides SNMP support in several ways. First and foremost is the ability to poll SNMP devices and receive SNMP traps, as described previously in Section 3.3. However, in order to configure polls on the management information (MIB) variables of various devices, one must first know what those variables are. Management platforms provide MIB "browsers" for this purpose. An MIB browser queries user-selected SNMP network devices and displays their MIB values. In addition, most platforms can display line or bar graphs of these MIB values, provided they are in numeric form (e.g., counters).

MIB browsers display raw and often cryptic, low-level device information. For this reason, platforms also provide MIB application builders that allow users to quickly create applications for displaying information on MIB objects in a more meaningful way. MIB applications may include graphing real-time information on selected network nodes. However, even MIB application builders are limited in supporting the high-level analyses more openly provided by third-party applications.

MIB compilers allow users to bring in third-party, device-specific MIBs (also called *private* or *extended* MIBs) and register them with the management platform. While most platforms ship with a number of third-party MIBs, they do not include all possible MIBs from all vendors. An MIB compiler is necessary for adding support for third parties whose MIBs are not shipped as part of the standard platform.

Some MIB compilers are more robust than others. Some will fail or abort processing if there is an error in the MIB being compiled. Unfortunately, errors in third-party MIBs are not rare. Therefore, it is desirable to have an MIB compiler that can flag errors and recover rather than stop immediately.

3.7.2.2.2 Core Management Services

The management framework is expected to offer core services and interfaces to other applications. The basic management applications to be provided are discovery/mapping, alarm management, and platform protection.

3.7.2.2.2.1 Discovery and Mapping Device discovery/network mapping discovery refers to the network management system's ability to automatically learn the identities and types of devices currently active on the network. At minimum, a management platform should be capable of discovering active IP devices by retrieving data from a router's IP tables and Address Resolution Protocol (ARP) tables.

However, even this capability does not guarantee that all IP devices on a given network will be detected. For example, relying solely on routing tables is inadequate in purely bridged networks where there are no routers. Thus, a more comprehensive discovery facility should also include other mechanisms such as broadcast messages (ping and others) that can reach out to any IP device and retrieve its address and other identifying information.

On the other hand, discovery mechanisms that rely completely on broadcasting (e.g., ping) will incur a tremendous amount of overhead related to the process of locating devices on the network. Ideally, a management platform should support a combination of ARP data retrieval and broadcasting.

A complete network discovery facility should be capable of detecting legacy system nodes as well. At present, most platforms rely on third-party applications or traffic monitoring applications to supply discovery data on non-TCP/IP devices.

Another desirable feature is the ability to run automatic or scheduled *dynamic discovery* processes after the initial discovery, with the goal of discerning any changes made to the network after the initial discovery has taken place. In large networks especially, overhead and consumed bandwidth for running a dynamic discovery process continually in background mode may be too great; therefore, the ability to schedule discovery at off-peak hours is important.

It is also important for the user to have the ability to set limits on the initial network discovery. Many corporate networks are now linked to the Internet, and without predefined limits a discovery application may cross corporate boundaries and begin discovering everything on the global Internet. Some management platforms allow users to run discovery on a segment-by-segment basis. This can prevent the discovery process from getting out of hand too fast.

Many management platforms are capable of automatically producing a topological map from the data collected during device discovery. However, these automatically generated maps rarely result in a useful graphical representation. When there are hundreds of devices, the resulting map can be very cluttered in appearance and be of little use.

Even when the discovery process operates on a limited or segment-by-segment basis, there will eventually come a time when the operator must edit the automatically generated network map to create a visual picture that is easier for human beings to relate to. Therefore, the ability to group objects on the map, and move them around in groups or perform other types of collective actions, can be a real time-saving feature.

3.7.2.2.2.2 Alarm Capabilities Management platforms act as a clearinghouse for critical status messages obtained from various devices and applications across the network. Messages arrive in the form of SNMP traps, alerts, or event reports when polling results indicate that thresholds have been exceeded.

The management platform supports setting of thresholds on any SNMP MIB variable. Typically, management platforms poll for device status by sending SNMP requests to devices with SNMP agents or Internet Control Message Protocol (ICMP) echo requests (pings) to any TCP/IP device.

The process of setting thresholds may be supported by third-party applications or by the management platform. Some platforms allow operators to configure polls on classes of devices; most require operators to configure a poll for each device individually.

A majority of platforms support some degree of alarm filtering. Rudimentary filtering allows operators to assign classifications to individual alarms, such as informational, warning, or critical, triggered

when thresholds are exceeded. Once classifications are assigned, the user can specify that only critical alarms are to be displayed on the screen, while, for example, all other alarms are to be logged.

More sophisticated alarm facilities support conditional alarms. An example of a conditional threshold is "errors on incoming packets from device B > 800 for more than 5 times in 25 minutes." Conditional alarms can account for periodic spikes in traffic or daily busy periods, for example. A further logical step is the capability to correlate alarms and display the most important, "root cause" alarms only. Finally, the platform should support the ability to automatically trigger scripts when specific alarms are received.

3.7.2.2.3 User Interface Services

The basic job of the graphical user interface (GUI) is to provide color-coded displays of management information, multiple windows into different core or management applications, and an iconic or menu-driven user interface. By offering a standardized interface between the user and the underlying tools, the GUI simplifies what a user needs to learn and provides a standard tool for application developers.

Most management operations are available from a menu bar; others are available from context menus. Point-and-click operations are standard features, as is context-sensitive help. Most platforms allow a certain degree of customization of maps and icons.

While most platform GUIs are the same, there can be a few subtle differences. Some GUIs have larger icons than others. While this makes it easier to read information on the icon and distinguish status changes more quickly, a screen can quickly become cluttered with just a few large icons. Icon size is strictly a matter of user preference. Web 2.0 applications may enrich GUI in the future.

3.7.2.2.4 Database Services

The database is the focal point for key data created and used by management applications, including MIB data, inventories, trouble tickets, configuration files, and performance data. Most platforms maintain event logs in flat-file ASCII format for performance reasons. However, this format limits the network manager's ability to search for information and manipulate the data. Therefore, links to relational database management systems (RDBMSs) are now important aspects of the framework architecture.

A RDBMS is essential for manipulating raw data and turning it into useful information. If the RDBMS schema of the application data is also published, users can obtain information from a RDBMS by writing requests, or queries, in Structured Query Language (SQL), a universally standard language for relational database communication. This is often the last resort and/or most powerful way of integrating applications.

While most management platforms also supply report configuration and generation facilities, these tools are generally not top-notch. However, high-quality third-party reporting applications can be used to extract data from a RDBMS.

3.7.2.2.5 Object Manipulation Services

Object-oriented and object-based technologies are helpful in relation to user interfaces, protocols, and databases. The use of object request brokers (ORBs) and Common Object Request Broker Architecture (CORBA) provides the glue needed to accomplish interoperability among heterogeneous systems. These services provide support for information exchange between objects as abstractions of physical and logical resources ranging from network devices computing systems resources to applications and management services. Included are operations on MIBs, object support services providing location transparency for objects exchanging requests and responses, persistent storage for MIBs, and support for object-oriented applications development.

In spite of its feature fullness, good performance, and architectural purity, CORBA is a legacy technology these days. In recent management software Corba APIs are replaced by Java or Web Services APIs.

3.7.2.2.6 Network Modeling Services

Network modeling is an artificial intelligence capability that can assist in automated fault isolation and diagnosis as well as performance and configuration management. Modeling allows a management system to infer the status of one object from the status of other objects.

Network modeling is facilitated by object-oriented programming techniques and languages such as C++. The goal of modeling is to simplify the representation of complex networks, creating a layer of abstraction that shields management applications from underlying details.

The building block of this technology is the model, which describes a network element such as a router. A model consists of data (attributes) describing the element as well as its relationships with other elements. Abstract elements, such as organizations and protocols, can also be modeled, as can nonintelligent devices such as cables. A model may use information from other models to determine its own state; modeling can reduce the complexity of management data and highlight the most important information. In this way, fault isolation and diagnosis can be automated. In addition, models can be used to depict traffic patterns, trends, topologies, or distributions to assist in performance and configuration management.

3.7.2.3 APIs and Development Toolkits

API and developer toolkit platform vendors encourage third-party applications by providing published APIs, toolkits that include libraries of software routines, and documentation to assist applications developers. Another aspect of this effort is the *partner program*—the marketing angle of encouraging third-party applications development.

An API shields applications developers from the details of the management platform's underlying data implementation and functional architecture. Management platform vendors generally include in their developer's kits several coded examples of how APIs can be used, as well as the APIs themselves.

Most modern platforms offer an API conforming to the Soap/Web Services (WS) technology. This offers some kind of neutrality regarding integration platform and implementation language. However, most WS APIs of today are strikingly slow, which often makes developers request and use further mode low-level APIs as well.

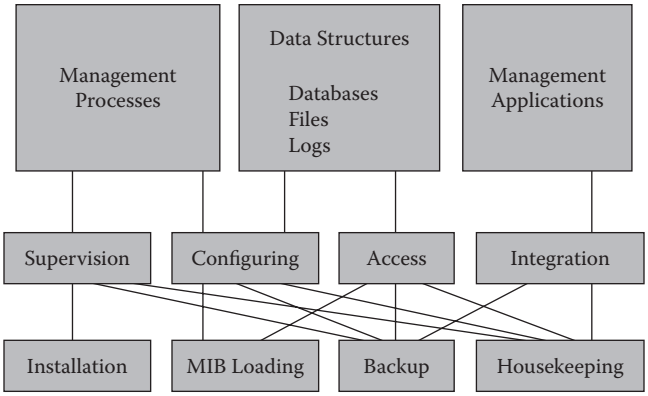
In most cases, if an application takes advantage of platform APIs, it must be recompiled with the platform code, resulting in a tightly integrated end product. Many independent software vendors (ISVs) and other third-party developers lack the resources necessary to pursue this level of integration. Or, perhaps a more accurate way of stating this is that ISVs are not convinced that expending the extra effort to fully integrate their applications with all leading management platforms will result in a proportionally larger revenue stream. ISVs and other third-party developers face a choice: tightly integrate their products with one management platform vendor or loosely integrate them with all leading platform providers. Most third parties have chosen the latter route, as they are unwilling to turn off prospective customers who may have chosen a different platform vendor as their strategic management provider.

As a result, at least 80% of the third-party applications available today are only loosely integrated with the underlying management platform—at the menu bar—and completely ignore APIs and other environment libraries. This is expected to change as the market matures and as platform vendors begin to offer high-level APIs that make porting applications from one management platform to another an almost trivial exercise.

In summary, published APIs and libraries make it possible for ISVs and other third parties to write applications that take advantage of other basic services provided by the management platform. To date, few third parties have taken full advantage of platform APIs, although this is expected to change over the next several years.

3.7.2.4 Management Operations Support Services

All management frameworks consist of framework services and management applications. Services are implemented as a set of related processes, databases, and file sets. The basic thrust of management implies



FIGu RE 3.7.3 Overview of management operations support services.

collection and processing of management-related information. All framework processes, including those that are part of the development environment, are coordinated through additional framework components commonly called *management operations support services*. These services are also responsible for application integration with framework services and for multiple national language systems support.

Management frameworks are basically sets of interconnected software programs that run on one or more computing platforms. Management operations support services provide supervision, coordination, maintenance, and management of the processes, applications, and databases that are part of the management framework. The requirements of management operations support services are as follows (MORR00):

- Facilitating interactions between framework services
- Allowing overall coordination and supervision of background processes
- Supporting integration between management services
- Allowing configuration and customization of framework services and associated processes
- Supporting registration of management applications that run on management platforms
- Providing easy integration of management applications with framework services
- Supporting multiple national language systems
- Facilitating incorporation of management information models into frameworks
- Supporting installation of framework services and management applications
- Supporting MIB loading, backup, and cleanup facilities
- Supporting distribution of management framework services and associated databases

This list of requirements indicates that management operations support services play a critical role in monitoring, administration, and management of the management framework itself.

Management operations support services are characterized by a layered architecture. The upper layer consists of management processes, data structures, and management applications (MORR00). The middle layer includes important support functions such as supervision and synchronization of management processes, configuring processes and databases, access to databases and files, and integration between framework services and management applications. The lower layer consists of tools, supporting installation, MIB loading, backups, and other general housekeeping functions. Figure 3.7.3 shows these layers.

3.7.3 Management Applications

Application platforms are powerless without management applications. These applications, which are provided by equipment vendors or by ISVs, serve various purposes.

3.7.3.1 Device-Dependent Applications

Equipment vendors develop and deploy management applications in order to promote sales of their equipment. It is no longer possible to sell networking gear without element management systems—in other words, without management applications. These applications are offered and sold at reasonable prices. Equipment vendors do not accrue much revenue with these element management systems because they must support multiple frameworks. Web-based management will bring relief by offering a unified interface to management applications. This interface is expected to be supported by all framework vendors.

3.7.3.2 Device-Independent Applications

Device-independent applications are designed, developed, and deployed to work in different environments. Typically, they address the following management areas:

- Trouble ticketing to better support fault management
- Performance analysis and reporting to support service assurance
- Security management to provide a protection umbrella
- Modeling to improve resource planning and utilization
- Dashboards and balanced scorecards to help managers judge performance
- KVM to improve local and remote operations
- Location services using mashups on geographical maps
- Processing and rating CDRs (Call Detail Records)
- Infrastructure access points for collecting data to support lawful intercepts

Also, these management applications can be integrated into frameworks using Web-based technology. The big benefit is that management applications can be loosely coupled with the framework and with each other.

3.7.4 Vendor Profiles

A few traditional big players are portrayed in greater detail. Later, the market research section tries to give a more complete picture.

3.7.4.1 Telcordia

Telcordia (original name: Bellcore) was established in 1984 as the joint R&D facility of the seven then-new Bell companies. It is a good example of an OSS vendor company that initially focused on serving U.S. wireline operators and then gradually switched its portfolio from legacy technologies to new data- and IP-centric services, including wireless. The Telcordia product portfolio is built around the four primary OSS functions: planning/engineering, fulfillment, service delivery, and service management.

Planning and engineering is supported by the Network Engineer suite. It is basically a GIS-backed inventory of the physical network (i.e., inside and outside cable assets and related hardware components). The applications offered as part of this suite provide access to data from several different perspectives:

- *Design Assistant:* Helps network designers create plans that comply with the company's build-out policies. Also, this tool provides functions that improve design efficiency and offers automatic checks and calculations.
- *Field Assistant:* Improves the efficiency of field workers by providing access to network drawings overlaid by GIS data. In addition, technicians are able to enter updates and changes to the network database, thus improving its accuracy.
- *Schematic Assistant:* Used to visualize network objects from different perspectives.

- *Analyst:* Provides query, analysis, and reporting functions on network data used for executive reports and for sales and marketing operations. Data may support high-level decisions on network development in certain geographical areas.
- *Integration Assistant:* Used to make the network engineering database available for external, higher-level provisioning functions or other inventories.

The fulfillment function is centered on the Granite technical inventory product, which offers a comprehensive view of central physical assets and the services implemented over them. This is also the basis for the automated design and implementation of new services (see below). As might be expected, Granite also offers data partitioning for access control and a federated mode of operation wherein authoritative inventory data is not stored locally, but is rather imported on the fly from other systems. Furthermore, optional Granite inventory modules provide:

- Network discovery and reconciliation with inventory
- Automation for repetitive tasks such as adding new equipment, configuring new services, and so forth
- Activation, that is, comprehensive implementation of new services that may include configuration of both the transport and circuit layers
- AutoRouter, which locates and scores alternative data routes across the network
- Number pool and IP address management (Range Manager and IPAM)
- Simple work order and task management functions (Works)
- View and query functions for analyzing data (e.g., Web Toolkit and View)
- Web-service wrapper product (Assign and Design WS) that provides access to inventory functions for service-oriented architecture (SOA)-style integrations and clients (e.g., Business Process Execution Language [BPEL] business processes)
- Exception manager that helps resolve failed automatic service implementations

Other core products related to fulfillment are Expediter, an order manager that includes an order entry module, and product catalog, which stores blueprints of marketed products and services. The Customer Number Manager is used for assigning new addresses or numbers during the processing of orders for new services.

The service delivery function is built around the Telcordia Converged Application Server, an IMS-style service delivery platform that can host both wireline and wireless services, meaning that services can typically be managed and customized by users themselves.

One readily available application is Real-Time Charging, a rating solution designed for sophisticated packages and products that also includes prepaid services and fixed-mobile-cable converged packages. Real-Time Charging is a key enabler for adding new products, such as location-based services or music and video downloads, to an operator's portfolio.

Real-Time Policy can be used for implementing customer-centric controls such as charge control, content filtering, and parental or corporate control. Usage-based service limitation is another possible application aimed at protecting provider revenue by disabling fraud or excessive use.

Telcordia is also offering the above functionality as a hosted application service for Managed Virtual Network Operators (MVNOs).

Solutions for the service management function focus on maximizing service levels through proactive monitoring and immediate action.

The Service Director product helps operators define and measure customer service-level agreements (SLAs), and by monitoring SLAs it can generate trend-based preemptive or proactive alarms and thus trigger timely responses. The product can also be used for SLA reporting for customers and service quality management (SQM) reporting for internal organizations. Analyses and reports can correlate fault, performance, and customer-originated trouble-ticket data.

The Device Director measures and analyzes the performance issues and problems related to certain device types (typically Customer Premises Equipment [CPE] devices). Analyses are based on real transactions rather than probe-simulated tests.

Data collected and processed in this way helps providers select and advertise devices that work best for certain uses. If an issue is detected, corrective actions (i.e., firmware updates) may be arranged, or the issues that cause the majority of CPE-related service problems can simply be eliminated.

The Telcordia products listed above are primarily aimed at new, next-generation services. Telcordia also offers a number of minor OSS systems as part of its core OSS suite. The tools are used for activating, testing, supervising, and servicing different technologies including telephony, xDSL, Sonet/SDH, DWDM, DLC, and Ethernet. These applications were originally developed for legacy technologies but were gradually extended to support the technological basis of new services as well.

3.7.4.2 Amdocs

Amdocs, a Chesterfield, Missouri–based company with approximately 16,000 employees, is another emblematic player in the telecommunications management market. Originally the vendor offered its own billing solution. Later on Amdocs extended its professional services and became the market leader at the largest Tier-1 service providers. Through the acquisition of significant companies such as Clarify, a customer relationship management (CRM) firm (in 2001), and more recently Cramer, an OSS inventory vendor, Amdocs is now able to offer a compelling wide range of OSS/BSS solutions to its customers. Amdocs products now fit into a portfolio labeled *integrated customer management* (ICM) by the company.

3.7.4.2.1 Revenue Management

The Amdocs "flagship" offering, the Convergent Billing Solution, comprises a set of modular products that include the following.

- Online charging contains several vital components of a billing solution:
  - The product catalog is a database that reflects the products and services offered by the provider and keeps track of dynamic changes in those offerings.
  - Billing mediation refers to the process of receiving, formatting, and validating event data from various sources that require event-based charges (e.g., long-distance calls or video on demand).
  - Events are analyzed by the rating module, whose operation may be influenced by customer- or service-specific attributes that justify discount or price variations. The rating process may take place either online or offline. The advantage of online rating is immediate feedback (necessary for prepaid systems), while offline rating may lead to implementation of more sophisticated rating schemes.
  - The charging process is completed by the invoicing module, and the result is a consolidated bill for all services used by any customer. Invoicing may provide additional adjustment logic to total and subtotal fee calculation processes.
- Recurring subscription-based charges (e.g., telephony subscriptions, IPTV subscriptions) are handled by a separate module. The charging process here is similarly complex and includes multiple services, management of bundles, promotional campaigns, and production of clear and detailed online or offline bills.
- The settlement modules enable providers to accomplish clearing with provider partners and other business-to-business (B2B) scenarios. There is a separate product specially designed to deal with the complexity of roaming settlements.

3.7.4.2.2 Customer Management

Customer management covers sales, customer relationships, help-desk functions, and so forth. In short, it refers to supporting the full "target-sell-order-bill-support-grow" life cycle of a service.

- Real-Time Decisioning is a sophisticated data warehousing solution used for improving the hit rates of targeted sales actions. The solution enables near-real-time browsing and analysis of customer-specific historical data from databases such as the billing record store or CRM contact history.
- The Contact Center product suite is designed to streamline multichannel (phone, e-mail, chat, and Web) CRM interactions with customers. This solution includes script engines that guide agents through various interactions (selling, up-selling, troubleshooting, complaints), tight integration to billing and other databases, and measurement of agent efficiency. A principal component of the suite is the Smart Agent Desktop solution, the application used by agents while interacting with the customer.
- The Customer Service and Support product is the tool set used to build customer-centric problem resolution services, whether in the case of service centers or field services. This process is facilitated by applications that allow operators to access their contracts, SLA commitments and records, and spare parts inventories.
- Self-service is clearly a hot topic in customer management, given that it is a way to reduce customer center operational costs and that, in many cases, customers prefer this alternative. Modules in the Self-Service solution include e-billing presentment and payments, self-service support, self-service reporting, and self-service service order. Self-service is not only about residential lines: Some features (e.g., allocation of charges to various customer cost centers) are specifically designed for enterprise customers. Targeted toward mobile operators, Personal Device Portal is a Web-based simulation engine allowing customers to try out advanced devices and services, with options for immediate purchase.
- Amdocs Sales is the key system through which the sale is actually made. Amdocs offers targeted solutions that provide salespeople with all of the comprehensive information they need to sell efficiently, including history of past interactions, pipeline management, forecasting, and so forth.
- Amdocs Ordering helps improve and speed up order processes by offering flow-through provisioning and billing activation, management of orders in jeopardy because of technical issues, and efficient management of service-specific procedures and knowledge bases.

Obviously, implementing and using these products efficiently is a challenge; however, this challenge is alleviated by the Amdocs professional services available for all of the modules mentioned.

3.7.4.2.3 OSS Solutions

Amdocs OSS solutions are centered on the Cramer portfolio, through its Service and Resource Management suite (Resource Manager and Service Manager), which is expected to provide a complete and accurate view of services through regular synchronization and automatic discovery. On the basis of this foundation, Amdocs offers:

- A complete Service and Resource Planner tool set that allows creation and simulation of planned services and forecasts of future resource demands
- With respect to fulfillment and provisioning, a full order-to-activation solution including the Service Catalog, order decomposition (Order Manager), and order activation (Activation Manager) for a large number of network element and service platform types
- In terms of assurance and support, a Customer Service and Support solution (see above) attached to the Amdocs SLA Manager, offering real-time views of SLA targets and results

3.7.4.2.4 Other Solutions from Amdocs

For the tasks of mobile and wireline service delivery and control, Amdocs also offers a service platform solution that can be used to quickly create and provide intelligent Session Initiation Protocol (SIP), IN, or Value Added Services (VAS) services for both prepaid and postpaid customers.

As an extension of the traditional Amdocs OSS/BSS portfolio, the company now also offers a range of solutions for digital/mobile commerce and advertising, including storefronts, portals, content delivery solutions, and related management applications.

3.7.4.3 Hewlett-Packard

HP, founded in 1939, is a name known well beyond the telecommunications market. Although active in telecommunications since its beginning, the company has changed fundamentally several times, marking distinct periods in its telecom presence.

HP first became known in telecommunications as a vendor of advanced testing and measurement products, some of them dedicated to the telecommunications field (e.g., transport network measurement equipment). Around 1970, the company gradually shifted its focus toward computers and peripherals, with measurement gradually losing weight over the ensuing years. This process culminated in 1999, when the test and measurement divisions separated from the company and started up as a new business entity, Agilent.

At approximately the same time, HP deliberately extended its portfolio in the software area with its HP-UX operating system and Openview network and systems management product suite. Although originally developed for enterprise networks, the product family soon included several high-end modules specifically dedicated to provider environments.

The most well-known Openview component is Network Node Manager, a network visualization tool and relatively simple fault/alarm center. With respect to telecom operators, it is recommended that they use an extended version of this product, Openview Operations. Operations, in addition to discovery and ping-based availability testing, also employs agents deployed on managed nodes (servers and network devices), thereby delivering much more detailed information and enabling an in-depth analysis of device status. This information is combined with traps, log messages, and other notifications generated by the devices themselves, thus making a comprehensive umbrella management system feasible. The key concept here is correlation of events through a rule-based system, which makes it possible to identify the root cause of the symptoms reported by various parts of the network. Without such a correlation-based aggregation, operations would face a swarm of alarms that would hinder them in identifying the basic problems they should handle.

On the basis of the success of these two flagship Openview products, HP has been extending its portfolio in the past decade through acquisitions. Many of these acquisitions have been applicable or directly dedicated to the telecommunications area.

OpenView TeMIP was originally the prestigious Telecommunications Management Information Platform product of the Digital Equipment Corporation, acquired by HP through its acquisition of Compaq. Basically an object-oriented framework for building different types of OSS applications, it is most typically used as a platform for fault management. The most valuable component of TeMIP is its wide base of robust integration modules for different technologies (SNMP, XML, message buses), including technologies typically not supported by later products (e.g., TL1, CMIP, Corba).

Although both TeMIP and Openview Operations focus on fault management, interestingly HP is not merging the two products into a single offering. Rather, TeMIP is seemingly being offered for the largest telcos, where it is a core operation that is carefully maintained and constantly improved to accommodate new technologies, while Operations, along with the other new products, are offered for smaller customers, the "masses" that require products they themselves can operate and configure without a great deal of vendor support.

HP has recently increased the momentum in its acquisition activities, with new products appearing practically every quarter. In the service management/support/help-desk area, HP OpenView again has two competing products: HP OpenView Service Desk and HP OpenView Service Center, a product acquired with the Peregrine acquisition in 2005. Although it is a successful and popular product, Service Desk is expected to be discontinued to promote the more higher-end Service Center. Both of these prod-

ucts provide ITIL-compatible service management functions; in particular, Asset Center offers incident and task management for smaller telecom operations.

In 2007, HP acquired another company, Opsware, and its product offering, a leading configuration and provisioning management solution for network devices and server-based software.

HP OpenView Internet Usage Manager (IUM) is a type of mediation software used to process call and event records for voice and data services. This consolidated data can be used for billing, customer behavior analysis, or reporting from various other perspectives.

The acquisition of Mercury Interactive in 2006 marks HP's commitment to management software. This acquisition enhances the HP portfolio with a comprehensive monitoring, diagnostic, and diagnostic framework. Again, obviously, the framework needs to be integrated with the remainder of the company's portfolio, and this work is currently well in progress.

OpenView Performance Manager is a partially agent-based software solution used to ensure continuous availability and performance monitoring of services and networks and to provide historical, troubleshooting, or trend analyses based on this data. Performance Manger includes an intuitive GUI that displays comprehensible overviews and aggregates but also allows drilling down to individual data samples for an accurate inspection of service history.

3.7.4.4 IBM

IBM initially entered the computer business in the 1950s, as the chief contractor developing computers for the U.S. Air Force's automated defense systems. By the 1960s, its mainframe computer systems had become well known and widely used, and by the 1980s the company had established itself as one of the main players in the information revolution through the introduction of its PC/AT and PC/XT personal computers.

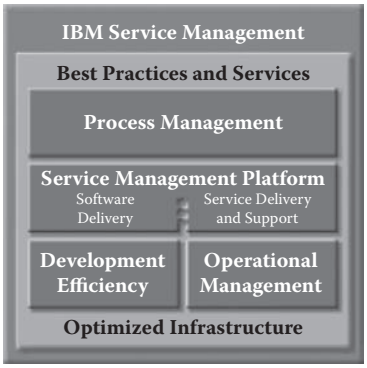
In the mid-1990s, IBM almost went out of business. This near disaster forced the company to change its focus: The decision was to shift away from components and hardware and toward software and services. In 2002, IBM strengthened its business advisory capabilities by acquiring the consulting arm of professional services firm PricewaterhouseCoopers. The company has increasingly focused on solution-driven consulting, services, and software, with an emphasis as well on high-value chips and hardware technologies. The new IBM has enhanced global delivery capabilities in consulting, software, and technology-based process services.

The basis for IBM IT services is the Tivoli Management Platform (a product of the former Tivoli Systems Inc., which was acquired by IBM in 1996 and assigned to the company's Software Group division). The platform's framework, a Corba-based architecture that allows the platform to manage large numbers of remote locations or devices, will be replaced by SOA techniques over time.

With the advent of convergent telecommunications services (and the new challenges posed by ICT and Next-Generation Networks), IBM entered this market by providing advanced support systems and services for service providers. Figure 3.7.4 shows the IBM service management architecture.

Telecommunications service providers have begun to implement next-generation operational support system solutions (NGOSSs) to enable the efficient provisioning, deployment, and implementation of new types of multimedia services. There will be an increasing demand for products and solutions supporting these NGOSS requirements. IBM offers communications service providers a broad range of hardware, software, applications, and services solutions through its Tivoli, WebSphere, and Global Services organizations. In addition, IBM has further expanded its OSS/BSS solution capabilities through the acquisition of Micromuse and Vallent. These acquisitions, when combined with IBM's service-oriented architecture (SOA) leadership and broad partnership and alliance network, position the company for a strong presence in the NGOSS arena. However, IBM's key challenge is addressing competition from end-to-end solution providers as well as vendors with lower-cost solutions.

Micromuse was a leading provider of network management software used by banks, telecommunications carriers, governments, retailers, and other organizations to monitor and manage their sophisticated technology infrastructures. The software helps customers manage increasingly complex IT systems that support the proliferation of voice and video traffic, in addition to data, as a result of the growing



FIGu RE 3.7.4 IBM service management architecture.

adoption of voice over IP (VoIP), audio, and video services delivered over the Internet. The combination of Micromuse's software and IBM's IT service management technology is expected to provide customers with a comprehensive approach to reducing the complexity of their IT environments, lowering their operational costs, and addressing compliance mandates.

Vallent used to be a leading supplier of network performance monitoring and service management software for wireless service providers worldwide. Vallent's software helps service providers manage the performance of their network infrastructure through monitoring and reporting problem areas such as dropped calls and traffic bottlenecks. It also helps operators improve wireless service quality and identify network problems before they impact a customer's experience. As a leading service assurance provider with over 300 installations worldwide, Vallent has been formed from numerous mergers and acquisitions, including those of WatchMark, Comnitel and ADC Metrica.

Vallent is integrated into IBM's Software Group as a part of its Tivoli Software unit, the same unit into which the Netcool assets from Micromuse acquisition have been integrated.

IBM's service assurance management strategy is to provide comprehensive solutions, across both IT and operations, that enable carriers to speed time to market, ensure quality of service, and reduce operational costs. The combination of IBM's Tivoli, Micromuse's Netcool, and Vallent's wireless service and performance management products amalgamate to form a powerful and comprehensive suite of service assurance functionality across network domains and discrete network environments.

- The combination of Micromuse's software and IBM's IT service management technology is expected to provide customers with a comprehensive approach for reducing the complexity of their IT environments, lowering operational costs, and addressing compliance mandates.
- With the acquisition of Vallent, IBM complements its real-time and event-based service assurance approach, mainly rooted in the fixed line and IP markets, with strong wireless functionality and customer-centric service assurance capabilities. The combination of Vallent's software and IBM's comprehensive management capabilities will enable service providers to deliver and manage end-to-end high-quality services across wireline, wireless, IP and converged fixed/mobile network infrastructures

These acquisitions, when combined with IBM's service-oriented architecture (SOA) leadership and broad partnership and alliance network, position the company for a strong presence in the NGOSS area. However, IBM's key challenge is to address competition from end-to-end solution providers as well as vendors with lower-cost solutions.

3.7.4.4.1 Tivoli Product Portfolio

Server, Network, and Device Management: Server, network, and device management is an integral part of IBM IT service management. It consists of the following parts.

- Tivoli Netcool/OMNIbus
- Tivoli Netcool/Webtop
- Netcool/Portal Tivoli
- Netcool/Reporter
- Tivoli Network Manager IP Edition (Formerly Netcool/Precision IP)
- Tivoli Network Manager Transmission Edition (Formerly Netcool/Precision TN) Netcool/Proviso

Business Application Management

- Tivoli Netcool/Impact
- Netcool/Realtime Active Dashboards
- Tivoli Monitoring
- Tivoli Composite Application Manager for Internet Service Monitoring (Formerly Netcool ISMs)

Security Management

- Tivoli Security Operations Manager

Wireless Service Management

- Tivoli Netcool Performance Manager for Wireless Technology Packs
- Tivoli Netcool Service Quality Manager

Tivoli Netcool Customer Experience Management Solution

3.7.4.4.1.1 Tivoli Netcool/OMNIbus Tivoli Netcool/OMNIbus provides consolidated operations management for real-time service management and performs the following functions:

- Centralizes and manages across complex data centers, network operation centers, and IT domains in real time to increase efficiency and productivity, reduce costs, and improve service availability and resiliency
- Delivers a central point of management for complex IT and network operations, including business applications, network devices, Internet protocols, security devices, and more
- Optimizes service availability and reliability through automated event correlation, isolation, and resolution capabilities that allow quick identification and resolution of the most critical problems across operational silos
- Consolidates data across operational silos into real-time Web dashboard views with customizable displays of events, service views, and operational indicators
- Provides tight integration with the IBM Tivoli monitoring family to proactively measure performance and user experiences across business applications
- Offers seamless integration with the broader Tivoli portfolio

Tivoli Netcool/OMNIbus is basically a type of event collection and correlation software.

3.7.4.4.1.2 Tivoli Netcool/Webtop Tivoli Netcool/Webtop delivers graphical maps, tables, and event lists to remote operators via HTML and Java. It enables users to manage Netcool alerts with the Netcool/Webtop flexible interface and advanced management capabilities. It extends Netcool/OMNIbus capabilities by adding a new set of graphical views as well as flexible management and administrative functions.

3.7.4.4.1.3 Netcool/Portal IBM Netcool/Portal enables organizations to provide secure access to interactive back-end Web applications. It also allows them to implement consolidated single sign-ons (SSOs) and centrally coordinate user access management. The portal integrates and consolidates disparate business applications and management tools; provides a single, real-time, Web-based console for visualization and centralized management of distributed data from across the enterprise; centrally manages all of the Netcool applications, as well as any mix of Web-enabled enterprise, OSS, and point management

applications already in use across the distributed operation; and provides a customizable Web dashboard that serves as a Web page proxy for applications and other Web sites.

3.7.4.4.1.4 Tivoli Netcool/Reporter Tivoli Netcool/Reporter captures, analyzes, and presents event data generated over various timeframes. It provides IT managers with long-term, retrospective information about the behavior of devices, links, and services within their networks and presents data generated over time in meaningful reports.

3.7.4.4.1.5 Tivoli Network Manager IP Edition (formerly Netcool/Precision IP) The Tivoli Network Manager IP Edition provides the network visibility needed to visualize and manage complex networks. It can be used to collect and distribute Layer 2 and Layer 3 network data and to build and maintain knowledge about physical and logical network connectivity. In addition, it provides physical, port-to-port connectivity between devices and captures logical connectivity information, including virtual private network (VPN), virtual local area network, asynchronous transfer mode (ATM), frame relay, and multiprotocol label switching (MPLS) services.

3.7.4.4.1.6 Tivoli Network Manager Transmission Edition (formerly Netcool/Precision TN) The Tivoli Network Manager Transmission Edition can be used to collect and distribute layered network data and build and maintain knowledge about physical and logical network connectivity. It assists in visualizing and managing complex networks—and, more importantly, the services they deliver—efficiently and effectively. Also, it provides valuable advanced fault correlation and diagnosis capabilities, includes real-time root-cause analysis to help operations personnel quickly identify the source of network faults and resolve problems, and delivers highly accurate, real-time information on network connectivity, availability, performance, usage, and inventory.

3.7.4.4.1.7 Netcool/Proviso Netcool/Proviso (formerly a product of Quallaby, which was acquired by Micromuse) provides a complete view of service quality and usage for both operations and customers, enabling them to proactively avoid problems or detect and rapidly resolve them. Informative, on-demand reports help operations and engineering improve service quality and reduce operating costs. Netcool/Proviso's flexibility and scalability have been proven in service provider deployments, particularly in next-generation IP network and OSS consolidation projects, allowing service providers to offer value-added reports to their customers.

Netcool/Proviso's consolidated, service-centric reports of network resource, server, and application performance illustrate the business and service effects of problems while making it easy to drill down to individual resources and service paths for troubleshooting. On-demand detailed trend reports, second-by-second real-time monitoring, and proactive forecast reports provide the context needed to make informed decisions quickly.

Designed for change and growth, Netcool/Proviso rapidly supports new services such as VoIP and MPLS VPNs, and it even automates the process of keeping up with daily network and customer changes to ensure data collection and report accuracy.

3.7.4.4.1.8 Tivoli Netcool/Impact IBM Tivoli Netcool/Impact provides a common platform for data access that circumvents organizational boundaries. Armed with data from virtually any source, one can correlate, calculate, enrich, deliver, notify, escalate, visualize, and perform a wide range of automated actions:

- Automatically correlating and prioritizing event responses according to business impact
- Maximizing productivity with a common platform for universal access to real-time data, events, and changes
- Improving time-to-resolution through real-time monitoring and analysis of business events
- Protecting existing data stores by virtualizing access to distributed information
- Delivering actionable information in user-specific, Web-based views

- Optimizing customer experiences by effectively tracking business and operational performance indicators.
- Automate workflow and maximize operator efficiency and productivity

3.7.4.4.1.9 Netcool/Realtime Active Dashboards IBM Tivoli Netcool/Realtime Active Dashboards provides advanced, real-time visualization of services and processes. Through the use of its service dependency model, business and operations staff can understand the complex relationships between business services and supporting technology. The tool performs the following functions:

- Incorporates data from a broad array of IT resources and business support systems that contribute to defining a service into a real-time, federated service model
- Addresses problem resolution with automated service impact, root-cause analysis, and integrated access to service support and delivery systems
- Helps improve service visibility for operations and business stakeholders—along with their customers—by offering real-time service scorecards and key performance indicators
- Automates service model definitions and maintenance with true real-time federated data access
- Provides real-time service-level tracking
- Measures and reflects—in real time—the quality of services and processes

3.7.4.4.1.10 Tivoli Monitoring Tivoli Monitoring is an enterprise-class solution that optimizes IT infrastructure performance and availability. It manages the IT infrastructure, including operating systems, databases, and servers across distributed and host environments, through a single customizable workspace portal. In addition:

- It provides common, flexible, and easy-to-use browser interfaces and customizable workspaces.
- It detects bottlenecks and potential problems in essential system resources and automatically recovers from critical situations to ensure that critical business applications are up and running.

3.7.4.4.1.11 Tivoli Composite Application Manager for Internet Service Monitoring (formerly Netcool ISMs) The Tivoli Composite Application Manager for Internet Service Monitoring tests Internet services from the user's perspective. It is a highly scaleable suite of 23 monitors that measure the availability, performance, and content of services through periodic polling from distributed points of presence. It performs the following functions:

- Produces both real-time alerts on service responses and Web-based reports on historical service performance relative to a flexibly definable SLA
- Uses these Web-based reports and analysis tools to confirm SLA adherence for end customers
- Enables operations staff to intelligently plan changes to the infrastructure and demonstrate the resulting effects on service performance

3.7.4.4.1.12 Tivoli Security Operations Manager The Tivoli Security Operations Manager is a security information and event management (SIEM) platform that centralizes and stores security data from throughout the technology infrastructure to improve security operations and information risk management. In addition, the platform:

- Enables automation of log aggregation, correlation, and analysis; automatic recognition and investigation of and response to incidents; and streamlining of incident tracking and handling
- Improves efficiency through operational integration by facilitating the flow of incident management data between security, network, and systems management operations teams
- Deepens understanding through comprehensive reporting, including on-the-fly data mining, historical reporting, self-auditing, and tracking capabilities

- Offers multiple deployment options to suit the organizational environment thanks to a modular architecture that can adapt to—and grow with—the organization's security infrastructure
- Provides a platform for offering managed security services and delivering reduced operational costs through automation and speedy implementation

3.7.4.4.1.13 Tivoli Netcool Performance Manager for Wireless Technology Packs Provides performance metrics to manage all aspects of a service provider's wireless network infrastructure. It offers an efficient way to manage individual technologies or services in a cross-vendor environment. This off-the-shelf network interfaces seamlessly aggregate and correlate data from multiple vendors and technologies. It is available for technologies such as GSM, GPRS, UMTS and CDMA/1xRTT, and also provide existing support for key technical HSPA, IMS, MSC-S, MGW and more. It allows the service provider to

- analyze and compare information from multiple vendors and technologies—including all the relevant counters each creates;
- convert performance data into a common format that allows you to manipulate and report on it without overburdening your systems;
- leverage IBM expertise in cross-vendor, cross-technology environments to drive efficiency and deliver role-specific insights.

The pack supports technologies from leading wireless equipment vendors such as Ericsson, Nokia, Motorola, Huawei, Alcatel, Siemens, Nortel Networks.

3.7.4.4.1.14 Tivoli Netcool Service Quality Manager Combines service quality management (SQM) and service-level agreement (SLA) management to manage and improve telecommunications service quality. This software:

- Provides a real-time, end-to-end view of a service to enable service providers to understand service quality from the customer's perspective.
- Monitors and improves the quality of each customer experience, resulting in more effective customer care and increased customer satisfaction.
- Responds to network issues based on corporate directives such as revenue, profitability, service and customer impact.
- Provides product differentiation to your enterprise sales team by offering guaranteed service-level agreements (SLAs) to attract and retain high-value enterprise customers.
- Enables the successful, rapid introduction of new services that you can offer with confidence in their service quality.
- Identifies high-priority problems quickly and accurately with powerful root-cause and impact analysis.
- Offers comprehensive data collection capabilities and extensive preestablished service models with full Key Quality Indicator (KQI)/Key Performance Indicator (KPI) mapping and reporting.
- Manage the quality of individual services delivered by telecommunication providers.

3.7.4.4.1.15 Tivoli Netcool Customer Experience Management Solution The Tivoli Netcool Customer Experience Management Solution helps telecommunication service providers proactively manage each customer's experience. The software delivers the visibility needed to improve the customer experience as they interact with different aspects of the operator's business. The offering correlates data from fault management, performance management, probe systems, and other OSS/BSS for a true view of the customer's experience. It provides different perspectives—all on one dashboard—of the subscriber's interaction with the service provider, including views by service, network, location, subscriber group, and device type.

The solution helps increase customer loyalty and drive new revenue opportunities by helping to:

- Detect potential service degradations before they impact customers, and understanding which problems are affecting the customer experience
- Increase marketing and up-sell capabilities by proactively monitoring behavior
- Control costs and increase organizational efficiency by focusing on services that are important to the customer
- Improve customer retention through differentiation on quality in the marketplace
- Drive true customer-centric business through better customer understanding across the organization

3.7.4.4.2 WebSphere

IBM's WebSphere product family provides open standards–based communications and integration infrastructure for application development spanning from the enbedded space to the enterprise. Most well known are the WebSphere Application Server products. These J2EE platforms provide for service-oriented archictecture (SOA) application development using Web Services, Enterprise Java Beans (EJBs), database connections and support extensions for Struts, AJAX, DoJo, JSR-168 Portlets, SSL (Secure Socket Layer), SOAP (Simple Object Access Protocol), XML, XSD (XML Schema Definition), WSDL (Web Services Description Language), JDBC (Java Database Connection), BPEL (Business Process Execution Language), and ESB (Enterprise Service Bus). Message queuing infrastructure (MQ Series and MicroBroker (aka Lotus Expeditor Client) allow for reliable messaging infrastructures to support publish/subscribe-based communications. The WebSphere Process Server provides work flow for business processes, while WebSphere Portal Server provides collaborative tools for a rich, personalized user experience.

3.7.4.4.3 IBM and SOA

Service-oriented architecture (SOA) enables separation of functions as services with public interfaces that enable integration without exposing the actual implementation of functions or database design. This has the major advantage that companies can continue to refine the functions providing services to improve performance and/or to improve how data is processed without breaking the public interfaces supported by the service layer. By implementing a service-oriented architecture, companies can:

- create composite business processes through legacy parts consolidation
- adapt to change through improved flexibility and distribution/separation of functions
- automate business processes and be more responsive to customer needs
- interconnect business processes for improved, more timely interaction
- lower development cycles and costs

Service-oriented architecture follows four life cycle phases:

- Model: collect and analyze business requirements and processes to understand producer/consumer interaction patterns.
- Assemble: define and develop the service interfaces for existing assets like enterprise resource planning (ERP) and enterprise asset management (EAM) systems, while identifying the white space requiring development of integration or transformation services.
- Deploy: install and configure the execution environment to meet service-level requirements necessary for your business processes to flourish.
- Manage: monitor and adjust configurations for optimal performance and availability. Review actual consumer/publisher interaction patterns as the introduction of the SOA often identifies new opportunities for synergy between applications and improves overall application effectiveness.

The WebSphere suite of integration products provide high-performance process and data integration between internal customer applications or external business partner applications. The core products comprise:

- WebSphere Process Server
- WebSphere ESB
- WebSphere Message Broker
- WebSphere Process Modeler
- WebSphere Process Monitor

The Enterprise Service Bus (ESB) is at the core of a service-oriented architecture, providing reliable communications services to interconnect:

- Business Services: facilitates better decision making with real-time business information
- IT Service Management: manages and secures applications and resources
- Infrastructure Services: optimizes throughput, availability, and performance
- Development Services: integrated environment for design and creation of solution assets

To accomplish this, the SOA reference architecture provides the following services:

- Interaction Services: enables collaboration among people, processes, and information
- Process Services: orchestrate and automate business processes
- Information Services: manages diverse data and content in a unified manner
- Partner Services: connect with trading partners
- Business Application Services: build on a robust, saleable, and secure services environment
- Access Services: facilitates interactions with existing information and application assets

The WebSphere Message Broker provides the connection point to the Enterprise Service Bus and provides reliable transport of inherently transactional intraservice messaging by building on the WebSphere Message Queue capabilities. It is fast, with extensive capability for scaling through clustering, and possesses established monitoring capabilities (though these are often enhanced for company application specific needs).

The messaging infrastructure can feed business process workflow managed and implemented within the WebSphere Process Server. The process server provides choreography of events requiring action, which can be presented to people with appropriate notification and escalation policies, or to application services to automate responses.

Service-oriented architecture design and development has the benefit of building from the middle (and not the lowest level) upward, and being results oriented. Trying to standardize on commonality at the lowest levels may impose unnecessary restrictions on the operating environment (e.g., requiring the same database infrastructure or schema even when the application may not warrant it). Building from the middle upward allows actual implementation of services to be tailored to the service providers' capabilities and operating environment while still adhering to agreed-upon public service interfaces.

3.7.4.5 Oracle

Oracle is an interesting example of an emerging OSS/BSS supplier. The entry point to service providers is strong, and most likely, database applications are in use. Oracle recognized early that service providers' databases would become differentiators. Simplified, cleansed, accurate databases offer a distinct competitive advantage. In particular, the following databases play a principal role for service providers:

- Product portfolio
- Customer
- Inventory (hardware and software)
- Workforce
- Workflow
- Critical performance indicators
- Enterprise resource planning (ERP)

There are some overlaps and duplications among these databases. In advanced cases, a powerful ERP solution can embed inventory and workforce in addition to the traditional CFO-oriented modules. Although these databases can be physically distributed, logically they must be connected and synchronized.

Oracle has purposefully put a multilayer telco suite together. The company has acquired several suppliers with very strong point products and used the Fusion portfolio to connect these products with each other. The suite is composed as follows (TURI08):

Application infrastructure: Hotsip (SIP and Web applications). HotSip is a provider of telecommunications infrastructure software and Session Initiation Protocol (SIP)–enabled applications for IP telephony, presence, messaging and conferencing on new converged networks. With the addition of HotSip's technology, Oracle will be better able to build on its leadership in middleware and in carrier-grade communications infrastructures.

Service delivery: Net4Call (Software delivery and distribution). The acquisition of Net4Call, a leading provider of Parlay/OSA service delivery components for the telecommunications industry, supports Oracle's roadmap to a comprehensive, standards-based service delivery platform (SDP) for the telecommunications industry.

The acquisition, combined with Oracle SDP, will help communications service providers, network operators, and system integrators to evolve current silo-based network investments into a service-oriented architecture and shrink time and cost to deploy new services on existing and next-generation communications IP networks.

The new Oracle SDP also leverages service-oriented architecture to help customers shrink the time and cost to deploy new voice data and integrated multimedia services on existing and next-generation communication IP networks. Enterprises will be able to extend their communication infrastructures with the Oracle SDP, providing a strong foundation for new voice over IP, mobile, and real-time applications.

The Oracle SDP will embrace the convergence of IT and network technologies to deliver a scalable platform with carrier-grade reliability, real-time performance, connectivity to traditional and next-generation IP-based networks, and interfaces to operational and business support systems (OSS/BSS).

Business Support Systems: Siebel-CRM, PeopleSoft-CRM, Portal (Billing, ERP, and Business Intelligence).

Operations Support Systems: MetaSolv Software (Provisioning, Inventory, Activation).

Network Management: Netsure (physical and logical network management). By adding Netsure, Oracle extends the Oracle Communications Application Product Suite to include business intelligence and analytics for the network domain. Netsure's solutions are product based, and support open standards. Netsure's product capabilities combined with Oracle's network inventory, provisioning, and financial asset management applications, are expected to enable service providers to improve proactive network planning, modeling, and optimization, reducing operational costs and increasing the utilization and efficiency of both leased and owned network capacity.

In particular, the BSS/CRM layer is strong. The likelihood is very high that targeted Tier 1, Tier 2, and Tier 3 telcos had used Siebel, PeopleSoft, and/or Portal products for CRM prior to the acquisitions of these companies by Oracle. This suite is especially strong when combined with ERP capabilities and business intelligence applications.

The addition of BEA will accelerate innovation by bringing together two companies with a common vision of a modern service-oriented architecture (SOA) infrastructure. Together, Oracle and BEA will provide a series of complementary and well-engineered middleware products, allowing customers to more easily build, deploy, and manage applications in a secure environment.

The acquisition of Hyperion makes Oracle the category leader in the high-growth enterprise performance management (EPM) market. Hyperion's EPM software coupled with Oracle's Business Intelligence (BI) tools and analytic applications form an end-to-end performance management system that includes planning, budgeting, consolidation, operational analytics, and compliance reporting.

Requirements for Performance Management and Business Intelligence solutions are increasingly converging. Given the critical need for managers across the enterprise to align operational decisions with strategy, now is the right time for Hyperion to combine with a strategic partner like Oracle to deliver the first, integrated end-to-end Enterprise Performance Management System.

3.7.4.6 Sample of New Vendors on the OSS Market

The OSS marketplace is relatively dynamic, especially with respect to the rapid development of telecommunication services and technologies. While there has been some stability in terms of established vendors providing plain old telephone service (POTS) and Sonet/SDH services, many newcomers have arrived with an explicit focus on new IP-based services. Furthermore, the rapidly changing business models followed by providers (e.g., one-day provisioning, product bundles, virtual operators), have opened up opportunities for new vendors to enter with fresh, innovative products.

For example, after being a silent outsider in the telecommunications arena for many years, software giant Microsoft recently joined the competition with its relatively successful IPTV platform. In their role as mission-critical residential services, IPTV technologies require special efforts in the areas of service assurance and fulfillment. The company uses its Microsoft Systems Center Operations Manager (MOM) product for this purpose. This monitoring and operations management system targets end-to-end service management challenges. A key advantage to Microsoft here is that MOM can natively manage the different Windows categories (Server, Client, and CE) used in this system. The appearance of traditional computer operating systems (Unix, Linux, and now Windows as well) and software solutions in real-life services has opened up opportunities for new management technologies.

Another typical example of a telecommunications management solution vendor is Alcatel, the French telecom equipment manufacturer that also acquired NewBridge, once a leading vendor of data communications equipment. Alcatel's OSS and Network and Service Management (NSM) solutions provide end-to-end management of multiaccess, multicore technologies and multiple services, thus offering service providers rapid service deployment across multiple domains. The company has built on cornerstone products from its voice (fixed and mobile), optical, and data networking groups, which have been popular EMS/NMS management systems for many years.

Newly developed regions such as the Far East and Eastern Europe also are making use of the relatively modern technologies and networks they have built for themselves in recent years. Local companies first provided innovative management solutions for local providers, but now some of them have also stepped into the world market with their solutions.

India is already a global hub of software development for off-site, outsourced development efforts, and recently several self-confident Indian brands have also emerged. A good example is Clarity, whose single, integrated OSS product covers seventeen eTOM L2 processes. Another Indian company, Subex Azure, surprised the world by purchasing Syndesis, a well-known American service fulfillment product vendor. Even if some facets of these products are not yet world-class, their huge internal market has helped make these companies respectable and potentially successful in the long term as well.

A strong representative from China is Huawei, a manufacturer of network and telecom equipment with a portfolio strongly resembling that of Cisco in both range and technology. As is often the case with hardware vendors, Huawei first entered the management application business with a network/element management system, iManager, which was used for its hardware products. In recent years, the iManager platform has gradually been extended, with capabilities including unified network management, fast and visual fault location, real-time performance monitoring, and easy service deployment.

In Eastern Europe, the other region witnessing spectacular modernization in recent years, many innovative ideas have been implemented by the region's generally well-educated population. Comarch, a Polish company, has established itself as a new European vendor of flexible billing and OSS products that include inventory, fault management, and performance management.

Another exemplary vendor from the new EU countries, NETvisor in Hungary, has established itself as a vendor of flexible, new-generation solutions in areas including IP/VPN service discovery

(NETExplorer/IPExplorer), inventory (NetInv), and performance management (PVSR). On the basis of these foundations, NETvisor has also built interesting solutions such as the Service Level Assistance Suite, a multisource SLA management and reporting solution; the ITVSense DVB/IPTV/3Play end-to-end management system; and ANMS, a platform-independent management framework aimed at access technologies (e.g., xDSL, Gigabit Passive Optical Network [GPON], WiMAX).

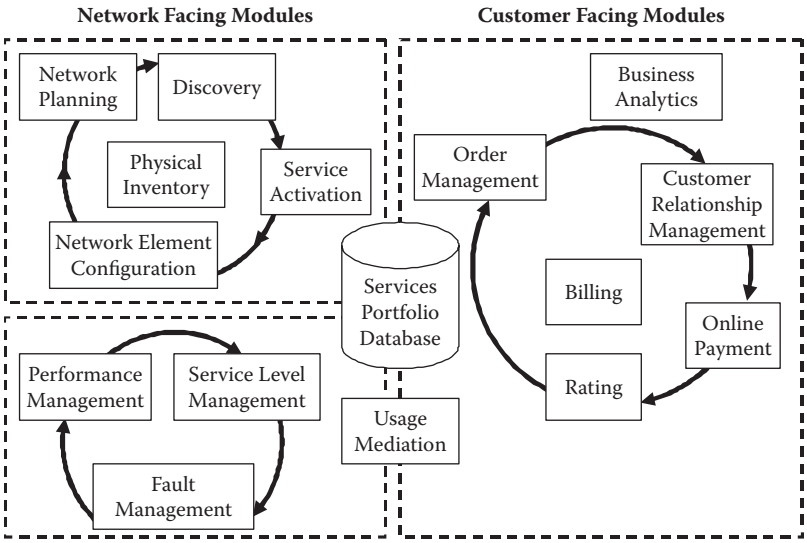
A particularly noteworthy market for these vendors consists of Tier 2/Tier 3 providers both in the region and in many other parts of the world (e.g., Asia and Africa). In the case of many of these providers, the complex systems of traditional vendors listed above are moderately useful (their new networks do not need to be integrated with extensive legacy technology) but prohibitively expensive.

3.7.5 Consolidation of Support Systems

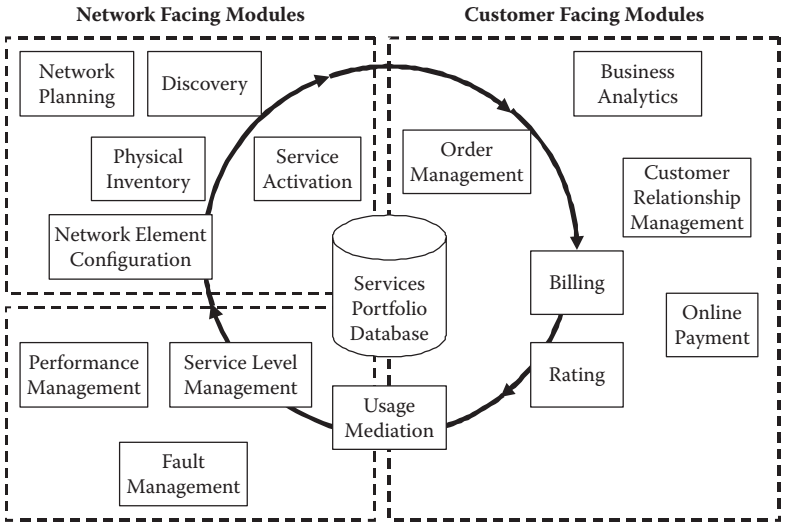
OSS and BSS continue to be consolidated. Usually, best-of-breed products are combined with each other to offer best-of-suite capabilities. This consolidation can occur in different ways:

- Suppliers extend the scope of their products to meet general and/or customer demand. Examples include Nakima Systems, Telcordia, BMC, Wily, and NetScout.
- System integrators and solution providers integrate products to meet specific customer demand. Examples include Cap Gemini Ernst & Young, Accenture, IBM, and Hewlett-Packard.
- Mergers and acquisitions lead to product portfolio cleansing, unification, and simplification of product offerings. Examples include IBM and Micromuse, Hewlett-Packard and Mercury, Amdocs and Craemer, Subex and Azure, CA and Wily, Telcordia and Granite, EMC and Smarts, CSG and Telution, Oracle and Siebel, and Metasolve and Portal.

Today, OSSs and BSSs are fragmented into the following functional groupings: service fulfillment, service assurance, and billing and revenue assurance. Figure 3.7.5 shows near-term consolidation of (1) network-facing modules with network-related service fulfillment (network planning, discovery, physical inventory, network element configuration, service activation) and assurance (fault management, performance management, service-level management) and (2) customer-facing modules with customer-related service fulfillment (order management, customer relationship management) and billing (rating, online payment).



FIGu RE 3.7.5 Near-term OSS consolidation.



FIGu RE 3.7.6 Long-term OSS consolidation.

It can be seen that, in the early stages of consolidation, the service portfolio database and usage mediation connect the three relatively independent principal functional modules. Mergers and acquisitions more or less follow this consolidation within the three functional modules.

Long-term consolidation goes a step further (Figure 3.7.6) and offers large-scale integration under one umbrella. The functional modules remain the same. This ambitious form of consolidation—one-stop shopping—is without doubt very beneficial for all three tiers of service providers. From a financial perspective, however, only a few system integrators and suppliers have the potential to become serious players in this long-term consolidation. To do so, companies must have very strong relationships with providers such as hardware, software, storage, database, and professional services. Companies like Oracle, IBM, Hewlett-Packard, Amdocs, Subex, and Telcordia could play this role in the near future.

3.7.6 Results of Market Research

Who will lead the near-term and long-term consolidation of OSS and BSS? Turino (TURI08) subdivides the strategic buyers into four principal groups: traditional players, newer entrants, potential entrants, and smaller, mostly private buyers.

3.7.6.1 Traditional Players

Companies in this group include Alcatel-Lucent, Amdocs, Cisco, Comverse, Convergys, CSG, Ericsson, Nokia, Motorola, and Tekelec. It is interesting to note the origins of these firms:

- The manufacturing companies, Motorola, Cisco, Nokia, Alcatel-Lucent, and Ericsson, include a number of strong element and network management systems.
- The billing companies, Convergys and CSG, are using their back-office position to sell more OSS/BSS and professional services.
- The measurement company, Tekelec, is seeking additional business (e.g., SIP in addition to SS7).
- The consulting company, Amdocs, is strengthening itself through acquisitions and concentrating fully on telecommunications service providers.

3.7.6.2 Newer Entrants

Newer entrants include Danaher, IBM, NeuStar, Nokia-Siemens Network, Oracle, Sonus Networks, and Syniverse. IBM and Oracle are interesting examples. Both are well-known and well-accepted suppliers

in their own domains. However, they want more business from telecommunications service providers. Using their strong positions, they can broaden their one-stop-shopping offers through strategic acquisitions—and they have been doing so over the past couple of years.

3.7.6.3 Potential Entrants

Potential entrants include Computer Associates, Hewlett-Packard, Nortel, SAP, Sun, and Telcordia. Some of these companies come from the enterprise network management market. Telcordia maintains a strong OSS/BSS portfolio but needs new, innovative, and lean products in addition. HP has been very ambitious under its new leadership. This acquisition of EDS places them in the front position with respect to professional services. Surprisingly, SAP has been very passive in terms of solution offerings for service providers. Maintaining its strong ERP position will not be enough over the long term.

3.7.6.4 Smaller, Mostly Private Buyers

Companies in this group include Aricent, Comptel, Exfo, InfoVista, Intec, Ixia, and Subex. Each of these companies has excellent niche products and, as a result, good entry positions with telcos. Subex is in the process of putting together a portfolio through acquisitions. This portfolio will consist of a management framework, provisioning software, and billing and revenue assurance applications. This offer can be completed by outsourcing professional services to India.

Table 3.7.1 gives an overview on available products and solutions. This table must be considered dynamic given that new entrants and suppliers will appear, portfolio changes will be made, and certain products and services will be retired. The list is recommended only as a starting point when evaluating suppliers and their offers.

3.7.7 Summary and Trends

Management frameworks are key for successfully managing communication infrastructures. The frameworks of the future will include very strong core components and a rich set of management applications. These management applications will be provided by independent software vendors and will address management process areas that are important for telecommunications services suppliers and enterprise users. Integration is most likely deeper in the telecommunications industry than in the enterprise environment. Some management applications are the same for both areas. Web technologies will take over certain integration tasks. OSS/BSS framework suppliers will follow the business process trends expected with service providers, such as opening their infrastructures to content providers, supporting external application development, and engaging in careful use of open source software. Most likely, framework suppliers will be differentiated according to the field they support best. This categorization will follow industry trends and will include the following groups: infrastructure, domain management, service delivery platforms, customer care, service fulfillment, service assurance, and billing and revenue assurance. OSS/BSS framework suppliers are expected to support strong and secure information exchange alternatives with customers, workforces, and partners of service providers.

TABl E 3.7.1 Categorization of Suppliers

Suppliers	Products and Services
	OSS/BSS Frameworks
Telcordia	Network engineering suite; Granite technical inventory
Amdocs	Revenue manager (bill formatter) for Any-Play services; Customer management (ClarifyCRM); Service and resource planner; Service catalogue; Order manager; Activation manager; ECare and fraud disvovery manager; Xacct mediation; Customer Experience Manager; Qpass for digital commerce; Cramer Systems for inventory and product management
Hewlett-Packard	TeMIP
IBM	Tivoli suite for network management; Tivoli suite for security management; Netcool suite for performance management; Service quality management
Oracle	Web/SIP applications (Hotsip); Subscriber and service management (Net4Call); Service fulfillment (provisioning, inventory) (Metasolve); Service activation (Metasolve); Billing and revenue management (Portal); Infrastructure management (Netsure); CRM (Siebel); Business intelligence and analytics (Hyperion)
Subex	Revenue Operations Center (Certo, RevMax); Service fulfillment and assurance (NetProvision, TrueSource)
Nakina Systems	Network OS; Adapter SDK; Common EMS (universal mediation, performance monitoring data collector, network element configurator, inventory browser, network element discovery)
Intec Telecom	InterActiveE ; Content Partner Management ; InterMediate Convergent mediation and billing; Customer management Product catalogue; Interconnect billing; Carrier Access Billing; Service activation; Charging, rating and active mediation
TTI Telecom	Netrac Traffic Guard; Netrac Revenue Assurance (billing verification, best cost routing, mediation, call expert); Netrac Performance Management (wireline, wireless)
Comarch	OSS suite includes network management, inventory management, trouble ticketing, and fault management on the basis of SOA, IMS, eTOM and Parlay
Comptel	Mobile activation solutions; Broadband service fulfillment
	Emerging Solutions from the Enterprise
Computer Associates	Unicenter (asset management, life-cycle management, service desk); BrightStor storage management; Network Health; ETrust and CleverPath Portal; Risk and security management
Hewlett-Packard	OpenView (service dek, service center, performance manager); Internet Usage Manager; Mercury Interactive solution suite
IBM	Tivoli for server and device management; Tivoli suite for server management
BMC	Service Impact Manager; IT Service and Application Management; Remedy IT Service Manager; IT operations and infrastructure management (Patrol, Mainview, Event Manager, Control-M, Control-SA); Enterprise data management; Resource modeling; BSM Software Performance Assurance
Paessler	PRTG Network Monitor using over 35 sensor types • IPCheck Server Monitor • Traffic Grapher • Network Discovery • Remote control using remote sensors • History database • WEB user interface for inquiries and reporting
	Application-Aware Monitoring
EMC	Smart
InfoVista	VistaInsight for servers, networks, and IP VPN VistaView for router and switch monitoring, help desks, LAN segments, network service levels, and Internet applications (BMC Patrol agent)

TABl E 3.7.1 Categorization of Suppliers (*Continued*)

Suppliers	Products and Services
NetScout	Ngenius probes; Ngenius real-time monitor; Ngenius application service level manager; Ngenius capacity planner; Ngenius synchetic transactions
CA-Wily	Introscope application performance monitor
	Device Dependent Solutions
Cisco	Cisco TelePresence Readiness Assessment Manager; Cisco Unified Operations Manager; Cisco Unified Provisioning Manager; Cisco Unified Service Monitor; Cisco Unified Service Statistics Manager; CiscoWorks Voice Manager; CiscoWorks Quality of Service Policy Manager; Cisco Network Connectivity Monitor; CiscoWorks Assistant CiscoWorks Campus Manager; CiscoWorks CiscoView; CiscoWorks Common Services; CiscoWorks Device Fault Manager; CiscoWorks Internetwork Performance Monitor CiscoWorks LAN Management Solution; CiscoWorks Health and Utilization Monitor; CiscoWorks Interface Configuration Manager; CiscoWorks for Mobile Wireless; Cisco Mobile Wireless Transport Manager; CiscoWorks Quality of Service Policy Manager; Cisco Router and Security Device Manager; Cisco Multicast Manager; Cisco Route Manager; Cisco WAN Manager; Cisco NetFlow Cisco Secure Access Control Server; Cisco Extensible Provisioning and Operations Manager; Cisco Media Gateway Control (MGC) Node Manager; Cisco Voice Services Provisioning Tool; Cisco Application Networking Manager; Cisco Data Center Network Manager; Cisco VFrame Data Center; Cisco Bandwidth Quality Manager Cisco Performance Visibility Manager; CiscoWorks Network Compliance Manager; Cisco Branch Routers Series Network Analysis Module; Cisco Catalyst 6500 Series Network Analysis Module; Cisco Network Analysis Module Software; Cisco Transport Manager; Cisco Router Web SetUp Tool; Cisco Active Network Abstraction (Network Resource Management); Cisco ANA Carrier Ethernet Activation; Cisco Assurance Management Solution; Cisco Video Assurance Management Solution; Cisco Info Center; Cisco Info Center for Security Monitoring; Cisco Info Center VPN Policy Manager; Cisco IP Solution Center; Cisco IP Solution Center L2 VPN Management; Cisco IP Solution Center MPLS VPN Management; Cisco IP Solution Center Security Management; Cisco IP Solution Center Traffic Engineering Management; Cisco Network Connectivity Monitor; Cisco Network Connectivity Center Business Impact Manager; Element Manager; Call Manager
Nokia	NetAct with network and domain management; NetAct is the basis of electronics manufacturing services (EMS); Intellisync Device management ; Intellisync Mobile Suite ; Nokia Configuration Tool; Network Security Manager
Alcate-Lucent	Alcatel-Lucent Optical Management (formerly Navis® Optical Management); Alcatel-Lucent 1300 Convergent Network Management Center (CMC); VitalSuite® Network Management Software ; VitalSuite® Network Performance Management; VitalSuite® Application Performance Management; VitalSuite® Network Traffic Management ;VitalSuite® Performance Management; VitalQIP® DNS/DHCP IP Address Management; Alcatel-Lucent 5020 Element Manager; Alcatel-Lucent 5620 Network Manager; Alcatel-Lucent 1330 Billing and Performance Management; Alcatel-Lucent Customer Service Manager; Alcatel-Lucent 5580 Home Network Manager; Alcatel-Lucent 9753 Operations & Maintenance Center; Alcatel-Lucent 9353 Wireless Management System; Alcatel-Lucent 1353 Litespan Management System; AnyMedia® Element Management System; DSL Element Management System; Alcatel-Lucent 1626 Light Manager; Alcatel-Lucent 5526 Access Management System
Ericsson	ServiceOn Broadband Operational Support System; ServiceOn Data Wireless OSS; Ericsson Mobile Operational Support Systems; Ericsson Network Inventory Management Solution; Ericsson Performance Management Solution
Nortel	Nortel Optical Network Manager; Metro Ethernet Manager QRadar Network Security Management; Communication Server 1000 Telephony Manager; Integrated Element Management System; Proactive Voice Quality; Management (PVQM) Solution; Network Resource Manager; Wireless Network Management System (W-NMS); Nortel WLAN IP Telephony Manager 2245

Continued

TABl E 3.7.1 Categorization of Suppliers (*Continued*)

Suppliers	Products and Services
Microsoft	Systems Management Server (SMS)
Siemens	HiPatch 4000¸ Openscape Unified Communication Server
Motorola	MeshManager Wireless Element Manager; CCE Content Manager; CCE Service Broker; Mobility Services Platform (Centralized Management for Mobile Services); Motorola's Expedience NetProvision™, Provisioning Management System; Motorola's Expedience NetManage™ Element Management System; AXSvision Element Management System; Motorola NBBS management platform; ERM 1000 Edge Resource Manager; SVOM 1000 Switched Video Operations Manager
Sun	Sun Management Center ; Solstice Enterprise Manager
	Service Delivery Platforms
Sigma Systems	All-Play-Solutions to define, manage and diagnose subscribed, on-demand and real-time broadband and wireless services Sigma Service Management Portfolio: • Sigma Service Profile Manager • Sigma Self Service Manager • Sigma Service Diagnostics Manager • Sigma Service Topology Manager • Sigma Service Creation Toolkit Sigma ServiceBroker Workflow Engine
NetCracker	Network inventory; Auto design and assign; Discover and reconciliation; Outside plant management; Order management; Asset management; Telecom cost management; Service layer transformation
Microsoft	Microsoft Provisioning Services (MPS); Microsoft Operations Framework (MOF)
Telcordia	Converged application server for service delivery, including real-time charging, real-time policy, and hosted solutions (voice and data)
	System Integrators
Accenture	They work with enterprises and service providers to build, manage and run all types of narrowband and broadband network in the wireline, wireless and cable industries. Business processes include network engineeirng, service integration, service management, operations and operational outsourcing
Amdocs	Enabler platform, Amdocs Foundation, B/OSS Manager
IBM	WebSphere
Oracle	Fusion; Enterprise Integration Architecture (BEA)
Hewlett-Packard	ISM (Integrated Service Management)
Microsoft	.net framework; XML Web services; BizTalk suite
Tibco	Enterprise Service Bus; TIB/RendezVous Active Enterprise (TIB/In Concert, TIB/Hawk,; TIB/message Broker); Active Exchange; Active Portal
Vitria	BusinessWare message bus and hub-and-spoke integration using the following integration layers: • Management layer • Enterprise Application Integration • Busines to Business • Real-time analysis • Common services Products: Order accelerator and Smart gateway
JacobsRimell	Single operational platform and customer view for all customers and services

Acronyms

API	Application Programming Interface
BI	Business Intelligence
BPEL	Business Process Execution Language
BSS	Business Support System
CDR	Call Detail Record
CMIP	Common Management Information Protocol
CMOT	CMIP Over TCP/IP
CORBA	Common Object Request Broker Architecture
CPE	Customer Premises Equipment
CRM	Customer Relationship Management
DMI	Desktop Management Interface
EAM	Enterprise Asset Management
EJB	Enterprise Java Beans
ESB	Enterprise Service Bus
ESP	Enterprise Service Provider
GPON	Gigabit Passive Optical Network
GUI	Graphical User Interface
HSPA	High Speed Packet Access
ICMP	Internet Control Message Protocol
ICT	Information and Communication Technologies
IMS	IP Multimedia Subsystem
ISS	Intelligence Support System
ISV	Independent Software Vendor
ITIL	IT Infrastructure Library
KPI	Key Performance Indicator
KQI	Key Quality Indicator
MIB	Management Information Base
MO	Managed Object
MPLS	Multiprotocol Label Switching
MSC	Mobile Switching Center
MVNO	Managed Virtual Network Operator
NGOSS	Next-Generation OSS
NMVT	Network Management Vector Transport
NTP	Network Time Protocol
ORB	Object Request Broker
OSS	Operations Support System
RDBMS	Relational Database Management System
RMON	ReMONitoring
SIP	Session Initiation Protocol
SLA	Service-Level Agreement
SNMP	Simple Network Management Protocol
SOA	Service-Oriented Architecture
SQL	Structured Query Language
SQM	Service Quality Management
SSO	Single Sign On
UTC	Universal Time Clock
VPN	Virtual Private Network
VAS	Value Added Services
WS	Web Services

References

BALL94: Ball, L. L. 1994. *Network Management with Smart Systems,* McGraw-Hill Series on Computer Communications. New York: McGraw-Hill.

DORF93: Dorf, C. R. 1993. *Handbook—Electrical Engineering*. Boca Raton, FL: CRC Press.

GARE95: Gareis, R., and Heywood, P. 1995. Tomorrow's Networks Today, *Data Communications,* September, 55–65.

GHET97: Ghetie, I. G. 1997. *Networks and Systems Management—Platforms, Analysis and Evaluations*. Kluwer Academic, Norwell, USA, 1997.

MORR00: Morreale, P., and Terplan, K. 2000. *The CRC Handbook of Modern Telecommunications.* Boca Raton, FL: CRC Press.

NMF95: Network Management Forum. 1995. *Network Management Forum: Discovering OMNIPoint 1 and OMNIPoint 2—A Common Approach to the Integrated Management of Networked Information Systems*. Englewood Cliffs, NJ: Prentice-Hall.

STAL96: Stalling, W. 1996. *SNMP, SNMP2 and RMON—The Practical Guide to Network Management Standards*. Reading, MA: Addison-Wesley Publishing Company.

TERP92: Terplan, K. 1992. *Communication Networks Management,* Second Edition. Englewood Cliffs, NJ: Prentice-Hall.

TOWL95: Towle, T. T. 1995. TMN as Applied to the GSM Network, *IEEE Communications Magazine,* March, 68–73.

TURI08: Turino, J. 2008. Mergers and Acquisitions in the B/OSS and in the Network Management Sectors, B/OSS Billing and OSS World Conference, Chicago Virgo Publishing (Phoenix, AZ), May 1, 2008.

YAMA95: Yamagishi, K. et al. 1995. An Implementation of a TMN-Based SDH Management System in Japan, *IEEE Communications Magazine,* March, 80–88.

3.8 Intelligence Support Systems

Paul Hoffmann and Kornel Terplan

ISS (Intelligence Support System) is about expanded infrastructure requirements of telecommunications service providers (TSP), which are basically no different than OSS/BSS (operations support system)/(business support system) requirements. Intelligence plays two principal roles. On one hand, it is about surveillance through collection of information on illegal activities, such as terrorism, criminality, fraud, money laundering, and the other hand, it provides the basic data that improves the bottom line, such as through revenue assurance, business intelligence, and fighting telecommunications fraud. In short, ISSs are those software elements or units that interface with or are part of billing and ordering, provisioning and authentication systems as well as with outside parties such as low enforcement agencies (Lucas 2003e).

TSP will be used as a generic term throughout the book for a number of different service providers, such as access providers, network operators, communications service providers, electronic communications service providers, and licensed operators for telecommunications services. Terms differ by county, by law enforcement agencies, and by standards for lawful interception.

3.8.1 Positioning Lawful Intercepts and Surveillance

Information and intelligence are two different things. Information, in the context of surveillance, consists of knowledge, data, objects, events, or facts that are sought or observed. It is the raw material from which intelligence is derived (Petersen 2003).

Intelligence is information that has been processed and assessed within a given context. Intelligence includes many categories (Petersen 2003). For the context of this book, communications intelligence

plays the key role. Communications intelligence is derived from communications that are intercepted or derived by an agent other than the expected or intended recipient or which are not known by the sender to be of significance if overheard or intercepted by the agent. Oral or written communications, whether traditional or electronic, are the most common objects of surveillance for communications intelligence, but it may broadly include letters, radio transmissions, e-mail, phone conversations, face-to-face communications, semaphore, sign language, and others. In practice, the original data that form a body of communications intelligence may or may not reach the intended recipient. Data may be intercepted, may reach the recipient at a later date than intended, or may be intercepted, changed, and then forwarded on. However, the definition of communications intelligence does not include the process of relaying delayed or changed information, but rather focuses on intelligence that can be derived from the detection, location, processing, decryption, translation, or interpretation of the information in a social, economic, defense, or other context (Petersen 2003).

Information collection usually supports surveillance activities. Surveillance is the keeping of watch over someone or something. Technological surveillance is the use of technological techniques or devices to aid in detecting attributes, activities, people, trends, or events (Petersen 2003).

There are three typical types of surveillance that are relevant to lawful intercepts:

- Covert surveillance: Surveillance in which the surveillance is not intended to be known to the surveillee. Covert wire traps, hidden cameras, cell phone intercepts, and unauthorized snooping in drawers or correspondence are examples. Most covert surveillance is unlawful and requires special permission, a warrant, or other authorization for its execution. Covert surveillance is commonly used in law enforcement, espionage, and unlawful activities.
- Overt surveillance: Surveillance in which the surveillee has been informed of the nature and the scope of the surveillance. This happens when the telecommunications service provider informs subscribers about the surveillance.
- Clandestine surveillance: Surveillance in which the surveilling system or its functioning is in the open, but is not obvious to the surveillee.

Finally, there are various categories of surveillance devices (Petersen 2003):

1. Acoustic surveillance (audio, infra and ultrasound, sonar)
2. Electromagnetic surveillance (radio, infrared, visible, ultraviolet, X-ray)
3. Biochemical surveillance (chemical, biological, biometrics)
4. Miscellaneous surveillance (magnetic, cryptologic, computer)

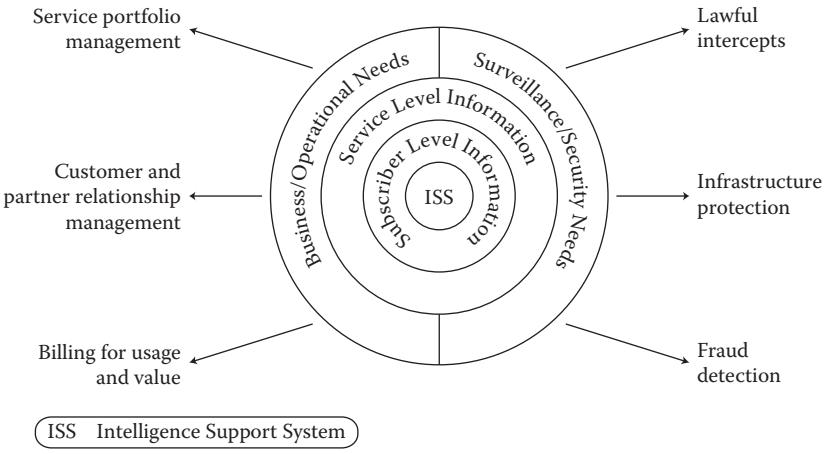
In the context of this book, a combination of surveillance devices in categories 1, 2, and 4 might be used. The appropriate sections will clearly highlight the technologies and devices in use.

3.8.2 ISS Basics and Application Areas

ISS (Intelligence Support Systems) is not about security, but about intelligence. Security provides firewalls, antivirus protection, intrusion detection and prevention; in summary, security is about guarding against loss. Intelligence in ISS is about gathering information about illegal activities and applying that knowledge to increase security where applicable. ISSs are those software elements or units that interface with or are part of billing, ordering, provisioning, and authentication systems as well as interface with or are part of law enforcement systems (LES).

Unlike "point" solutions of intercepts and security that cover small portions of the networking infrastructure, are costly to implement, and may slow down the network, an ISS has a low operational impact, a low cost to operate, and is able to proactively provide intelligence on any size networks. ISSs are feasible today based on the communications technologies and their support systems in use.

All ISS-based processes must ultimately provide comprehensive surveillance in a lawful manner. This includes comprehensive information from any network (e.g., wireline, wireless, access, transport,



FIGu RE 3.8.1 Vision for Intelligence Support Systems.

broadband), on any information, and on any scale. The ISS-based process should provide comprehensive information on a real-time basis providing proactive intelligence.

An ISS provides, in addition to surveillance, general business information on networks for different purposes:

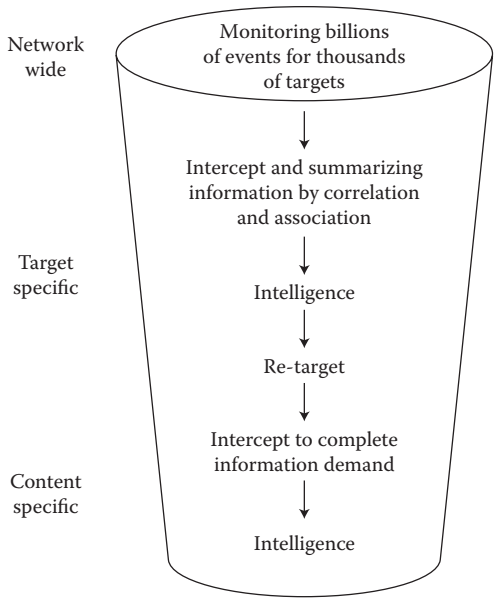
- Service usage information
- Definition and deployment of value-based services
- Subscriber-level information
- Information on network abuse or infrastructure attacks (security)
- Fraud detection
- Layer 4 to 7 information for more accurate value- and/or usage-based billing
- Carrier-grade tier 1 network coverage for any network type (mobile, broadband, backbone)
- Customer and partner relationship management

The vision for ISS is shown in Figure 3.8.1. ISS is the core surrounded by various layers, such as:

- Subscriber-level information (names, addresses, reach numbers)
- Service-level information (metrics, compliance, service portfolios)
- Business and operational needs (usage- and value-based billing, relation management)
- Surveillance and security needs (infrastructure protection, fraud detection)

Proactive intelligence requires that nationwide, even global networks be constructed in such a manner that all communications can be monitored on a grand scale to identify potential targets with summary intelligence while respecting privacy laws. Once these targets have been identified, further monitoring can be done and intelligence can be obtained as lawful authorization is received. An ISS that provides this level of information needs to capture all key summary information in a manner that is lawful and protects the rights of individuals. For instance, an ISS may provide "capture of everyone" that visits a particular suspect Web site but without capturing the individuals' names. The ISS may then take this information and see if any of the IP addresses visiting the Web site have also been communicating via e-mail or chatting with another known target. If so, a legal authorization may then be obtained to look at the individual in more detail.

Government law enforcement agencies have developed the ability to deal with one-dimensional communication on a limited scale. But what is now needed is the ability to deal with multidimensional communication on a global scale if countries are to prevent terrorist attacks and other criminal acts. This can be achieved today with Intelligence Support Systems.



FIGu RE 3.8.2 Monitoring the network for intelligence.

Figure 3.8.2 shows the funnel of data capturing and processing to support law enforcement. The top of the funnel needs to provide summary intelligence information. The middle of the funnel needs to provide intelligence on specific targets. The end of the funnel needs to provide detailed intelligence on specific content.

The recurring steps are:

- Networkwide level
  - Monitoring with billions of events for thousands of targets
  - Summary intercept information
- Target-specific level
  - Collecting intelligence
  - Retarget monitoring
- Content-specific level
  - Complete information demand for intercept
- Restructure intelligence

There are three different types of intelligence as indicated in Figure 1.2: (Cohen 2003).

- Summary intelligence: An ISS that provides this level of information needs to capture all key summary information in a manner that is lawful and protects the rights of individuals. For instance, an ISS may provide "capture of everyone" that visits a particular suspect Web site but without capturing the individuals' names. The ISS may then take this information and see if any of the IP addresses visiting the Web site have also been communicating via e-mail or chatting with another known target. If so, a legal authorization may then be obtained to look at this individual in more detail.
- Target intelligence: Once a target has been identified based on summary intelligence or other information sources and lawful authorization has been received, it may be necessary to look at any and all communications on any network for that particular individual. For instance, an ISS could then look specifically at all types of communication for the individual including e-mail, Web sites visited, chatting, instant messaging, Short Message Service (SMS), Multimedia Message

Service (MMS) messages by mobile phone, VoIP calls by a broadband connection, etc. This information leads then to specific details.

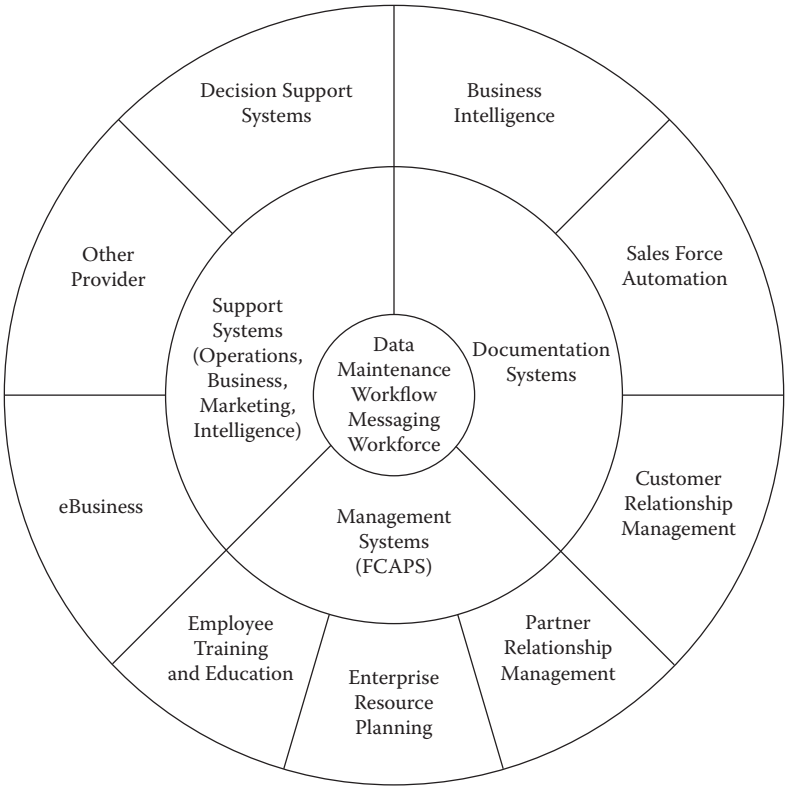
- Content intelligence: At this level it may be necessary to lawfully review the specific content, for example, all the e-mail communication of the lawful target. The ISS should provide the ability to look at this detailed content information by the type of communication (e.g., e-mail, VoIP call, Web site replay, chatting replay, etc.).

3.8.3 ISS Positioning among Other Support and Security Systems

ISS is positioned next to the OSS and BSS of telecommunications service providers. Figure 1.3 shows the structure and hierarchy of the most widely used support, documentation, and management systems together with other important enterprise applications.

Support, documentation, and management systems are not isolated from other service provider business systems. Figure 3.8.3 positions their relationships to each other. This display illustrates the shared role of frameworks at the core, supporting data maintenance, workflow, messaging, and workforce management. Frameworks are also challenged for future applications. They are expected to add value by supporting the following attributes:

- Flexibility to support new communication services, convergence networks, voice, and data
- Adaptability to implement to new pricing schemes (e.g., new services, bundles, new metrics, and thresholds)
- Interoperability with numerous best-of-breed OSS systems and where applicable, existing legacy solutions



FIGu RE 3.8.3 High-level interfaces for service providers.

- Scalability to support rapid carrier growth
- Expediency to facilitate rapid time to market

The outside layer represents many other enabling processes and functions of service providers that have not been addressed explicitly in this publication. These include among others:

- Enterprise Resource Planning (ERP) represents a well-known set of functions and services including asset management, maintenance, general ledger, accounts payable, procurement and purchasing, bill verification, and commissions management.
- Customer Relationship Management (CRM) offers emerging customer-facing services including trouble management, account management, cross-selling, bill inquiry, and bill adjustment functions.
- Partner Relationship Management (PRM) represents a new emerging area for well-organized collaboration among business partners. A highly flexible infrastructure is required to support mergers and acquisitions, and the various depths of partnerships.
- Sales force automation represents a new emerging area of account, sales force, opportunity, contract and contact management.
- Business intelligence (BI) is a special area of providing tailored business rules for operations metrics, SLA management, data warehousing, product management, marketing, and CRM.
- Decision Support Systems (DSS) are based on business intelligence, and business rules are implemented for a higher-level of automation, in particular to operate the underlying network infrastructure.
- Support of eBusiness: This new and emerging area could play a significant, even survival role for service provider. It includes Web-based order entry and returns, Web-based trouble reporting and status inquiries, electronic bill delivery and payment, and Web-based customer profile and product information. This is the basis for business-to-business (B2B) and business-to-consumer (B2C) on behalf of service providers.
- Interconnection between multiple service providers: All the technologies implemented with enterprises, small businesses, and residential customers may be implemented among multiple providers, retailers, and wholesalers. Besides traditional techniques for supporting settlements, eCommerce techniques are also expected to be implemented.
- Employee training and education: Besides powerful workforce solutions to support workforce dispatch optimization, training, education, and cross-education of all employees of the service provider are extremely important. This includes knowledge distribution about the service portfolio, sales techniques, support systems, documentation systems, management systems, basic financials, and the strategic position of the company among other competitors. State-of-the-art Internet-based technologies may help to increase educational efficiency.

It is obvious that ISSs have tight connections to:

- Mediation systems
- Inventory and documentation systems
- Provisioning solutions
- Billing products

These systems and products are the core components of OSSs and BSSs.

Security management is considered as part of management systems. Usually, TSPs structure their management solutions around FCAPS (fault, configuration, accounting, performance, and security management), whereby accounting and security receive special attention. Accounting is becoming part of OSSs with the exception of capturing raw data in networking equipment. Security is supported in different areas by different subject matter experts. Subgroups may include:

- Securing the networking infrastructure using, among others, security frameworks, intrusion detection and prevention, firewalls, and virus protection.
- Protecting customer privacy of TSPs.

- Securing links to partners and other service providers.
- Authentication and authorization of the own employees.

3.8.4 Basic Requirements for Lawful Intercepts

Telecommunications service providers are being asked to meet lawful intercept requirements for voice, data, and video in a varieties of countries worldwide. The requirements vary from country to country but some requirements remain common even though details such as delivery formats may differ. Baker (2003) gives an excellent basis for streamlining requirements.

Generic strategic requirements are somehow contradictory, based on "more access for less money":

- Telecommunications service providers need return on investment (ROI) for their ISS deployment
- Government agencies need information, but do not have ready access to networks
- Telecommunications service providers need systems that fit business requirements without undue burden
- Governments need cost-effective solutions with economies of scale
- Telecommunications service providers and governments need to address privacy challenges (e.g., separate content from signal)

Generic functional requirements include:

- Comprehensive IP monitoring
- Scalable, tier 1 networks
- Any data, any network (mobile, broadband, access, backbone transport)
- Leverage commercial off-the-shelf software (COTS)
- Availability of real-time information
- Business or surveillance policy enforcement

Generic legal requirements include:

- Lawful Intercepts (LI) must be undetectable by the intercept subject.
- Mechanisms must be in place to limit unauthorized personnel from performing or knowing about lawfully authorized intercepts.
- If multiple law enforcement agencies (LEAs) are intercepting the same subject, they must not be aware of each other.
- There is often a requirement to provide intercept-related information (IRI) separately from the actual content of interest.
- If IRI is delivered separately from content, there must be some means to correlate the IRI and the content with each other.
- If the information being intercepted is encrypted by the telecommunications service provider and the provider has access to the keys, then the information must be decrypted before delivery to the LEA or the encryption keys must be passed to the LEA to allow them to decrypt the information.
- If the information being intercepted is encrypted by the intercept subject and its associate and the service provider has access to the keys, then the telecommunications service provider may deliver the keys to the LEA.

In terms of requests by LEAs, there are four fundamental types:

- Past billing and statistical traffic records of communications: These records must be maintained by telecommunications service providers for a certain period of time. This duration depends on the country. Usually, there are strict guidelines about the storage media with the result that TSPs may innovate their billing systems (e.g., Electronic Bill Presentment and Payment [EBPP]) and storage devices without violating any data-retention rules.

- Contents of computer long-term storage: LEAs can usually search and/or seize computers and storage media. Even damaged devices may be searched. This falls into computer forensics. It is a combination of science and art. LEAs today process software tools that search PCs, servers, and networks for evidence, such as text files, images, and e-mails that can be used to ferret out criminals. Requests by LEAs can be triggered by noncompliance issues, such as violation of company policies, circulation of inappropriate content, or misappropriation of information.
- Current billing and statistical records of communications, desirable in real time. This includes "pen register" and/or "trap and trace" data, usually defined under IRI (intercept-related information).
- Delivery of content: This includes the collection of the full content of communications, which is offered by telecommunications service providers as communication and not as an information service.

3.8.5 Principal Functions of Interception

The principal functions of accessing, delivering, and collecting data, which must be supported by all service providers, are summarized here. Identifying basic functions as early as possible will facilitate the allocation of functions to various standards.

3.8.5.1 Functional Group: Accessing Data

The following functions are important:

- Review technology to be intercepted
- Identify network elements, such as multiplexers, switches, routers, load balancers, and others with built-in capturing features
- Check collaboration willingness of manufacturers of network elements
- Identify Incept Access Points (IAPs) for both IRI and content
- Provision IAPs for both IRI and content
- Provision data channels to central processing (or monitoring) facility
- Select and train human resources

3.8.5.2 Functional Group: Delivering Data

The following functions are important:

- Evaluate processing needs
- Evaluate mediation solutions
- Decide about build or buy processing applications
- Select applications in case of buy decision
- Quantify storage requirements
- Define interfaces to data sources
- Define interfaces to hand-over points
- Maintain data, information, and intelligence
- Provision data channels from/to LEAs
- Secure data channels from/to LEAs
- Select and train human resources

3.8.5.3 Functional Group: Collecting Data

The following functions are important:

- Define process to authenticate requests by LEAs
- Select correct information and/or intelligence to be transferred to LEAs
- Select and train human resources

These lists of principal functions should have a high level of matches with lawful intercept functions as defined and recommended by the standards bodies of various continents.

3.8.5 Reference Models for Lawful Intercepts

3.8.6.1 CALEA Reference Model with the J-STD-025 Standard

CALEA definitions are related to the Telecommunications Act. The most important ones are extracted as follows:

Telecommunications: the transmission, between or among points specified by the user, of information of the user's choosing, without change in the form or content of the information as sent and received.

Telecommunications service: the offering of the telecommunications for a fee directly to the public, or to such classes of users as to be effectively available directly to the public, regardless of the facilities used.

Information services: the offering of a capability for generating, storing, transforming, retrieving, utilizing, or making available information via telecommunications, and includes electronic publishing, but does not include any use of any such capability for the management, control, or operation of a telecommunications system or the management of a telecommunications service.

Telecommunication carrier (The term telecommunications service provider is preferred by the authors in this handbook.): entity engaged in the transmission or switching of wire or electronic communications as a common carrier for hire.

- includes commercial mobile radio service
- includes an entity engaged in the transmission or switching of wire or electronic communications to the extent that the FCC finds that such service is a replacement for a substantial portion of the local telephone exchange service and that is in the public interest to deem (it) to be a telecommunications carrier for (these) purposes
- does not include entities insofar as they are engaged in providing information services

Call-identifying information: Dialing or signaling information that identifies the origin, direction, destination or termination of each communication generated by means of any equipment, facility, service or a telecommunications carrier.

3.8.6.1.1 CALEA Interfaces

The principal interfaces with the CALEA reference model are:

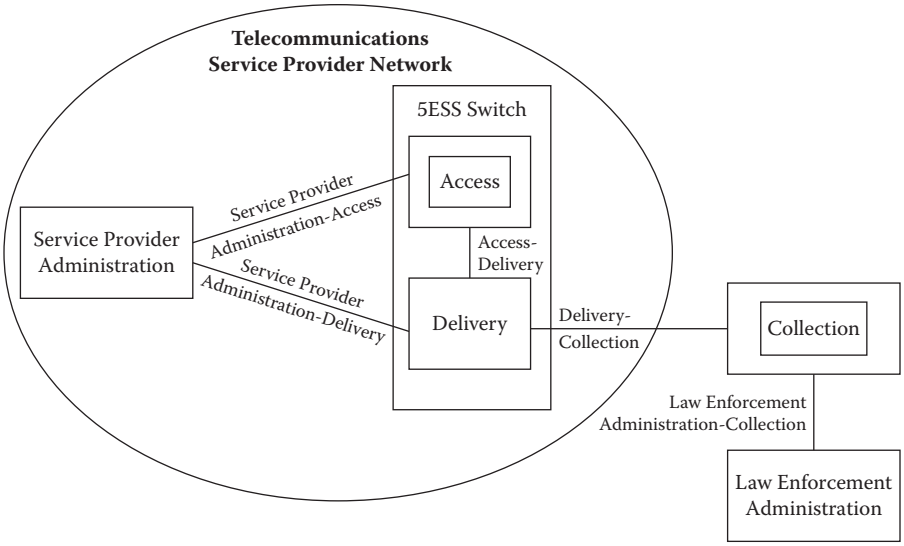
- Surveillance Administration System (SAS): Performs subject provisioning and receive alarms related to CALEA interfaces
- Call Data Channel (CDC): A network connection reporting call-identifying information—CDC messages—from the switch to the LEA
- Call Content Channel (CCC): A network connection delivering call content from the switch to the LEA

The reference model is shown in Figure 3.8.4. It offers a generic view of the lawful intercept architecture (access, delivery, and collection functions).

3.8.6.1.2 CALEA Principal Functions

There are basically three principal functions:

- Access Functions (AF)
  - Network elements (CO switches, MSC, HLR, AAA, PDSN, SGSN, GGSN, routers, trunking gateways, soft switches, CMTS) that provide access to and replication of intercepted traffic
  - Sniffers and splitters that can passively monitor network traffic



FIGu RE 3.8.4 CALEA reference model.

- Delivery Function (DF)
  - Database of target and warrant information
  - Provisioning interface
  - Proprietary interfaces to AFs
  - Standards based (J-STD-025, ETSI, TIIT, PacketCable) delivery of intercepted traffic to CFs
  - Runs on off the shelf UNIX based platforms and programmable switch machines
- Collection Function (CF)
- Collects and records lawfully authorized intercepted communications (e.g., call content) and call-identifying information for law enforcement agencies

Figure 3.8.5 shows from another perspective the principal functions and interfaces of the lawful intercept architecture.

3.8.6.2 ETSI Reference Model for the European Community

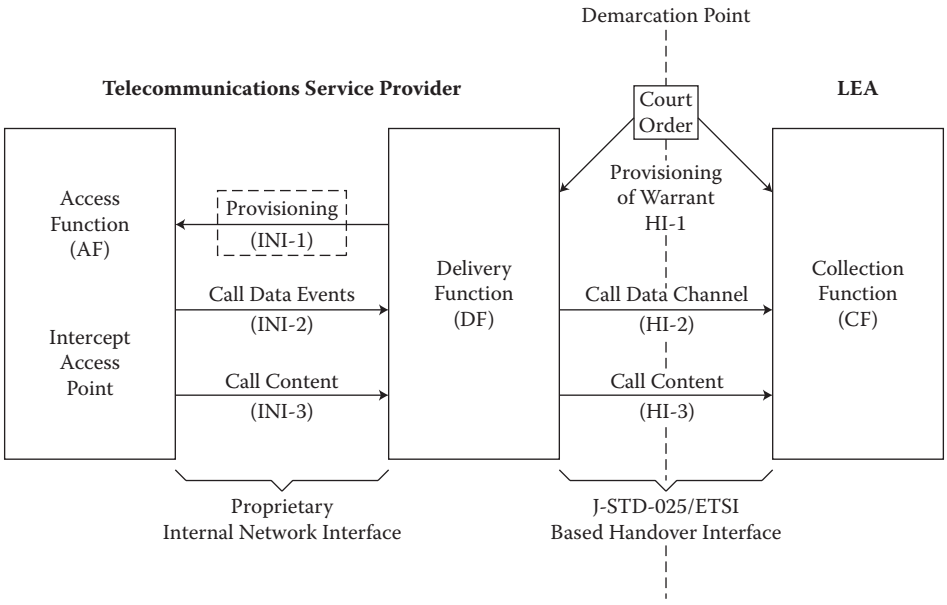
The European Telecommunications Standard Institute (ETSI) definition of lawful interception is the following: Interception is an action based on the law, performed by a network operator or access provider or service provider (NOW/AP/SVP, the term telecommunications service provider is preferred by the authors in this book), of making available certain information and providing that information to a law enforcement monitoring facility.

Figure 3.8.6 shows the basic flow of information exchange between service providers and LEAs.

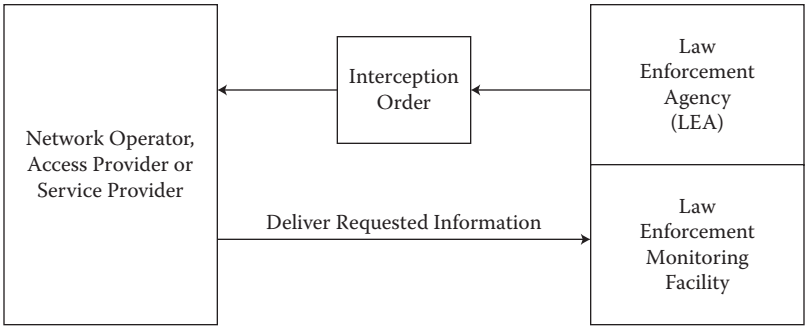
3.8.6.2.1 Basics of This Standard

The chosen solution to the requirements of the LEAs is a three-ported interface. Such an interface structure is shown in Figure 3.8.7.

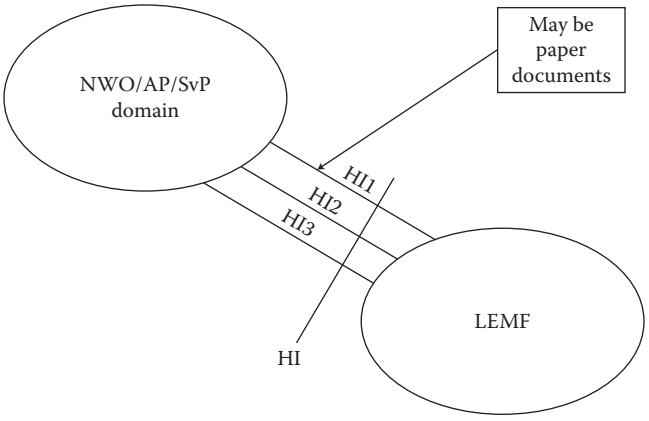
The first HI port shall transport various kinds of administrative information from/to LEA and NOW/AP/SVP. There shall be a complete separation between the administrative interface (HI1) and the technical interfaces (HI2 and HI3) of the NOW/AP/SVP, in order not to give the Law Enforcement Monitoring Facility (LEMF) the possibility to establish or modify an interception without an action of a mandated agent of the NOW/AP/SVP. In case of a nonautomatic administrative interaction, this interface may also be manual rather than electronic.



FIGu RE 3.8.5 Generic view of the LI architecture.



FIGu RE 3.8.6 Basic information exchange.

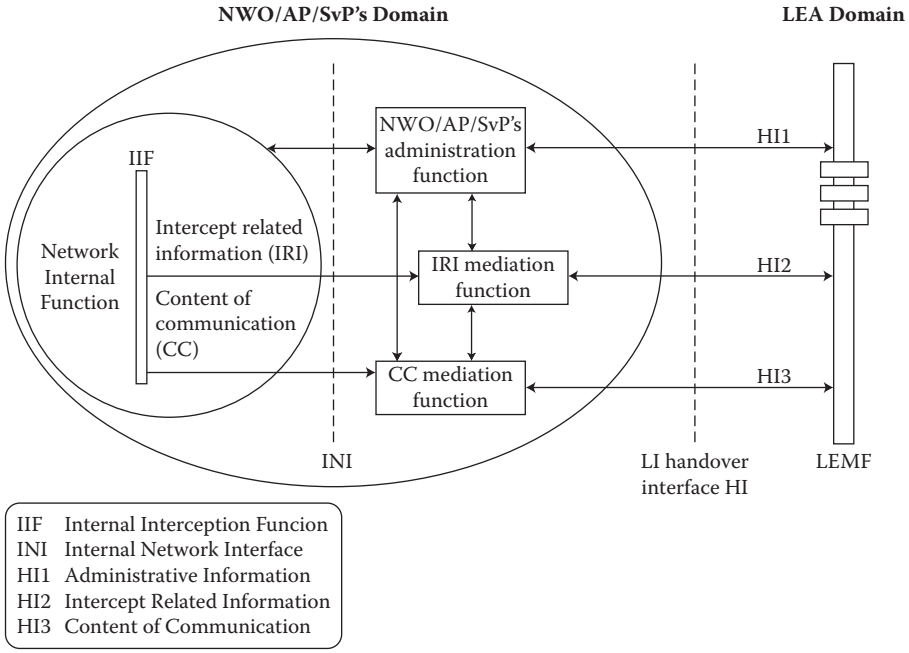


FIGu RE 3.8.7 Diagram showing hand-over interfaces between NWO/AP/SVP and LEMF.

The functional modules include the following domains:

- Telecommunications service provider domain:
  - Network internal functions
  - Internal interception function (IIF)
  - Administrative function of telecommunications service providers
  - IRI mediation function for the intercept-related information
  - CC mediation function for the content of information
- LEA domain:
  - Law enforcement monitoring function
  - Law enforcement analysis function
- This reference model supports various interfaces:
  - INI = Internal network interface
  - HI1 = administrative information
  - HI2 = intercept-related information
  - HI3 = content of communications
- The HI2 and HI3 logical ports could, for example, be physically mapped to:
  - a single channel-oriented channel
  - a single packet-oriented channel
  - several circuit-oriented channels
  - several packet-oriented channels
  - several circuit-oriented channels and one or more packet-oriented channels

The functional block diagram is shown in Figure 3.8.8. The functional components are defined in Table 3.8.1.



FIGu RE 3.8.8 NWO/AP/SvP functional block diagram showing handover interfaces.

TABl E 3.8.1 Functional Block Diagram Components

Component	Description
IIF	An IIF within the NWO/AP/SVP domain. There may be more than one IIF involved in the provision of interception.
INI	An INI within the NWO/AP/SVP domain, which exists between an IIF and the mediation function
NOW/AP/SVP Administration Center	The administration center contacted via the port HI1 (which may be partly electronic and partly paper based depending on circumstances) is used to set up the interception action on the LEA request.
Mediation Function	A function that selects sequences and transforms information, including CC when necessary, between a number of IIFs and the HI. Sometimes the mediation function may be a null function, e.g., direct delivery of CC to the LEMF via HI3 with no changes.
Delivery Mechanism to LEA/LEMF	Intercept requests, status, and alarm reports are transmitted between the administration center and the LEA/LEMF IRI is transmitted through the mediation function (may be transparent) to the LEMF CC is transmitted through the mediation function (may be transparent) to the LEMF

3.8.6.2.2 Hand-Over Interfaces

3.8.6.2.2.1 HI1: Interface for Administrative Information The HI1 shall transport all kinds of administrative information from/to the LEA and NWO/AP/SVP. This port shall be used for the transmission of the request to establish or to remove the interception action from the LEA to the NWO/AP/SVP and the acknowledgment message back to the LEA. The transmission between these parties should support manual and/or electronic transmission from/to the LEMF and the NWO/AP/SVP facility.

The status report should cover all kinds of alarms, reports, or information related to the intercept function. The status reports and the alarm reports are transmitted via HI1 to the LEMF or LEA if necessary. Alarms being not specific for a certain target identity can be received by all LEAs, other alarms (e.g., LEMF busy or no answer from LEMF) should only be transmitted to a specific LEA to which the alarms apply.

The generic status reports can typically be:

- Target identity removed from service
- Target identity has changed within the network
- Bulk modification of subscriber numbers
- Individual modification of subscriber numbers
- New multiple subscriber number (MSN) creation
- LI database lost (e.g., software replacement, recovery, fall back)
- General setup failure

Status reports indicating transmission problems between NWO/AP/SVP and the LEMF can typically be:

- Transmission problems to LEMF
- LEMF is busy
- No answer from LEMF

3.8.6.2.2.2 HI2: Interface for IRI HI2 shall transport all IRI. This interface shall be used to transmit information or data associated with the telecommunications services of the target identity apparent to the network. It includes signaling information used to establish the telecommunication service and to control its progress (e.g., target identification, identification of the other parties of a communication, basic services used, direction of the call or the event, answer identification and/or release causes, and

TABl E 3.8.2 Possible Correlation Mechanisms

Group	CC (Communication Content)	IRI (Interception-Related Information)
A	Time of arrival of call at LEMF	Time stamp in information record
B	Unique number sent in an associated signaling channel	Unique number in information record
B	LEMF address	LEMF address in information record
D	Particular physical channel	Particular physical channel

time stamps). If available, further information, such as supplementary service information or location information may be included.

Sending on of the IRI to the LEMF shall in general take place as soon as possible, usually in the range of a few seconds. In exceptional cases (e.g., data link problems), the IRI may be buffered for later transmission for a specified period of time.

IRI shall be structured as a sequence of records. Record types are:

1. IRI-BEGIN: record at the first event of a communication attempt, opening the IRI transaction
2. IRI-END: record at the end of a communication or communication attempt, closing the IRI transaction
3. IRI-CONTINUE: record at any time during a communication or communication attempt within the IRI transaction
4. IRI-REPORT: record used in general for noncommunication-related events

IRI will be transmitted to the LEMF with no translation of information content. This has the following advantages:

- There is minimum of translation to be kept up to date
- The mediation functionality is minimized
- The amendment required when introducing new services is minimized

But information may require enveloping before being passed to the LEMF.

3.8.6.2.2.3 HI3: Interface for Communication Content (CC) The port HI3 shall transport the CC of the intercepted telecommunication service to the LEMF. The CC shall be presented as a transparent clear copy of the information flow during an established, frequently bidirectional communication of the interception subject. It may contain voice and data.

The transmission media used to support the HI3 port will usually be those associated with a telecommunications network or its access arrangements. In cases of failure, CC is lost. The network does not provide any recording function.

3.8.6.2.2.4 Correlation of HI2 and HI3 When an HI3 port is established, the target identification of the target identity shall be passed across to enable the LEMF to correlate the CC on HI3 with the IRI on HI2.

In situations, where an LEMF may be connected to more than one source of the result of interception, it is necessary to ensure reliable correlation between the CC and IRI. Several mechanisms used at the same time will ensure correct correlation. Possible mechanisms are given in Table 3.8.2. The use of a given circumstance will be dependent on national rules and technical considerations.

3.8.6.2.3 Security Recommendations from ETSI

There is a general requirement that the operation of interception facilities should be discreet, confidential, and efficient. For prevention of unauthorized administration, as well as unauthorized use, appropriate security features are necessary. General security requirements are:

- A security management system should be established.

- There should be physical and logical access controls.
- Any necessary key, passwords, and user identifications for the authorization and the logical access to the interception function should be securely stored.
- Any transmission of passwords and user identifications for access to interception functions should be secure.
- Physical interfaces should be secured mechanically and/or logically against unauthorized use.

Transmission of all information between the NOW/AP/SVP and the LEMF across HI1, HI2, and HI3 shall be confidential. During communication between systems that are not based on leased lines, appropriate mechanisms should ensure that the recipient is in the position to verify or authenticate the identity of the sender while the connection is set up. During communication between systems that are not based on leased lines, appropriate mechanisms should ensure that the sender is in the position to verify or authenticate the identity of the recipient at the start of a connection.

Only specifically authorized personnel should be able to control interceptions. In general, the LEA should not have any direct access to any network element.

The entire communication between the administration system and the interception function should be confidential. All internal interfaces must be secured.

The interception functions shall be implemented in such a manner that:

- The interception subject and his correspondents do not know that a lawful interception is active.
- During the intercepted communication itself, the quality of the communication shall remain the same as usual and the service shall be unchanged, including all supplementary services such as call forwarding, etc.
- When there is no intercepted communication, the quality of communication shall remain the same and the service shall be unchanged; in other words, there is no modification to services supplied or information received either by the interception subject or by some other party.

An employee of NWO/AP/SVP who has been duly authorized may be permitted to know that interception is in progress, or that a subscriber is an interception target. But an employee of NWO/AP/SVP who has not been duly authorized may not be permitted to know that interception is in progress, or that a subscriber is an interception target.

3.8.7 Principles of Monitoring and Intercepts (Hardware and Software Probes)

State-of-the-art technology permits monitoring of everything in networking infrastructures. But everything has a price tag. The following challenges must be addressed by telecommunications service providers and LEAs:

- How to identify dynamic targets
- How to deal with roaming in/out subscribers
- How to intercept compressed and encrypted traffic
- How to capture a call in progress
- How to meet real time constraints
- How to deal with identity management
- How to identify target locations
- How to identify prepaid targets

Lawful intercepts require full monitoring capabilities in networks. This requirement causes several concerns on behalf of the telecommunications service providers:

- Optimizing the combination of internal and external lawful intercept capabilities

- Data throughput is growing exponentially requiring the monitoring of high-bandwidth channels
- Packets to be monitored and intercepted packets must not be lost
- Highly distributing IAPs is very expensive
- Real-time data association and correlation need high-speed processing power when multiple sources must be considered

3.8.7.1 Internal and External Lawful Interception

Depending on accessibility to network system components, LEAs request IP interception through processes internal or external to the networks that presumably support the traffic and applications of a target under surveillance.

Internal interception enables the LEA, via the mediation platform and hand-over interfaces to extract interception-related information (IRI, otherwise known as call data) and the target's content data directly from application servers (e.g., e-mail, Web, chat), network access systems (e.g., RADIUS server system), DSL/cable modem termination points, routers, switches, etc., which are all part of the NWO's or SP's infrastructure. Internal interception of application platforms has the obvious advantage of *directly* delivering target data to the mediation platform because the application is inherently known, and the interception data are explicitly provided. Interception of internal network transport elements also narrows the network traffic originating from or going to specific targets. Common Wi-Fi network "sniffing" is, in effect, a form of internal interception since it focuses on a specific wireless LAN with highly localized targets, that is, the targeted users contained within the coverage zone of the wireless base station.

Nevertheless, internal interception carries two strong assumptions that might not be valid. First, we assume that targeted IRI and content data from selected network and applications systems are available to the LEA, perhaps as mandated by local/national regulations. Second, the network and applications systems must support secure data paths to the mediation platform (e.g., mail servers must output targeted header and content information directly to the interception mediation platform). However, such assumptions may not hold. In many developed countries, ISPs are often reluctant to open their networks to LEAs without considerable legal fighting; hence, their operations are not readily adaptable to systematic lawful interception. Perhaps even more problematic are the current applications systems in place, which by their design and implementation are not readily conducive to interception. For example, most e-mail servers for handling large volumes of e-mail still must be modified if they are to provide systematic delivery of targeted IRI and content through purpose-built ports dedicated to interception data conveyance. This is not a trivial undertaking, especially when interception ports have to also accommodate requisite network security to protect the transport of interception data and prevent "back-door" attacks into the system. Finally, mechanisms must be in place to prevent potential targets from detecting that their data flows are being intercepted; this implied need for secure application design.

When the availability of internal interception fails, or when LEAs desire to conduct clandestine surveillance, interception needs to take place at network levels outside the realm of the target's immediate application service or network provider. In other words, external interception must be performed. Such interception is performed on Internet circuits outside the target's immediate network, typically at adjacent networks or major public network concentration points. The core equipment typically consists of a probe made of a physical tap and/or a router with filtering capabilities. This probe typically replicates traffic flow through a network point at the physical layer; the filter targets packets containing specified IP addresses or IP address ranges and routes them to a port dedicated to interception purposes. From there, packets are routed to the mediation platform and ultimately to the LEA for analysis of datagram headers and content. Systems that perform external interception tend to be sophisticated and not officially publicized. Where traffic is light, open source programs can assist in analyzing the protocols and content of data traversing a given path.

Targets must not be able to know that they are the subject of surveillance. Minimally sophisticated targets could at least suspect that interception of some kind is underway through:

- Trace route commands. These display the router hops that a subject's Internet traffic traverses to/from a given destination. Any change from the ordinary could imply the introduction of an interception router or other device. However, the proper use of interception probes can avoid the introduction on new router hops.
- Unusual signaling activity in their modem, voice-over-IP interface box, or other hardware. These devices carry important identification and traffic information associated with the user, but can reveal interception activity to the interception target. Therefore it is not recommended that LI probe customer premises equipment (CPE); this process poses risks for the LEAs especially when the devices are tampered with by the users.
- Degradation or interruptions of service. These are obvious factors in arousing suspicion by the targets that surveillance might be taking place.

3.8.7.2 Access Function Implementation Approaches

The basic choices for access are:

- Network or service element as data source
- Probes as data source

If the network or service element is the choice, the following issues should be addressed in greater detail:

- Restricted in location within the network to where access, routing, or service is performed
- May be limited to seeing only compressed/encrypted traffic
- May require interception in multiple elements
- May require sophisticated association data exchange
- Provisioning and delivery require multiple different interfaces

In addition, any additional function, hardware or software, may impact the network or service element in delivering the expected performance. Observations show that incorporating data collection functions for usage-based billing onto routers, causes their performance to be significantly impacted.

The issues with probes are:

- Potential for reuse for other applications
- An additional non-service-element device to put in the network
- More flexibly able to comply with future requirements

3.8.7.3 Use of Probes

When the decision is for probes, four questions should be addressed:

- Active or passive probes
- Software or hardware probes
- Dedicated or shared probes
- Flow-based analysis probes

3.8.7.3.1 Active versus Passive Probes

When an active probe becomes part of the network, the consequences are:

- Can be guaranteed to capture all traffic that flows through it
- Must be costly to deploy as it must be engineered to avoid impact on the network
- Alternatively, impact on the network is likely affecting reliability, latency and jitter

A passive probe is just listening to the traffic. Even in this passive role, sufficient speed of processing is required. The consequences are:

- No impact on service network and consumers
- Requires statistical method to prove reliability

3.8.7.3.2 Software versus Hardware Probes

Not only in this case, but in general: software is more flexible with some overhead, but hardware is faster without overhead. If software is the choice, the consequences are:

- Easier to reconfigure and to add new capabilities such as
  - Extraction of content intercept information
  - Decoding of tunnels
  - New protocol metadata
  - Potential for recycling
- If hardware is preferred, the consequences are:
  - Can the probe scale much more cost effectively?
  - Has very limited upgrade path
  - Potential for better availability/reliability
- The only viable solution for active probes

Service MONitoring (SMON) is just a simple way of standardizing the controlling of port-mirroring sessions in a switched environment. In the hub environment, port-mirroring is not necessary because every LAN connection receives all the data. In a switched environment, however, switch vendors have been implementing port-mirroring, which allows the switch to copy all the data from specific ports to a monitoring device in addition to forwarding to the real targeted destination. The way to control port-mirroring is mostly vendor specific. SMON allows a standard, SNMP-based method for setting up and clearing port-mirroring sessions. While this has been implemented by various vendors to support SNMP-based port-mirroring on the switch and SMON-based port-mirroring control from the management station, the generic monitoring market did not change significantly.

3.8.7.3.3 Dedicated versus Shared Probes

Dedicated probes with single functionality are great performers. They will definitively fulfill the lawful intercept job, but cost justification remains very problematic.

When shared among multiple functionalities toward a "mediation" probe, the consequences are:

- Investment can be leveraged
- High risk of impact across application boundary
- Security risk

But ISS requires sharing up to a certain extent. The priorities must be set by telecommunications service providers.

3.8.7.3.4 Flow-Based Analysis Probes

Flow-based analysis is a rather interesting alternative or complementary solution to probe-based network analysis. While typically flow-based analysis lacks the granularity of the potential deep packet analysis in probe-based solutions, it can cover a much larger infrastructure at a lower cost.

One would argue that in a highly meshed networking environment, probe-based traffic analysis is very expensive, and flow-based analysis can scale much better at reasonable costs. Networking hardware vendors can present the statistics in any way they prefer, while the probe vendor can offer vendor-independent traffic statistics.

Probe vendors try to integrate flow-based solutions into their products to provide a combined solution instead of being forced out of the monitoring market. Some vendors go as far as claiming that a large number of probes are necessary to be able to efficiently process a large amount of traffic flow information.

It is proven that flow-based statistics result in large amounts of data—approximately 20 gigabytes per month on a 100,000-port network—which is still less than the amount of data that NetFlow generates for the same environment. The ratio is expected to be 50 to 1. Collection of flow statistics for large networking infrastructures may require a distributed management architecture. Non-probe vendors have proved

that flow-based statistics may be collected and presented using one collection station for up to 500 router/switch interfaces. Probe vendors work usually with one probe for up to 16 router/switch interfaces.

The most prominent management software vendor for sFlow is InMon. The most prominent hardware vendor supporting sFlow is Foundry in cooperation with Hewlett-Packard. NetFlow is implemented and maintained by Cisco.

The IP Flow Information Export (IPFIX) protocol is a template-based flow reporting method that supports flow aggregation, quality of service (QoS), Border Gateway Protocol (BGP) next hop, VLANs, multicast, network address translation (NAT), Multiprotocol Label Switching (MPLS), and IPv6 among others. The Internet Engineering Task Force (IETF) is getting closer ratifying IPFIX as a new standard for flow reporting. This standardization process is highly supported and driven by Cisco due to the fact that IPFIX is based on a high NetFlow version.

The biggest challenge with all flow-based solutions is reporting, assuming that data can be collected without impacting performance. For lawful intercepts, it will be key to find the right data reduction and information reporting solutions.

Although the IETF sFlow draft standard has been available for some time, few vendors have implemented it. But as network traffic speeds grow to Gigabit and to 10 G in some infrastructures, sFlow will become a more important technology for tracking network performance and providing network security.

sFlow is a technology that uses random sampling of LAN's and WAN's data packet flows across an entire network to give users a detailed, real-time view of network traffic performance, trends, and problems. sFlow is deployed through network management information bases (MIBs)—either hardware- or software-based agents—running on the actual switches and routers in the network. This allows for a broader picture of network performance. sFlow backers say that monitoring happens on every port of every sFlow-enabled switch, rather than on just the port or segment to which a probe is attached. Proponents of sFlow say that the technology allows for more widespread network monitoring because mirroring every port would be difficult and expensive for both network staff and LAN bandwidth. Up to half a switch or router would have to be dedicated to port mirroring to achieve this (Hochmuth 2004).

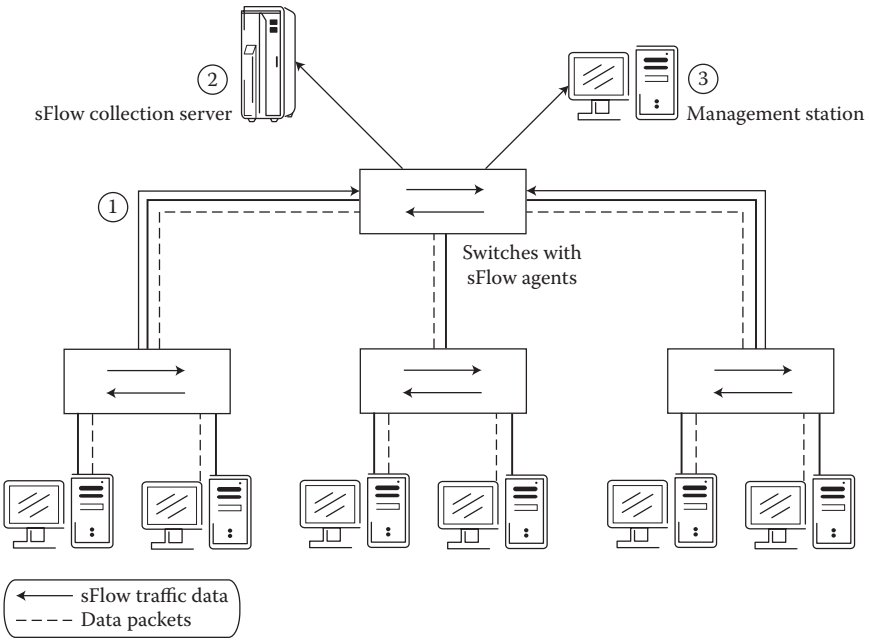
Figure 3.8.9 shows the principal components of sFlow solutions. The most important functions are:

1. Switches with sFlow agents take random samples of traffic from all ports on the switch.
2. Sample data is sent to an sFlow collection server, where sFlow samples from the network are calculated.
3. Management workstations can tap the sFlow server to view an overall picture.

Instead of capturing and logging every packet on a switch or router port, sFlow MIBs take random samples of packets traveling through ports. These so-called sFlow datagrams are forwarded to an sFlow collection server on a network—a DCN or the production network of the service provider. On this box the datagrams are run through an algorithm that generates a complete model of network traffic based on the sampled data. The technology behind sFlow was developed jointly by engineers at InMon, a manufacturer of switch monitoring software, and developers at HP and Foundry Networks. Support of sFlow is included in products such as HP OpenView, nGenius Performance Manager from NetScout, and Traffic Server from InMon.

In addition to providing real-time snapshots of network performance, sFlow can be used as a network security tool. An example is in the direction of unauthorized network devices acting as network address translation (NAT) boxes. This could include a commodity NAT-enabled wireless router. While NAT devices attached to a network might appear as legitimate end nodes, these could serve as backdoors, allowing access to unauthorized connections from wired or wireless users. Because sFlow samples traffic from every port in a network, sFlow data analyzers can identify nodes that are acting as NAT devices on a network by comparing subnet data among switches/routers and NAT devices.

Lawful intercept requires more than sFlow-based techniques may provide. But, in particular for strategic surveillance, statistical techniques are very useful. Based on suspicious traffic flows, on-the-fly



FIGu RE 3.8.9 Functional overview of sFlow.

provisioning may help to activate monitoring solutions that are able to provide both intercept-related and content-related data for LEAs.

Prior to selection of the optimal solution for lawful intercepts, the following questions should be answered:

- Does the network/service element really do the functions or are they delegated to a probe?
- Does the support in the network/service element meet all the legal requirements?
- Does the (active) solution affect quality of service for all users?
- Does the solution prevent packet loss?
- Is security really maintained across application boundaries for multiuse deployments?
- Can a software solution ultimately compete with a hardware-centric solution?
- Can multiple users really use the IAP or can the impact adversely affect other applications unintentionally? Can this be avoided?

3.8.7.4 Intelligence Transmission

Telecommunications service providers usually operate a separate data communication network (DCN). If so, intelligence data may be routed to this network and sent to the hand-over interface with LEAs. This out-of-band solution has the following attributes:

- Implement out-of-band infrastructure with signal splitters
- User proximity improves selectivity in terms of dynamic address changes and multicast
- Supports heterogeneous vendor environments
- Sometimes preferred for operational isolation

If no DCN is available, traffic sharing cannot be avoided. This in-band solution has got the following attributes:

- Use existing network elements
- Independent of access link technology, such as POTS, ISDN, xDSL, cable, wireless

- Provides cost reduction, implementation speed
- Preferred when above criteria are relevant

In case of in-band, traffic separation should be supported. Traffic shapers could take this responsibility. A dedicated band is always preferable for better security solutions.

3.8.8 Receiver Applications

LEAs or outsourcers are the receivers. They are expected to take advantage of collected data. It is assumed that they may access all necessary data collected and maintained by telecommunications service providers. The nature of applications under consideration, the time of activating, executing and evaluating data, and interpreting results are very different. Both targeted surveillance and strategic observations are supported. This is a very dynamically changing area with many new applications. But in most cases, suppliers of such applications do not market them aggressively. They are rather interested in playing the role of a professional outsourcer to LEAs.

3.8.8.1 Support for Recognizing Criminal Activities

3.8.8.1.1 Search for Criminal Activities

This activity belongs to strategic surveillance due to the fact that there are no specific targets at the beginning. The observation of communication directions and paths, and evaluating contents of various applications may narrow down potential targets. Altogether, large amounts of data must be evaluated, filtered, analyzed, sorted, classified, and selected in order to find conclusions about individuals, groups, locations, and suspect activities.

3.8.8.1.2 Communication Analysis

In order to collect actionable intelligence, in-depth analysis of communication activities is required. It does involve the correlation of time stamps, locations, communication relationships, authentications, directions, communications forms, and volumes. As techniques for support, location tracking, geographical information systems, and data mining are under consideration.

3.8.8.1.3 Content Analysis

In order to collect actionable intelligence, in-depth analysis of communicated content is necessary. It does involve the analysis and correlation of application identification, language recognition, speaker recognition and identification, word spotting, topic recognition, optical character recognition, logo recognition, and image recognition. Both text-based and audio-based analysis is frequently used.

3.8.8.1.4 Automated Intelligence Support

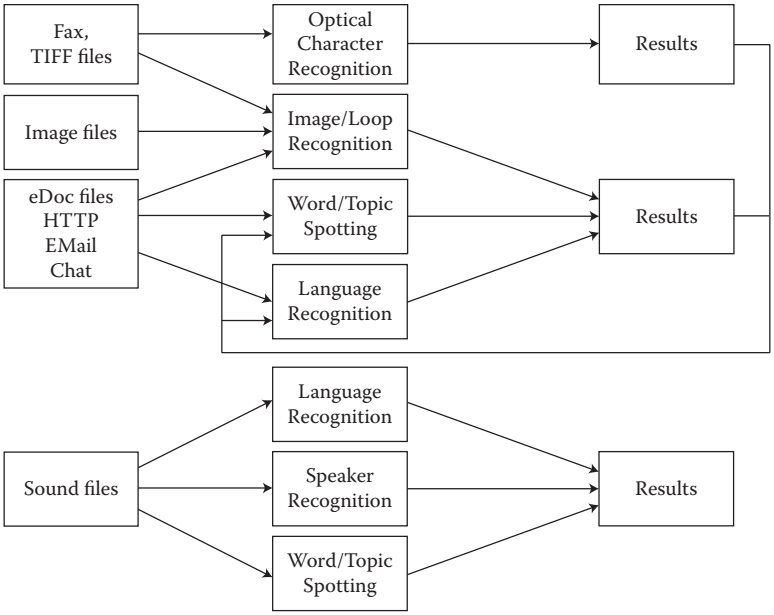
Content is created from both text- and audio-based documentation elements. Results are generated by combining and correlating multiple inputs, such as faxes, TIFF files, image files, eDoc files, HTTP, e-mail, chat, and sound files. Figure 3.8.10 shows such a combination and correlation example (Axland 2004).

In a further step, languages and speakers may be recognized by using combined intelligence.

3.8.8.2 Analysis Procedures and Tools

Intercepting and collecting communications is just the beginning. This raw information must be analyzed and turned into actionable intelligence that can be used to further the investigation and seek conviction of the offenders. Intelligent analysis tools help both telecommunications service providers and LEAs extract the important information from the intercepted data, find patterns and relationships, and apply it more effectively to decisions affecting the entire criminal case.

The following types of tools are important for intelligently analyzing intercepted data.



FIGu RE 3.8.10 Applying and combining automated intelligence support.

3.8.8.2.1 Free Search

Free search tools reduce costs and save time by finding the significance in huge amounts of information from e-mail, chat, instant messaging, Web sites, and transferred files. Advanced call data search can find specific information such as a person's name or a specific telephone number. Sophisticated text search by call content allows service providers and LEAs to search all locations and all forms of information with just a keyword, phrase, or synonym. Fuzzy searches are used to improve the effectiveness of the search activities. Reports of all call details can also enlighten investigators with important information.

3.8.8.2.2 Visual Analysis

Another category of tools that help service providers and LEAs uncover instances and relationships is visual analysis. These tools help to put all the intercepted data into graphic presentations, pictures, and charts to identify otherwise hidden patterns and relationships.

3.8.8.2.3 Location Tracking

Knowing the criminal's location as the call is being intercepted can be as important as the call itself. Location tracking solutions, for example those from Verint with its RELIANT product, provide a graphic positioning of the target using a cell phone, drawing the route and direction of the target directly onto maps of the geographical area.

3.8.8.2.4 Voice Verification

Verifying the speaker's identity in a conversation can be accomplished with a voice verification function. A voice bank stores voice samples so the user can compare the intercepted caller's voice to the samples. This helps ensure accuracy of evidence as well as enhancing the effectiveness of the investigation.

3.8.8.2.5 Court Evidence

Successful prosecutions require comprehensive collection and strict adherence to the rules of evidence handling. Intelligent analysis tools provide law enforcement and prosecutors with the basis to develop and manage intercepted communications throughout the process from collection to prosecution.

Storage and archiving are critical parts of the process that is expected to be automated to a certain extent to prevent errors and to reduce labor expenses.

A portable court playback solution replays original evidence for courtroom presentations.

3.8.9 Summary and Trends

Lawful interception will play an important role with telecommunications service providers. Honest support is expected regardless of whether lawful interception is helping to generate additional revenue or not. Combining lawful interception with powerful ISS is the optimal path for telecommunications service providers. In addition to providing the basic functionality for lawful interception, an ISS can support other important functions, such as real-time and usage-based billing, fraud avoidance, churn avoidance, providing government agencies with the necessary data, and real-time traffic control. Cost recovery is guaranteed somehow this way, assuming a certain level of reimbursement by governments.

Critical success factors for lawful interception are:

- Process: the functional steps of lawful interception, including request for lawful interception, provisioning resources, accessing, delivering and collection of data, maintaining data, and converting data into intelligence by telecommunications service providers, by LEAs, and by outsourcers.
- Products: all existing and future applications supporting process steps, including active and passive hardware or software probes, built-in intercept access points, management software, in-band or out-of-band data communication networks, receiver applications, forensics, evidence collection, and court replays.
- People: necessary skills and experiences of subject matter experts of telecommunications service providers, LEAs, and outsourcers to support process steps and available lawful intercept products and tools.

Lawful intercept (LI) also referred to as *wiretapping* or *communications interception*, is the identification, isolation, delivery, and collection of communication sessions (voice, e-mail, packet data, etc.) for use by law enforcement. This critical law enforcement tool is used by many authorized government agencies to investigate criminal activities. Law enforcement agencies (LEAs), telecommunications service providers, and equipment manufacturers are continuously working together to develop products and define technical standards for lawful intercept in the quest to aid LEAs in their role of protecting the public.

At this time, the operating principles may be summarized as follows:

- The technology of intercepting practically all telecommunications services is available, but the price tag may be high to pursue interceptions in particular for SMS, MMS, VPNs, VoIP, and for encrypted traffic.
- Existing laws are usually for telephone systems and they actually do not work with Internet-based voice, data, and video technologies; there is still argument about the border line between *information* and *communications* services.
- There are practically no laws, rules, or guidelines for data retention, which causes problems with very large volumes of stored e-mails, intercepted voice communications, video teleconferences, and other communications-related applications.
- There are very few examples of direct dialogue among LEAs, outsourcers, and telecommunications service providers addressing technical, economic, and privacy challenges.
- Telecommunications service providers are usually not yet deploying ISSs with the result of offering segmented applications around otherwise powerful mediation solutions.
- In particular, the hand-over between LEAs and TSPs is not supported by the latest technology with the result that state-of-the-art security solutions cannot be fully implemented to protect sensitive data and intelligence.

- Educational work is needed to separate and integrate solutions for lawful interception, security, and forensics; they are very different, but still they can and should collaborate with each other.
- Data management is key for telecommunications service providers; not only lawful interception, but also other information requests by LEAs, SEC, HIPAA, Sarbanes-Oxley, Gramm-Leach-Bliley, and other government agencies can be met today and also in the future may be met by querying internal and external databases.
- Some technological challenges will remain; for example, targeting prepaid customers, location tracking, identity management, searches, IPv4 migration toward IPv6, emerging new products and services; telecommunications service providers and network equipment manufacturers are advised to embed IAPs in their equipment from the very beginning.

International and domestic lawful interception standards for IRI, content, and hand-over are not expected to change much in the near future. Thus, the support of the access, delivery, and collection functions remains mandatory.

In terms of the access function, the expectations for the future are:

- More embedded IAPs
- Flexible provisioning of IAPs
- Easy physical and logical configuration of probes
- Collaboration between strategic and targeted surveillance techniques
- Easier target identification via IPv6-based addressing mechanisms
- Resolution of dynamic identifiers (IP addresses) with static identifiers (MSID) username
- Correlation of data flows from multiple network nodes
- Collect data on all activity at all nodes in the voice/data/video networks
- Inform network routers to stream specific flows to session/packet reconstructors
- Flag specific users on content gateways and servers

In terms of the delivery function, the expectations for the future are:

- Mediation systems play the role of a coordinator for receiving data from all kinds of IAPs
- Mediation systems are in charge of distributing data to LEAs and other ISS applications
- Flexible conversion of mediated data into intelligence
- State-of-the-art peering with monitoring centers, outsourcers, and LEAs
- Filter data in real-time based on intercept warrants
- Provide reports on all filtered usage (real time or batch)
- Dynamically provide alarms for Layer 7 (HTTP) access of targeted Web sites, bulletin boards, or by restricted users

In terms of the collection function, the expectations for the future are:

- State-of-the-art connections among LEAs, outsourcers, and telecommunications service providers
- Flexible collaboration among LEAs, outsourcers, and telecommunications service providers
- More secure hand-over solutions
- New applications for back-office functions
- Excellent, unified, and simplified user interfaces for LEAs to easily access intelligence at multiple telecommunications service providers and outsourcers

Also, the administration function should be improved with new features. Examples are:

- Secure local or remote access to intercept rules base and data sets
- Dynamic deployment of rules base and data sets to collectors and correlators
- Automated addition of target subjects to watch lists
- Capability of querying any external databases as required

Real-time, active mediation for lawful intercept requires the following points of consideration:

- Lawful intercept is a regulatory mandate, not a profit generator: the right active mediation deployment can provide a solution for many of the regulators within a software component that is used for other profitable purposes.
- Mediation is already in the network: Most likely, the mediation solution is already communicating with all the network nodes as part of its existing business.
- Restricted use group access: The mediation solution should have a security subsystem to provide a secure method for delivering access to only those users who are authorized to view the rule base.
- Extensible rules base: Intercept rules and target data sets (cell number, IP address, username) should be able to be easily added to the system.
- Dynamic deployment rules: Rules deployment should be dynamic, eliminating the need to take the system offline to update datasets.

New technologies and new telecommunications service providers will become subject to lawful intercept regulations. One of the latest FCC rulings in the United States has decided that communications services offered over broadband pipes, including VoIP, are subject to CALEA requirements to comply with LEA requests for IRI and content surveillance. The tentative rules also would cover managed communications services offered over broadband connections, including managed instant message or video services. Nonmanaged peer-to-peer services, including consumer-grade instant messaging service and noncommercial VoIP services, likely would not be subject to CALEA regulations under the proposed order.

New applications for dynamic surveillance rules, correlations between intelligence sources and the use of external databases will help to improve targeting and identifying criminals and terrorists.

Future-proof lawful intercept strategies must consider the following attributes:

- Operates seamlessly in PSTN, IP, next-generation, converged, and hybrid networks
- Offers maximum flexibility for defining rules for multiple data sources:
  - Enhances capabilities to dynamically search through live data
  - Handles location attributes
  - Recognizes multiple identification points for the trace
- Guarantees powerful correlation
  - Collects data from multiple sources
  - Combines and correlates data
- Integrates with new data sources and feeds additional downstream systems
- Enables cost-effective implementation leveraging a revenue-generating platform
- Uses a state-of-the-art technology, starting with a core platform
- Allows authorized operators to actively manage warrants
- Coordinates secure delivery functions between the carrier's network and LEAs
- Indirectly provisions core network element for level 1 or level 2 surveillance
- Utilizes secure Web services framework
- Provides secured tunnels for transmitting CDC and CCC to LEAs

In most democratic countries, the lawfulness of surveillance has been constantly supervised by human rights and privacy organizations. As a result, the number of illegal interceptions could be reduced to a reasonable minimum. True anonymity while using telecommunications products and services remains very important. Politically engaged persons of oppressive regimes can pursue political activities without fear of revealing their identity. But the entertainment industry sees anonymity as an obstacle to protecting digital rights, tracking violations, and collecting due revenues. Furthermore, anonymity protects criminals and terrorists who violate the law.

Optimal compromises are needed among law enforcement agencies, privacy protectors, investments in surveillance technologies, and emerging telecommunications services on the basis of mutually trusted dialogue.

Acronyms

AP	Access Provider
BI	Business Intelligence
BSS	Business Support System
CRM	Customer Relationship Management
DSS	Decision Support System
EBPP	Electronic Bill Presentment and Payment
ETSI	European Telecommunications Standard Institute
ERP	Enterprise Resource Planning
FCC	Federal Communication Committee
IAP	Incept Access Point
IPDIR	IP Detail Record
IPFIX	IP Flow Information Export
ISS	Intelligence Support System
LEMF	Law Enforcement Monitoring Facility
LI	Lawful Intercepts
MMS	Multimedia Message Service
MSS	Marketing Support System
NOW	Network Operator
OSS	Operations Support System
PRM	Partner Relationship Management
SFA	Sale Force Automation
SMON	Service MONitoring
SMS	Short Message Service
SVP	Service Provider

References

American National Standards Institute (ANSI). 2003. *Lawfully Authorized Electronic Surveillance (LAES) for Voice over Packet Technologies in Wireline Telecommunications Networks*, TI.678.

Axland, J. 2004. Keep your eyes open, ISS World Conference, Washington, DC, May.

Baker, F. 2003. *Cisco support for Lawful Intercept in IP networks*, Internet Engineering Task Force.

Cohen, O. 2003. Internet surveillance via an Intelligence Support System, Narus White Paper, Palo Alto, CA, 2003-12-24.

Cable Television Laboratories, Inc. (CTL). 2004. *PacketCable Electronic Surveillance Specification*, PKT-SP-ESP-Io3-040113, January.

European Telecommunications Standards Institute (ETSI). 2001. *Handover Interface for the Lawful Interception of Telecommunications Traffic*, ETSI ES-201-671, Telecommunications Security.

European Telecommunications Standards Institute (ETSI). 2002. *Lawful Interceptions (LI) Requirements for Network Functions*. ETSI ES 201 158, Telecommunications security.

European Telecommunications Standards Institute (ETSI). *Lawful Interception (LI) Handover Interface for Lawful Interception of Telecommunication Traffic*. 2001. ETSI ES 201 671 Edition 2, Draft 13: Telecommunication security.

European Telecommunications Standards Institute (ETSI). 2004a. *Service Specific Details for Email Services, Technical Specification*, ETSI TS 102 233, Telecommunications Security, Lawful Interception, version 1.1.1.

European Telecommunications Standards Institute (ETSI). 2004b. *Service Specific Details for Internet Access Services, Technical Specification*, ETSI TS 102 234 Telecommunications Security, Lawful Interception, version 1.1.1.

European Telecommunications Standards Institute (ETSI). 2004c. *Handover Specification for IP Delivery*, ETSI TS-102-232, Telecommunications Security.
Haynes, I. 2003. Using Tools to Manage Sourcing, Cutter Consortium Advisory Service Executive Report.
Hochmuth, P. 2004. Users Tap Network Monitoring Technology, *Network World*, February 16, 17–18.
International Telecommunications Union (ITU). 2002. Recommendation X.690, Information Technology: ASN.1 Encoding Rules: Specification of Basic Encoding Rules (BER), Canonial Encoding Rules (CER), and Distinguished Encoding Rules (DER), July.
Lucas, J. 2003a. Circuit Switched Voice and VoIP Basics, Tutorial, ISS World Conference, Washington, DC, November.
Lucas, J. 2003b. Internet Technology Basics, Tutorial, ISS World Conference, Washington, DC, November.
Lucas, J. 2003c. Introduction to Wireless Networks, Tutorial, ISS World Conference, Washington, DC, November.
Lucas, J. 2003d. Lawful Interception for Cable/VoIP Networks, Tutorial, ISS World Conference, Washington, DC, November.
Lucas, J. 2003e. ISS ROI Revisited, *BillingWorld & OSS Today*, December, 6, 43.
Lucas, J. 2003f. Making Money with ISS, *Billing World & OSS Today*, July, 4–5.
Lunetta, L. 2003. Frameworks Coordinate Security, *Network World*, November 24, 31.
Packet Cable. 2004. *PacketCable Electronic Surveillance Specification*, PKT-SP-ESP-I03-040113, Cable Television Laboratories, Inc.
Petersen, J. K. 2001. *Understanding Surveillance Technologies Handbook*. Boca Raton, FL: CRC Press.
Rapoza, J. 2004. Who Am I—Net Anonymity, Technical Directions, *eWeek*, August 16.
Rutkowsky, Tony. 2003. Handover Interfaces and Standards, ISS Conference, Washington, DC.
Society of Cable Television Engineers (SCTE). 2001. *IPCableComm Electronic Surveillance Standard*, ANSI/SCTE 24-13, May.
Terplan K. 1999. *Intranet Performance Management*. Boca Raton, FL: CRC Press.
Terplan, K. 2001. *OSS Essentials—Support System Solutions for Service Providers*. New York: John Wiley & Sons.
Terplan, K. 2003. *Electronic Bill Presentment and Payment*. Boca Raton, FL: CRC Press.

3.9 Management of Sensor Networks

Jim Frey

An essential part of operating today's advanced telecommunications services networks is the need for tools, technologies, and processes that improve an operator's ability to assure service integrity. Beyond careful capacity planning and quality assurance during the fulfillment phase, operators must pay close attention to their subscribers' quality of experience (QoE) during the live delivery of services. And with the growing diversity of methods that communications customers demand over which to receive their services, aggravated by stiffening price competition between carriers, QoE becomes an essential focal point for building and protecting customer loyalty. This pressure provides a strong business incentive to embrace and deploy sensory networks of purpose-built monitoring technologies that can provide real-time and predictive indications of QoE plus an informational base for rapidly and definitively trouble-shooting problems when they occur.

This chapter will focus on the objectives, challenges, technology alternatives, and deployment of sensory systems as a basis for service assurance and QoE optimization. And, since the world's service delivery networks are moving steadily and unwaveringly toward pure IP-based architectures, the discussion centers heavily on the sensor networks required to support connectionless, IP-based service delivery infrastructures.

3.9.1 Objectives of Sensory Monitoring Systems

According to the TeleManagement Forum (TMF)'s *SLA Management Handbook* [TELE05], the three main sources of service performance information are network measurements, customer interviews, and customer complaints (service center calls). While service center calls can and should be tallied as a retrospective means for tracking service quality, the only of the three sources that offers the opportunity to proactively assure services is network measurement, or more broadly, service delivery infrastructure measurement.

Sensory networks are defined here as groups of tools and technologies that either actively or passively test, sample, and/or measure the current activity in a communications service delivery environment. The emphasis here is on monitoring of the live services network, as opposed to test and measurement systems, which are used during manufacture, lab test, or initial deployment of such environments. Where deployed, sensory networks can play a key role in many operational/functional domains within Operational and Business Support Systems (OSS/BSS). Using the Telemanagement Forum's Telecom Applications Map [TELE07] as a guide, sensory networks directly underpin the following service assurance areas:

- Resource status monitoring
- Resource performance monitoring/management
- (Resource) Correlation and root cause analysis
- Service performance management
- Service quality monitoring and impact analysis

Sensory networks can also contribute indirectly to a broad range of other BSS/OSS functional areas within the broader categories of service fulfillment and assurance, including:

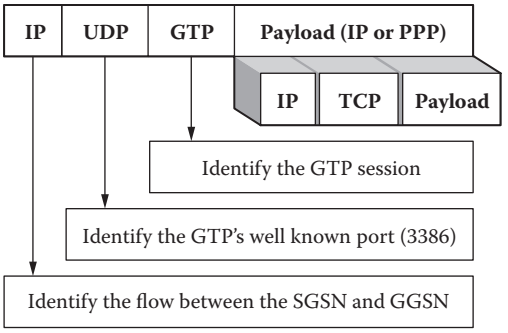
- Resource problem management
- Resource planning/optimization
- Resource design/assign
- Service design/assign
- Service problem management
- Customer contact management, retention, and loyalty
- Customer self-management
- Customer QoS/SLA management
- Customer service/account problem resolution
- Product performance management

With such broad potential impact, sensory systems hold great promise for improving the operational efficiency, cross-functional collaboration, service integrity, and ultimately, customer retention and service revenue growth.

3.9.2 Challenges of Monitoring

There are several significant hurdles to be faced when considering, designing, and commissioning a sensor network strategy. In particular, prudence dictates that sensor networks not be created for each and every individual service or delivery technology, and so any approach needs to accommodate as many existing and potential services and technologies as possible. As noted above, these networks should be able to contribute to a broad range of operational needs and objectives, and hence they should receive proper attention as long-term, service-protecting investments that will provide significant returns on investment via long-term customer satisfaction (and hence retention) and higher levels of sustained adoption of high-value next-generation services.

Following are some of the more significant technological and domain-specific challenges faced when planning a sensory network.



FIGu RE 3.9.1 GPRS Tunneling Protocol example of multi-level IP addressing.

3.9.2.1 New Technologies—IP Services

The worldwide communication infrastructure is moving rapidly to Internet Protocol (IP) as a technology basis. Traditional fixed and mobile voice services are being replaced with IP-based equivalents across the board, and newly enabled services are adding to the mix. IPTV adoption is early but growing fast. Mobile IP has pushed the mobile operators to the forefront in terms of revenue growth, competition, and focus on customer service quality as a result of inexorable demand for easier nonfixed access to multimedia content and applications. Greater and greater exploitation of full multimedia by content providers is enabled via IP transport, and is directly driving substantial growth in total traffic volumes, stressing capacity.

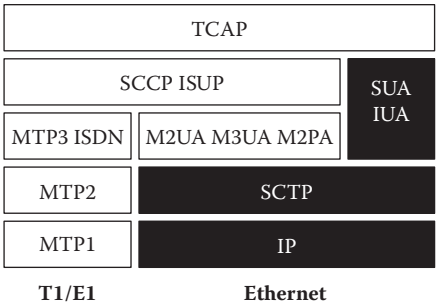
One of the greatest challenges facing operators who are deploying new, next-generation IP-based services is the fact that with new service flexibility comes new protocols and delivery chain complexity. This new technology brings with it an entirely new set of protocols that are used in the middle protocol layers (Layers 2 through 4) of the delivery network. More concerning than that, there are also a virtually limitless potential variety of types of layer 7 (application) traffic that will be traversing the infrastructure. To make matters more interesting, there can be multiple layers of encapsulation and addressing within IP traffic. For example, see Figure 3.9.1, where IP tunneling is used to transport IP traffic across the core bearer network of a mobile operator using GPRS Tunneling Protocol (GTP). This same concept can be applied multiple times, meaning that understanding just what is contained within flows of packets transiting an IP network can be very difficult without directly and deeply inspecting the live traffic.

Most of the actual service traffic will be predefined and offered as revenue-based services; however, as soon as IP-based transport is made available, subscribers find ways to transport whatever traffic they deem desirable. An example of this is Skype—a bulk toll-bypass form of IP Telephony that uses a variety of adaptive approaches to identify means of establishing and delivering voice calls over public IP networks.

Multitier service delivery architectures now commonly include off-net content sources. Reliance on these external providers raises a need for selective direct testing to assure availability and responsiveness of noncontrolled elements, especially if they are perceived to be an integral part of the service experience. Ultimately, measurements and sensory system input must be brought together across multiple domains in order to provide promised operational value.

3.9.2.2 Merging of Traditional Signaling onto IP Networks

Signaling Transport (SIGTRAN), which enables the transport of traditional public switched telephone network (PSTN) signaling over IP networks (see Figure 3.9.2), is enjoying a rapid embrace by providers who need to meet capacity demands and lower operational cost. SIGTRAN standards are defined by the Internet Engineering Task Force (IETF), and allow providers to operate smoothly in hybrid environments, where services may originate from or be terminated within the PSTN, hence requiring both



FIGu RE 3.9.2 Protocol stacks for SS7 over TDM and IP/Ethernet.

traditional SS7 control traffic as well as next-generation IP-based control traffic in order to establish (and hence monitor and troubleshoot) service sessions or transactions. Operations groups will want to deploy sensory monitoring of SIGTRAN to identify route configuration errors, troubleshoot calls/sessions, and recognize overutilized links.

3.9.2.3 New Architectures: IMS

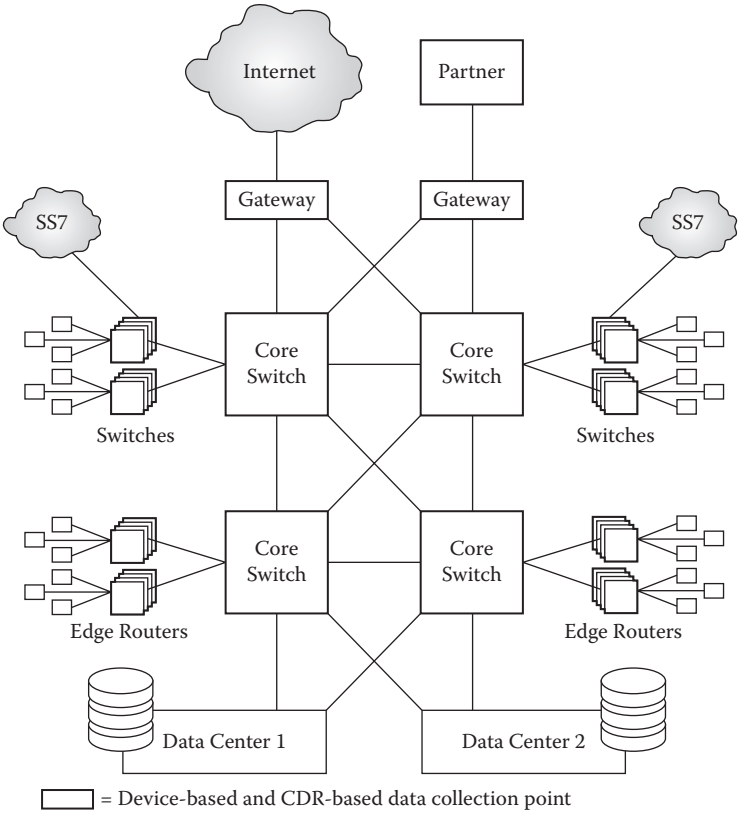
With the introduction of purely packet-based next-generation network architectures such as IP Multimedia Subsystem (IMS), an appropriate adaption in monitoring technology and strategy becomes essential. IMS offers a much more flexible, adaptable, and lower cost-control layer for both wireless and wireline services by providing access to a rich set of service enablers. Within IMS networks, a primary purpose for sensory monitoring systems will be the need to identify and isolate problems ranging from congestion (such as mass calling events or massive video event) to component degradation (for example, a failure in an enabling service application like DNS or partial equipment failure) to configuration errors (often caused by improper traffic classification at the edge of the network) to peering problems (which might be control traffic floods due to malfunctioning session border controllers). Malicious activities can also create service-impacting performance issues that must be definitively resolved in order to restore services. Decoding messages and determining the point of impact is an essential requirement and expectation of sensory systems. Particularly important is that within IMS networks, the signaling and control flows share the exact same topological paths as the service payloads. Consequently, the interplay, interaction, and interrelationships between these two important but different traffic types must be constantly monitored to ensure healthy coexistence.

3.9.3 Sensory Monitoring Technologies and Alternatives

There are many options from which to choose when considering tools and technologies for your sensor network. All have advantages and disadvantages, and all vary in terms of total deployment and operating cost. Some have specific advantages for certain types of services. Ultimately, most operators will choose a combination of these technologies as they outfit their monitoring architecture. All of the following categories of sensory approaches represent unique sources of sensory data, and each requires some form of sensory system to harvest, process, analyze, and present that data.

3.9.3.1 Device-Based Sensor Networks

Historically, the vast majority of infrastructure performance monitoring has focused on gathering metrics and statistics from the interconnected equipment or device nodes on the network, such as switches, routers, service nodes, firewalls, servers, etc. A wide range of protocols and data models have been used to serve and gather this information, with some degree of standardization in management data communications, such as Common Management Information Protocol (CMIP), TL1, and Simple Network



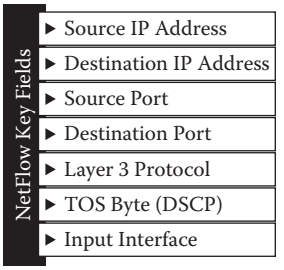
FIGu RE 3.9.3 Deploying device-based and CDR sensor networks.

Management Protocol (SNMP), and data models such as the IETF (Internet Engineering Task Force) RFCs and the Telemanagement Forum's NGOSS-SID. Also common is a wide degree of variation and proprietary extension introduced by device manufacturers, often as a means of market/feature differentiation. With the move to IP as a primary transport and service technology, device management interface variety is declining and converging on SNMP. The most recent iteration of SNMP, version 3 (SNMPv3) includes adaptations for scale and security that are necessary to assure integrity within large communications services environments.

In order to gather device-based data from SNMP, CMIP, TL1, or other device management interfaces, sensor approaches need to include polling systems to regularly harvest data and make it available for aggregation and analysis. Deployment of device-based sensory monitoring is presented in Figure 3.9.3.

The primary advantages of the device viewpoint for service assurance are the broad availability of SNMP-based data provided by IP infrastructure vendors. Much of this is standards based in terms of MIB (Management Information Base) data structures, though most all equipment manufacturers also have important proprietary extensions that must be referenced to gather a complete view of the health of the device and the aggregate measures of traffic flowing through it.

While these sensory data sources are prevalent and relatively low cost, the level of information that can be interpreted from them is inherently limited to summary statistics regarding the flow of service traffic to/from/through them, along with internal device/node health measurements. Consequently, while they can be used to gauge service activity and quality on an aggregate basis, they cannot be used for recognizing or troubleshooting individual service sessions or experiences.



FIGu RE 3.9.4 Sensory data available within NetFlow records.

3.9.3.2 CDR-Based Sensor Networks

A substantial amount of valuable operational intelligence can be gathered by processing postsession CDR (Call Detail Record) types of data. This is something of a special case for device-based sensor networks, as the source of the data is usually the infrastructure nodes themselves. In IP networks, this is often in service activity records such as IP Detail Record (IPDR) (which is also the basis of service billing), flow records like NetFlow/IPFIX (from Cisco core routers or service platforms), JFlow (from Juniper core routers), or statistically sampled equivalents such as sFlow. By their very nature, they provide information about each subscriber session, including what application or service was invoked and some basic statistical measures regarding the aggregate session, such as total traffic volume and session duration and some basic quality metrics. Key information available from NetFlow records is indicated in Figure 3.9.4.

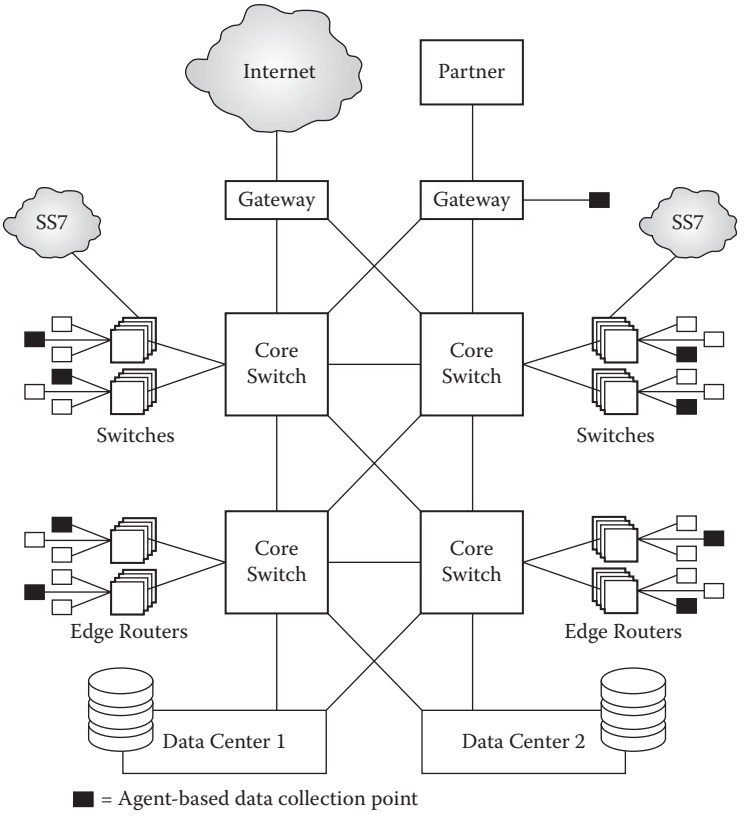
CDR-based sensory deployments are very similar in nature to device-based deployments (see Figure 3.9.4), in that all of the sensory data is drawn from network devices. The basic difference in this case is that CDRs are generated by the devices and sent to a defined/configured destination; hence, instead of using a polling collector to proactively gather the data, a receiver collector approach is used.

The primary advantage of using these types of data sources for service assurance is the fact that they are being generated directly on the basis of each individual service invocation, and thus are true reflections of service session activities. Additionally, these are commonly available from IP device manufacturers at no additional cost (for the generation of the data records), although there is investment required in tools for collecting and processing the records.

Disadvantages include the load that they place on the delivery infrastructure elements (though it is small for most devices, even a small additional load may be deemed an unacceptable performance risk), and the fact that this data set leaves out some of the most critical information that is required to discretely recognize and understand individual subscriber sessions, complex multi-element content services, and key quality metrics such as responsiveness and resolution within Web-based transport protocols. Also, since CDRs are postsession records, they are not a practical basis for troubleshooting live services in real time, or for recognizing intrasession anomalies.

3.9.3.3 Agent-Based Sensor Networks

The next category of technologies are those that measure service quality by using independent software or hardware agents to generate and introduce synthetic test traffic into the service delivery network on a periodic basis to assess service availability and responsiveness. This may be accomplished by deployment of agentry directly by the operator, or by using a third-party service that executes the tests and reports the results back to the operator. The test traffic can either be a select set of trials that exercise key service components, representative approximations of complete service sessions, or replays of actual captured subscriber sessions. There are also agent-based solutions that directly instrument live subscriber sessions from the end device or at the edge of the network and passively monitor actual service sessions, though



FIGu RE 3.9.5 Deploying agent-based sensor networks.

these are not often found deployed within communications service provider environments. Often, a mix of these approaches is used to best deliver a comprehensive coverage of potential risks and objectives.

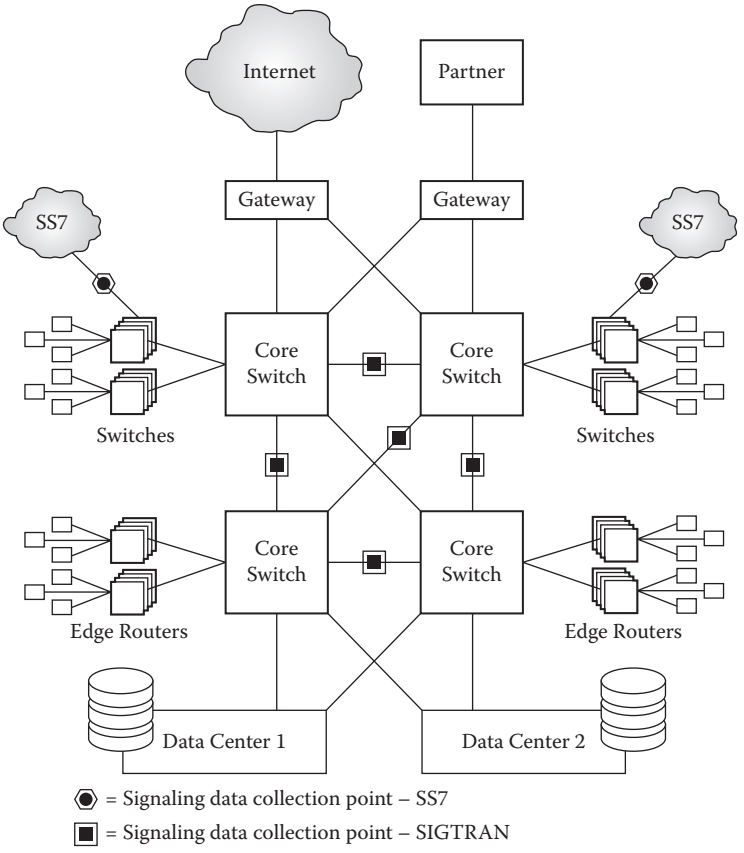
Deployment of agent-based sensing is typically focused just at or just outside the service provider edge, or from within the provider's network and focused on testing of third-party partner content sources. See Figure 3.9.5 for an example of agent-based sensory deployment.

The primary advantage of agent-based measurement is that the result is a very close approximation of what a typical subscriber will experience, and thus provides key indications of likely customer sentiment. Additionally, the use of synthetic test traffic allows operators to assess the delivery infrastructure's readiness and ability to deliver quality services even when there is no revenue traffic.

The downsides of agent-based measurements are the extra load that they can introduce into the delivery infrastructure (even though it may be minimal, operational policies may strictly limit any nonrevenue traffic) and the incomplete viewpoint that they provide, while the metrics gathered from agents are good indicators of service experience (especially if they are monitoring real subscriber sessions) they cannot often help with understanding where to look to uncover the source of a service degradation or failure. In particular, regarding the latter point, agentry cannot recognize any of the macro traffic congestion issues or multitier, multicomponent delivery architecture problems that might lurk behind a particular service session being poorly responsive.

3.9.3.4 Signaling-Based Sensor Networks

All service sessions utilize some form of control protocols, in order to establish and authorize the request and use of a service. In the legacy circuit-switched telephony world, this was the independent (off-net)

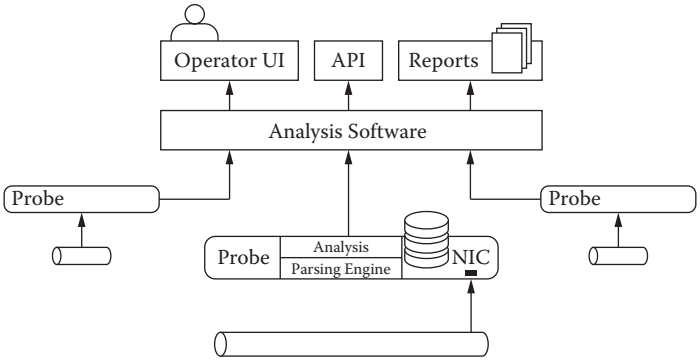


FIGu RE 3.9.6 Deploying signaling-based sensor networks.

SS7 network and control protocol suite. In the emerging IP-based world, there are multiple levels of control protocols, starting at the lowest layers with protocol handshakes and session control, and ultimately culminating in authentication and authorization followed by flow controls during the service session and tear-downs following service completion. With the advent of IP, this signaling traffic is moving to IP-based transport and is now running over the same transport infrastructure as service payloads. The control protocols are very similar, even identical, but the protocol stack has changed and the point of measurement has also changed.

Sensor networks for signaling data commonly consist of dedicated *probe* instrumentation, which is passively connected to the access points into the signaling network. Since this is dedicated instrumentation, these sensory networks provide real-time information on customers/subscribers and the services they are invoking, as well as session details regarding service setup, tear-down, and summary results. Instrumentation points are normally in the access layers of the network, where switches connected to the out-of-band SS7 network, although with the convergence onto IP-based signaling these instrumentation points now often include in-band backhaul and core network links. See Figure 3.9.6 for an example of signaling-based sensory deployment.

Signaling sensor networks are less well suited for providing detailed characterization of service quality during the actual delivery process, intrasession or intratransaction. Since they do not have a view into the actual service payload, signaling sensor networks are also unable to support session playbacks for reconstructive analysis. Lastly, since they focus exclusively on control data, they do not recognize nor track congestion or errors within the transport network.



FIGu RE 3.9.7 Packet-flow sensory product architecture.

3.9.3.5 Packet-Flow-Based Sensor Networks

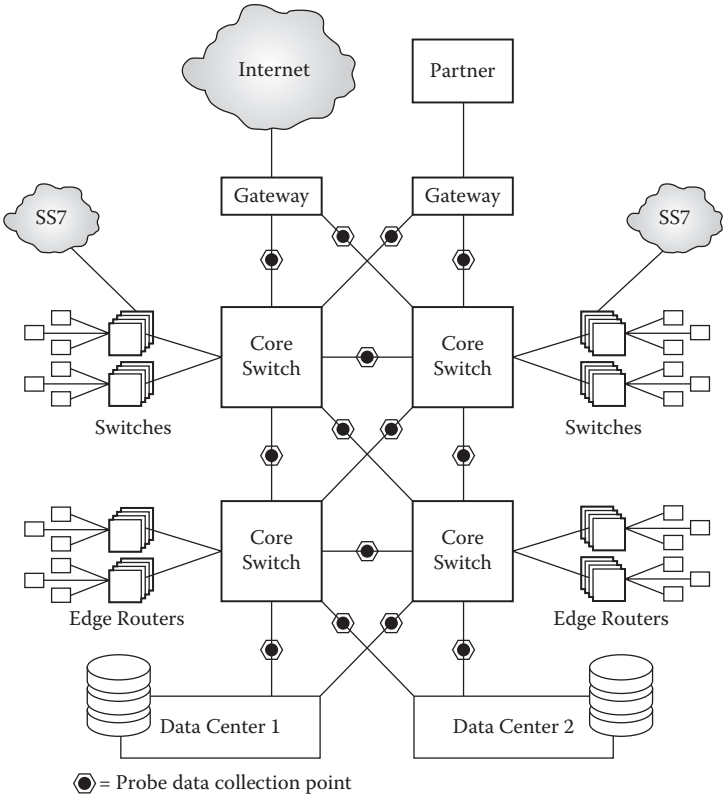
The final category of sensor technologies are deep packet inspection (DPI) products that watch the flow of actual revenue traffic across the delivery network and build a finely granular and complete view of all services and control traffic flows. These solutions are typically provided via purpose-built *probe* appliances that attach passively to key points of aggregation in the service delivery network and inspect each and every stream of packets that comprise connectionless services. Since these sensory systems start with packets, but ultimately tie those packets together into common threads, or *flows*, they are referred to here as *packet-flow* solutions. While this type of technology has long been available for test and troubleshooting purposes, and has been broadly deployed in the predominantly IP realms of government and enterprise networks, its application to communications service provider operating environments is relatively recent.

The basic elements of packet-flow sensory systems include (see Figure 3.9.7) high-speed network interface cards that are responsible for providing a live, real-time window into the packet streams, combined with a parsing engine that inspects each packet and categorizes it in terms of service user, service type, service volume, and key quality indicators, and analysis software for harvesting and presenting the collected metrics.

The more advanced systems of this type will include additional functional elements such as:

- Stream-to-disk for complete reconstructive forensic analysis of extended packet sequences
- Detailed packet decode and expert analysis functions (may reside directly on the instrumentation devices)
- Automated recognition and tracking of traffic types within core traffic categories (such as individual Web URLs or sub-URLs versus aggregate HTTP or HTTPS-based Web traffic), individual application transaction types, and bit sequence pattern matching
- Advanced, predictive analysis of key performance metrics for recognizing relevant early indications of service quality degradations
- Integration with other OSS products for alarm forwarding and data sharing

The primary challenge with architecting packet-flow solutions is in handling the speeds at which service delivery networks operate. Unlike other sensory approaches that deal with summary or sampled data, or with only control traffic, packet-flow sensors must watch every bit and byte of the service traffic traversing the instrumented links and select the appropriate information from which operational information will be interpreted. Current state-of-the-art sensors can accomplish this goal at line rates of up to 10 Gbps, and vendor suppliers continue to working on raising that rate to keep pace with the global growth in IP traffic and the commensurate steady increase in transport technology speeds.



FIGu RE 3.9.8 Deploying packet-flow sensor networks.

Deployment of packet-flow sensory systems is similar in many ways to signaling sensory systems, in that passive direct connections to network links are the primary points of instrumentation. The difference comes in the focus of instrumentation. Whereas signaling sensors are normally located in the access layer, packet-flow sensors are more often placed with the core bearer network, on key external roaming or content partnering connections, and within the data centers that host enabling services and content services. An example of deployment of packet-flow sensory systems is presented in Figure 3.9.8.

Packet-flow sensory systems hold some distinct advantages over other technologies, in that they provide a true record of actual services being delivered in the native delivery infrastructure. This goes beyond the initiation/termination data that could be gleaned from monitoring only the control traffic, in the way that signaling sensory systems work, to include the ability to monitor service quality intrasession. This basis of information also provides a means for detailed retrospective analysis of problem service sessions, going beyond mere call trace analysis to allow full session reconstruction and enabling service personnel to directly witness the service experience.

Shortcomings of the packet-flow approach are primarily based on their point of measurement. Because they rely on instrumenting high-speed delivery infrastructure links, they will always be under pressure to expand the speeds at which they are able to operate. Also, packet-flow instrumentation appliances are not broadly available at a cost point that would allow deployment directly to all types of customer/subscriber end points. As a result, service responsiveness measurements are taken from the provider edge (at best) or at other points of aggregation in the delivery path and may only be a proxy for the actual customer/subscriber experience. Finally, packet-flow sensory systems are useful only on packet-based infrastructure, and cannot provide full visibility into legacy circuit or hybrid delivery environments.

3.9.4 Selecting and Applying Sensory Systems

Each of the sensory technologies we have covered has advantages and disadvantages. While agent-based and device-based sensory technologies contribute valuable information to provider operations, due to their limited viewpoints they can only play a secondary role in sensory network architecture. Similarly, CDR-based sensory technologies operate in a retrospective manner, making them valuable for offline capacity planning and service product planning, but less useful for real-time operations and problem troubleshooting. Probe-based systems for signaling or packet-flow monitoring provide extensive data, but cannot tell you what is happening inside the service delivery network elements themselves. Consequently, a mix of these technologies will be required to meet all potential operational needs, and should be selected based on current gaps, operational priorities, and service objectives.

3.9.4.1 Visibility

The primary function of a sensory network is to provide a clear image of the health and activity in a communications delivery infrastructure. Importantly, while detailed and somewhat unique views and presentations of sensory data are needed to satisfy each of the many functional areas described above, the sensory network technology and architecture decisions can be driven by focusing on three key viewpoints—network-centric, service-centric, and customer-centric [STRA07].

3.9.4.1.1 Network Visibility

In a circuit-switched network, there are predefined channels (circuits) that carry the message depending on the destination, with each channel operating at a fixed bandwidth. Regardless of message size, information is transported to its intended destination at the allocated transport level. For IP, bandwidth allocation is variable according to the number of messages to be processed (incoming or outgoing).

In the PSTN, an operator monitors the percent utilization of each channel so it knows when a channel is filling up and more bandwidth is needed. Traffic on an IP network often fluctuates, and contention issues arise when there are more messages than bandwidth, causing new messages to be refused or blocked. If a message is blocked, the network is designed to retransmit the message, however the retransmitted message can also be blocked if bandwidth still is not available. This process can be repeated several times resulting in message delays that often generate customer complaints.

IP network virtualization presents a different problem as providers migrate toward a more integrated, if not entirely converged, service offering environment. Virtual IP service networks can include the network itself (e.g., virtual network operators), the data storage layer, any number of network-supplied services (e.g., VPN, virtual voice network), and the application delivery layer (e.g., Web services, SLAs) to name a few. These multiple abstraction layers make it difficult to determine, for example, the root cause of a reported problem or to measure the quality of service (QoS) associated with a customer offering.

Virtualization problems are not visible through device-based sensory monitoring, signaling sensory monitoring, agent-based sensing, or CDR-based sensing—only packet-flow sensory systems have the depth of detail to see through the abstraction fog of virtualization and help deliver understanding of the flow of different services and service components across complex IP networks.

3.9.4.1.2 Service Visibility

Services can take on many forms including pure network-based services such as voice mail, e-mail, three-way calling, and broadband access. They can also involve combinations such as broadband access with third-party applications or digital content. Due to the number of potential network and partner-supplied capabilities that can be used to define a service, and the eventual goal of delivering services targeted to each individual customer-subscriber, assuring such combinations is a dynamic and sometimes monumental undertaking. To create an integrated view of service quality, and to monitor service levels experienced by individual subscribers, a number of service quality indicators defined around each component of a service should be considered. For example:

- Device-based Monitoring: Metrics defining traditional network connectivity relative to broadband access or message delivery are an essential part of understanding service availability. Traditional network fault and performance management information is one essential part of understanding service availability.
- Signaling-based Monitoring: SS7 signaling metrics can provide insight about quality and performance issues concerning the flow of messages throughout the PSTN infrastructure.
- Packet-Flow Monitoring: Gathering details of the IP packet header and sometimes payload can help to identify IP traffic flow bottlenecks, failures, or degradations of IP-based service enablers or service components, and conditions when IP service activity are not normal.

While after-the-fact trending and analysis is effective for service planning and tracking customer utilization metrics, an additional requirement rests with data monitoring for those operations work groups that must deal in real time with customer complaints and network problems. Real-time data is necessary for isolating and recovering from service problems before they become serious. Such a strategy sounds straightforward but is difficult to implement due to limited access of partner-supplied information, data disparities among nonintegrated systems, and often less than optimized processes. The ability to isolate service problems and to correlate them with network issues is critical to the security/stability of the network and in maintaining good customer relations. In most cases, service quality issues are expected to be identified, isolated, and often resolved in real time to prevent SLA violations and service disruptions.

3.9.4.1.3 Customer Visibility

Maintaining the reliability and quality of the network and developing a better understanding the status of services remain critical; however they alone are not sufficient for ensuring satisfactory customer experience. The complexity of collecting and correlating sensory data from within a multiservice, multi-infrastructure environment requires a broader OSS/BSS strategy that involves not only sensory data collection, but also data correlation and filtering using specific business definitions with the flexibility to change as business needs change.

To facilitate a closer focus on the customer, key performance indicators (KPIs) should be directly collected from sensory networks plus additional sources—services platforms, call centers, and business support systems. These KPIs can be combined to create customer- and operator-specific key quality indicators (KQIs), including measures of speed, availability, reliability, security, simplicity, and flexibility in the following broad categories.

- Access: The time to establish a connection and the availability of sufficient bandwidth to use a service includes metrics that point out:
  - Security problems, retries, or poor response times. These are geared to provide warning about possible network, service, application, or user device problems.
  - Areas where service availability, coverage, or bandwidth may be insufficient. These are geared to identify when and where a user is connected to the network, what services were used, and how long each service was accessed.
- Usage: Beyond tracking usage for billing, there is need to understand what services, devices, applications, and access networks are being used, when, and for how long, including metrics that:
  - Identify services or applications that are performing poorly
  - Show which services customers are not using, and which might be linked to poor service definition, lack of flexibility, inadequate support, or simplicity issues
  - Point to where potential sales opportunities might exist.
- Disconnect: Determining when a user has failed to connect to a service or the connection has failed during the user session.

- Faults: In addition to reliability or availability faults in the above categories, there are user device faults, roaming faults, and application faults that are customer specific and not detectable by traditional fault management systems.

Sensory network systems not only play an important role in providing customer visibility, they are increasingly essential as it becomes increasingly difficult to recognize and aggregate customer experience data within and across multitier, multicomponent IP service delivery architectures. Device-based sensory systems do not have the resolution to provide KPIs for customer visibility; however all of the others do, to varying degrees and extents.

3.9.4.2 Signaling vs. Packet-Flow

In order to deliver the full promise of sensory networks and their potential ability to positively influence a broad range of operational functions, a sensory architecture must leverage direct, dedicated instrumentation that can deliver both service and customer visibility in real time. The two technologies that offer this level of visibility are signaling probes and packet-flow probes. As outlined above, each has their primary advantages and disadvantages, and due to the fact that many operators still operate in mixed/hybrid service delivery models, where traditional TDM and next-generation IP delivery models are both present, there are elements of each that will be required. Table 3.9.1 provides an overview of the primary strengths of each approach.

3.9.5 Summary and Trends

With the move to connectionless, IP-based service delivery infrastructures and the exponentially increasing complexity of content origination, sensory networks are a requisite element in carriers' service assurance architectures. Many options exist for sensing technologies, and given the state of sensory network technology today, operators will need to examine and deploy a combination of sensory technologies and collection engines in order to fully cover the service delivery environment. The two technologies that hold out the best promise for real-time and predictive protection for customer quality of experience are those based on high-speed probe sensing of signaling or payload IP traffic.

TABl E 3.9.1 Comparison of Signaling Sensory versus Packet-Flow Sensory Approaches

	Signaling Sensory Monitoring	Packet-Flow Sensory Monitoring
Viewpoint	**Call and Session Control:** Management of the fixed connection TDM network but also includes packet services for wireless, e.g., GPRS. Setup and tear-down of fixed and mobile sessions, with normal or abnormal notification/closure.	**Session Delivery:** Monitoring actual delivery of service content and payload. Service quality monitoring during a session and not at the setup or tear-down of a session. Intrasession service quality.
Orientation	**Connection-based:** Show end-to-end call path involving all network elements for a service. Topology aware of what network elements are delivering what from the network. Setup or tear-down of connection paths within the network.	**Connectionless:** Tracking of packet flows on a session-by-session basis regardless of what network elements are used. Can show adjacent devices, but concentrates on the flow of packets across a point of the network. Topology agnostic because there is not a specific path involved due to rerouting capabilities in mesh networks.
Service Perception	**Aggregate Service:** Looking at the flow of a revenue service to the end user. Noting the path that the revenue service takes from origination to termination (at the end user point). Usually does not see actual content of the revenue service. Sees setup, sees data flow during the session, and session tear-down. Includes both successful and failed access attempts to deliver the revenue service.	**All Service Components**: Looks at all the individual enabling services (i.e., DNS, DHCP, AAA, LDAP) as well as all of the content components. Service enablers are the essential parts that work together to allow the secure flow of content to the end user. Only shows the flow of IP content (not voice content) whether payload is encrypted or not. Payload is noticed but not decrypted. Tells packet loss, tells delay of packet transport.
Planning Value	**Trending traffic volume for call capacity.** Fixed per service and per session. Tracking the number of port assignments in use at any given point in time. Sizing of connection capacity (ports, transmission paths, backhaul need) for fixed capacity content. Monitoring the flow of connection capacity, rate, trending of number of connections. Good at tracking call setup and tear-down, not understanding the total volume of information flowing across the network.	**Trending volume for data capacity.** Understanding the volume of data/IP traffic as variable per service and per session. Tracking of the volume of data flowing between two points that may involve a variable number of access connections. Assists capacity planning (sizing of IP delivery infrastructure, routers, links, switches, servers). Monitors mixed service types across a portion of a bearer network. Trends around content volume regardless of the number of sessions. Not good at understanding total number of connections needed to get data from origination to destination points.
Errors Recognized	**Network Transport Errors:** Failure to connect, but not failure to properly deliver service components. Only errors resulting in a loss of the service, when the service is abnormal termination. If loss of packets creates poor customer services, customer may abruptly stop the service. Cannot tell this type of termination (poor-quality termination) vs. customer needing to abort from other reasons. **Session Errors:** Focus on service failure errors (abnormal termination of service) e.g., service fault.	**Network Transport Errors** including those that happen during process of delivering a service. Retransmission, packet loss, packet errors. Service transport errors seen if signaling is over IP. **Session + Network Errors:** Service degradation errors. Response time increase to unacceptability for service quality impact.

Continued

TABl E 3.9.1 Comparison of Signaling Sensory versus Packet-Flow Sensory Approaches (*Continued*)

	Signaling Sensory Monitoring	Packet-Flow Sensory Monitoring
Primary Troubleshooting Features	**Call Trace:** Focused on TDM-based voice traffic and IP data session setup and tear down. Content of the voice session is not traced. Deals more with "channelized" IP sessions rather than true connectionless environment.	**Protocol Decode / Packet Analysis / Payload Analysis:** Ability to study structure and payload of packets in the IP stream. Able to definitively troubleshoot service design issues and application design issues for performance. **Packet Analysis:** Deepest level detail study of the information delivered. Looking at individual fields in the header and in the payload. If the payload is not encrypted then content can be looked at too. Includes reconstruction (e.g., repaint of HTML pages) based on packets delivered or replaying of an IP-based voice call. **Deals only with IP-based messages**; does not address anything that is TDM oriented.

Note 1: A unique number may be devised in various ways.
Note 2: Table 3.3 is not exhaustive.

Acronyms

CDR	Call Detail Record
DPI	Deep Packet Inspection
GTP	GPRS Tunneling Protocol
GPRS	General Packet Radio Service
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
IPDR	IP Detail Record
KPI	Key Performance Indicator
KQI	Key Quality Indicator
SIGTRAN	Signaling Transport
SS7	Signaling System 7
TDM	Time Division Multiplexing
VPN	Virtual Private Network

References

STRA07 Whitelock, K., and Ruzicka, N. 2007. *Sorting Out the Signals—Gathering Assurance Data That Service Providers Need*. San Antonio, TX: Stratecast Frost & Sullivan, June.

TELE05 Telemanagement Forum. 2005. *SLA Management Handbook, Vol. 3, Service and Technology Examples, GB917-3*, v2.1, Morristown, NJ, January.

TELE07 Telemanagement Forum. 2007. *Telecom Operations Map, GB929 v2.5*, Morristown, NJ, August.

3.10 Solution Architectures

Norman Kincl

3.10.1 Business Strategy Drives Technology

A simple perusal of industry news is all one needs to see that the big thing in Operational Support Systems is NGOSS. Whether one considers NGOSS as the formal TeleManagement Forum (TMF) New Generation Operations Systems and Software program or just as the common usage of Next-Generation Operational Support Systems, it is clear that there is a strong industry trend to define a new approach to managing the infrastructure and services of today's communications, media, and entertainment service providers. We provide the Hewlett-Packard (HP) perspective of NGOSS. With HP positioned as the NGOSS Solution Company, this overview will take a comprehensive architectural view of the complete system.

An Operational Support System (OSS) can be loosely defined as the set of processes and systems that deal with the revenue-generating infrastructure of a service provider. Traditionally the OSS has addressed the telecom network. As service providers move to next-generation networks and services that increasingly rely on value-added services supplied by traditional IT equipment, the OSS has had to expand its scope beyond the network to also address the IT systems that are in the call path.

Like any other business tool or entity, the OSS needs to derive its existence by the support it gives to the business strategy. While this may seem obvious, the reality is that the discussions often start at the technology level, having the technology attempt to drive the business processes. The result is that while the system implemented may be a technological wonder, it is a business failure, or at best a mediocre success, since it does not address key business drivers.

A complete system will include people, processes, and technologies working together to accomplish the common task. While the people and organizations involved with OSS systems are critical, we will limit our discussions of OSS to the processes and technologies that comprise it.

3.10.2 A Comprehensive Architectural Approach

To ensure that the resultant system does meet the business goals, Hewlett-Packard uses a system architecture methodology that drives the technology from the business. This is accomplished by taking an approach that drives distinct architectural views.

3.10.2.1 Architectural View

The concept of architectural views is not new or unique. Conceptually, it is similar to the different types of drawings and models that a traditional architect may need to develop when designing a building. The views provide a representation of the system from the perspective of a stakeholder (or a related set of concerns).

The IEEE has formalized the concept of architectural descriptions, including the need for different architectural views, in IEEE 1471. While IEEE leaves the selection of the viewpoints to the using organizations, it does specify that all resulting views are equal.

3.10.2.2 The Four Critical Views

The HP Global Methodology for IT Systems Architecture defines four fundamental views: business, functional, technical, and implementation. Table 3.10.1 provides an overview of these four fundamental views.

While as IEEE 1471 states, all the views are equal, there is value in proceeding in a top–down manner (i.e., business, functional, technical, implementation). While other approaches can also provide the required consistency across the views, an approach that is primarily top–down ensures that the resultant system meets the business requirements.

TABl E 3.10.1 HP Global Methodology for IT Systems Architecture: Four Fundamental Views

View	Primary stakeholders	Content
Business	Business manager	Why are we doing this?
	System acquirer	What are the internal and external drivers?
	Business analyst	What are the business models and processes?
		Who participates in the business processes?
		What are the project goals?
		How will the success of the solution be measured?
Functional	System users	What should the solution do?
	Business process designers	What will the completed solution do?
	Information modelers	How will it be used and what services will it provide?
		What information will it provide? To whom?
		What qualities must the solution have?
Technical	System developers	How should the solution work?
	Technical consultants	How will the system be structured and constructed?
	Subsystem suppliers	What are the interfaces and other constraints?
		What applications and data are needed?
		What does the infrastructure look like?
		What standards will apply?
		How will the system qualities be achieved?
Implementation	Project managers	With what will the solution be built?
	System developers	What specific products and components, from which vendors, are needed to build the system?
	System testers	
	System deployers	How will the system be developed and deployed?
	Operators/Managers	What validation methods will be used?
		How will it be managed?
		What is the source of funding?

We will use these concepts and views to describe the HP view of an OSS. First, we take a brief look at the business drivers for a new approach to OSS—answering the *why* questions about the system. We then look at the job of the OSS—what it needs to do, both from the business processes it needs to support and the functionality it needs to deliver. The section on how to build an OSS describes the technical approaches that are needed to support the required functionality. Finally, since the details of a technology change quickly, rather than looking at the components needed to build an OSS, we look at some of the considerations that should be taken into account when specifying the implementation view.

3.10.3 Business Drivers Require a New Approach to OSS

The first thing we want to understand is the business view: why do we need an OSS, and more importantly, why can't we just keep the OSS approaches and implementations that have served us so well for so many years?

3.10.3.1 Convergence

The industry is undergoing massive changes that can best be categorized as a multifaceted convergence. Convergence is happening across business, customer, service, infrastructure, and other dimensions, each of which has implications for the service provider business and the how the OSS needs to support these changes.

- Traditional telecommunications network architectures consisted of multiple networks, sometimes interconnected or dependent on each other, but mostly autonomous. Through network convergence, these multiple separate networks are converging on a single IP/MPLS network that supports multiple access methods. Interestingly, this sometimes results in customer circuits being provisioned through nondeterministic packed-switched networks. Not only does the OSS need to

understand the new IP/MPLS technologies and how the services are provisioned on top of them, it also needs to make sure that the nondeterministic aspects of the network do not affect the customers' perceived service quality.

- Not only are the core networks converging, but the value-added services that are being introduced rely heavily on a traditional IT infrastructure. This is driving a convergence between telecom and IT. This introduces problems more fundamental than just needing the OSS to now manage IT resources. The language and culture of the OSS world and the IT management world are not the same; the NGOSS solution must be able to transcend the two domains.
- A specific difference between the telecom and IT worlds that needs highlighting is the difference in process models. The telecom world tends to base its business process models on Enhanced Telecom Operations Map (eTOM) from TMF since this was specifically defined with the telecommunications enterprise in mind, while the IT world tends to use the generic IT Infrastructure Library (ITIL) approach.
- The distinction between fixed and mobile services is being erased as fixed–mobile convergence takes hold. Not only can you no longer rely on a fixed service access point, the device the customer is using and the capabilities it has may change during a session. The OSS is still responsible for understanding the service levels delivered.
- The whole concept of differentiating OSS and BSS is disappearing, with many people already talking about B/OSS or even BOSS. To introduce further complexity, the new services are being built using Service Delivery Platforms (SDP). The functionality provided by the SDP includes some functionality that was traditionally considered the domain of the OSS or BSS. This drives the need for the OSS to be able to tightly integrate with the SDP.
- The mergers, acquisitions, and divestitures trend continues. Since a justification for mergers and acquisitions is reduced costs through consolidating systems, the impact for the OSS is clear: after a merger or acquisition, the two separate systems will need to quickly converge and become one OSS.

3.10.3.2 Cost Reduction

While the OSS needs to step up to the task of managing this new converging world, it is also looked to as a major contributor of required cost savings. All companies are under pressures to reduce expenses and improve profitability. While telecommunications companies may have been an exception in the days of telecom monopolies, this is no longer the case. The OSS needs to contribute to the cost reductions by reducing the cost of the OSS itself as well as by reducing the ongoing operational costs.

Beyond contributing to cost reduction, the OSS is increasingly being looked to for improving profitability. This needs to be achieved by making sure that the OSS has the necessary flexibility to allow new services to be quickly introduced and reduce the order to bill delay.

3.10.3.3 Business Measures

In addition to understanding why we need an OSS, the business view also needs to consider how success (of the OSS) will be measured. While each company will need to have its own specific measures that reflect its unique business model, the measures need to be sufficiently broad to be able to provide architectural guidance in properly balancing trade-offs. For example, there is a fundamental "balance point" between operational spending and customer loss.

As part of its Business Benchmarking Service, the TeleManagement Forum has defined a Business Metrics Framework (see [GB922] and [GB935]) that can provide the foundation for measurements that can be used to indicate OSS success.

3.10.4 The Job of an OSS

Once we have a business view of a system and understand why it is needed, the next step is to look at the functional view to understand what the system needs to do.

The traditional answer to the functional view, what should an OSS do, is to either answer with the TMN pyramid or with FCAPS (Fault, Configuration, Accounting, Performance, and Security management). The TMN pyramid comes from the International Telecommunications Union (ITU) "TMN Logical Layered Architecture within the TMN Functional Architecture" [M3010]. This hierarchy is usually drawn as a pyramid (but does not appear as a pyramid in M.3010). At the base is the element management layer, followed by the network management layer, the service management layer, with the business management layer at the top.

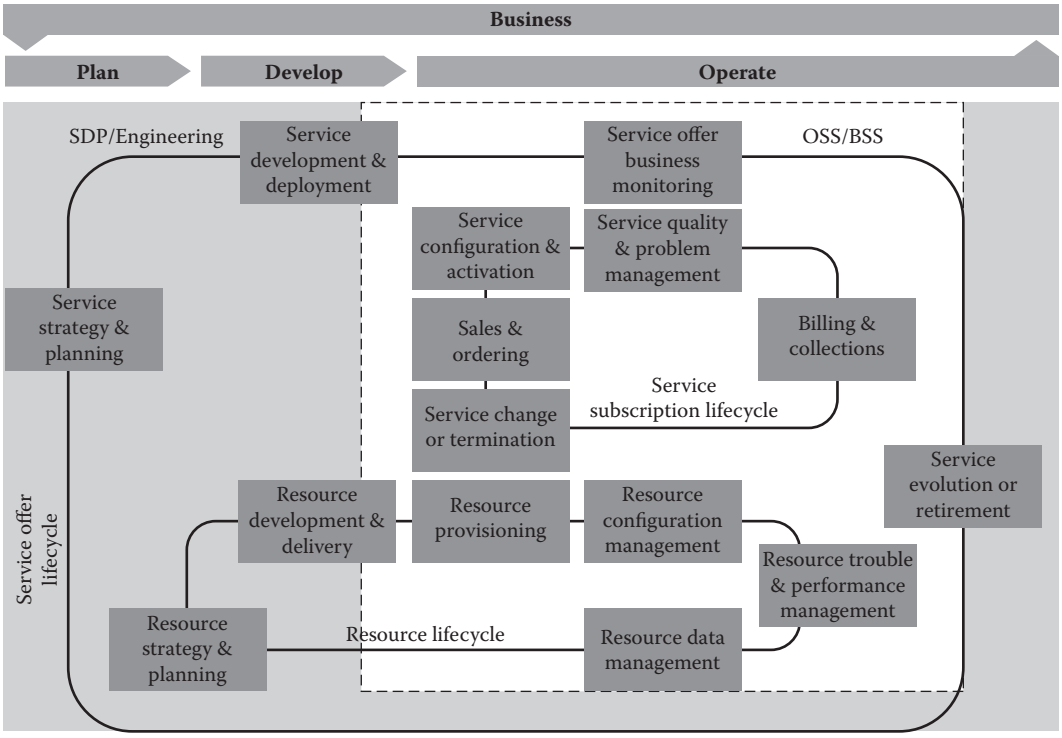
While the TMN pyramid can provide valuable insight into functional segmentation, it does not really say that much about what needs to be accomplished.

The ITU also defines the concept of five functional management areas. These are introduced in M.3010 and elaborated in M.3400 [M3400]. The five areas are usually referred to as FCAPS. While FCAPS does address some aspects of the "what" question, it addresses functionality from a technical perspective. First we need to understand the processes that the OSS needs to support and implement. Then we can look at the technical functionality needed to support the business processes.

3.10.4.1 Service Lifecycles

The approach we take to looking at what an OSS should do is to start by understanding the business processes that need to be supported. One way of doing this would be to use the eTOM standard as the base to define the functional view. While eTOM does provide a comprehensive process model for telecom companies, what we want to understand is the dynamic view of how the processes work together to accomplish the business goals.

For this, we take a lifecycle view that focuses on the three main lifecycles (Figure 3.10.1) that are involved in an OSS: the resource, service subscription, and service offer lifecycles.



FIGu RE 3.10.1 Resource, service subscription, and service offer lifecycles.

3.10.4.1.1 Resource Lifecycle

The resource lifecycle addresses the management of the revenue-generating networks and systems—those that are directly involved in delivering the services. It focuses on domain and element management, including network management for managing the network components and IT management for managing the systems and applications that are part of the revenue-generating infrastructure (VAS platforms and applications, SDP, etc.). Different domain-specific resource lifecycles may exist. The resource lifecycle is driven by both network engineering and operations, with heavy involvement from the OSS systems. Typical business scenarios include:

- Understanding the resource needs and planning for new resources. This can be driven by the need to support new services or adjust capacity to changes in demand.
- Engineering the details of the deployment of the resources. This could include determining where the equipment will be located, how it will be interconnected, etc.
- Deploying and configuring the equipment to meet the design.
- Providing for ongoing configuration management that is independent of service instances such as installing software patches.
- Monitoring the resources for any faults and performance issues, and repairing as necessary.
- Managing the operational data of the resources, in particular collecting the usage information.

3.10.4.1.2 Service Subscription

The service subscription lifecycle covers the use cases that provide a specific customer with an instance of a service. The service lifecycle must be traversed once for each service being used by a particular customer. This lifecycle is often driven by the account manager, customer service representative (CSR), and customer. The primary systems involved with this lifecycle are OSS and BSS. Typical business scenarios involved with service subscription are as follows:

- Sell the service to a customer and take the order.
- Provision the service (determine the changes and configurations required to provide the service) and activate it (make the necessary changes in the infrastructure).
- Monitor the service to ensure that the customer's expectations are being met; resolve any incidents or problems that may occur.
- Determine usage of the service, calculate bills, and collect payments.
- Change or terminate the service to the customer based on customer request or other business factors.

3.10.4.1.3 Service Offer

The service offer lifecycle covers the activities involved in the management of bringing a new service to market, monitoring its business performance, and evolving or withdrawing the service. The service offer lifecycle, also called *product lifecycle management*, is not tied to any particular customer; it provides the business view of a service offer or product. It is typically driven by a line-of-business manager. While the OSS and BSS are involved in aspects of this lifecycle, the business planning, SDP, and engineering systems are the primary systems involved. Typical business scenarios are:

- Decide on a new service to provide and plan how to deliver it.
- Develop, test, and deploy the service by making necessary changes in the infrastructure as well as in the required support systems.
- Monitor the service to ensure that business requirements are met; this will include general quality delivered, revenue, capacity, churn, etc.
- Evolve or retire the service.

3.10.4.1.4 End-to-End Service View

Within the three lifecycles, Operations looks at three major business process areas: fulfillment, assurance, and usage. Fulfillment supports the processes responsible for getting everything ready to take

customer orders to fulfilling the orders and maintaining the resources. Assurance ensures that the services and resources are working as they should and the business and customer expectations are being met. Usage supports the processes responsible for collecting information about, and understanding how, services are used and creating the billing and business intelligence information from that data.

3.10.4.1.5 Other Lifecycles

There are other lifecycles that occur within a business and within a service provider beyond the three we have considered so far. For example, the relationship that a service provider has with their customers goes beyond what is captured by the service subscription lifecycle. One could talk about a customer lifecycle that drives the broader relationship and may contain within it multiple instances of the service subscription lifecycle (i.e., a customer subscribing to multiple independent services).

On a different level, we can also look at lifecycles such as the service session lifecycle: the initiation, use, and termination of a particular session by a customer (e.g., placing a phone call or watching a movie). A typical service subscription lifecycle would include multiple service sessions. This provides the distinction between, for example, having telephone service (service subscription) and making a telephone call (service session).

While these other lifecycles could be vital for the enterprise, we do not consider them here because they only loosely affect the OSS.

3.10.4.1.6 Interactions among the Lifecycles

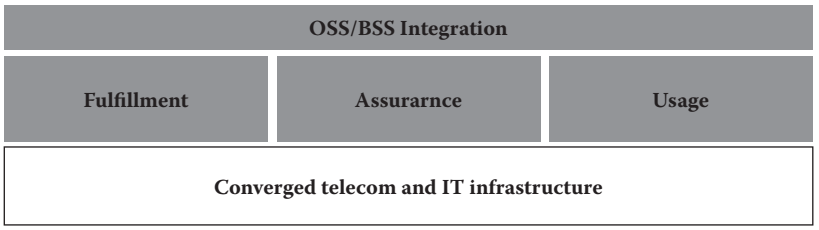
The three lifecycles are not independent of each other—links exist between them. For example, the decision to introduce a new service offer may require major enhancements to the infrastructure in order to support the service. That would drive requirements that would be implemented thought the resource lifecycle.

An interesting link is that between the service instance and the resources. Traditional services have a very direct link between these two since the services have dedicated resources. As the world evolves to include new services provided through shared infrastructure, such as packet-switched networks and VAS services provided through traditional IT systems, the link between the service instance and the resource is becoming increasingly virtualized. This, in turn, introduces abstractions to the links between the service instance lifecycle and the resource lifecycle.

3.10.4.2 Fulfillment, Assurance, and Usage

The ITU defines five management functional areas [M3010]: performance management, fault management, configuration management, accounting management, and security management (usually abbreviated as FCAPS). While this view of the functionality provides value at the element management layer, the abstractions defined by the TMF are more appropriate at the business process layer. The TMF talks about fulfillment, assurance, and billing (FAB). As we see in the following, it makes sense to expand this to fulfillment, assurance, and usage.

In addition to the three functional areas, we need to introduce functionality that is common across all three areas. We call this the OSS/BSS Integration area. Figure 3.10.2 illustrates OSS functionality.



FIGu RE 3.10.2 A high-level view of the OSS functionality.

3.10.4.3 OSS/BSS Integration Functions

The OSS/BSS Integration area contains two distinct types of things. First, we have the definitions of the business processes and policies that apply to the OSS, along with the definitions of the information that needs to be shared across the different processes and applications. Second, we have the technology that is required to support the integration and that is common across all the areas.

3.10.4.3.1 Business Processes and Policy

Since we want the business processes to drive the technology, and not the other way around, we clearly need to capture the definition of the processes. This needs to be done whether or not the processes are automated. By dealing with these up front, we can

- maintain separation between process and application, providing the flexibility to change one without the other;
- have an easier understanding of the relationship between the OSS/BSS processes and other business processes in the enterprise;
- model and simulate the processes, allowing for continual process improvement; and
- load the processes into process management engines for automation and monitoring.

In an analogous manner, we also need to understand the general business policies defined by the enterprise in, and the specific policies defined for, the OSS; understand and model how high-level policies are implemented through lower-level policies; and monitor the policies for compliance. While policy-based management has been implemented in some specific areas, such controlling access to network resources, the definition, management, and monitoring of policies in a structured and auditable way is an emerging field. The Autonomic Communications Forum (ACF) is an industry and academia group that, among other things, is working to understand this broader role for policies.

Finally, we need consistent information models for data that is common across processes or applications. As we will see later, common syntax and semantics are crucial for communication to take place. By defining a consistent model based on specifications such as TMF's Shared Information/Data Model (SID) [GB922], information models can be easily extended and any process or application that needs to use the information can easily understand it.

3.10.4.3.2 Integration Technology Layer

The integration technology layer provides the technology that enables common functionality and the integration of the various OSS/BSS components. While some people may think that this would only be a framework or middleware, much more is required here than those terms generally imply.

Operational support and readiness is the general set of functionality that supports the fulfillment, assurance, and usage areas. Included here are functions such as workforce management, contract management (which is needed for both customers and suppliers), and customer contact management (allowing a common way to communicate with customers based on their preferences). Unlike the TMF, which includes inventory as part of operational support and readiness, as discussed above, because of its importance we split inventory into its own area: unified data management.

Unified data management is one of the most critical components of a successful NGOSS. In the IT Infrastructure Library Version 3 (ITILv3), the IT industry has come to realize something that the OSS industry has known for a while: a federated repository of the management information is the only way to provide the necessary balance among completeness of information, access to information, and accuracy of information while allowing the focus to move from technology to business outcomes. This move by ITIL comes at an opportune time, allowing the network focus of traditional telecom inventory systems to be enhanced with the application and services focus of IT Configuration Management Data Base (CMDB) systems. By federating these together, it becomes possible to have a comprehensive view of the

end-to-end service, augmenting the network inventory with the complex relationships introduced by value-added services.

Unified data management needs to provide the following:

- Traditional inventory, capturing information about all the network and IT resources, and how they are connected. This includes the traditional telecom network inventory as well as information on the traditional IT equipment that is in the call path.
- Services inventory, which maintains information on how the individual service instances are provisioned. It also needs to provide information on the sequence of interactions among the components that define the service.
- Product catalog, which captures information on the service offers that are available. This would include information on how to decompose a service offer into the service elements that make up the service.
- Discovery, reconciliation, and data-loading facilities. Discovery is necessary since the sheer amount of data and its dynamic nature means that the only way it can remain current is through automatic discovery processes. Reconciliation is necessary since any discrepancies between the discovered as-is and as-planned configurations need to be understood: the network is not always right. Data loading needs to be performed to provide the necessary data to applications that require local caches for performance reasons.

While theoretically, security management is no different than any other form of management, we pull out the management of the security mechanisms because of the importance, and oftentimes the requirements for closer auditing. Note that security management is not the security mechanisms, but the rather the management of those mechanisms. For example, a firewall protecting systems from external threats is a security mechanism. Configuring the firewall to implement the business policies and monitoring it for any breaches in security is part of security management.

Finally, we need to provide support for the actual communications mechanisms. Foremost here is some form of technology to carry the information from application to application. This could be provided through middleware, a communications bus, or other mechanisms. Note that there is no requirement that this be a single communication infrastructure. Sometimes interlining different communication mechanisms provides a better solution; we will consider later why this may make sense and how to do it so it does not increase complexity. It is necessary, however, to ensure that proper governance takes place over the interfaces (or IT services) that are exposed to the communication mechanisms.

Within the communications and orchestration functions we also need to provide support for the process and policy engines that need to handle the orchestration of the processes across OSS components. Since processes and policies are integral parts of many applications, it is not expected, or even desirable, to have a single overarching process or policy engine that is used everywhere.

Figure 3.10.3 provides an illustration of the OSS/BSS integration functions.

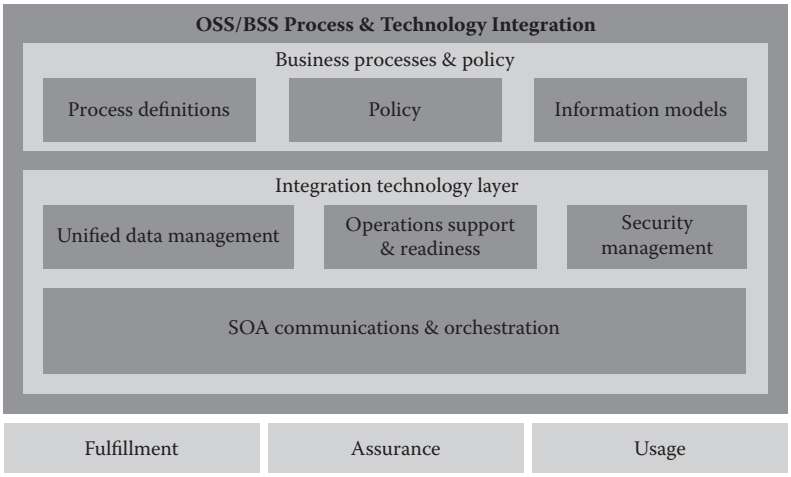
3.10.4.4 Fulfillment Functions

As we look to understand the functionality that needs to be supplied by the fulfillment area, we need to remember that all three of the service lifecycles that characterize our business processes need to be supported. Figure 3.10.4 provides an overview of fulfillment functions.

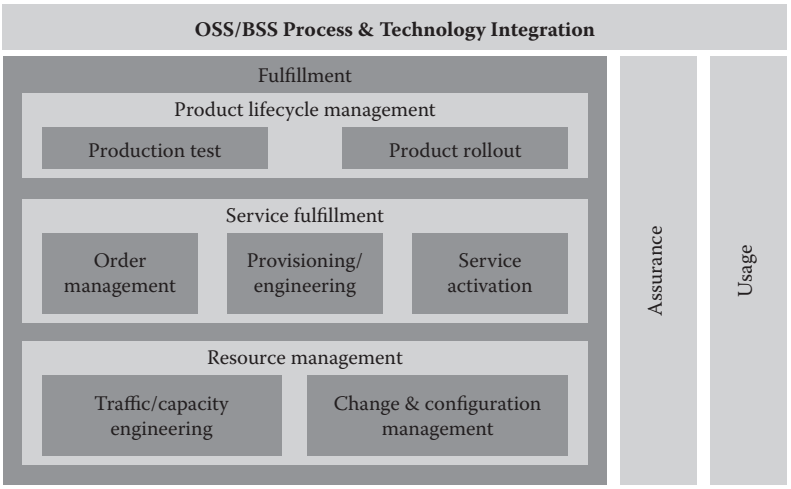
3.10.4.4.1 Service Subscription Lifecycle Support

The functionality most people initially think of when considering fulfillment is that of service fulfillment. This is the functionality that fulfillment adds to support the service subscription lifecycle.

We can look at three main functional areas within service fulfillment: order management, provisioning/engineering, and service activation. In addition, the inventory functionality that is part of the OSS/BSS integration area is heavily involved in service fulfillment, to the point of sometimes being considered part of fulfillment.



FIGu RE 3.10.3 OSS/BSS integration functions.



FIGu RE 3.10.4 Fulfillment functions.

Order management provides the technical order management (interfacing with the customer is part of the customer relationship management [CRM] systems). The first step is a prequalification check of the customer order to ensure technical feasibility. Part of the feasibility check is to provide order validation for completeness and consistency. Once the order is determined to be feasible, complete, and consistent, processing can start. Order management will decompose the order into the constituent services, provisioning and activating them using the provisioning and activation functions.

Provisioning/engineering has the task of verifying the availability, suitability, and reservation of resources prior to changing their state. For some services, design and analysis may be required to determine which resources to assign to the service. In addition, if specific service-level commitments are required of the service, the provisioning and engineering function needs to ascertain that the resources used can provide the necessary service levels. For other services, the provisioning and engineering function may be relatively trivial (e.g., assign a phone number or a free Digital Subscriber Line Access Multiplexer [DSLAM] port) and might even be performed as part of the order management or service activation function.

Service activation coordinates and executes the steps that need to be taken to enable, change, or disable a service for a particular customer. The focus is on interfacing, communicating, and changing the state of resources. Service activation needs to ensure atomicity of its actions—in the end, either the service is fully activated or the activation fails and all resources are returned to their pre-activation state.

For complex services, the design of the service instance is done by a provisioning function; for services that do not require significant planning for each instance, service activation is fully responsible for providing the service.

3.10.4.4.2 Resource Lifecycle Support

When we look at the resource lifecycle aspects of fulfillment, we need to consider the functionality required to support the resources independently of any specific service instance.

Resource change and configuration management deals with the required changes to maintain the resources. The first responsibility is to configure the resources for preproduction testing and then for production. However, change and configuration management needs to also deal with changes after the resource is in production. These changes may come about because of engineering changes (e.g., upgrade to a new version of the OS), they may be proactive measures (e.g., change the password on all routers), or reactive measures (e.g., install a patch to resolve an incident or a problem). In all cases, change and configuration management needs to verify dependencies and conflicts on the change with the other resources.

While strictly speaking traffic and capacity engineering is part of change and configuration management, it has very specific goals and operational processes to warrant considering it separately. Traffic and capacity engineering has a goal of rebalancing how the resources are being used, without affecting the services, to optimize the utilization of the resources. For example, it may reroute an Label Switched Path (LSP) across a different route or change the weights in routing calculations to better balance throughput or it may move applications to different/additional servers to improve response times.

3.10.4.4.3 Service Offer Lifecycle Support

Finally, looking at fulfillment from the perspective of the service offer lifecycle, we get to the functionalities required to deploy a new product into production. The clean and efficient handoff of a new service from the service development team to the operations team becomes more critical as the product lifespan shortens. Two functionalities are especially important.

Before a new product or service goes into production, it needs to be tested to ensure that it works properly in a production environment. This production testing may involve friendly trails, establishing performance baselines for the product, and generally validating that all operational and billing systems can handle the new service.

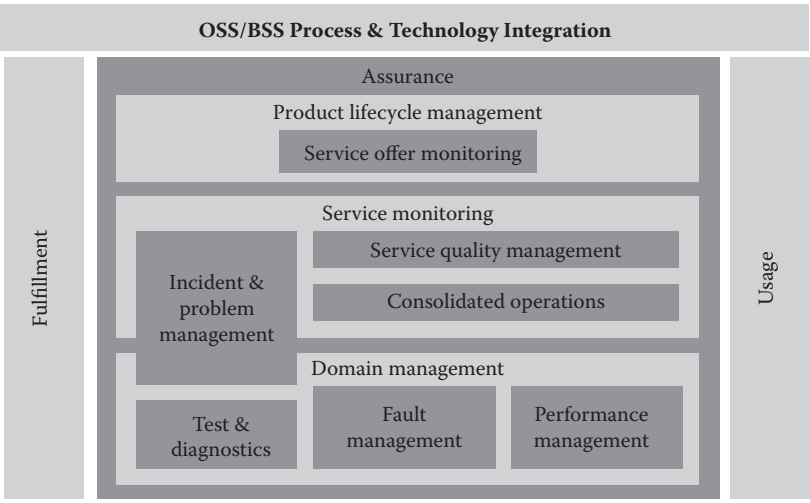
The second aspect is actually rolling out the new service into production. If the new service is an enhancement to or replacement for an existing service, production rollout will include making any necessary changes to customers of the old service. Fortunately, some of the convergence aspects between SDP, OSS, and BSS can facilitate the processes. By properly integrating the OSS with the SDP platform, it becomes possible for the OSS to automatically discover new services that have been introduced into the SDP and load that information into the OSS systems. This is why it is important to look at the service offer lifecycle across all systems that are affected by it, and not just at the systems individually.

3.10.4.5 Assurance Functions

As with the fulfillment area, we can look to see how assurance supports the three lifecycles. We consider these in a bottom–up manner, starting closest to the infrastructure. Figure 3.10.5 provides an overview of assurance functions.

3.10.4.5.1 Resource Lifecycle Support

The resource lifecycle is addressed through what is traditionally known as *domain management* or, using the TMN model, *network and element management*. Domain management focuses on ensuring



FIGu RE 3.10.5 Assurance functions.

that a particular domain is performing as expected. Domains are typically structured by the technology being managed, but may also be separated organizationally or for regulatory reasons.

The most fundamental aspect of resource management is knowing whether the various infrastructure components are working or not working. This is the role of *fault management*. It provides for the collection and correlation of alarms and other relevant events to provide an accurate view of the health of the infrastructure.

The lack of faults does not necessarily mean that the infrastructure is running properly. Though the infrastructure may be functioning, it may not be performing; the load on it may be such that it is being asked to do more than it can. This is where performance management comes in.

Performance management involves the collection and analysis of performance data. The data can be collected from performance counters in the equipment or in element managers. It may also be collected from instrumentation added in the form of probes. These probes could be passive (monitoring activities and taking measurements) or active (simulating a demand for service and measuring the pertinent response times).

Real-time performance management monitors the collected data, making sure it is within prespecified parameters. Whenever the data crosses the predefined thresholds, an event is generated to the fault management system. In addition to the real-time nature of performance management, the data is also used to identify trends and create forecasts.

In addition to monitoring the state and performance of the infrastructure, it is also important to be able to run tests and diagnostics, both to get further information that can be used to better understand faults or poor performance as well as to verify that all the components respond as they should. This verification process can be a desirable last step of a provisioning process.

Finally, *incident and problem management* supports the resolution of any incidents and problems. We use the ITIL terminology here rather than a more traditional telecom terminology since it brings additional clarity. Incidents are the events that cause or may cause a disruption in a service or its quality. A problem is the root cause behind one or more actual or potential incidents. For example, an incident could be "excessive retransmissions on a link"; the problem could be "failing board" or "heavy rain causes microwave signal degradation." The goal of incident management is to resolve the incident as quickly as possible. The goal of problem management is to resolve the problem so that it no longer exists or, if this is not possible, capture the necessary knowledge so that any incidents that do occur can be identified and resolved faster.

3.10.4.5.2 Service Subscription Lifecycle Support

Assurance support for the service subscription lifecycle provides us with what is often called *service monitoring* or *service management*. The focus here is to monitor the end-to-end services being delivered to customers, whether individually per customer or aggregated across customer groups.

While fault and performance management need to have components that are domain-specific, we also need to have a consolidated view across all domains. This is important since a fault in one domain can cause faults in other domains. Through facilities such as cross-domain root cause analysis, consolidated operations provides the ability to understand the underlying cause of a number of related failures. Through service impact analysis, it provides a mechanism to identify which services are affected by a fault.

Service quality management monitors whether the services delivered match the customer expectations. This goes beyond Service-Level Agreement (SLA) monitoring. Not all customers will have a formal SLA. However, every customer has expectations about the service quality. These will be set either by a formal SLA, by the service provider's marketing department, or by the competition. In any case, not meeting the service expectations results in dissatisfied customers and churn.

This does not imply that every customer's individual service level delivered for every service session must be monitored. While this may be practical for some services, it may be prohibitively expensive and thus not cost effective for other types of services. In many cases, especially with mass-market consumer services, the requirement is to have an aggregate view across the customer base, with the ability to focus on particular customers when required.

Finally, the service subscription lifecycle also needs to deal with incidents and problems. While the tools for resource and service-level incident/problem management may differ, they need to work together.

3.10.4.5.3 Service Offer Lifecycle Support

The assurance aspects of the service offer lifecycle provide information that can help ensure that the business expectations are being met. This information is generally not directed toward the operations department but rather toward the line of business managers responsible for the various products.

For example, service offer monitoring will monitor if the service expectations being negotiated in SLAs or set by the marketing department are reasonable. If the expectations are too high, then either the services or infrastructure needs to be improved to meet the expectations or the expectations need to be lowered. If the expectations are too low, then the service provider may be able to market more aggressive service expectations (thus gaining competitive advantage) or reduce costs by scaling back the requirements on the infrastructure.

3.10.4.6 Usage

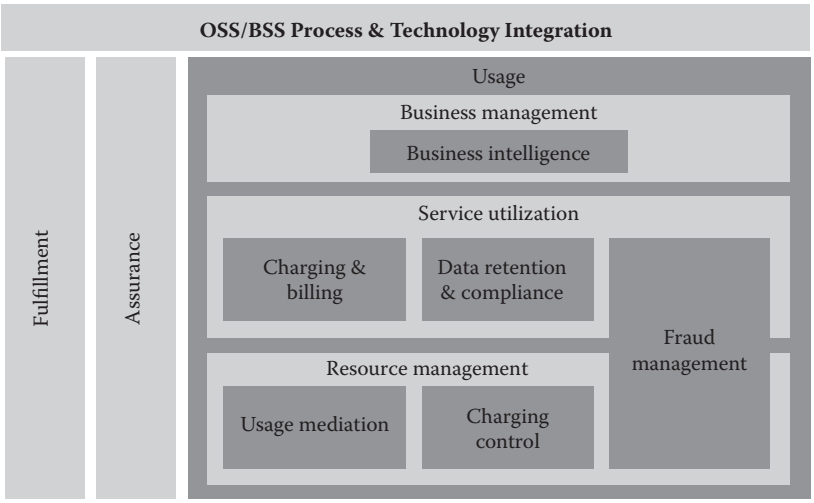
The TMF talks about FAB, or fulfillment, assurance, and billing. While that makes for a nice acronym, the term *billing* is too limiting for what is really involved. While billing for the services delivered is critical, much more needs to be provided than just billing. This is captured by using the broader term *usage*. Figure 3.10.6 provides an overview of usage functions.

The usage area is traditionally handled by the BSS and not the OSS. However, since there is a move toward convergence of OSS and BSS, we provide an overview of the usage functionality here.

Usage mediation collects usage data from the infrastructure and correlates this into usage data for each service session. While the primary use of this information is billing for usage, the usage records often contain additional valuable information, such as session quality, which can be used by the assurance processes.

While usage mediation is sufficient for postpaid billing, prepaid and pay now billing schemes require charge control mechanisms that can authorize session initiation based on account balance, terminating the session if necessary when a balance is exhausted.

For postpaid billing, charging and billing translates the session usage information into bills for individual customers. This includes rating the usage or each session and aggregating it into a periodic bill.



FIGu RE 3.10.6 Usage functionality.

Beyond billing, the usage functionalities introduce two additional functions that need to understand the individual service instances. The data retention and compliance facility addresses the increasing set of regulations that require both the collection of data for law enforcement uses coupled with increasing protection of the data privacy. Fraud management, which also stretches into the resource management lifecycle, seeks to minimize fraudulent use of services or infrastructure resources.

Finally, the business intelligence functions collect information from disparate sources into a business-focused data warehouse. Information that was previously only used for billing can now provide rich information about competitive opportunities.

3.10.5 How to Build an OSS

Over the years, many approaches have been taken to try to standardize how OSS should be built to enable and simplify interoperability among the various components. These have been primarily focused around methods of communication rather than the content of the communication. The primary area where a concerted industry effort had taken place in defining content is in the standardized definition for some of the managed resources. While standardizing the Management Information Base (MIB), which describes the managed resources, allows applications to relatively easily manage resources from different vendors, it does not address the issue of different OSS applications communicating and cooperating to address a business process.

The current trend these days is Web services. Build everything using Web services and it will be magically interoperable. Unfortunately, life is not that simple. To understand why, let us first look at how people communicate with each other.

In human communication, we often speak of the method, structure, meaning, and effect of communication. The method (sometimes called the *empirics*) discusses how the message transfers from one person to another: spoken word, written words, signs, etc. The structure of the communication is defined by the syntax and usually tied to a language. The meaning is conveyed by the semantics of the communication. Ultimately, the intended goal or desired effect of the conversation is covered by the pragmatics. For example, if you call me on the telephone (method of communication), I might say in English (syntax) "I am busy" (which has a semantic meaning to it). What I really mean is, "please call back" (intended meaning). While all the steps leading up to the pragmatics are necessary for the communication to take place, the goal of human-to-human communication is at this layer.

Similar considerations come into play when we look at application-to-application communication. First, we need to establish a method of communication. While this is important and can be complex, it is far from sufficient; sadly, many specifications stop here. The more astute will realize that they need to specify a common syntax. These two layers are also where much of the standardization efforts seem to focus. Yet even when a common syntax is defined between applications, there is no guarantee that the applications will be able to really communicate.

To understand why, we can look at an example. Two applications can share a common syntax for "contact" and thus know how to exchange contact information. Yet unless there is common semantics defined that indicate what contact means, application A may think contact is the billing contact while application B may think that contact is the person reporting a fault—often very different people. So defining a common semantics will ensure that the two applications apply the same meaning to the information they are exchanging. But unless we define a common pragmatics, the two applications may not have a common understanding of why they are exchanging the information in the first place. Often we are asked how much it will cost to integrate two applications. The answer is always the same: "Anywhere from $0 to millions of dollars, it all depends on what you want the integration to accomplish."

Fortunately, methodology for building systems exists that, if used properly, provides the necessary approach. This is what has come to be known as *Service-Oriented Architecture*.

3.10.5.1 SOA Approach

Service-Oriented Architecture (SOA) is an approach that considers the services provided by the different organizations in an enterprise and the parallel services offered by the systems that support those organizations. Properly done, the service decomposition is always based on the objectives and desired business outcomes of the enterprise. By recognizing that any system can be decomposed as a system of systems, this decomposition can be recursively refined to the necessary level of depth.

While the term *SOA* is relatively new, the concepts behind SOA have been around for a while. The TMF's NGOSS is defined using SOA principles, formalized through well-defined "contracts" or interfaces providing the service specification. The HP OSS solution, which was developed simultaneously with TMF NGOSS (with some of the same people involved in both efforts), also follows the SOA approach.

These days, SOA is often equated with Web services. While Web services do provide a good mechanism for implementing a SOA, it is important to recognize that SOA is a technology-neutral approach. As we saw in our discussion of human-to-human communication above, the method of the communication, though important, is not the critical aspect of the communication. Similarly, while Web services might make the implementation of SOA easier, there are other technical approaches that may make more sense in some situations.

In its implementation of SOA interfaces for the OSS applications, HP has developed an approach that allows us to create different integration profiles. These integration profiles provide the communication mechanisms without affecting the semantics or intentions behind the communications. This approach allows different technologies to be used, and even to change the technology used for communication, without affecting the business processes or even the applications implementing the processes.

3.10.5.1.1 Governance and Standardized Interfaces (SOA Services)

Many years ago it was not uncommon to hear "Oh, you are building an application using DCE? So am I. That's great, they will be able to work together as a distributed system!" That thinking has remained over the years, with only the technology changing from DCE to CORBA, EJB, JMS, XML, and so forth. The same sentiments now exist with the technology discussed being Web services. While it is true that these technologies facilitate building distributed systems, and each technology brings particular value, these technologies, on their own, are insufficient to allow the definition of a distributed system. It does not take too much effort to come up with examples of independently defined interfaces for similar functionality that use sufficiently different syntax and semantics to make mapping between the two problematic at best.

What is needed is not just a specification of the technology to use and the communication patterns, but the syntax and semantics of the interface. Two groups had been working on this within the scope of the TMF: OSS through Java (OSS/J) and Multi-Technology Operations System Interface (MTOSI). These two groups are now joining as the TMF Interface Program (TIP). While the success of TIP will need to be evaluated over time, it has set out in the right direction.

By using these standardized interfaces, HP has successfully integrated its applications with those developed by other vendors, cutting the integration costs by about half.

While the technology and communication patterns remain important, if interfaces are defined at the syntactical and semantic level, it is possible to have the same interface with different integration profiles. While OSS/J started defining interfaces only for EJB, two additional integration profiles were added: XML/JMS and Web services. HP now implements their OSS interfaces using technology that allows additional integration profiles to be easily created. A single adapter can communicate through multiple integration profiles simultaneously, allowing the most appropriate technology to be used. Different technologies can be chosen to integrate two applications that are part of the same operations domain as those that are controlled by different departments, or even those between the service provider and a customer, yet the interfaces remain consistent.

3.10.5.1.2 Considerations for the Implementation View

After we determine the why, what, and how of a system, we finally come to the "with what" questions of the architecture. While the business, functional, and technical views of an architecture provide critical and necessary structure to the final system, the implementation view provides the bricks and mortar that ultimately makes up the system. Two things need to be especially considered here: the deployment steps that need to take place and the actual products that will be used.

3.10.5.1.3 Deployment Steps

Whether one is looking at an NGOSS, a regular OSS, or any other complex system, there is a basic set of steps that need to be followed to ensure proper diligence. These can be summarized as define, design, develop, deploy, and operate. The steps need to be brought together through program management.

3.10.5.1.3.1 Define The first step is to define or plan what needs to be done. This includes defining the policies, principles, and architecture that will guide the implementation of the OSS. HP Services uses the methodology outlined here to help service providers develop the necessary plans. While many aspects of the architecture may match what is described here, every service provider has a different business environment and business goals, resulting in each company's architecture needing to be different.

An important aspect of planning is defining the sequence of phases that need to be followed to transform an existing OSS into an NGOSS. Big-bang replacements do not work. The phasing needs to determine which processes and applications need to be replaced and when. The complete OSS needs to remain a viable and working system throughout the phases and rework needs to be minimized.

The definition stage should also validate that the plans are financially viable. Through processes like HP's Financial Returns Assessment, it is possible to understand the financial impact of each phase and the complete transformation.

3.10.5.1.3.2 Design The design step takes the defined plans and provides the specifications needed to be able to implement the system or phase. A valuable approach to use for the design step is to take the business and functional views defined in the architecture and elaborate these into business scenarios, or use cases, that can drive the design.

An important aspect of the design step is to select the components that will be used. Business rules may mandate competitive selection of vendors, so the design step may need to include creating a Request for Proposal and evaluating the submissions of the vendors.

The design needs to take into account not just the technology, but also the business processes that need to be supported by the technology. To aid in keeping the technology aligned with the processes,

HP has developed a tool, called the HP Process Framework. The Process Framework is a tool that brings together operational processes, information models, applications, and application interfaces. By incorporating standards like eTOM and ITIL with HP best practices into its multidimensional object modeling techniques, it is possible to define and model processes, and the information flowing in the processes, from the highest level all the way to the interfaces to the applications that implement them.

3.10.5.1.3.3 Develop The develop step used to include a lot of custom development of bespoke applications. While this allows the exact desired functionality to be implemented, in most cases this is not a cost-effective approach. Even if the development does not introduce significant costs, the maintenance and support of the applications can quickly become prohibitive.

These days, the develop step does not refer to developing new applications through coding but rather developing the complete system by configuring and customizing off-the-shelf applications.

A significant effort during the development step is to integrate different applications. Two approaches can be taken to minimize this step. The first is to insist that all applications come from a single vendor already integrated. While this may provide some short-term cost reduction for greenfield environments, this approach does not work when there are legacy applications present, nor does it provide future flexibility. Rather, a better approach is to use standardized interfaces between applications, such as those defined by OSS/J or the TeleManagement Forum Interface Program (TIP). HP has successfully used this approach to reduce integration costs by 40 to 50%.

Since the complete system is more than the technology, the development step also needs to take into account the full specification of any new processes that need to be specified.

3.10.5.1.3.4 Deploy The deployment step takes the solution and rolls it out into production. The technology aspects of deployment are generally well understood. What is not always remembered is that the system includes not just the technology, but also the people and processes. The deployment needs to also include deploying the new processes and, even more importantly, training the people involved in both the new technology and the new processes.

In those cases where the transformation to an NGOSS can include changes to organizational structures, deployment also needs to include management of change: helping the people change from one set of organizations, roles, and responsibilities to a new one.

3.10.5.1.3.5 Operate Finally, once the system is deployed, it needs to be operated. For an OSS, there are two aspects to this. First, there is "operate" from the perspective of the services and infrastructure that the OSS is monitoring. From this perspective, operate means being involved in the OSS aspects three lifecycles. This is the responsibility of the operations department.

The other way to look at "operate" is from the perspective of the OSS itself. Somebody needs to make sure the systems that make up the OSS are up and running, performing as needed, and maintained. While this is sometimes done by the operations department, it can also be done by the IT department or even outsourced. From this perspective, the OSS systems are no different from any other internal business-critical system.

3.10.5.2 The Applications

While a discussion of the applications comes last so that we can keep them subservient to the business requirements, the OSS is a system with a heavy technology component, making the right selection of applications critical.

3.10.5.2.1 Telecom vs. IT applications

As with many things, when selecting OSS applications, there are extremes to avoid. One set of extremes is between specialized telecom applications and regular IT applications. In the past, the choice was obvious—regular IT applications were not adequate to the task, requiring telecom-specific applications.

With the IT world growing up and with more and more telecom services relying on traditional IT infrastructure, this is no longer the best approach.

However, just moving to standard IT applications will also not work. That is because there are some unique aspects that need to be provided by an OSS. First, telecom companies have types of equipment that they need to manage that are not generally seen by enterprises and regular IT software. For example, very few enterprises have a need to manage their own SDH network. Yet a facilities-based service provider with an SDH network will need to manage all aspects the network, something not doable through standard IT management software. Second, enterprises do not have a strong distinction between service offer and service subscription, nor between different customers. Because of this, regular IT management software generally is missing the functionality that can deal with these concepts. Finally, the scale that a telecom needs to manage is, in almost all aspects, much larger than the sale in an enterprise. While there are some very large enterprises, most regular IT management software has different scalability characteristics than that needed by a large telecom company.

HP NGOSS includes both telecom-specific and general IT software. Where the HP Business Technology Optimization (BTO) software meets the telecom requirements, it is used. Sometimes, the HP BTO software core meets the requirements, but it needs telecom-specific configurations, such as using an SID-based data model. In those cases, HP NGOSS augments the HP BTO application with telecom-specific configurations and customizations. Where the functionality or scalability requirements of the telecom industry go beyond what is needed for enterprises, HP NGOSS uses HP software specialized for the industry or it partners with other companies.

3.10.5.2.2 Integrated Monolith vs. Point Products

Another set of extremes to avoid is between a single monolithic application and multiple point products. A single monolithic application may work for managing a limited set of environments, but it introduces two main business problems. First, if there are any existing legacy applications (and this will be the case in most places, at least during some of the transformation phases), the monolithic application will have a harder time working with the legacy systems. Second, a single monolithic application will be less flexible. Future new services or infrastructure may require functionality not supported by the monolithic application vendor.

The opposite extreme is building the OSS from multiple different point products, all selected independently. The value here is that no company has all the required software functionality. The problem is that, if each application comes from a different vendor, this will increase costs.

HP NGOSS brings a whole portfolio of HP and partner applications that can act together as a unified system, but allows any application to be replaced. That allows legacy applications to easily fill in where they provide the necessary functionality and retain business value. This avoids the monolithic application problem. Because HP has a broad set of applications, bringing in and pre-integrating partners where necessary, this avoids the problems with multiple point products.

3.10.6 Final Words

We have looked at the four major views of the HP NGOSS: business, functional, technical, and implementation. These are not the full architectural views, nor are they exactly applicable to any one company. Rather, they are sketches of the solution that can be adjusted and filled in to meet the specific business requirements of any service provider. The focus of HP NGOSS is to support service providers as they offer new convergent services that rely on a combination of traditional telecom service elements together with new service elements provided through traditional IT systems. The approach is equally valuable to service providers that are remaining with just providing traditional telecom services, as well as new types of service providers that are only relying on IT systems for their services.

Several years ago, a CEO of a national telecommunications company asked the vice president in charge of networks, "I am already building you an NGN, why do you also want an NGOSS?" It is hoped that with this paper, that question will be unnecessary.

Acronyms

ACF	Autonomic Communications Forum
BSS	Business Support System
CMDB	Configuration Management Data Base
CORBA	Common Object Request Broker Architecture
DCE	Distributed Computing Environment
DSLAM	Digital Subscriber Line Access Multiplexer
EJB	Enterprise Java Beans
eTOM	Enhanced Telecom Operations Map
ITIL	IT Infrastructure Library
JMS	Java Message Service
LSP	Label Switched Path
MTOSI	Multi-Technology Operations System Interface
NGOSS	New Generation Operations Systems and Software (TMF) Next Generation Operational Support Systems (HP, common usage)
OSS/J	Operational Support Systems through Java
SDP	Service Delivery Platform
SOA	Service-Oriented Architecture
TAM	Telecom Applications Map
TIP	TMF Interface Program
TMF	TeleManagement Forum
XML	eXtensible Markup Language

References

[GB922]: TeleManagement Forum. 2005. *Shared Information/Data (SID) Model*, GB922, Release 6.0, November.

[GB935]: TeleManagement Forum. 2006. *Business Metrics Framework*, GB935, Version 2.2, August.

[IEEE1471]: ANSI/IEEE. 1471-2000. 2000. *Recommended Practice for Architecture Description of Software-Intensive Systems* (also adopted by ISO as ISO/IEC 42010:2007).

[M3010]: International Telecommunications Union. 2000. *Principles for a Telecommunications Management Network*, M.3010, Section 9.5, February.

[M3400]: International Telecommunications Union. 2000. *TMN Management Functions*, M.3400, February.

Summary and Trends

Kornel Terplan

As a result of competitive pressures, service providers are accelerating the process of service and product creation. This fruitful strategy allows existing customers to be retained and new ones to be acquired. In most cases, however, service creation and infrastructure innovation are not in synch with respect to speed. The same physical infrastructure will be used by many services, many customers, and many business partners. It is extremely important to identify the potential business impact of infrastructure components. However, managing the product portfolio and the networking infrastructure is, in many cases, an afterthought. Service providers must remedy this situation in the near future.

Service providers are on the right track to streamlining many service offerings using IP as a basis. Everything will be offered over IP; triple-play, quad-play, and more or less "any-play" will use IP as an aggregation platform. The transition to IPv6 will lift all address-space limitations. However, this transitioning process may take many years.

Service provider support systems—OSS, BSS, CRM—will be interfaced more tightly with each other, showing a much desired and expected convergence. The most important prerequisite is cleansing of databases. The starting point is the unification of customer databases. SOA may help considerably in connecting legacy and new applications. SOA and Web 2.0 may even be combined with each other to increase effectiveness and efficiency.

The use of open source software is no longer a trend but a presence. It is expected that proprietary solutions will be seamlessly combined with open source. Although open source software can be considered as freeware, it is not free of charge. Careful evaluation of technical and legal risks is absolutely necessary. Open source software is a fresh and new technology that represents an incredible tool with respect to motivating and retaining IT talents and personnel of service providers.

Service providers will take advantage of collected call detail and performance data and will use them for various purposes such as mediation for billing, behavior analysis for highlighting security violations, lawful intercepts for intelligence consolidation, service assurance, and business intelligence.

Many service provider business processes in the areas of service fulfillment, service assurance, billing, and revenue assurance are moving to the Web, taking advantage of the capabilities of Web 2.0 (Enterprise 2.0). Application performance management for these processes will have a key influence on customer experience and user satisfaction.

Best-of-suite solutions are becoming very popular with service providers. In addition to leading OSS/BSS/CRM vendors such as Amdocs, Telcordia, Hewlett-Packard, Convergys, and Subex, enterprise management platform suppliers such as IBM, BMC, CA, and Oracle are entering this field and want their market shares with high agility.

In summary, service providers are expected to measure service quality, optimize performance, and continuously plan for capacity. They will extend their scope of services by measuring and managing the user experience.

4

Network Organization and Governance

Introduction **4**-2

4.1 Information Life Cycle Management **4**-3
Terms, Standards, and Statistics • Document Life Cycle • Hot Topics • Critical Success Factors of Document Management • Summary and Trends

4.2 Information Technology Alignment with Businesses **4**-17
Does IT Matter? • Support of Business by IT • Baseline Information Technology (IT) plan • Real-Time Enterprise (RTE) • Directions for Service Orientation • Role of the IT Infrastructure • Future Trends

4.3 Business Intelligence and Analytics **4**-27
The 4 Cs of the Telecommunications Industry • Customer Is King • Business Intelligence in the Telecommunications Industry • Strategy at Work • Disappointments from the Past and Barriers to Business Intelligence • Overcoming Barriers • Customer Intelligence • Impact of BI and BPM on the Telecommunications Industry • Summary and Trends

4.4 Service-Level Management **4**-51
Introduction • Principal Terms and Metrics • Process of Service-Level Management • Sample SLA • Certification of SLAs • Role of SLAs in Settlements between Service Providers • Summary and Trends

4.5 Management Services and Outsourcing **4**-77
Introduction • Policies and Tasks of Governance • Multiprovider Collaboration and Peering • Management Services and Outsourcing • Contract Management • Summary and Trends

4.6 Network Management Organization **4**-107
Organization Structure of the Average Provider • Assigning Subject Matter Experts to Processes and Support Systems • Building the Management Teams • Keeping the Management Teams • Job Profiles for Human Resources of Telecommunications Service Providers • Summary and Trends

4.7 Best Practices Benchmarks for Service Providers **4**-123
Benchmarking • ITIL (Information Technology Infrastructure Library) • ISO/IEC 17799 • CoBIT (Control Objectives for Information and Related Technology) • Dashboards and Balanced Scorecards • Summary and Trends

Summary and Trends **4**-147

Patricia Morreale
Kean University

Deepak Pareek
Consultant

Kornel Terplan
Industry Consultant and Professor

Christian Voigt
Siemens AG

Introduction

All service providers need positioning. Executives want to know where they stand in comparison with the industry average and to best practices. Benchmarking is well known and well accepted in many industries. This chapter devotes particular attention to benchmarking indicators, myths, advantages, and risks. Product and service portfolio benchmarks are very popular with service providers. But differentiation is difficult because neither customers nor service providers know the difference due to the high number of product and service options.

This chapter offers other opportunities for differentiation by addressing IT alignment with business, information life-cycle management, meaningful use of business intelligence, and implementing governance to the enterprise as a whole.

How important is IT to the business processes of service providers? This chapter will give concrete, actionable directions for improving the effectiveness, efficiency, and reputation of IT. In particular, state-of-the-art business applications–related technologies, such as SOA (Service-Oriented Architecture, SOBA (Service-Oriented Business Application), SODA (Service-Oriented Development Application), Business Impact Management (BIM), Business Service Management (BSM), and the requirements of the Real-Time Enterprise (RTE) are addressed in some depth. Their proper use—in-house or outsourced—can ensure the expected competitive advantage.

IT managers are challenged by the need for innovation, virtualization of servers and data centers, supporting agile and real-time business processes, by budget cuts and by overaverage turnover of human resources. As a result, there is little time left for them to address document management. But the number of files, records, images, video clips, e-mails, PowerPoint presentations, collaboration logs, and blogs to be managed is increasing considerably. Even worse, meeting compliance demands by the enterprise may extend the life cycle of documents requiring more sophisticated storage resource management.

Business intelligence is a discipline of developing information that is conclusive, fact based, and actionable. Business intelligence gives enterprises the ability to discover and utilize information they already own, and turn it into the knowledge that directly affects corporate performance. It is an umbrella term that ties together other closely related data disciplines including data mining, statistical analysis, forecasting, and decision support.

Service-level management (SLM) regulates the relationship between customers and service providers. After an introduction to the principal phases of SLM, key performance indicators (KPIs) are addressed in greater depth. A proven template for service-level agreements (SLAs) is also included. The popularity of managed services for customer support, integration, and reporting is growing continuously.

Actually, the culture of an enterprise depends to a large degree on governance. Five different alternatives will be discussed: hard management, centralizing model, centrifugal model, loosely coupled organizations, and the laissez-faire approach. Workforce is one of the critical success factors in building, maintaining, and operating communication networks. Motivated people are the differentiating factor between a well run and a badly run business in the service provider area. One section focuses on human resources and their management by introducing a sample organization. Human resources are assigned to principal business processes and support tools of the service provider. Typical job profiles will also be introduced.

Outsourcing is always a viable option for enterprise executives. Various business models have been implemented with mixed results and experiences. In any case, new responsibility areas are on the way to be implemented. Contract management and managing a distributed workforce are gaining momentum.

Standardization is an important way to approach best practice performance. Three management and control specifications (as part of standards) will be addressed: Control Objectives for Information and Related Technology (CoBIT), IT Infrastructure Library (ITIL), and International Standard Organization (ISO) 17799. CoBIT tells management what to monitor and control. ITIL describes how to go about implementing the processes for doing that. ISO 17799 lays out a process for securing these services and addressing legal requirements.

The final part of this chapter focuses on the use of dashboards and scorecards to improve visibility and effectivity of the overall performance of the enterprise.

4.1 Information Life Cycle Management

Kornel Terplan

IT managers are challenged by the need for innovation, virtualization of servers and data centers, supporting agile and real-time business processes, by budget cuts, and by higher than average turnover of human resources. As a result, little time is left for them to address document management. But the number of files, records, images, video clips, e-mails, PowerPoint presentations, collaboration logs, and blogs to be managed, is increasing considerably. Even worse, meeting compliance demands by the enterprise may extend the life cycle of documents requiring more sophisticated storage resource management. This section addresses the principal life cycle of documents, including the following steps: create, distribute, archive, and manage. Particular emphasis is on creating good documents, data leak prevention, criteria for deleting and retaining documents, impacts of compliance, and links to other support systems, such as Enterprise Resource Management (ERP), Customer Relationship Management (CRM), Partner Relationship Management (PRM) and Enterprise Search Engine (ESE), and outsourcing.

4.1.1 Terms, Standards, and Statistics

4.1.1.1 Terms

According to *Merriam-Webster's Collegiate Dictionary*, a document is an original or official paper relied on as the basis of proof or support of something or a writing conveying information. Using this basic definition, anything put on paper becomes a document. But the document has evolved beyond this point. It has become more complex based on technology and components, such as pictures, graphics, and charts. Variable data can reside in different locations and not really become a document until composed into a meaningful format conveying information.

Records were defined from early on as single pieces of information, not a complete history as the term implies today. Also the term *content* is frequently used. Any content, regardless of source and type should be storable, searchable, and routable according to business rules. Content Management Systems (CMS) require a new way of thinking about records and documents. Once transactions are complete, customer service and retention become priorities, and secure access to documents and information becomes a key concern. In addition, companies must comply with far-reaching privacy and reporting regulations. Organizations must effectively manage content from the time they are created or received, through distribution, use, and maintenance, until they are finally destroyed or permanently archived.

Every enterprise has records, such as expense reports, tax bills, invoices from vendors and suppliers, copies of outbound communications, and internal memos. Some of the records may arrive on paper with an electronic follow-up, while other records may arrive only electronically or only on paper. Electronic records may be copied and printed, and may be used as a tracking document through a reconciliation process. Records include not only the primary documents, such as bills and statements, but also secondary documents, such as e-mails and even paper notes related to business activity (GREW04).

Developments have made human resources accessible 24 hours a day, anywhere in the world. It has provided instant information at our fingertips. We can now view documents of world news, get driving directions, information on enterprises and persons on a laptop, PDA, or cell phone. PCs and IP technologies enable users to do research, and stay in touch with employers and families without interruption. Computing and communication technologies are changing the way we are utilizing documents.

Businesses may be called upon to produce documents according to Rule 34 of the Federal Rules of Civil Procedure (Source United States):

A party (e.g., enterprise) must produce upon proper request any designated documents, including writings, drawings, graphs, charts, photographs, phone records and other data compilations from which information can be obtained, translated, if necessary, by the respondent through detection devices into reasonable usable form.

A document informs and/or entertains. All documents have some type of information, whether pertinent or not. Some documents simply entertain. It is hard to find documents that do both.

4.1.1.2 Document Standards

While a Document Management System (DMS) can store any digital object, a typical DMS for an office application is primarily concerned with the following document types:

1. Microsoft Office: Word, Excel, PowerPoint
2. Scanned document images that typically use Tagged Image File Format (TIFF)
3. Graphic or photographic files that use Joint Photographic Experts Group (JPEG), Graphic Interchange Format (GIF), and Portable Network Graphics (PNG)
4. Adobe Portable Document Format (PDF)

4.1.1.3 Statistics

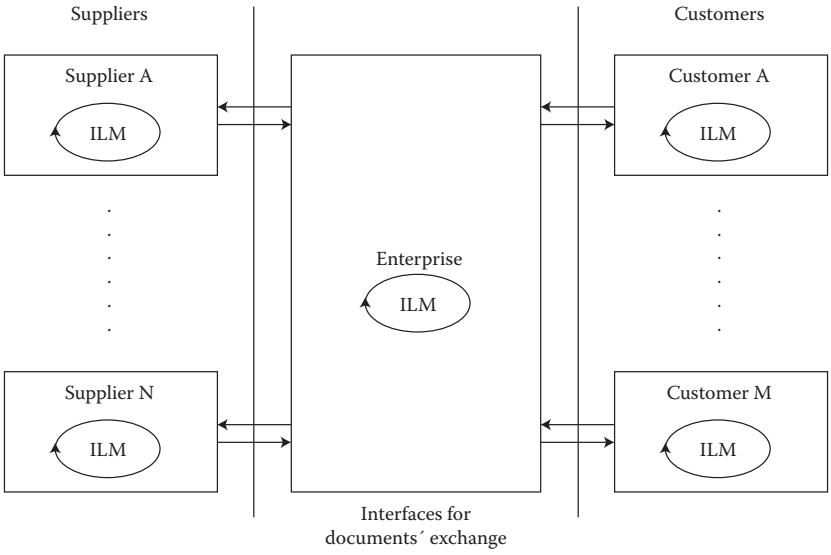
The following statistics will stimulate the thinking process for many document managers in enterprises.

- Large companies spend between 7 and 9% of their annual revenue on executing the life cycle of customer facing documents (invoices, marketing, questionnaires, etc.).
- The largest internal/mail finishing center can require between $20 and $50 million per year to operate (CapEx [capital expenditures] and Opex [operating expenses]).
- Best-of-class operations can save between 10 and 20% of life-cycle management costs per year, when optimizing processes, by selecting the right tools and when human resources are well assigned to processes and tools.
- Each typical business day there are 84 billion e-mails being sent and received worldwide.
- Billion-dollar companies in the United States face an average of 550 lawsuits in a random sample in last 3–5 years.
- Review costs can range from $1,200 to $2,500 per gigabyte.
- $2 review costs per e-mail.
- 63% of users have not yet analyzed the risks they face from mismanagement of electronic information.
- The projected size of the e-mail retention market by 2010 is close to $8 billion; up from $800 million in 2006.
- Estimated amount of digital information for an average Fortune 2000 company: 160 billion gigabytes (2006); 990 billion gigabytes (2010)

4.1.2 Document Life Cycle

The concept of document life cycle recognizes the varying aspects of documents regardless of whether they are presented in paper or digital form. Using a document life cycle to frame a document strategy recognizes the fact that no single medium can satisfy all life cycle requirements for every document application. Paper and electronic formats each have distinctive strengths and limitations.

Descriptions may vary according to specific perspectives, but most document processes include, as separate but linked steps, the creation, distribution, and archiving of a specific document. Ideally, all of these activities are managed in some way that assures quality and security. And even more ideally, somewhere between the receipt and the archiving of the document, the process is refreshed via some kind of analytics or business intelligence tools. The updating of customer information serves to extend the life of the document into the next go-around so that the process is dynamic and continuous.



FIGu RE 4.1.1 Document supply chain.

In addressing the life cycle of documents, we must assume that the enterprise is in the middle; it has relationships with customers by communicating with them through different kinds of documents. The same is true with their suppliers, who are actually doing the same from another perspective. In addition, there are the internal documents that have almost the same life cycle, but without external transactional processes. Figure 4.1.1 shows these relationships.

Workflow typically controls the way in which *workpackets* (e.g., scanned paper documents, uploaded electronic documents, data files, voice notes, videos, etc.) are received, indexed, quality assured, routed, acted upon, linked to other systems, routed again, decided upon, finalized, archived, retained, and deleted. Workflow can automate high-level processes and detailed process steps. Workflow also includes workforce management by assigning workflow steps to human resources.

Life-cycle management of documents usually includes the phases shown in Figure 4.1.2 (GERS03).

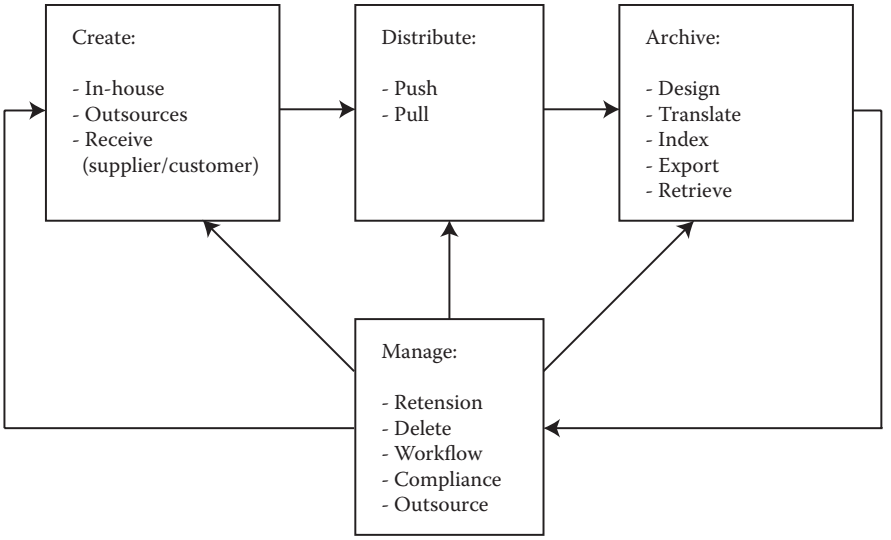
4.1.2.1 Create

Key to successfully integrating the output architecture is establishing the means to create a document that can be delivered in a number of ways. The process begins with the creation of a print file using a document composition tool. Customer data is extracted from a data repository and merged with a variable document format to create a print image.

Print-stream engineering software that operates post-process can be used to modify the print stream, either to clean up old formats, perform page counts or add barcodes to each page to enable the document to interact with file-based processing control systems to assure mailpiece integrity. These modifications need not be extensive, but the document print stream must be viewed as flexible enough if the company is to capitalize fully on the available print/mail finishing technology to accomplish processing efficiency and mailpiece integrity.

In the next step, data quality should be analyzed to prevent returned mail, a hidden and costly aspect of customer messaging, and advice given on how to improve customer data and address quality and in particular how to stay in contact with customers who change addresses.

Print/finish operations traditionally distributed documents in isolation, with little interaction with other parts of the enterprise. Some of the more advanced producers of high-volume document output are moving to file-based processing to facilitate the integration of their output methodologies.



FIGu RE 4.1.2 Management of documents life cycle.

File-based processing asserts control over inserting equipment through reading of a barcode that accesses a data file. Since the file contains more information than any number of barcodes on a document, file-based processing gives better control over the insertion process. Since the barcode is only used to point to a data file, the page itself is generally cleaner, leading to a more precise reading, which drives greater integrity. Since there are fewer mistakes, companies experience new levels of productivity. Since file-based systems are driven by data records, typically one per envelope or document, it is a more cost-effective method of document distribution. In fact, operations that have recently introduced file-based processing systems have found that they can swap out two older devices for a single file-based system. File-based processing provides numerous other benefits, such as an ability to create an electronic document file that can be used to send documents in a digital format or the ability to implement systems that track documents as they move through the life cycle. The next logical step is an interface with a track-and-trace solution that extends tracking throughout the entire mail stream using, for instance, postal service databases.

4.1.2.2 Distribute

Intelligent mail may reenergize the mail channel as its power is exploited by businesses that relay on the mail. With piece-level visibility into the mail stream, companies can more proactively manage cash flow, mitigate risk factors from fraud, prevent unnecessary dunning notices, and make each customer interaction both more rewarding and more relevant to the recipient. Intelligent mail can empower a print/mail executive to proactively create customer and shareholder value and be a positive influence on the revenue-generating activity of his/her organization.

The tracking of return reply envelopes can be used to proactively help colleagues in treasury, accounts receivable, and payment processing by allowing the financial experts to more intelligently forecast incoming dollar volumes based upon the latest status reports on mail in transit. Integrating output also allows call center agents to see a caller's statement as it was rendered, which results in decreasing call duration, higher first call resolution, and a significant increase in effectiveness as a dedicated sales channel.

4.1.2.3 Archive

Once the document is safely in-house, it should be archived in a manner that allows the marketing team to convert data into information, which can then be published within the enterprise for more efficient decision making. For example, aggregating data on when and how quickly a certain customer tends to

reply to a bill with a payment can improve cash flow management. According to that data, billing cycles may be rescheduled.

Archiving documents in a data repository allows the storage of several years worth of documents on powerful servers eliminating the need for expensive computer output to laser disk storage and makes retrieval easier for customer service agents and others who may need to see the exact document as it was rendered to the customer. Whereas microfiche and CD-ROMs have many positive attributes, the time and expense of seeking the correct document, pulling it, and reprinting it is counterproductive. Archiving documents in a flexible data repository facilitates multichannel communication. Front-end integration into customer service software and marketing automation systems is useful for campaign development and response tracking. Each mailpiece then becomes a vital customer touch point.

In reality, document imaging is a very straightforward operation. The process requires these five simple steps: (ESPI03)

1. Design: Create a database with a defined file structure
2. Capture and translate: Scan, fax, or import documents into the database while interpreting handwriting, machine print, barcodes, and data entry fields.
3. Index: Index and store documents to CD, DVD, or magnetic media.
4. Export: Distribute images via the corporate network, the Internet, or CD-ROMs.
5. Retrieve: Quickly search and retrieve any stored documents.

4.1.2.4 Manage

Since the goal of the integrated output architecture is to build a tightly integrated and information-rich environment, it is of importance that a system-to-system workflow management tool be implemented to monitor and measure the stream of data as it moves across the many front- and back-end systems.

Valuable information is often underutilized or even overlooked for integration because it exists in a format such as e-mail or in Web pages that make it seem unobtainable. Unstructured or loosely structured data formats, such as in e-mail or electronic reports, hold huge amounts of potential information that can enrich business decision-making processes. The solution must manage both structured and unstructured content equally. To ensure success, a homogeneous framework is required with the following principal attributes:

- Robust imaging capabilities that provide efficient document and content capture
- Customizable indexing fields allow for intercompany search and rapid document retrieval
- Carrier-specific indexing fields to allow for intercompany search and rapid document retrieval
- Records management capabilities to manage archiving and audit trail retention required for regulatory compliance
- A secure and scalable object repository that keeps all content in order

To be successful, solutions should meld content and process management, workflow, and enterprise and Web connectivity.

4.1.3 Hot Topics

4.1.3.1 Documents for Optimal Communications

There are three kinds of transactional documents: the good, the bad, and the ugly. Good statements tend to belong to well-respected organizations, whilst the bad tend to come from organizations that have less than stellar reputations, and the ugly tend to be either from regulated monopolies or by once-famous organizations now under the control of cost accountants. The companies that produce good documents understand that no matter how accurate or complete the information in a document is, the reader will dismiss it if it is impenetrable, contains too many fonts or the type is too small or too large to read. In short, it is ugly and consumers have better things to do than decipher such documents. These

organizations also understand that bad statements display a clean image, are uncluttered and might appear attractive, but with key information missing, sequence awry, and the what-to-do-next hidden, readers will struggle with this form, may default to expensive dialogs with customer service, probably make mistakes with responses or payments, and generally wish they had chosen another product. The good gets it right—the correct information is in the right order, important messages are highlighted, it is easy to read and remains uncluttered.

Design is an important aspect of creating a good document, but only if as much attention is paid to the content as well. It is the combination of the two that results in spectacular output. The following is advice for creating good documents (BROD06):

- Contact information: Ensure that contact details are clearly indicated. Position them in their own section on the page and distinguish between general customer service and specific contacts, if necessary. Keep corporate addresses well hidden if you do not want your customers to call. Remember, make it easy to contact the right number, the first time.
- Key information: Any key information should be highlighted so that it stands out from the rest of the text or numbers. This can apply to dates for payment, amounts, call numbers, advantages of responding, and even the consequences of not responding. Tools are available, such as emboldening, italicizing, increasing font size, decreasing font size, white space, and color.
- Sequence: Get the information into the right sequence; otherwise, you will annoy those who have to follow it. They will make mistakes, liabilities will be tested, and there will be no repeat business.
- Educational messages: Take a little more time to explain to the readers what certain terms mean, or what various options are open to them, or what to do next. It saves time, improves the customer service, and preempts calls, which can cost a lot.
- Focused messages: If possible, address the customer using personalization. Identify the customer. Target by age, ethnicity, marital status, and geography for a more focused message, and you will get a more focused response.
- Corporate identity: There is no reason why the elements of a corporate identity cannot be incorporated in company statements and invoices, which will be seen by many more customers. Select the right fonts and logo. Altogether, apply the same rigors to your variable data documents as you do to the offset material so that the output will look inviting.

4.1.3.2 Data Leak Prevention

In most enterprises, there are several access control mechanisms, such as firewalls, encryption, permission, and access control lists. But data and information still get lost due to authorized users. Data leak prevention products give IT visibility into where protected data rest and how it moves through the enterprise, letting organizations make informed choices about security infrastructure implementations. But they may also create a honey pot of sorts, so access to them must be closely monitored. Businesses will benefit from the ability to enforce policy proactively, and these products provide significant regulatory compliance solution alternatives. But they may become data bottlenecks at the perimeter as policy violations are quarantined until somebody responds and releases these violations. Prevention policies and tools protect intellectual property, a vital part of any organization, but one that varies by company.

New research indicates that while most companies are investing in technology and policy to secure sensitive data from attacks, the threat of data loss at the hands of their own employees is what should have the attention of most companies. In addition to severely damaging a company's reputation, leaked customer or corporate data can potentially result in legal action if the business violates regulations.

Information today flows through and beyond organizations more easily than ever before. Content is proliferating at an amazing rate; the volume of unstructured business content is estimated as doubling every three months. Relying on end users to enforce information retention and security policies is no longer sufficient. The leakage of privacy-sensitive personal information has been shown to have significant negative consequences on an organization's brand, reputation, and customer trust as well as legal,

operational, and financial implications. For organizations to protect such information, they must put policies and monitoring/filtering tools in place to monitor and control privacy-sensitive identity information from data leakage.

The question is: What is better—monitoring the data or the people? Monitoring company data itself is a strong control of "security-in-depth" and gives enterprises a significant and valid means of data security. However, not all threats are data-centric. The challenge of the insider threat is that it covers a variety of behaviors that put business at risk—people might do bad things. When management decides to put more emphasis on the employees' side, new roles and responsibilities must be created and implemented. Examples are:

- Responsibilities for protecting company and customer data
- The value of understanding employee behavior and the need for situational context
- Behaviors that put businesses at risk: accidental or process failure, to the malicious, hacks, sabotage, IP theft, customer theft, and fraud
- Where policies fail and breaches take place
- How to enact company policies that will curtail malicious behavior

Most of the vendors primarily address the data-leakage issue from the network perimeter as their products are designed to sit at the network edge and scan multiple communication protocols. These protocols are used for supporting various applications, such as e-mail, Web browsing, IM, and FTP, to determine whether sensitive content is wrongly communicated outside the boundaries of the enterprise network. A monitor that typically hangs off a network switch captures traffic and passes information about it back to the administrative console for analysis and storage purposes. Most enterprises initially install these products and run them for several months in a simple monitoring mode (instead of immediately blocking suspicious outgoing traffic) to watch employee work activities so they can identify trends that will assist in establishing appropriate policies. Many products offer policy wizards that help define the keywords or patterns to look for in addition to monitoring for specific user behavior, such as altering certain documents. When these attributes are used in conjunction with policy rules, administrators reduce the risk of false positives.

Once administrators have imported specific data formats, such as social security numbers, credit card numbers, and intellectual property brand identifications, into monitoring and filtering products, they can create policies that will notify them whenever data has left the corporate boundary with these patterns. Some products combine filtering and monitoring with regulatory compliance and security.

How a company uses these products is unique to the internal culture of the organization, the industry it plays in, and what it ultimately hopes to gain from using these products.

Content filtering and monitoring technology should be one component of an overall internal or external auditing process, as it keeps an eye toward improving operational efficiencies by identifying internal policy violations, providing more accurate financial reporting, limiting exposure to class-action lawsuits, and complying with applicable industry, local, and federal regulations. But can the audit logs generated by these products help in legal situations involving employees who criminally violate company policy? While noting that the privacy laws have not yet been tested in the courts, consultants say the logs and reports generated by these tools indicate that a corporation is taking effective, efficient actions to maintain privacy practices required to avoid the courtroom.

In general, these products are costly. Pricing varies greatly, but most vendors will charge per user/workstation, per appliance, or per the exit points at which information can leave the corporate network, such as through e-mail attachments, IM, and data uploading to a FTP server.

When selection of tools is under consideration, the following attributes could become important:

- Can documents be defined based on categories (e.g., internal names, draft press releases, price lists, etc.)
- Can documents be assigned different access controls?

- Can the product be "trained" to ignore common, nonconfidential content?
- Does the product have build-in templates for outgoing messages with governmental regulations (e.g., Health Insurance Portability and Accountability Act (HIPAA), Gramm-Leach-Bliley Act, and Sarbanes-Oxley Act (SOX), European Acts, the Communications Assistance for Law Enforcement Act (CALEA), European Telecommunications Standards Institute (ETSI), state regulations, etc.)?
- What is the total number of file/document types supported by the product?
- Can the product identify specific data elements (e.g., social security number, credit card number, account numbers)?
- Can the product learn key phrases that distinguish private from public documents?
- What are the financial conditions (e.g., price, maintenance, etc.)

The return on investment (ROI) regarding filtering and monitoring products is closely aligned with business risk, and they are often purchased under the umbrella of compliance and risk management.

Usually, the following protections are common:

- Word and e-mail attachments
- Copied Word content onto e-mail body
- Zipped documents transferred via FTP
- Encrypted documents transferred via standard protocols
- Data sent via HTTPS and HTTP
- Data copied to USB devices
- Image data transfers
- Altered image data transfers

Steps to prevent data loss include:

1. Guard against human error: Use security technologies, such as data encryption, as a safety net for honest mistakes.
2. When in doubt, encrypt: All laptop hard drives should be encrypted.
3. Monitor outgoing messages: Use software to block e-mail messages or the file transfers with confidential data.
4. Ensure that security is easy to use: Otherwise, employees will find ways to get around it.
5. Audit security practices regularly: Experts say such reviews should happen at least monthly.

4.1.3.3 Deleting and Retaining Documents

Enterprises are challenged with respect to the decision on deleting or retaining data. Mistakes either way may be very painful. There following provides some advice on each:

4.1.3.3.1 Deleting Documents

- Be consistent and objective about what you want to delete
- Know, what, when, and why data was deleted and by whom
- Get rid of all copies when deleting documents
- Wipe computers when switching owners or when decommissioning
- Make sure data is completely gone

4.1.3.3.2 Retaining Documents

- Know your data and organize it well
- Identify laws, acts, and regulations that affect your company and keep data as long as required
- Hold onto data that could be subject to investigation
- Educate employees on what to keep
- Avoid creating risky e-mails

Enterprises need a records retention methodology that is testable, scalable, and reliable. That means the enterprise will either build it or buy it. In any case, the following items are important (GREW04):

- An archival system
- An indexing and search system
- Transformation to normalized records and to legal formats
- Authentication as well as digital encryption facilities
- Firewall and anti-intrusion systems
- Data mining facilities
- Analytics and reporting systems
- Interfaces to back-office systems like SAP or Oracle

4.1.3.3.3 Data Destruction

There are a few methods to destroy data together with storage media. They include:

- Bash it: The goal is to smash the heads into the disk platter, which probably will break the mechanism, letting them access the data. A sledgehammer is usually the tool that should be applied to the large side of the drive.
- Heat it: Heating over the Curie point, hard disk metal loses its magnetic properties.
- Smelt it: Melting hard disk drives dawn to slag.
- Microwave it: CDs can be destroyed by putting them into a microwave for 2 to 3 seconds.
- Shred it: Use this method for hard drives, CDs, and even cell phones. Shredded material can be sent to metal recycling.
- Dismantle and demagnetize it: Hard drives are taken apart carefully and demagnetized individually.

4.1.3.4 E-Mail Management

The key in building a business case for e-mail management consists of meeting three objectives:

- Reducing legal and compliance risks
- Managing the ever-increasing costs of e-discovery
- Ensuring the availability and security of e-mail and associated records

The sources of e-mail, its attachments, and embedded files' retention and management problems are the result of the confluence of five unprecedented operational, legal, and technical factors, whose impact, both individually and collectively, have grown over time (MARK06).

- *Growth in volume:* The rapid growth in e-mail usage and subsequent storage requirements literally overwhelmed the functional capacities of many initial e-records management systems. The stunning increase in volume, much of which consists of duplicate copies and low-value records, has created a comparable increase in the effort required to search, retrieve, and review relevant records in legal and regulatory discovery activities.
- *Informality of e-mail communications:* Early use of e-mail manifested a false belief that this method of communication, like telephones, was private and available only to the communicating parties. In fact, just the opposite is true. With the metadata that is intrinsically associated with e-records, much more information can be found out about the communicating parties than just the contents of the records.
- *Growth in litigation:* Growth in the volume of litigation has become an indigenous part of doing business in several countries—in particular in the United States of America. Therefore, with this trend, a newfound emphasis on the process and power of legal discovery has come to prominence. This, in turn, compels improvement in the way e-mail, specifically, and e-records, generally, are managed throughout their respective life cycles.
- *Focus on compliance and enforcement:* Greater private sector accountability requires more regulations for business conduct. This is reflected in (1) more compliance regulations, (2) greater specificity in the regulations, and (3) increased compliance enforcement.

- *Increased tenacity of litigation:* The significant increase in legal compensation associated with verdicts and out-of-court settlements has created a financial incentive for greater tenacity by the legal profession in its pursuit of "justice." An important manifestation has been in records discovery—with special focus on e-records and e-mail—where extremely difficult-to-fulfill requests have been filed with the objective of creating situations where pretrial settlement becomes preferable to the cost and burden of complying with discovery orders, regardless of the merits of the dispute.

Perfection is not required in considering actionable solutions. Reasonable measures do not include keeping every single e-mail for an indefinite period of time. Instead, reasonable measures mean the development and maintenance of organizational and technical processes that (MARK06):

- Retain electronic information as long as it is needed for legal and ongoing business reasons and in a manner that allows for efficient search and retrieval.
- Permit the destruction of such information thereafter in accordance with the company's established retention policies and practices.
- Demonstrate that all actions taken in the life cycle of the organization's electronic records were in accordance with established policies and procedures.
- Document audit trails of key activities performed, including management oversight, including who did what and when.
- Provide assurances that the accuracy, reliability, and trustworthiness of records are preserved as the electronic records are managed over time and through any successive technology upgrades of migrations.

4.1.3.5 Impact of Compliance

Achieving compliance is a straightforward process that can be easily managed through careful planning. There are usually four steps to guide companies through the compliance process (SUFF04).

- **Step 1**: Audit your current system: The first step is to evaluate the current internal control policies your company has in place and identify any gaps in processes like document retention or approvals. Use comparison with best practices; the Committee of Sponsoring Organizations of the Treadway Commission (COSO) can help. COSO is dedicated to improving the quality of financial reporting and is a valuable resource for companies to meet Sarbanes-Oxley Act (SOX) compliance.
- **Step 2**: Set a new strategy for internal controls: Step 2 gives you the opportunity to improve your processes already in effect, investigate ways to simplify procedures and reduce costs, and create new controls when needed. An effective internal control must strike a balance among people, process, and technology. As you develop your new internal control strategy, be sure to consider factors like the structure and policies supporting your controls, the risks particular to your industry, internal communication methodology, and monitoring tools like checklists and approval slips. Successful internal controls touch all aspects of the organization, so make sure to include your execution team, financial controllers, and systems engineers in the planning process. Be specific: spell out which managers need to approve each type of document, outline which documents employees need to keep, detail the method they should use to store this content (whether physical paperwork or electronic files), and specify how long these documents or files need to be retained.
- **Step 3**: Adopt the right technology to support your new controls: Companies should consider implementing technology, such as an enterprise content management (ECM) solution as a central part of any compliance strategy and to accelerate compliance efforts. An ECM workflow solution enables companies to turn the newly created rules into an easy-to-follow process that help guide and educate employees through the new internal control rules. ECM tools are invaluable because they empower companies to archive data, retrieve it easily, based on any number of criteria, and specify permanent deletion dates when documents are no longer needed. In case of an audit, ECM tools give companies the power to locate records immediately, removing the high

TABl E 4.1.1 Important Regulations for Information Life-Cycle Management

SOX	Sarbanes-Oxley Act
HIPAA	Health Insurance Portability and Accountability Act
CALEA	Communications Assistance for Law Enforcement Act
GLBA	Gramm-Leach-Bliley Act
NASD	National Association of Securities Dealers
ETSI	European Telecommunications Standards Institute
PIPED	Canada's Personal Information Protection and Electronic Documents Act
BASEL	Rewrites banking safety rules with a new focus of operational risks, such as losses from fraud, computer failure, and acts of negligence.

cost and uncertainty of manual searching. Manual discovery is time consuming, expensive, and unreliable. Selecting the right technology framework allows a company to set up its control environment according to processes, cycles, accounts, etc. and provide the correct reporting requirements across all key criteria.

- **Step 4**: Educate your staff: Once you have solidified your new internal guidelines, take the time to walk your employees through the new process and provide a written overview of the internal controls and document retention policies. Hold a training session for your staff to educate employees and communicate the importance of everyone adhering to the regulations.

Table 4.1.1 summarizes the most important regulations that are relevant for life-cycle management of documents. The list is not complete; many states and many countries have specific regulations that are not listed.

SOX is the most general compliance challenge representing requirements for honest business conducts. Table 4.1.2 summarizes document-related technologies for SOX compliance support.

TABl E 4.1.2 Document-Related Technologies for SOX Compliance Support

SOX Section Number	Section Title and Expectations	Document-Related Technologies
103	Auditing, quality control, and independence standards and rules	Document management Message archiving
104	Inspections of registered public accounting firms	Document management Records management Message archiving
105	Investigations and disciplinary proceedings	Document management Records management Message archiving
301	Public company audit committees	Message archiving CRM or compliant management software
302	Corporate responsibility for financial reports	Financial reporting and disclosure Workflow BPM
404	Management assessment of internal controls	Internal control and audit Workflow BPM
409	Real-time issuer disclosures	Financial reporting Workflow BPM Web content management
501	Treatment of securities analysts by registered securities associations and national securities exchanges	Message archiving
801	Corporate and criminal fraud accountability	Message archiving
802	Criminal penalties for altering documents	Document management Records management Web content management Message archiving

4.1.3.6 Links to Other Applications

Enterprises operate multiple support systems. As examples, four support systems are defined.

Enterprise Resource Planning (ERP): Usually, the core for running enterprises from the financial perspective. All inventory, accounts/receivable, accounts/payable, orders, deliveries, general ledger related data, information, and documents are maintained here. Document life-cycle management must work ERP-systems on the basis of mutual document exchange agreement.

Customer Relationship Management (CRM): Effective document management greatly supports customer relationship management. In the life cycle, the delivery phase of document management is the initiator of customer dialogs. The dialogs themselves are managed, however, by CRM applications. Collected and collapsed data then flow back to document management.

Partner Relationship Management (PRM): Effective document management greatly supports partner relationship management. In the life cycle, both the creation and delivery phases of document management play a key role. Partner documents enter document management in the create phase; the delivery phase distributes documents to partners. In order to facilitate exchange of existing documents, agreements on formats, indexing, digitalizing, and electronic exchange are extremely important.

Enterprise Search Engine (ESE): Search capabilities are considered very important for easy viewing and retrieval of existing documents. The search engines usually reach documents, but no data that serve applications of support systems. The trend is to open these applications of support systems to enterprise search solutions, such as Google Enterprise.

4.1.3.7 Outsourcing

Enterprises are challenged by improving their services to their customers while reducing capital and operating expenses. The tools that support the life cycle of document management are constantly improving, and new business models are being created. These modes involve an English-speaking, low-cost, highly motivated, and technically competent workforce in practically all geographies. India is the first target for the document creating step.

4.1.4 Critical Success Factors of Document Management

There are many indicators that determine the quality of document management. Not all of them can be evaluated in a conference paper, but the critical success factors are be described as follows:

- Quality checks for each life-cycle phase: Each phase of the life cycle should show proven high-quality solutions regarding procedures, rules, data cleansing, security, and performance. Once the whole process is in operation, periodic audits are recommended.
- Best practice benchmarks: Each enterprise should compare its document management solutions in terms of processes, people, and tools with the industry average and with best practices. The basis of this comparison is the proper use of key performance indicators (KPIs).
- Vulnerability analysis: Each phase of the life cycle is subject to security breaches. The vulnerability may be anywhere and the weakest element in the life cycle is the quality indicator. Vulnerability analysis is recommended to be conducted periodically, and if affordable by an external consulting company.
- Timely adaption to compliance requests: Compliance requests are expected to grow in volume and in severity. It is absolutely necessary to analyze, interpret, and understand these requests as early as possible and make the necessary changes to processes, tools, and to assignments of human resources.
- Fine-tuning of document sharing strategies and rules: Documents represent a considerable value to enterprises. The proper use is helping business processes to increase their effectivity and efficiency. Data, e-mail attachments, records, video clips, and others may be searched, combined, and

analyzed as part of business intelligence ventures. But the eligibility to access, authorization, and authentication of users should be governed by enterprise rules.
- Continuous education on new processes and tools: The business and IT world is in steady change. Continuous education should be planned, budgeted, and executed for new document management processes and tools.
- Concentrate on efficient processing, effective messaging, electronic delivery, and enterprisewide data access: The 4 Es will play a significant role in every phase of the document management life cycle. Each step is expected to be efficient on its own; processes, people, and tools should collaborate by effective information exchange; transition to electronic document presentment and delivery is a must; and finally, the most economical utilization of documents should be guaranteed to each eligible user in the enterprise.

A successful ILM deployment depends on buy-in and brainstorming from three main players. They are (FORR07):

- *Information manager:* The information manager must ensure information is searchable and that only the right users have access. This person also needs to weigh privacy concerns along with still-evolving rules on data retention. Then there is the uncertainty: Will this information be needed if there is a litigation?
- *Storage manager:* If the information manager determines that data might be sensitive with respect to possible litigation or audit, then the storage manager must decide whether it needs to be kept on more expensive tamper-resistant storage media. And if the call comes to retrieve the data, how fast is fast enough?
- *Security manager:* When moving this data to storage, does it need to be encrypted? The security manager also assesses whether storage policies for each piece of information meet appropriate legal and contractual requirements without overburdening the business and employees.

4.1.5 Summary and Trends

The composition of the document will continue to change, but its purpose will remain constant. Technology will continue to make consumers more accessible, promoting continued growth in the number of documents available. But it will be the consumers and providers of documents that will determine their relevance. Enterprises must be aware of the considerable growth rates of the future due to data volume growth in key application systems, such as ERP, due to richer data sets about customers that are maintained by many suppliers, due to the reading of sensors, such as radio frequency identification (RFID), and of course the almost exponential growth of unstructured data, such as emails.

ILM helps match the business value of corporate information with appropriate retention policies and storage systems. Faced with new federal rules for e-discovery, companies can save millions in litigation costs by using ILM. ILM can help in the following areas:

- Keep storage and data management costs in check by limiting companies' long-term storage only to data with lasting business value.
- Prevent costly legal judgments stemming from the inability to produce electronic evidence in a timely manner.
- Keep legal discovery costs down by making it easier to pull relevant data from corporate archives.

Critical implementation steps include:

- Form a steering committee including employees from IT, legal, compliance, finance, document management, and human resources.
- Perform a comprehensive audit of data and business processes to determine what constitutes business records, and assess the business value of different types of records over time.

- Assess document-management and e-mail archiving systems, and evaluate products and options to improve capability.
- Evaluate greater use of document management and collaboration tools to reduce reliance on e-mail and other unstructured communications.
- Develop a data management and retention policy with CXO-level support.
- Train records managers and employees on policy and promote it as critical to the business.

Acronyms

ADF	Automated Data Factory
BPM	Business Process Management
CMS	Content Management System
COSO	Committee of Sponsoring Organizations of the Treadway Commission
CRM	Customer Relationship Management
DIM	Document Image Management
DMS	Document Management System
ECM	Enterprise Content Management
EDM	Electronic Document Management
EMM	E-Mail Management
EOMS	Enterprise Output Management System
ERP	Enterprise Resource Planning
ESE	Enterprise Search Engine
GIF	Graphic Interchange Format
GLBA	Graham Leach Bliley Act
HIPAA	Healthcare Insurance Portability and Accountability Act
ILM	Information Lifecycle Management
JPEG	Joint Photographic Experts Group
KPI	Key Performance Indicator
OCC	Outbound Content Compliance
OCR	Optical Character Recognition
PDF	Portable Document Format
PNG	Portable Network Graphics
PRM	Partner Relationship Management
RDF	Resource Description Framework
RSS	Really Simple Syndication
RIM	Record and Information Management
SOX	Sarbanes-Oxley Act
TIFF	Tagged Image File Format

References

BROD06: Broddy, W. 2006. Designing for optimal communication; *Document*, October.

ESPI03: Espinola, M. 2003. Automate your documents now, *Document*, October.

FORR07: Forrester Research, 2007. Information classification must reach beyond knowledge management, *Research Note*, October 2.

GERS03: Gerschwer, S. 2003. Managing the document life cycle, *Document*, August.

GREW04: McGrew, P. C. 2004. Think your records over, *Document*, December.

MARK06: Markham, R. 2006. Under scrutiny—e-mail and e-records management and archival solution, *Document*, October.

SUFF04: Suffoletto, M. 2004. Solving the compliance puzzle, *Document*, December.

4.2 Information Technology Alignment with Businesses

Kornel Terplan

IT budgets have been cut continuously and significantly over the last 4 years. "Do more with less" is the guideline by many business managers for their IT departments. But really, how important is IT for the business? This section begins with interesting survey results, provided by *CIO Magazine* in the United States. Then, the paper will give concrete, actionable directions for improving effectivity, efficiency, and reputation of IT. In particular, the roles of SOA (Service-Oriented Architecture), SOBA (Service-Oriented Business Application), SODA (Service-Oriented Development Application), Business Impact Management (BIM), Business Service Management (BSM) and the requirements of the Real-Time Enterprise are addressed. When these initiatives are deployed properly, the enterprise will benefit significantly. The final part of the presentation goes into details of focusing on the role of the IT infrastructure for business success by addressing flexibility, collaboration with partners, security, scalability, and compliance with standards.

4.2.1 Does IT Matter?

An agile enterprise reacts quickly and efficiently to changes, such as market opportunities or mergers and acquisitions. The ability to react quickly translates to top-line revenue, and the ability to adjust efficiently equals lower costs. In order to make an enterprise agile, several actions may be initiated, such as: consideration of wireless technology, maintenance of multiple service provider relationships, design of redundancy, and use of virtual networking. Agility is not something that can be forklifted into a company, but using it as a guiding principle can yield measurable benefits.

IT could be the key to agility of an enterprise. IT resources properly planned, deployed, and operated will guarantee the leading edge.

The IT operations department within the IT organization is responsible for ensuring the integrity and quality of service of the production environment, which consists of business applications and dependent infrastructure (for example, servers, networks, desktops, storage, facilities, databases, and systems software). It manages constant change and ensures that mission-critical applications supporting business processes are running and performing. Many business processes are extremely dependent on IT operations processes; however, they are not always considered mission-critical to the business.

4.2.1.1 Survey Results: (from *CIO Magazine*—Alignment Special, Fall 2004)

Finding 1: Most companies have their house in order when it comes to alignment, but there are cracks in the foundation. About 80% of both IT and business executives say that IT is well aligned with the business, that alignment has improved, and that alignment has boosted their bottom line. But there are shortcomings as well. Only 37% of IT executives and 30% business executives say IT is "very effective" at working with the business. Business executives give the IT department lower marks for alignment than they do to most other functions. And CIOs are often more optimistic in their assessments than either business executives or the IT executives below them.

Finding 2: The underlying problem is weak IT leadership and unclear business strategy. When alignment is not working, CIOs often blame lack of strategic clarity, while business executives are most likely to point to weak IT leadership and execution. Who is right? Both sides. It is telling that 26% of lower-level IT executives cite weak IT leadership as a problem, while just 4% of CIOs do. Companies often fail to hit their specific alignment goals, a sign of execution problems. But agreement about goals is also an issue: business executives are less focused on cost savings and productivity and more focused customer retention. The same is true with alignment techniques: beyond good communication, there is no consensus on which techniques are particularly effective.

Finding 3: Without a supportive corporate culture, alignment is more difficult to achieve. Executives at firms where IT is well aligned are significantly more likely to describe their corporate culture as being collaborative, adaptive to changes in the marketplace, and well suited to the company's strategy. Business executives, even more than IT executives, appreciate that alignment is most likely to succeed when alignment is a companywide concern, not a concern just in the IT department. The message to CIOs with alignment problems is clear: Fixing the problem often requires cultural change throughout the company. Focusing on the IT department alone will seldom work.

Finding 4: Alignment processes are not working as well as CIOs think. Nearly all CIOs believe their department works with others. If that were true in practice, more business-side CXOs and vice presidents would agree; however, 43% of director-level business managers say that the IT culture clashes with the business. Nor would the CIOs' employees consistently give far lower marks on creating a collaborative culture, identifying alignment-fostering processes, and other activities that are part of an alignment culture than their bosses do. CIOs need to do a serious reality check.

Finding 5: The CIO's background is less important for alignment than the company's culture. Many business and IT executives assume companies with CIOs from business background are more likely to be aligned. It turns out, however, that there is no correlation at all. What does matter is whether the CIO participates in setting corporate strategy. Interestingly, companies where that occurs are also likely to have a culture that supports alignment. Does the opportunity to work with other executives on strategy give CIOs a chance to influence their company's culture?

Finding 6: Alignment requires effective governance. Whether IT is centralized, decentralized, or follows a shared-services model has little correlation with alignment. But effective governance is critical for alignment—in large part, it appears, because good governance goes hand-in-hand with a collaborative management style. That is why it is odd that IT executives who claim they are "IT monarchists" and have little need to share IT decisions with business say their approach to governance is effective.

4.2.2 Support of Business by IT

In response to the pressures to increase quality of service while reducing costs, many IT organizations are transforming themselves to be more business and customer centric and operate like an independent business—that is, as a service provider to business customers. Although a small percentage of IT organizations pursued this strategy for years, it did not reach mainstream until between 2002 and 2003. At that time, IT organizations realized that to provide the highest service at the lowest costs, they needed to invest systematically in architecture, product/service definition, and repeatable processes. In pursuing this strategy, IT organizations need to reengineer their operational processes, invest in automation to increase agility and service quality, and reduce costs. Although this transformation process often takes two to five years, it yields significant benefits for organizations, regardless of whether their IT operations are in-house, outsourced, or a combination of both.

To better align business and IT planning, many enterprises have established an IT strategy council, whose role it is to ensure that the business strategy is supported by an IT plan that is timely, feasible, and affordable. In support of this effort, IT and business planners are asked to develop a baseline IT plan to be adopted by the IT strategy council. The plan should lay out the business strategies to be supported for the strategic planning period and a comprehensive yet high-level view of the IT support required.

4.2.3 Baseline Information Technology (IT) plan

The part of the baseline plan that covers the coming year can be extracted and updated, as part of the annual planning process, to incorporate the annual business operating plan. Then, throughout the business year, the plan can be updated to reflect ad hoc business projects driven by unplanned, short-term needs. As part of the updating process, the plan is fleshed out in much greater detail and becomes, in fact, the IT annual operating plan. These are the recommended steps for producing the baseline IT plan:

4.2.3.1 What Business Strategies Will Be Supported, and across What Time Period?

Business and IT leaders must determine the organization's business strategy across the strategic planning time period by defining a target vision for the end of the period and the strategies the organization will pursue to achieve it. A key complication for many multidivisional businesses is that many corporate level "strategic plans" for these organizations consist of financial and other quantitative projections. The operational strategies and plans to achieve them are developed at the business unit level and are less well articulated and specific as the plans extend over time.

4.2.3.2 What Elements of the Strategic Plan Will Depend on the IT Organization?

Identify the business functionality and capabilities that will have to be developed, delivered, and supported by IT to fulfill the strategies. Define the implications of developing these for IT infrastructure, applications, data, and (especially if e-business strategies are to be supported) business processes. For many enterprises, a significant operational and organizational transformation might be on the horizon with significant impact on IT planning. For instance, an electric company that goes from selling power in its local area to reselling energy on the international market can expect to make significant investments in IT to support the endeavor.

4.2.3.3 What Is the Scope and Scale of the Investments Required in IT to Support the Plan?

Using the information developed in Step 2, define "big picture" development programs that will be required for infrastructure, as well as applications and data management needed to deliver the required functionality and capabilities.

4.2.3.4 Are These Investment Programs Manageable?

Standard project management work breakdown processes can be used to divide these programs into more manageable project-sized pieces, identifying specific key deliverables and dependencies. This creates several smaller projects that are more visible (as opposed to one or two large ones) and give the enterprise greater flexibility.

4.2.3.5 Is It Feasible to Assign or Acquire the Necessary Resources To Do All the Work in the Required Time Period?

Develop high-level estimates of time and resource requirements for implementing each project and develop a sourcing strategy to collectively support the efforts. One of the purposes of this process is to establish the feasibility of executing the required development workload in the specified time period. To determine this, the enterprise must consider the specific requirement for skills, competencies, and manpower and decide whether resources should, and could, be sourced from inside or outside the organization. If outsourcing or using external services providers is planned, the IS organization must also identify and evaluate the likely suppliers.

4.2.3.6 Is the Strategy Affordable?

Given the development of the resource requirements, develop high-level capital and expense estimates. These provide the necessary feedback to business planners and managers to allow them to address the question, "Is this affordable?" If it is not, an iterative process can be used to either trim functionality or extend the time periods until an affordable spending level is attained.

4.2.4 Real-Time Enterprise (RTE)

Real-time enterprise initiatives remove delays from critical business processes. IT plays a crucial part in this since only electronic systems can work with the necessary speed. A variety of specific technologies

can be used and most of these, such as database management systems, mobile phones, and Internet, have been available for many years. Leaders in the race to real time have energetically exploited these technologies, often at considerable cost and risk.

Many emerging technologies will become essential parts of RTE. The most promising technologies are:

a. Electronic tags based on radio frequency identification (RFID), allowing easier and more precise tracking of objects and packages.
b. Web services, which will ease the linking of applications and enterprises in support of business processes.
c. Enterprise instant messaging (IM), which will add a real-time, customer service option and link people internally and with partners.
d. E-mail response systems (ERS), which will make e-mail a real-time process.
e. Just-in-time e-learning, which will enable enterprises to continuously train their staff in real time.
f. Emerging applications, such as decision engines and business activity monitoring (BAM).

The 10 critical steps toward a real-time enterprise are:

- **Real-time visibility**: It is important to understand what the status of the process is before enterprises can improve it. All critical metrics should be presented in real time.
- **Real-time management**: Embedding key controls, alerts, and metrics into existing legacy, BPM, CRM, and supply chain systems allows business management the chance to see real exceptions and take action.
- **Business process modeling**: With time to produce new products going from years to days, staying with the status you have actually means failure. Simulation tools may help let a company understand a process by cost, time, and other metrics.
- **Process automation**: A number of processes and models can have very robust decision rules that allow enterprises to follow multiple paths and have each of the process models included so that the business areas can be as unique as they need to be in order to satisfy customers.
- **Rapid deployment of solutions**: Companies cannot afford to rip and replace technology. New BPM and enterprise integration architecture (EIA)-based solution templates from various vendors can provide anywhere from 20 to 80% of solutions to problems such as Health Insurance Portability and Accountability Act (HIPAA) and Sarbanes-Oxley compliance.
- **End-to-end integration**: Integration is targeted among computing platforms, applications, people, and processes.
- **Flexible infrastructure**: Real-time enterprises need to react quickly to both business and technology changes.
- **Service-based architecture**: Component- and service-oriented architectures provide maximum agility and flexible deployment options that isolate and enable infrastructure changes.
- **Support for standards**: Electronic commerce demands have sparked more rapid adoption of standards, such as RosettaNet, XML, SOAP, WSDL, UDDI, and UML. They can reduce the cost of maintaining skill sets within an organization.
- **Organizational agility**: Most people resist change just as a matter of fact; new metrics are needed that reward employees according to their project's contribution to real-time agility and long-term return on investment (ROI).

4.2.5 Directions for Service Orientation

Driven by Web services momentum, service-oriented architecture moves from leading-edge software projects to the mainstream. However, most enterprises are confused about its benefits and do not understand its risks.

4.2.5.1 Service-Oriented Architecture (SOA)

Service-oriented architecture (SOA) is a client/server software design approach in which an application consists of *software services* and *software service consumers* (also known as clients or service requesters). SOA differs from the more general client/server model in its definitive emphasis on loose coupling between software components, and in its use of separately standing interfaces. SOA principles are rendered during application design, development, and deployment. These renditions share the essential principles of encapsulation and flexible coupling, but they differ in detail. The fundamental intent of SOA is the nonintrusive reuse of software components (services) in new runtime contexts. The design and development of SOA is performed for the purpose of achieving such an agile runtime environment.

SOA provides a framework for software on one system to securely and reliably request and receive computational resources on other systems. SOA is the latest answer to the perennial problem of reducing the complexity and interdependencies of componentized software systems. SOAs today are built on Web services, delivered primarily as Simple Object Access Protocol/XML interfaces, Web Services Description Language definitions, and on Universal Description, Discovery, and Integration Protocol. Services are building blocks that can be used to construct complex systems across a distributed network. SOA represents a fundamental break with client/server. Instead of a master–slave relationship between servers and clients, systems now can communicate in a distributed peer-to-peer relationship similar to routers in IP networks.

The Service-Oriented Business Application (SOBA) perspective is that applications are delivered as a set of packaged software services, as opposed to monolithic application modules. These applications will have process orientation and flexibility that goes beyond the traditional monolithic application suites. The platform for such applications combines the infrastructure of application server and integration broker with portal, analytics, content management, and collaboration support technology. SOA and integration with established application systems are prerequisites for these products.

The Service-Oriented Development Application (SODA) is the development paradigm for building applications on an SOA. Using services as the primary unit of modularity requires a new approach that involves composing applications from sets of loosely integrated processes. This is the assembly list approach, rather than a code first approach used in traditional development. Code is necessary in a SODA world, but it is hidden behind the service interfaces. This means that consumers do not need to be concerned with the structure or technology of the underlying program logic and platform.

4.2.5.2 Business Service Management

End-to-end IT service management embodies the people, processes, and technologies that an IT service provider must employ to develop, deliver, and support appropriate and competitive IT services at the levels necessary to support business goals. IT operations departments are moving from a component orientation (such as networks, servers, storage, database, and applications) to managing business-oriented, end-to-end IT services. Business units that lack interest in individual IT components are driving this evolution. They want the end-to-end IT services that support their business processes to be available and perform to meet their needs. Yet because IT operations departments have been so segregated in component-oriented organizational structure and metrics, this transition is challenging for most of them.

By taking an end-to-end IT service management approach, IT organizations can achieve many benefits. Aligning IT infrastructure components with business-oriented IT services helps IT operations staff understand the business impact of IT problems. This results in improved availability, reduced downtime, and shorter problem resolution time because IT support focuses on solving the correct, high-priority, business-relevant issues. Because there is less "firefighting" and more repeatable processes, labor costs are reduced. In addition, taking an IT service management approach enables more productive communication between the IT organization that delivers the services and the business users who consume them. This raises the IT organization's credibility with the business by demonstrating its understanding of how IT services support business processes.

Business Impact Management (BIM) reacts at the time of a performance problem or network failure to identify the affected applications, end users, and customers. BIM is reactive by correlating past and present events with business impacts.

Business Service Management (BSM) means linking business and technical information into a logical whole. BSM has given rise to the next generation of management tools. BSM tools are aimed at helping network executives prioritize IT projects and address their fixes based on policies that align IT with business goals, processes, and services. With their management products already collecting volumes of data on network, system, and application health and performance, vendors propose the next step is correlating network health with business performance. BSM is proactive with alerting IT staff ahead of business impacts due to infrastructure shortages.

4.2.6 Role of the IT Infrastructure

The infrastructure of a telecommunications service provider consists basically of three components: applications and services, computers, and networks and network equipment.

4.2.6.1 Applications and Services

The concepts of applications and service are not yet clearly separated from each other. The following differentiations are recommended.

- Business applications
- Services and products
- Business applications of customers

In several cases, services are identical to targeted applications. This is the case, for instance, with e-mail and instant messaging.

In general, service intelligence levels are increasing, with the result that customers do not need to engage in a significant amount of in-house development. Communications-related services can be customized and extended to satisfy user needs.

4.2.6.2 Computers

This group of components contains all of the computing resources associated with service providers' data centers. Application hosts and storage area resources, as well as DMZ firewalls, are typical components. Of course, there are many more processors, but they are traditionally assigned to network equipment. Further differentiation will assist in the process of subdividing infrastructure components into software and hardware categories.

The software infrastructure comprises application- and process-independent software components and basic services. These services are also called *horizontal services*. This software, built on the hardware and network infrastructure, is the basis for applications. Examples are as follows:

- Office components and services
- Location and directory services
- Data management services
- Data interchange services

Operating systems and operating system extensions are seen as part of the hardware infrastructure.

In contrast to the logical tiers of the software infrastructure, the hardware infrastructure is divided into physical tiers: (1) client systems, (2) server systems, and (3) storage systems. In addition, each of these tiers is divided into hardware, operating system, and operating system extensions.

It is getting more and more difficult to draw a line between computers and networks. Due to distributed processing capabilities, grid computing, on-demand computing, adaptive computing, virtualization of computing resources, and utility computing, networks are practically becoming the "computer."

4.2.6.3 Networks and Network Equipment

Networking infrastructures differentiate between transport and access networks. The media are identical in both cases: wire, coax, fiber, and airwaves.

Typical examples of transport network equipment are: switches, bridges, multiplexers, routers, amplifiers, activators, fault diagnostic tools, and element managers.

The access segment consists of two principal groups: provider equipment and customer equipment. The line of demarcation is usually drawn by ownership. Often, Customer Premises Equipment (CPEs) are managed by telecommunications service providers.

On the provider side, the following equipment is common: edge routers traffic shapers, load balancers, gatekeepers, gateways, firewalls, and Web switches.

4.2.6.4 Real-Time Infrastructure (RTI) solutions

Real-time infrastructures can bring substantial IT benefits, but many issues remain. When selecting vendors, the following ten questions may be very helpful.

A real-time infrastructure (RTI) is an IT infrastructure shared across customers, business units, or applications, where business policies and service-level agreements drive its dynamic and automatic optimization. A benefit promised by automating IT systems includes a self-optimizing, self-healing, and self-configuring shared infrastructure.

4.2.6.4.1 What will the RTI implementation cost? How will its return on investment be measured?

The various pricing methodologies that vendors apply to leading management products present a challenge to many operations and procurement departments. An RTI is likely to be even more difficult to calculate. A return on investment analysis of an RTI needs to look at alternatives. Compare the cost of current solutions (such as less-expensive commodity hardware that relies on expensive human labor) to the cost of complex and potentially fragile software that likely will prove to be of higher value over time. An RTI implementation will consist of many parts, including monitoring systems, resource allocation technologies, data repositories, policy engines, and developer tools. Keep in mind what all these elements will cost. Ensure that you also include product, service, operational, and maintenance costs.

4.2.6.4.2 What is the vendor's road map for deliverables?

A vendor may require years to create and evolve a fully implemented RTI strategy. All of the components may not be in place when you evaluate a vendor's products and services. Nevertheless, as early as possible, you should ask the vendor for specific information regarding its intended deliverables and release schedule. You cannot plan an RTI road map for your enterprise or manage the risks involved in this expensive, elaborate undertaking unless you know how the vendor intends to develop and release the technology to market. To ascertain whether the vendor understands what will be necessary to fulfill its vision of the RTI, you should also discuss how much it plans to deliver through organic development and acquisitions.

4.2.6.4.3 How will the vendor's product integrate with RTI initiatives from other suppliers?

Few, if any, enterprises rely on a single vendor for their entire IT infrastructure. For example, many companies combine offerings from IBM and Microsoft. Yet these two vendors tout competing visions and technologies for the RTI. If you use both vendors' products and services, then ensure that their policies and automation actions are able to flow seamlessly across each other's RTI fabrics. The dilemma posed by such integration is reminiscent of the problems that enterprises encountered with management frameworks. To this day, these frameworks fail to interoperate effectively. In addition, the architectural complexity embedded within the RTI furthers "lock in" with the vendors you initially select. Thus, you need to weigh the benefits and drawbacks of having fewer options with regard to third-party management vendors.

4.2.6.4.4 What role will standards play in a vendor's strategy?

No broad-based standards have been established yet for an RTI management schema, although several efforts are underway:

- The Data Center Markup Language (DCML) organization announced that it has formed a vendor group charged with developing a utility computing model. Unfortunately, many enterprises and major RTI vendors remain absent from this group, including HP, IBM, Microsoft, and Sun.
- Other standards bodies, such as the Organization for the Advancement of Structured Information Standards (OASIS) and the Open Grid Services Architecture (OGSA), have efforts under way that touch on RTI technologies and, thus, may overlap with the efforts of the DCML consortium.
- Some vendors are developing their own internal standards. For example, IBM's Common Event Format is meant to improve automation capabilities.

With so many overlapping, and in several cases uncoordinated, efforts occurring throughout various vendors and groups, the likelihood of attaining a common standard for RTI interoperability is remote. In response, vendors will likely seek to protect their proprietary interests first. Ask your vendor what standards it intends to follow; however, don't rely too heavily on standards to ensure integration between vendor RTI offerings.

4.2.6.4.5 How will application vendors support the usage-based pricing model?

One of the tenets behind the RTI is the notion of building virtualized infrastructures that can rapidly reallocate resources to higher-priority business needs. Such efficiency is good news for companies. However, one of the ugly truths within the IT industry is that vendors make a large profit on inefficiency; for example, when an enterprise buys more hardware capacity or bandwidth than is strictly needed, or purchases more end-user licenses than necessary. If the RTI enables enterprises to gain improved control over their IT assets, it will likely have a serious impact on vendors' revenue. Thus, many independent software vendors are reluctant to support the RTI model—at least in the form that it is presented by large hardware vendors such as HP, IBM, and Sun. Some vendors, such as Oracle and SAP, are looking into somewhat competitive RTI offerings in which their own software dominates. Although they may offer competitive RTI-based technology, they are not offering usage-based pricing because of the potential impact on their key revenue streams.

4.2.6.4.6 Is RTI "one size fits all," or will it vary by IT organizational maturity?

After you have deployed an RTI architecture, don't assume that the technology will do most of the hard work for you. Your company's ability to effectively use RTI products and services will depend largely on the process maturity level achieved by your IT organization. You will still have to overcome the typical hurdles centering on process development, interorganizational communication, and technological readiness. Your vendor can help by recommending best practices for implementing and maintaining its technology and services. Such directions should take into account the varying levels of enterprise preparedness. In general, the best way to implement an RTI is over time, in stages. Without strong change management and planning, most RTI implementations will likely fail.

4.2.6.4.7 How will RTI affect IT operations?

Implementing an RTI will not end the IT organization's responsibility for supervising operations. The self-tuning systems promised by an RTI likely will increase the demand for skilled operations personnel who are fluent in the new language of *autonomics*. Rather than managing events, IT staff will focus more on providing services. The increasing emphasis on services will make collaboration across the IT organization more critical. Ask your vendor to recommend best practices in operations and maintenance.

4.2.6.4.8 What role will application development play in the vendor's RTI strategies?

Vendors have different visions regarding the role of application development in RTI:

- Some vendors (for example, Microsoft) believe that RTI begins when applications are being developed. Their plans state that instrumentation that will help control the behavior of environmental resource allocation must be added during program development.
- Other vendors (for example, IBM) suggest that development happens mostly during production operations. Such vendors will likely emphasize just-in-time instrumentation techniques—for example, byte code instrumentation, which is found in many of today's Java 2 Platform, Enterprise Edition–based management products.

Companies will need to adopt both approaches. You should look for vendors that can satisfy a spectrum of needs. Expect to make trade-offs between development complexity and the management of information granularity in production.

4.2.6.4.9 How will the roles of testing departments in IT organizations change after RTI is adopted?

Testing for the RTI environment will become more complex because the IT system is now expected to be self-configuring and optimizing. Expect to assemble a more sophisticated quality assurance team than you've relied on in the past. The team will have to understand how the system reacts and reprioritizes infrastructure capacity due to demand fluctuations. By contrast, single-application stress and load testing is likely to become less important. Your new team will test policies rather than applications and observe whether the system behaves in a compliant manner. The team also will need to be familiar with the preplanned orchestration of many interrelated infrastructure components. Your vendor should offer you guidance on best practices in testing.

4.2.6.4.10 How will enterprise management technology evolve to support the needs of the RTI?

Configuration data, understanding relationships, and dependencies between and among components on different levels of the stack are critical to the success of RTI. To realize the promise of the RTI, systems management science will have to evolve beyond currently available management capabilities. Problems that vendors must address include the following:

- Root-cause analysis technology is inadequate and must become smarter, faster, and more adaptable.
- Hard-coded, rule-based systems suffer limitations (such as with maintenance) in complex, heterogeneous scenarios.
- No unified management schema is available to enable an end-to-end evaluation of the managed object infrastructure.
- Real-time modeling and simulation capabilities are nonexistent. Capacity planning tools are available, but most are designed to work offline rather than in real time.

Because potential limitations in areas such as bandwidth will affect decision latency, some corrective actions may have to be preauthorized. RTI will likely have to become more of a federated system of automation "cells." Expect to make decisions regarding whether control will be local or centralized. Ask RTI vendors how they will overcome these technical challenges.

4.2.7 Future Trends

IT management must strive to seamlessly integrate IT planning into the business planning process by providing practical answers to key business planning questions.

Achieving end-to-end service management benefits requires significant IT process reengineering, including changing workplace cultures and behavior, which is often a difficult task. The principal challenge is moving toward a customer-centric service orientation, where standard products are delivered

to the business, with agreed-on SLAs. Not only is every function of IT affected, but also the relationship between IT and the business, where expectations are clearly set for IT services delivered.

One of the most important critical success factors is ensuring that IT metrics are optimized toward service management (and not suboptimized toward a particular functional group). For example, measuring all functional (as well as business) groups on end-to-end availability will drive behavior toward optimization of the end-to-end service, rather than for the benefit of one group (but to the detriment of others). A culture of teamwork is needed. Those with a "blame" culture will fail. As a result, IT operations departments are redesigning their IT management processes at a record rate. Additional investments need to be made in people (for example, process managers), training, and tools.

"Digital organizations" seem to get the most out of IT. The following are seven primary attributes of successful digital organizations:

- Migrate analog business processes to digital.
- Foster open information access.
- Establish a distributed decision-making process.
- Link incentives to performance.
- Maintain focus and communicate corporate goals.
- Hire the best people.
- Invest in human capital.

Acronyms

BAM	Business Activity Monitoring
BIM	Business Impact Management
BPM	Business Process Management
BSM	Business Service Management
CIO	Chief Information Officer
CPE	Customer Premises Equipment
CRM	Customer Relationship Management
DCML	Data Center Markup Language
EIA	Enterprise Integration Architecture
ERS	E-Mail Response System
IM	Instant Messaging
OASIS	Organization for the Advancement of Structured Information Standards
OGSA	Open Grid Services Architecture
RFID	Radio Frequency Identification
RTE	Real-Time Enterprise
RTI	Real-Time Infrastructure
SLA	Service-Level Agreement
SOA	Service-Oriented Architecture
SOAP	Simple Object Access Protocol
SOBA	Service-Oriented Business Application
SODA	Service-Oriented Development Application
UDDI	Universal Description Discovery and Integration
UML	Universal Modeling Language
WSDL	Web Services Description Language
XML	Extended Markup Language

4.3 Business Intelligence and Analytics

Patricia Morreale and Deepak Pareek

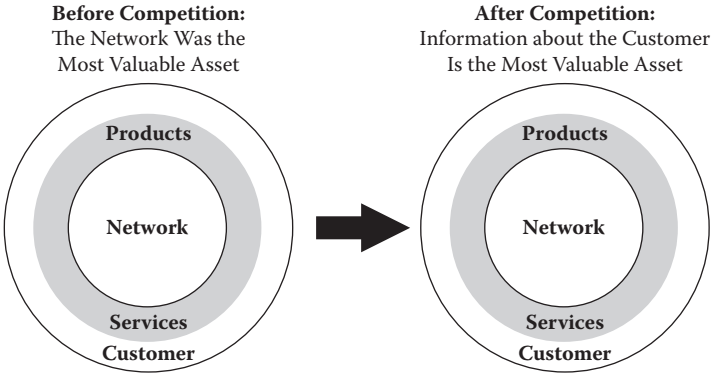
The telecommunications industry expanded substantially in the past decade. Technology advancement along with the liberalization of once closed markets and privatization of government-held monopolies changed the nature of the industry in the 1990s and continues to shake up the industry every now and then. In early 2000, the industry scaled new highs with respect to market capitalization. Both business and technology disruptions have introduced significant expansion and innovation (see Figure 4.3.1).

Since then, however, the telecommunications industry has been overwhelmed by a series of events that have led to a depressed industry segment. Over the past few years, it has suffered from severe debt, overcapacity, customer churn, and service commoditization. The earnings declines and flight of capital have driven the industry to reevaluate and reassess fundamental business practices and devise survival strategies that will lead it back to competitiveness and consistent profitability.

The same forces that fed the development of new services and the entrance of new players also saw margins grow slimmer for most services as well as significant customer churn as competitors offered alternative choices. The expansion of the infrastructure has not been absorbed yet with an equivalent rise in demand and profitability. On the other hand, investments incurred for expansion and for acquisition of 3G licenses are becoming a backbreaking exercise for the once solvent and profitable communication service operators. Trillions of dollars in market capitalization have evaporated in the past couple of years and hundreds of thousands of jobs have disappeared from the sector. The telecom industry, which is undergoing a period of difficult change, is under a great deal of competitive and market pressure. The major challenges facing the industry are:

- Increased customer dissatisfaction with existing telecom services
- Market uncertainty and excessive debt
- Bandwidth commoditization
- Limited market capital
- Large, expensive, and inflexible IT infrastructures

In today's extremely challenging business environment, many telecommunications operators and carriers are measuring their success by the size and growth of their profit margins. As a result, carriers are under intense pressure to reduce or eliminate the major threats to these slim margins, including revenue leakages and frauds, inaccurate or missed intercarrier billing, churn, inefficient network usage, and least-cost routing plans. These competitive and market pressures are also making the telecom industry reassess its business model and redefine the path that will return it to competitiveness and profitability.



FIGu RE 4.3.1 The changing landscape of telecom.

4.3.1 The 4 Cs of the Telecommunications Industry

There are four key challenges faced by the telecommunications industry today, described as the 4Cs. As with financial services, deregulation is driving many of these currents as a metadynamic that influences all of these trends:

- Consolidation
- Competition
- Commoditization
- Customer service

Since deregulation, the ability to compete in a much wider array of markets has opened established telecommunications companies, and prospects of new entrants gaining market share have also become a reality. This has resulted in existing telecommunications companies buying the capabilities to enter new markets and new players gaining substantial market shares. This has resulted in tremendous merger and acquisition activity along with increased competition. As a result of these trends and regulatory pressure, there is a significant emphasis on customer service that did not exist before.

In competitive telecommunications environments, customers choose their service providers. Today, this is a reality for all sizes of telecommunications companies as well as all types of telecommunications companies, whether long distance, Internet Service Provider (ISP), wireless, or local plain old telephone service (POTS).

Under competitive conditions, the customer becomes the central focus of the carrier's activities. Customer requirements not only determine service offerings, but also shape the network and affect the organizational structure of the carrier's focus on particular types of customers. With the customer at the center of the telecom enterprise strategy, the key to survival for these providers and operators today is to focus on the very basics of business such as retaining and servicing existing customers as well as reducing the cost of operations. Nearly all telecommunications companies today are responding to a mission-critical need to compete more effectively as a result of:

- Rapidly changing, increasingly competitive, and global markets
- Increasingly volatile consumer and market behavior
- Rapidly shortening product life cycles

To do so, it is necessary to analyze accurate and timely information about operations, customers, and products using familiar business terms, in order to gain analytical insight into business problems and opportunities.

The business landscape of the telecommunications industry is quickly evolving. The previous model, shaped by a handful of competitors in each country, is being replaced by a model shaped by hundreds of competitors vying for a global presence. To survive in this environment, telecommunications companies can continue marketing their products to the masses. This market-share strategy has been very popular in the telecommunications industry in the past. To compete, companies are driven to increase advertising and marketing costs aggressively while discounting their products. Unfortunately, this strategy has driven customer loyalty to an all-time low. For example, it is not unusual for a consumer to switch service providers twice in the period of a single year.

This may be why companies can report a 40-percent disconnecting rate over the period of a year and still show an increase in market share. Clearly, this model for doing business presents significant challenges and over time threatens to drive profit margins unacceptably low. The ultimate evolution may be similar to what has been seen in the retail industry, where a good year produces profit margins in the two- to three-percent range.

Companies can focus on tailoring products to the individual customer. In this "share of customer" environment, customers are differentiated in addition to products. Corporate resources are efficiently

allocated to customer care in relation to the customer's lifetime value. Those customers whose loyalty can be earned and whose lifetime value to the company is high will receive a majority of the attention. In contrast, customers who are not loyal or whose lifetime value is low will receive a lower degree of attention. The result will be an environment that optimizes profits by nurturing valued customer relationships. Essential to this strategy will be the ability to leverage evolving technologies to accomplish the following:

- Understand the customers' needs and behaviors.
- Leverage this understanding to identify, develop, and deliver relevant products and services.

Choosing or integrating these strategies and migrating to this new environment is one of the most profound decisions facing telecommunications companies across the globe.

4.3.2 Customer Is King

As telecommunications markets become increasingly competitive, the ability to react quickly and decisively to market trends and to tailor products and services to individual customers is more critical than ever. Although data volumes continue to increase at an astounding rate, the problem is no longer simply one of quantity.

At the heart of the issue is how companies are using their information. Increasingly, particularly in the telecommunications industry, it is important to understand customer preferences and behaviors. It is imperative to understand all the parameters of a customer, whether individual or otherwise.

Although this sounds simple enough, a telecommunications company faces several hurdles in achieving this objective and in targeting its product lines to current or prospective customers. Ironically, the telecommunications industry is in a unique position to understand the customer because it can direct its energies in various channels to obtain customer information.

4.3.2.1 Strategic Shift—From Product to Customer

Many companies are aggressively moving (or have already moved) from a business model based on a product strategy to a business model based on a customer strategy. This environment is characterized by customer relationships, product customization, and profitability, and is in response to pressures transforming the business landscape throughout the telecommunications industry.

4.3.2.1.1 Growing Consumer Demand

Customers (or consumers) are expecting companies to understand and respect their needs and desires. In this world, the customer drives the relationship. It is the role of the business to hear what the customer has to say and respond by delivering relevant products and services (what they want) on their terms (how they want it). Companies can no longer expect to sell several products and services to the masses (mass marketing), but must tailor many products and services (i.e., mass customization) to the individual. This is generally referred to as *mass customization*.

4.3.2.1.2 Growing Competition

The ability to refocus a product mix in response to evolving competition is a critical success factor for any business. The key is to be able to anticipate the needs of the marketplace before one's competitors. It is this ability to outpace competitors that most companies find difficult or impossible to do, given today's amalgamation of technologies and architectures. Why is this important? Corporations today are facing more and more deregulation; mergers and acquisitions are blurring the relationships to customers; and globalization of the marketplace and consumer is opening up businesses to new avenues for expansion and, as a result, new competitors. Therefore, it is mandatory for a corporation to restructure itself quickly without losing the ability to compete.

4.3.2.1.3 Optimization

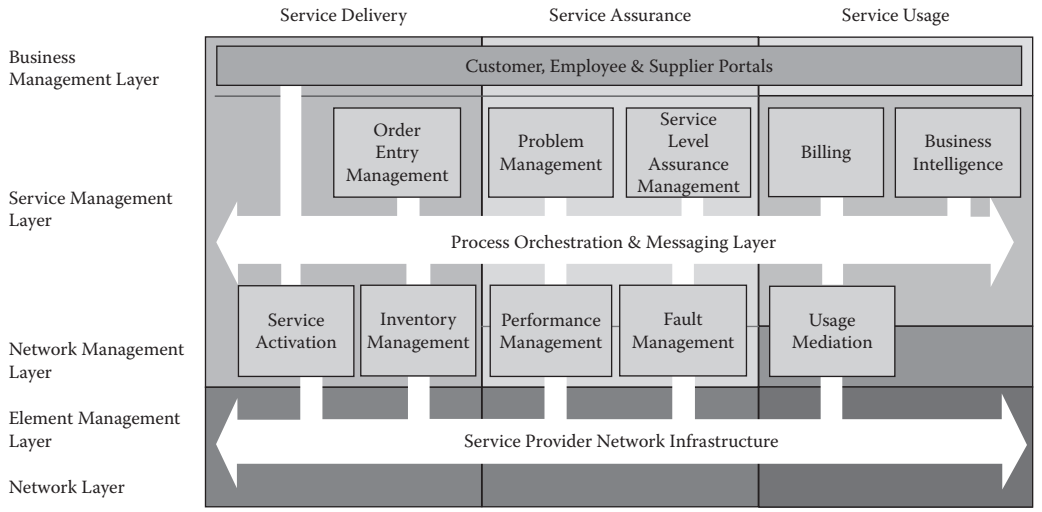
The ability to measure and predict return on investment (ROI) is something that corporations find difficult to perform rapidly. These measurements indicate the health of the corporation, and the ability to determine them rapidly allows a corporation to change its direction with minimal loss. Other examples of the need for optimization include the ability to determine the most efficient channels for contacting customers, target the appropriate customers for a corporation's product mix, and identify new product opportunities before the competition.

4.3.2.1.4 Technology Constraints

Unfortunately, most information systems at telecommunications companies are built around the product strategy business model. This has resulted in a variety of product-oriented systems that effectively run day-to-day operations. Carriers struggle with their existing patchworks of general-purpose data warehouse solutions to store and analyze the mountains of data they create every day. Large networks and their associated switches, billing systems, and service departments can generate hundreds of millions of terabytes daily. These terabytes of dynamic customer data will continue to grow exponentially as carriers add new services and as IP-based traffic increases. This ever-expanding volume of data puts a strain on the performance capabilities of today's traditional relational databases, servers, and storage systems, which provide the foundation for Business Intelligence (BI; see Figure 4.3.2).

Customer demand for new services, such as third-generation wireless networks, consolidated billing, and consistent and reliable service, has been affected by technology limitations. The difficulties created by legacy back-office systems, known as Operational Support Systems/Business Support Systems (OSS/BSS), are primarily rooted in their complexity, scale, rigid operational requirements, lack of interoperability, and lack of service focus. This has led to enormous challenges when attempting to deploy new services and adapt to rapidly changing customer needs.

In the past decade, the service providers have spent significant resources and energy installing systems for operations management and business process automation both for business and operational support. However, the complex questions that need to be answered go beyond any one operational system. Today telecommunications companies might know who their customers are, and are marginally effective in marketing new services to them. However, few of them are equipped to know who their profitable customers are, which services these top customers use that make them profitable, and which marketing campaigns can be targeted to this segment.



FIGu RE 4.3.2 Telecom enterprise architecture.

Most of the rudimentary systems available or used in the past by these telecommunications companies are those where data must be extracted from multiple systems and complex spreadsheets must be used to manipulate data. This exercise is time consuming and fraught with delays. If any parameter changes in the equation, the corporation finds itself at square one, without the ability to construct what-if scenarios and respond effectively to the changing market conditions.

Carriers rely on analyses of their terabytes of customer, product, and traffic data to help them make business-critical decisions that will positively affect their bottom line. High-end data warehouses and powerful business intelligence solutions are essential tools to help carriers meet profit goals. Analyzing and integrating in-depth data from multiple departments enable carriers to reduce revenue leakage and churn, mitigate fraud, optimize network usage, and increase profits.

Competition makes things more difficult for operators, and the analysis they need for operations, to stay ahead, needs to be much more granular to allow for true assessment of the profit contribution of the customer and the services, especially when most of the business questions they want an answer for go across multiple operational systems as business processes flow across many departments.

Telecom companies need to deploy new and more focused business services and create stronger customer relationships. This implies a high degree of responsiveness to customer needs and concerns. Telecom companies are, however, strapped with large legacy IT infrastructures that are extremely complex to manage. These infrastructures have rigid operational requirements and resist incremental change, thereby affecting the agility and responsiveness of the telecom companies to customer demands and the deployment of new and improved services.

Telecom companies are looking for ways to build IT flexibility and agility in their businesses, in an effort to lower the costs and risks involved in upgrading and evolving these OSS/BSS. The need to be flexible and nimble requires telecom companies to migrate to more open technology environments and to adopt an architectural approach to integration challenges.

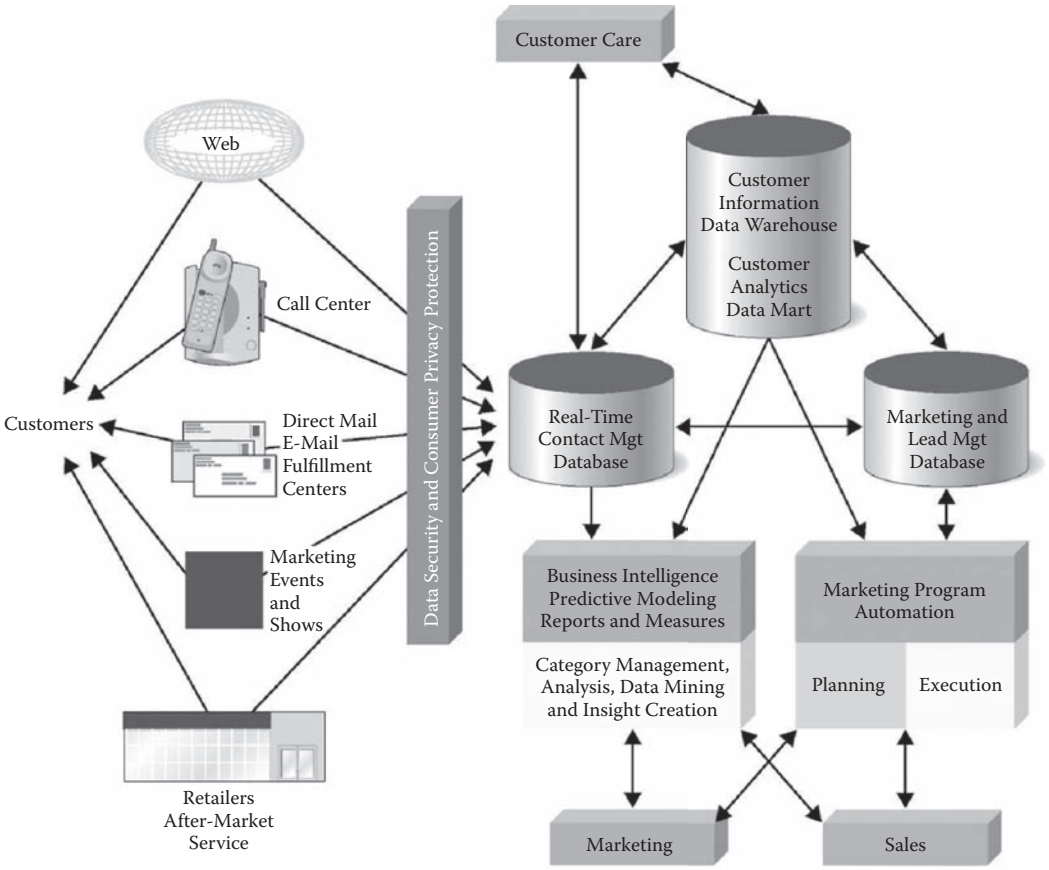
The need to deploy new services requires these companies to adopt a new generation of standards-based middleware that enables rapid and cost-effective service-oriented integration of OSS/BSS infrastructures. The new technology needs to adopt an incremental approach to the creation of service-oriented OSS/BSS architectures, thereby migrating to the IT base of telecom companies through phases of renovation and service consolidation while driving customer-focused service innovation.

Many telecommunications companies are organized first by customer segment—be it consumer or business—then by product or service, such as consumer long distance, business, or local. Managing data within these silos and across these silos can be a challenge, and data warehousing is an essential lynchpin in maintaining a profitable business model.

A robust BI solution can help telecommunications companies manage the complexities involved in this calculation. Telecommunications companies worldwide are exploring business intelligence solutions to achieve competitive advantage. The key solutions for which telecommunications companies are looking involve marketing, such as customer retention, target marketing, and campaign management, customer-relationship management, and network business intelligence, to streamline network assets. Moving forward, additional systems are needed with capabilities to deliver best-of-breed business intelligence and business management. These capabilities enable companies to accomplish the following:

- Understand the needs of their business (business intelligence)
- Manage actions based on those needs (business management)
- Effectively run day-to-day operations (business operations)

These capabilities will enable companies to realize the opportunity of a business landscape characterized by customer relationships, customized product delivery, and opportunity-driven profit. One of the key enabling technologies to this evolution is the data warehouse.



FIGu RE 4.3.3 Strategic BI architecture.

4.3.3 Business Intelligence in the Telecommunications Industry

Telecom was among the first industry verticals to experience the benefits that BI brings to the corporate table (see Figure 4.3.3). Telecom was also the first to experiment with how BI, or rather analytical capabilities in conjunction with Customer Relationship Management (CRM) solutions, can improve customer experience and thereby the business. Among the first applications in this area were the telecom industry's BI initiatives to reduce customer churn. The use of BI connected to operational CRM systems helped identify customers who were most likely to shift to another service provider, by analyzing the number and nature of grievances registered by users.

Although the success of these initiatives has resulted in CRM products with analytical capabilities, the case is still strong for a dedicated BI system connected to an operational CRM system, provided the linking is done optimally. This is because the new CRM products still do not match up to a full-fledged BI system's analytical capabilities, not so far at least. Today telecommunications companies can provide basic answers. However, the complex questions they need to answer today go beyond any one operational system.

For example, today they know who their customers are and are marginally effective in marketing new services to them. However, if they want to know who their profitable customers are, which services the customers use that make them profitable, and which marketing campaigns should be targeted to this segment, they find themselves having to extract data from multiple systems and manipulate complex spreadsheets.

This exercise is time consuming and fraught with delays. If any parameter changes in the equation, they find themselves at square one, without the ability to construct what-if scenarios and respond effectively to the changing market conditions. The analysis needs to be much more granular to allow for true assessment of the profit contribution of the customer and the services. Most of the business questions they want to answer go across multiple operational systems as business processes flow across many departments.

The same holds true for many of the operational processes. For example, one area that has severe implications for customer satisfaction is the time it takes to activate a new customer. The activation of a new mobile customer can take several hours and for a new broadband customer it can take several days.

Activating a new customer touches many departments from customer and service support to accounting, credit approval, billing, network planning, network support, inventory management, scheduling a service person to configure and install the customer premises equipment, and turning the service on. Most operators have disparate systems managing the provisioning and customer activation process, and there is very little visibility of the bottlenecks and disconnects. The activation of new customers has revenue implications, and therefore affects the share price and market evaluation of the carrier. This is a key business performance parameter with visibility in the office of the CEO and to the board.

The same disconnects happen in many other areas as well: interconnect billing, network operations and management, service creation, and introduction of new services, determining the profitability of various products, services, and rate plans. The systems at most service operators have been built for accounting purposes and do not provide adequate information across processes and functional areas to support business users in making business decisions based on the power of information.

There are software solutions that are helping service providers to bridge this gap. The business intelligence applications extract and connect disjointed systems and data from disparate sources and enable the business user and decision maker to make an informed decision.

Generally, the IT organization at a service provider will consolidate data from multiple sources into large data warehouses or smaller data marts. The analytical applications and reporting tools are utilized to query the database and provide regularly scheduled production reports to the business users. Solutions have very easy-to-use user interfaces so the users can also create their own reports and make ad hoc queries as required by business conditions. The analytical applications and digital dashboards go a step further and let users keep their fingers on the pulse with prestructured guided analysis and sophisticated event-management engines. The user can set alerts based on predetermined business rules and thresholds.

It is important for service operators to be able to segment their customers more precisely than they have in the past. Not all services are attractive to all customers, and it is very costly to run outbound marketing campaigns that result in low response rates. Not only do they tax the resources of already overburdened call centers, but they also result in customer irritation or oftentimes acquisition of customers who do not add any value to the bottom line. Service operators are using business intelligence tools to obtain a better understanding of their customers.

Who are their high-value customers? How can they treat them differently?

This is the same strategy that some other industries, such as airlines and credit cards, have been using for years. Understand the impact of combining one service with another. Understand the basket mix of services being purchased by premier versus low-usage customers and the revenue impact of each service. Understand the impact of marketing campaigns sooner and be able to take corrective action in midstream. Without business intelligence applications, they may not know the results of a marketing campaign for three to four months. A product promotion that is run in July will have customer usage in August show up in the September billing cycle. A 30- to 60-day delay removes the ability of the product manager to take action during the campaign. With business intelligence tools, the product manager can see the customer activation and usage on a daily basis without having to wait three months for the billing records. This enables her to adjust the message and the product if the campaign is not meeting its targets.

There is no quick fix out of this market. It is imperative that the service providers dive into their plentiful data and really analyze the operational and financial parameters. To obtain the margin improvements they are seeking, the carriers will have to perform detailed analyses of cost of service, customer profitability, and product margins in order to survive and thrive.

4.3.3.1 BI Requirements of the Telecommunications Industry

Telecommunications service providers should pursue an integrated information strategy from a single platform without compromising security, user requirements, or future flexibility. There are several strategic benefits to this type of "define once, deploy everywhere" solution. This type of solution enables detailed analyses within business units while allowing cross-divisional analyses by enterprise users. Once the enterprise data warehouse is built, organizations can then define data marts that meet multiple business unit needs.

Further solutions should include a multilayer security model that authenticates and authorizes access to data, protects data transmission, and controls application functionality by associated privileges. This ensures that users see only the data intended for them as access is defined for each user against project, row, and object.

Finally, the solution's integrated and flexible platform allows enterprise analysts to survey the entire enterprise to measure ROI, cost of capital, internal rate of return (IRR), and additional financial analyses across units, campaigns, and customer segments (see Table 4.3.1).

4.3.3.2 BI Application Areas

A telecommunications company can use various BI tools for strategic as well as operational decision making. Furthermore, it can carry out various analyses to suit its unique requirements and position within the industry. Among the applications that play important roles in telecommunications companies' success are strategic decision support, scoring and segmentation, campaign assignment and management, traffic analysis, customer relationship analysis, corporate performance monitoring, and last but not least, financial analysis. Other than these central application areas, other areas key to telecommunications companies' strategy are risk analysis, fraud detection (or revenue assurance), and platform convergence.

4.3.3.2.1 Integrated Customer View

This is the first hurdle that any telecommunications company meets. It is important to have all customer/account/transaction data in one place and to be able to correlate customers across their product holdings. In all probability, every product line within a telecommunications company sits on a separate system. The Customer Information System (CIS) should be able to correlate all customers from various product systems and define the various relationships.

If the telecommunications company does not have a CIS system, then individual customers and accounts need to be gathered from various systems and a customer integration process begun. This process can be carried out using industry-standard tools. Individual customers can be parsed using name, date of birth, address, gender, social security number, and a distinct customer ID can be assigned to a customer across all his or her product holdings.

Householding too can be applied based on specific business requirements. The telecommunications company can identify that two customers listed in both the savings and mortgage systems are actually one customer. This provides the most crucial information for a telecommunications company upon which accurate and effective analyses can be run. The data warehouse will then provide a 360-degree view of a customer. Although this process is extremely cumbersome, it is critical for effective analysis and should be treated as the starting point.

4.3.3.2.2 Customer Life Cycle

Every telecommunications company develops a customer life-cycle model and maps its products to this model. This methodology reflects the thinking that customers require different products and services

TABl E 4.3.1 Telecom BI Applications

Fraud Management	Financial Analysis	Marketing Analysis
Fraud detection tool that helps management stop crime and operate efficiently. Drilling down into customer and employee contact records, it delivers insight that can reveal possible fraudulent activity, as well as identifying operational problems that can be fixed. Covers areas including: • Fraud analysis • Corrective action and notification • Product affinity/bundling • Pricing models • Discounting • Call volumes • Call times • Response times • Complaint logs • Employee productivity • Capacity forecasting • Contracts reporting	This vital BI tool enables telecommunications carriers to take the financial pulse of their business whenever needed. Examination of financial performance metrics from across the enterprise arms financial managers with intelligence to make the most profitable business decisions possible. Financial insight ultimately improves gross margins and bottom line performance. Covers areas including: • Revenue reporting • P&L reporting • Cost analysis • Margin analysis • Tariffs • Taxes • Budget variance analysis • Access and line charges • AR/AP reporting • Collections reporting	This analytical tool makes effective category management of telecommunications services possible by providing analytics across a wide range of marketing, planning, pricing, operations, and network variables, helping management determine what promotions and service plans are most effective for specific customer profiles. Covers areas including: • Up-sell analysis • Loyalty programs • Customer segmentation • Demographic analysis • Cross-sell analysis • Service history • Channel efficiency • Next to buy • Promo lift • Price points • Market share
Network Optimization	**Sales Analysis and Billing**	**Customer Care and Analytical CRM**
Growing and maintaining profit margins requires optimum network efficiency. Powerful analytics tool that allows carriers to compare a wide range of metrics across network operations, and create real-time reports that identify problems for immediate attention. Alerts can also be created for instant notification of emergency situations requiring rapid response. Covers areas including: • Traffic analysis • Network planning • Quality of service • Network utilization • Switch operations • Call routing • Capacity • Switch utilization • Volume management • Failure notification • Capacity analysis	A vital tool to gain effective insight from the terabytes of data associated with selling and billing for residential, business, bundled, and unbundled services. Leverages data analysis into competitive advantage by revealing more profitable sales opportunities and the path to more efficient back-office operations. Covers areas including: • Product sales and trends • Customer trends • Sales force performance • Commission reporting • Product affinity • Account balances • Utilization • Fraud • Telemarketing • EBPP/intelligent billing • Quota attainment	Fierce competition for customers across the telecommunications landscape demands advanced customer care efforts. This BI application enables telcos to segment customers by demographic, service plans, billing, and other criteria, delivering insight where it is needed, enabling managers to develop effective strategies that win and retain profitable customers while weeding out unprofitable ones. Covers areas including: • Customer scorecards • Churn analysis • Customer profitability • Customer plan migration • Service level agreement • Trouble ticket • Service complaints • Customer inquiry • Dispatch request • Service call monitoring • Preferences and permissions

at various points in their lives. For example, a customer using basic services may opt for mobile services and then DSL services in the future. Using this hypothesis, it is possible to segment customers in the data warehouse based on their demographics: age, location, annual income, occupation, usage, and the like. Such analysis is primarily based on a customer's position within the life cycle and the products that are applicable to that stage. This model is one of the critical factors that determine the telecommunications company's marketing campaign efforts. Every telecommunications company must have such a model and efforts should be made to integrate this model within the data warehouse analytical efforts.

4.3.3.2.3 Customer Financial Portfolio

Customer data can be enhanced by purchasing data from credit bureaus. Although only credit-related products are available from credit bureaus, it does provide the telecommunications company with crucial details on a customer's credit worthiness.

4.3.3.2.4 Customer Profitability

This is a standard analysis that any telecommunications company can carry out using a data warehouse. Analysis of a customer's profitability to the telecommunications company is crucial for campaign effectiveness. Customer profitability can be calculated in different ways. Whichever method is chosen, it should incorporate the telecommunications company's transfer-pricing mechanism at the account level. The accuracy of the profitability numbers is dependent on one factor: where will these numbers be used? If it is only for determining the type of products to be marketed to a customer, then the profitability numbers can incorporate some level of acceptable deviations.

Profitability should ideally be at the contribution level, as this number defines the net revenues that are a direct reflection of the customer's transaction behavior. Fixed costs, although allocable, are dependent on the allocation methods and can impinge on a customer's profitability.

4.3.3.2.5 Diversification Indicator

This strategy should accompany the customer profitability analysis. Customer profitability by itself does not signify much. First, the profitability numbers need to be compared with those of other customers. Assuming a customer's profitability is high, the telecommunications company must make the decision to either up-sell or cross-sell its products. The diversification indictor specifies the diversity of a customer's product portfolio.

The basic premise is that a customer with stable but average profitability is preferable in the long run to a customer who has high profitability but carries a less stable profitability. This indicator will have to be constructed using the product lines under which a customer holds specific products. A string field can be used to depict such holdings and a model needs to be created that assigns a specific ranking based on product-holding combinations. The ranking is based on the profitability that each distinct combination will accrue to the telecommunications company.

4.3.3.2.6 Product Profitability

This can be calculated as an extension to the customer profitability exercise. The idea is to determine the profitability of various products offered by the telecommunications company and to make product decisions based on such profitability. The process followed is very similar to that used for customer profitability. Again, specific focus should be placed on direct revenues and expenses. Fixed costs are important here (unlike with customer profitability) as a telecommunications company's ability to market its products is based significantly on its infrastructure, which is not a direct product expense.

4.3.3.2.7 Channel Profitability

This is another important aspect of the profitability exercise. A telecommunications company needs to determine which of its delivery channels are more profitable or cost-effective and should try to move its customers to the more profitable channels.

Channel profitability is a difficult exercise as most fixed expenses are usually not directly allocable to a channel. Another problem is the cost of campaigns; most campaigns go out through direct mail, branches, and the call centers. Such channels need to be allocated a higher proportion of the fixed expenses. All these profitability measures need relevant accounts and customers to be available within the data warehouse. Moreover, profitability needs to be analyzed over time and sufficient historic data will be required.

4.3.3.2.8 Event Triggers

Another important tool that can be used for analytics is event-triggered campaigns. Campaigns can be based on specific customer actions (nonavailability of credit limit, significant high-cost services usage, etc.).

4.3.3.2.9 Other Analyses

There are several other analyses that can be carried out using BI tools, as follows.

4.3.3.2.9.1 Fraud Analysis Transactions can be used for fraud detection. Such analysis uses historic data about customer, usage, payment record, and so on, which can be validated within the customer life-cycle framework, and fraud detection triggers can be constructed.

4.3.3.2.9.2 CRM Components A lot of the historic data within the data warehouse can be used to support the telecommunications company's CRM initiatives. A data warehouse provides a 360-degree view of a customer and enables a telecommunications company to study and reasonably predict customer behavior. A data warehouse integrates well with all campaign channels and provides a framework to generate integrated campaigns.

4.3.3.2.9.3 Predictive Models A very important aspect of a data warehouse is its ability to provide integrated and historic information on customers, accounts, transactions, delivery channels, and the underlying data. All these can be used by a data-mining group to understand customer behavior over time, carry out trend analysis, and construct statistical predictive models. Models can range from attrition models to other predictive models that determine the probability of a customer's behavior.

4.3.4 Strategy at Work

A BI solution can jump-start the required impetus to a telecommunications company's strategic efforts. It is important for a telecommunications company to sequence these monitoring, organizing, and analysis efforts, in order to effectively utilize various resources. A great deal of time needs to be devoted to planning the various initiatives to maximize benefits and major coordination between various departments that are involved in these initiatives is required.

There are several critical areas and tools that any telecommunications company should undertake, although each telecommunications company might differ in the method of analysis and implementation.

4.3.4.1 Strategic Decision Support

This is the cornerstone of business intelligence. In this model, end users are provided with intuitive tools to distill information about corporate assets and their performance. Corporate assets include customers, products and services, network infrastructure, and employees. Typical performance measurements include profitability, availability, usage, sales, and lifetime value. Companies can now track key performance measurements, refine customer segments and scores, and optimize campaign strategies.

Some of the typical strategic decision support capabilities in the telecommunications industry include the following:

- Develop simple reporting capabilities that allow one to measure and trend key performance metrics; these metrics include the following:
  - Install and disconnect rates

  - Call center average sales per hour
  - Call center average talk time
  - Campaign performance
  - Customer segment lifetime value
  - Peak network volumes
  - Uncollected receivables
  - Customer satisfaction
- Develop complex reporting capabilities that allow one to uncover problems and discover new opportunities; typical areas for analysis include the following:
  - Market assessment
  - Channel planning
  - Competition assessment
  - Strategy and pricing
  - Customer penetration and profitability
  - Customer segmentation
  - Program definition
  - Recognition of patterns relative to customer behavior and needs
- Develop statistical models that predict customer needs and behaviors; for example, one can build models that predict a customer's likelihood to do the following:
  - Buy a new product
  - Generate high profitability
  - Respond to contacts through specific channels (e.g., direct mail, telemarketing, e-mail, etc.)
  - Not pay their bill

In addition, models can be built that predict network growth and fraud based on traffic patterns in the network.

4.3.4.2 Scoring and Segmentation

These provide the mechanisms for deploying score and segmentation rules developed through strategic decision support. Scoring provides processes that apply statistical models to each customer (or prospect). A score from 1 to 100 is then assigned to indicate how well the customer fits the model. For example, suppose that a model predicted who was likely to be a high-usage customer. This model would be applied to each customer and a resulting score would be assigned.

A score of 100 would indicate a near-perfect match to the model, as opposed to a score of 1, which would indicate that the customer did not fit the model at all. Segmentation provides a means for grouping similar customers. For example, one may segment the customer base between residential and commercial markets. In addition, one may decide to provide further granularity by defining segmentation within these subsegments. Defining customer segments is the first key step toward defining a customer management strategy.

4.3.4.3 Campaign Assignment and Management

These start where strategic decision support and scoring and segmentation leave off. Now that we understand what products to deliver, to whom, and how, it is time to set up a campaign to orchestrate the contact activity. Generally, campaigns contain six key elements:

1. The list of customers to be contacted as part of the campaign
2. The channel to be used in reaching the customer
3. The product, program, and service to be offered
4. The incentive to be used in selling
5. The relationship relative to other campaigns
6. The priority relative to other campaigns

Once the campaign has been defined, it is executed via the contact management capability.

BI can be a very effective means of analyzing, organizing, and monitoring the complex barrage of information generated in one's business and helping to generate a more effective business model for increasing revenue by keeping one's customer base happy and increasing profitability by cutting costs.

4.3.4.4 Customer Retention

BI tools can be applied to a variety of processes forming the telecom service provider's business. These business processes can be customer retention, cost cutting, or traffic management. For customer retention, strategic decision support BI tools would be used to track key performance metrics relative to customer install and disconnect activity and would assist telecommunications companies.

- This would provide early warning of increasing disconnect activity.
- If disconnect activity began to grow beyond acceptable limits, it would analyze why customers were disconnecting and extrapolate the impact on profitability.
- If the profitability impacts were not acceptable, it would formulate strategies for retention.

Once strategies were formulated, it would develop predictive models that would align retention strategies to the appropriate customers. Scoring and segmentation BI tools would assist telecommunications companies by applying predictive models from strategic decision support BI tools to the entire base of customers, assigning a score value.

The campaign assignment BI tool would assist telecommunications companies by applying scores from the scoring BI tool and other relevant data to assign customer lists to the appropriate retention campaigns.

Business management would initiate these campaigns and manage their execution. As feedback is returned from business management, input would be used by strategic decision support to refine retention strategies.

As may be deduced, a number of capabilities are needed to support a single business need (e.g., retention), and these capabilities are integrated through the business process. What may not be quite as evident is that these capabilities can be reused to support other business needs, such as customer care or fraud. Capabilities are essential to providing telecommunications companies with the ability to respond to the changing needs of their customers and the marketplace quickly and cost-effectively.

4.3.5 Disappointments from the Past and Barriers to Business Intelligence

Business intelligence solutions have had rather moderate success in terms of adding true business value to telecommunications companies. Notwithstanding the large investments in building BI solutions, telecommunications companies continue to face serious challenges in accessing data that is trustworthy, complete, and accurate, and data that makes business users self-sufficient.

Indeed, it is not uncommon to come across telecommunications companies where marketing and finance departments yield two different results on a critical input such as the success rate of campaigns. Also, if one adds up the time business users spend attempting to learn minute technical details in order to be self-sufficient and the costs of different departments acquiring their own technology pool to support their respective business needs, the drain on a telecommunications company's resources could be considerable.

On the other hand, though, technology departments tend to believe BI solutions have been extremely successful. Essentially, this is because technology teams assess success in terms of a very large database being implemented or a complex ETL (Extraction, Transformation, and Load) problem being addressed within the telecommunications company, as is frequently the case when a BI solution is implemented. Therefore, although BI solutions may have been technological success stories, their true value in terms of enhancing the trustworthiness of information or providing a holistic view has been limited.

BI efforts are taken up in some cases at the enterprise level, and in other cases at the department or function level. However, when building a BI solution, a common framework is usually adopted. The typical features of this framework are:

A common data repository with simplified structures to facilitate its consumption, periodic acquisition, and refining of data from many sources, loading this data into the repository, and extraction of data from these sources with a set of reporting tools.

BI solutions built using this approach have grown exponentially in size and scope, forcing the creation of more manageable subsets of data to suit the business requirements of different departments. Even as these smaller subsets of data became de facto sources, seeds for their uncontrolled proliferation were sown and, today, telecommunications companies have thousands of these smaller subsets of data that business users depend on for their day-to-day analysis and reporting. With this as the context, let's examine some of the challenges telecommunications companies face today with respect to their business intelligence infrastructure.

4.3.5.1 I Do Not Get the Full Picture

Telecommunications companies find it extremely difficult to acquire a cross-functional or holistic view of data. For example, how can a risk manager combine the profitability view of customers with their risk view, and analyze dependencies between profits and risky behavior, a common business need for managing the risk of a portfolio?

Cause: Risk data and profitability data reside in silos. The basic definitions, data structures, and granularity of representation are different. It is a systems integration nightmare to combine the two views of data.

4.3.5.2 I Do Not Trust the Data

The same question often elicits different responses from two different departments of a telecommunications company. Which is the data that the senior management should trust and why?

Cause: Answers come from different silos although both silos may have obtained their data from the same underlying data repository. Different definitions: What is the process to ensure that the silos are leveraging the same definition or calculations? No traceability: How were the estimates and figures arrived at? What were the bases? Can these figures be traced back to the source data? What were the transformations along the way? Invariably, a telecommunications company would draw a blank on all these questions.

4.3.5.3 I Am Not Empowered

Can a user gather simple business intelligence on, say, the total number of new customers or the customer attrition figures as of a specified date? Very often, this may be an involved exercise. Again, to gather this information, does a user have to be familiar with multiple query languages, tools, reporting interfaces, and databases? The answer from most business users is usually in the affirmative.

Cause: Users need to know the silo or data source to go to. Most users depend on "techies" to find them the answers. Most reporting tools require a basic understanding of syntax. Business users, again, depend on technical resources to delve into databases for relevant, often critical information. There is no common framework for leveraging definitions, and calculation reporting tools do not provide a common and consistent language for interaction.

4.3.5.4 I Cannot Close the Loop

To act, managers and analysts need current information proactively. For example, up-to-date information on high-end customers is critical to a customer relationship manager in determining customer satisfaction and probability of attrition. However, this kind of information seldom reaches line managers

on time and proactively. The ability to "close the loop" by sending proactive information and forcing the telecommunications company to act is a challenge that is mired in technological complexity.

Cause: Reporting solutions lack comprehensive alerting capabilities and are not flexible enough to allow users to define alerts themselves. Even if these facilities are available in some reporting solutions, business users require technical assistance to extract the information they require. Applications do not provide access to common business definitions and calculations that would allow business users to define seamless alerts and have the information delivered on multiple touch points.

4.3.5.5 It Takes Too Long to Build a BI Solution

Despite considerable spending on a BI infrastructure, most telecommunications companies find that what they have does not match up to their unique business needs. Building a BI solution for a telecommunications company requires extensive understanding of the telecommunications domain and how business users perform various analyses.

Cause: The effort to build a BI solution frequently begins from scratch, with teams trying to understand business needs and building suitable models. This can consume significant resources and time. Teams building a BI solution lack domain expertise. Often, teams of modelers and architects spend too much time understanding the business. Most vendors offer tools and technologies for point solutions. There are few complete, connected, and consistent analytic solutions.

4.3.6 Overcoming Barriers

Building a BI solution is more than an implementation of a bunch of technology tools. It requires a framework driven by a set of guiding principles. Some of these core principles of implementation are explained here:

1. *Establish the need for trustworthiness:* A BI solution has to deliver data that is trustworthy. Data becomes trustworthy when it can be substantiated. Ensure that data that is delivered to business users can be substantiated with a seamless traceability all the way to its source, without any additional effort. When a business user looks at a number and wants to authenticate it, the task should be as simple as a click of a button. Users should not have to depend on technical resources to draw out simple reports. In short, the traceability and lineage of data should be comprehensive and seamless in a BI solution.
2. *Establish the need for a holistic view:* Build a BI solution to provide a holistic view. Even though the effort may begin with a single department or division or a specific subject area, the solution should be designed such that additional subject areas can be added seamlessly. For example, a telecommunications company can begin by building the BI solution for risk. However, the telecommunications company must ensure that the underlying risk structures are created such that profitability, CRM, and the investment sides of the business can be accommodated at a later date. If efforts are decentralized and managed by individual departments, it is important to ensure that they coordinate the underlying structural aspects in ways that can facilitate extensibility and integration. Unless such strategic guidance is available, efforts tend to be independent.
3. *Establish the need and sanctity of consistent business terms:* A BI solution without this fundamental driver is set to fail. Facilitate and create a glossary of terms that is consistently applied across your enterprise. It is important to involve business users in this exercise. Also appreciate the diversity among departments, and ensure that everybody's requirements are accurately described.
4. *Enforce consistency at all costs:* Defining or establishing a business language in itself does not solve the problem. A mechanism to enforce this common language must be established. The ability to enforce this feature across the BI solution is a key component of the solution. Telecommunications companies must realize that departments would prefer managing their own subsets of data and would like to limit their access to those subsets. Departmentalized subsets are inevitable and are simple to manage and maintain. The challenge is to control their proliferation.

5. *Empower business users:* The ultimate test and one of the most uncompromising success factors of a BI solution is to empower business users and reduce their dependence on technology resources significantly. True empowerment comes when a business user interacts with the solutions using common and simple business language without worrying about where the data is stored and how the data is retrieved.
6. *Adopt connected and consistent solutions:* Guard against the tendency to go in for a point solution that supports your immediate business demands. When the risk department, for instance, chooses a specific Application Lifestyle Management (ALM) solution or an investment management solution without considering the need for holistic enterprise risk and CRM solutions, the scope to extend the point solution is limited. This is because point solutions are built for specific needs using varied technology paradigms and choices. Prefer, instead, an integrated, connected, future-proof solution: a solution that offers connected and consistent analytics that are derived from the same underlying business model, business definitions, and processing backbone.
7. *Reduce the implementation cycle:* Unless managed and controlled well, a BI solution implementation, particularly at the enterprise, division, or department level, can easily spin out of control. The best way to avoid this is to adopt prebuilt solutions that are, nevertheless, extensible. A good rule of thumb is to opt for an analytical solution suite that can jump-start your BI initiatives by 50% or so. This immediately reduces the implementation turnaround by about 60%.
8. *Close the loop and deliver proactively:* Another key feature of a successful BI solution is the capability to deliver business data proactively to users across multiple touch points. Business users should be able to define the necessary criteria for alerts and the BI solution should be capable of sending information as alerts on a variety of touch points. It is this that really determines the effectiveness of a BI solution: when information is offered in a manner that results in action.

Very few telecommunications companies have built BI solutions that have truly empowered them at all levels—strategic and tactical—to gain insight into their health and to act on this insight. Many an implementation story that is touted as a "model" centers on the technological rather the business aspects of BI. For a BI implementation to succeed in the business sense, it is important for top management to be committed and supportive, while providing the necessary direction and guidance. Also, it is important to adopt a strategic design and a common framework across the enterprise, implementing the solution in phases.

Furthermore, the management must keep in mind that point solutions that cater to specific functional needs may indeed be successful in the short term but, over the long haul, this may put the enterprise at considerable risk. Choosing prebuilt analytical solutions that are connected and consistent will improve the chances of success enormously.

4.3.7 Customer Intelligence

Out of 10,000 customers of an organization, 1,000 customers are providing 80% of the revenue and many of these most profitable customers reside in a sales region handled by that sales manager recently hired. Think of how the organization's strategy might change if it knew some simple, basic facts about its customers.

Enterprises should see customer activity—interactions with customer service, accounts payable, sales and marketing, and more—in its entirety. In recent years, businesses have frequently ignored, or at least paid only cursory attention to, one of the most fundamental keys to success: their relationship with their customers.

The sometimes paradoxical relationship between customers and businesses arises from fundamental spheres of influence. Consumers had little influence on how businesses responded to their needs, and businesses could derive little reward by distinguishing themselves through customer relationships and superior service. In effect, businesses dictated the relationships with their customers, and customers often accepted that standard. The status quo reigned. In today's competitive environment, the nature of customer

relationships has changed. Consumers have many choices to meet their needs, and aggressive advertising or access to the Internet increasingly broadens a consumer's horizons for competing products.

Companies are competing for the same customer, and successful businesses must provide a superior relationship with customers to stand out. In a way, the rise of CRM systems and methodologies that exploded in the late 1990s was merely a desire to return to "traditional" customer relationships. Successful corporations win and keep customers and prospects by establishing direct, sustainable, and manageable relationships.

4.3.7.1 Customer Data in the Telecommunications Industry

The telecommunications industry is a good example of how vital customer knowledge is to compete more effectively and nurture customer loyalty. Business managers and decision makers are constantly facing the need to answer fundamental business questions such as, "Who are our most profitable customers?" "What makes them profitable?" and "Which marketing campaigns should be developed to target this segment?" in order to monitor customer base evolution and behavior, define the most adequate marketing strategies, and assess the overall operational performance.

However, the intelligence required to answer this type of question is spread across several operational databases (billing, CRM, mediation, provisioning, enterprise resource planning [ERP], etc.) that do not support the ability to easily query, report, and analyze business data. Moreover, it is extremely complex to cross data coming from different systems. For instance, correlating traffic and customer data would be very useful to define the most competitive pricing plans.

4.3.7.2 Customer Relationship Management (CRM)

CRM is the alignment of business strategy, organizational structure and culture, and customer information and technology so that all customer interactions can be conducted to the long-term satisfaction of the customer and to the benefit and profit of the organization. To implement a coordinated, customer-focused business strategy, an organization must have business strategies that promote CRM across functional boundaries. Goals that include phrases such as "customer-focused" or "customer satisfaction" are indicators that CRM is important. However, if there are no underlying strategies in place that force a customer view across business functions, the organization is not likely to move far from the traditional product focus.

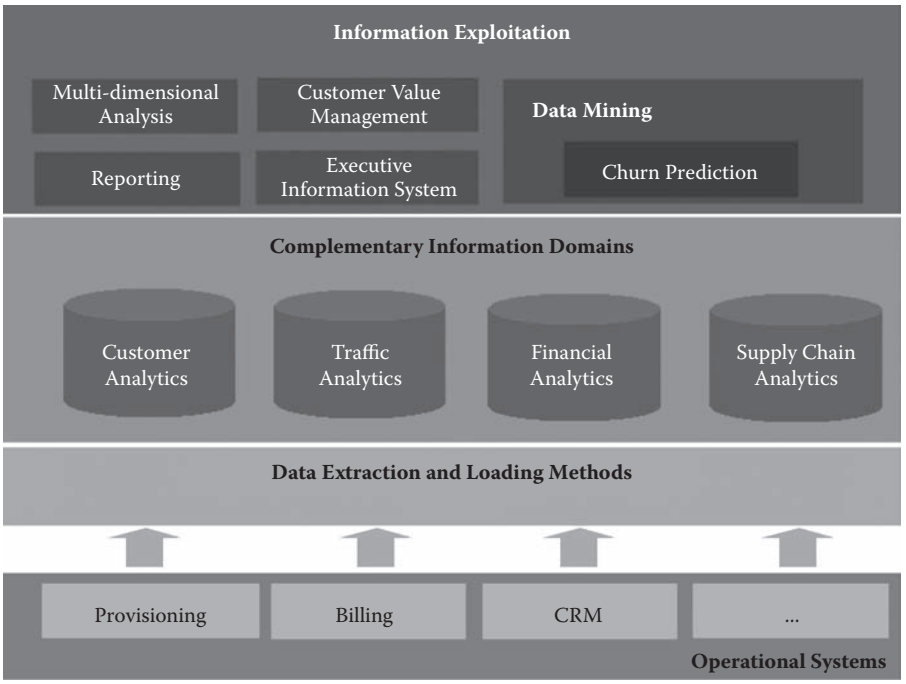
Telecommunications companies having the desire for customer-focused initiatives must implement enterprise strategies for moving the company in that direction. It should create a CRM-friendly organizational structure (see Figure 4.3.4). The overall organizational structure must promote cross-functional cooperation. Independent product-oriented business units, multiple marketing and sales organizations, and distributed customer care centers can all inhibit an organization's ability to determine and carry out the next promotion or service activity for the customer. With the autonomy and control possessed by each business unit executive, the telecommunications companies may lack the organizational structure required to implement cross-department initiatives.

4.3.7.2.1 CRM-Savvy Organizational Culture

Culture is a critical but often overlooked factor that can have a strong influence on the success or failure of any CRM endeavor. There are three predominant aspects to consider: (1) the organization's ability and willingness to effect change in business and thought processes, (2) the degree to which the business units work together, reach compromise, and facilitate shared strategies, and (3) it is important that executives support CRM.

4.3.7.2.2 Integrated Customer Information Environment

Customer information is the cornerstone of a successful CRM program. This information must provide a common customer view and must be distributed across the organization to facilitate both operational and analytical uses. This always requires a technology architecture that integrates multiple applications,



FIGu RE 4.3.4 Customer information system.

ranging from operational legacy applications, to call center systems, to the data warehouse and its associated data marts.

Although the technologists in the telecommunications industry understand the need for consolidated customer information, the business units of most companies across the industry until recently maintained their disparate systems. However, today's organizations are leveraging BI in conjunction with CRM solutions in numerous ways to derive hitherto unseen benefits and business opportunities. Using BI tools, the intelligence can be built into CRM applications as well.

Another point in favor of a BI–CRM combination is the fact that many organizations already have a CRM solution in place. It makes more sense for such enterprises to leverage existing CRM investments and deploy a full-fledged BI system. This will ensure that data from all the data sources (including the CRM) can be brought under one roof for a better all-round perspective of the business.

In today's extremely challenging business environment, telecommunications companies are under intense pressure to reduce or eliminate the major threats to these slim margins: revenue leakage, churn, inefficient network usage, and least-cost routing plans.

Telecommunications companies rely on analysis of their terabytes of call data record (CDR) data to help them make business-critical decisions that will positively affect their bottom line. High-end data warehouses and powerful business intelligence solutions are essential tools to help carriers meet profit goals.

Many critical telecommunications functions rely on fast complex analysis of CDR data. Key initiatives include analyzing behavior data using CRM programs to optimally target services and reduce churn, ensuring complete and accurate billing and modeling call behavior with revenue assurance programs, and optimizing network operations using operations management programs. These initiatives all benefit from improved access to CDR-level data, access to large quantities of historical information for trend analysis, and from the ability to quickly run complex BI queries.

These significant performance limitations force carriers to make a choice. They must decide whether to summarize or filter the data for analysis, or create a massive, often complex CDR warehouse to analyze call

detail information. Both of these options pose serious limitations and challenges, resulting in incomplete information for decision making or costly and time-consuming system development and maintenance.

There is tremendous value latent in call detail information for CRM (Customer Relationship Management), revenue assurance, fraud detection, and network usage analysis. Performing real-time analysis of voluminous call detail data with complex queries requires much more performance than legacy general-purpose systems can provide. Effective BI programs open the door for increased profitability, while eliminating the barriers to accessing and analyzing dynamic, detailed information, and offering carriers' performance, value, and simplicity in a data warehouse system. For the first time, carriers can leverage their terabytes of CDR data for real-time, better-informed, and more strategic business decisions.

4.3.7.3 Using Client Data for Intelligence

Before you can establish meaningful relationships, however, companies must be able to answer—with precision and confidence—one seemingly easy question. Exactly who are my customers? In an increasingly competitive world, using the client database smartly, to gain a better understanding of the organization's number one asset, customers, can make or break the success of the organization. BI answers questions such as the following:

1. Who are my most/least profitable customers and products?
2. To whom should I address my marketing action/campaign?
3. What are the sales performances of this period with respect to my objectives and with respect to the same period last year?
4. Which are currently my best performing products in Germany?
5. Which customers are about to churn; who are fraudulent?
6. At what price should I sell my service/product in Geneva?

Most enterprises use databases to store information about their current customers, previous customers, business partners, and potential customers. The challenge lies in finding a way to harness the useful information contained within these high-volume databases in order to produce intelligent business solutions.

Analyzing the information that an organization stores in connection with all customer interactions can reveal a lot of remarkable facts about the buying behavior of customers, what motivates them, and what might make them stop buying from you. It also provides a scientific method to monitor an organization's own business performance.

Detailed analysis of customer data will also provide insight into their needs and wants. The exercise will analyze and segment customers' buying patterns and identify potential services that are in demand. Organizations can use this information to shorten response times to market changes, which then allows for better alignment of products and services with customers' needs.

An in-depth understanding of customers, provided through comprehensive data analysis, will also allow picking and targeting better prospects, achieve a higher response rate from marketing programs, and at the same time identify reasons for customer attrition and create or alter programs and services accordingly.

An important point to consider is whether the analysis is guided by predefined questions. Predefined points of analysis are aimed at understanding certain types of behaviors by analyzing relationships between various predecided influencing factors. For example, a predefined analysis of customer service sales would illustrate the effect of good and bad customer service on sales, and would answer questions such as how important customer service is to customers and how much it influences future sales. On the contrary, the objective of an open-ended analysis is to discover trends that are not anticipated by ordinary immersion in day-to-day business. Performing an open-ended analysis internally is often impaired by the expectations brought on by individuals working within the organization.

The techniques used to analyze data are complex. In order for an organization to be able to use the results of the data analysis, it is crucial that the results not be clouded by the complexity of the calculations but are delivered in a straightforward manner. It is important for an organization to recognize that a good understanding of its customers is useful only to the extent to which this knowledge can be

translated into real business practices. Business intelligence refers not only to the data analysis in itself, but also to how you relate the results from the data analysis to everyday business decisions and how you translate the recommended actions stemming from the analysis into live campaigns.

It is therefore important to ensure that the marketing department interacts with the data analysts constantly throughout the process. That way, when the data analysis is complete, the marketing personnel will already be in tune with the issues the organization is facing, and will be able to develop campaigns to capitalize on opportunities and strategies to mend weaknesses quickly and effectively.

Understanding how external market conditions affect your business will enable you to react quickly to future changes in the market. Finally, understanding customer behavior and the way customers use products and services will enable the organization to improve its service to its current client base as well as to target new business more effectively.

True customer intelligence is an elusive goal, but it is becoming a reality for companies tired of the incomplete and slippery information that tries to pass as corporate truth. For years, companies of all sizes in every industry have searched for ways to discover the truth about their customers. But more often than not, businesses face uncertainties about even simple truths, such as how many customers they really have and which ones are worth keeping.

In fact, the sad truth is that many large enterprises still do not know their customers very well. In most cases, businesses have amassed wads of customer data. As organizations increasingly standardized on different data collection methods—CRM, ERP, Data Warehouses (DWs), and the like—customer data often was replicated in different systems. And each business unit or division may have its own systems. This viral spread in applications led to a confused and untenable view of the customer.

To maintain, manage, and track these critically important relationships and the associated customer activity, corporations are investing valuable time and resources into managing customer data with Customer Data Integration (CDI) systems. As a result, every department has a different piece of the customer puzzle and the business has no way to fit them all together. Enter customer data integration, or CDI. CDI will change all this.

4.3.8 Impact of BI and BPM on the Telecommunications Industry

High performance BI tools such as advanced analytics, reports, and executive dashboards are playing a vital role in the telecommunications industry today. The value for the enterprise lies in the challenge of aggressively retaining and growing the customer base and extracting more value from each customer relationship. Those departments receiving the greatest benefit from BI and business process management (BPM) include:

- *Marketing:* To oversee product marketing and product management activities
- *Customer care:* To improve customer retention and customer satisfaction levels
- *Call center:* To better manage up-sell, cross-sell, and outbound marketing campaigns
- *Network operations:* To retain and improve network and service quality management
- *Finance:* To monitor cost per gross addition, marketing spending, plus the impact of churn

BI and BPS tools enhance core functionalities such as customer care, sales, billing, and service delivery, as they:

- Provide analysis and visibility into business metrics for decision makers
- Facilitate improvement of operational processes
- Support one integrated view of the customer across all departments, processes, and products
- Give enterprisewide visibility of revenue, cost sources, and allocation
- Increase revenue from greater understanding of business processes and customer segments

With advanced BI tools and BPM, telecommunications companies are today addressing the key areas of customer insight and revenue growth including the following.

4.3.8.1 Average Revenue per User (ARPU) Boosting

Boosting ARPU provides operators with the capability to identify and deploy potential revenue-increasing programs within the customer base. BI and BPM tools help operators in the improvement of ARPU and revenue stream analysis, the ability to identify segments with the propensity to buy, market penetration analysis, and campaign effectiveness analysis.

In most mature markets, communications service providers are experiencing a decline in fixed access lines and are near saturation in terms of mobile customer penetration. Due to diminishing subscriber growth, network operators focus on retaining profitable customers as well as driving incremental revenue from the installed base.

4.3.8.2 Churn Prediction and Management

It is six times more costly to acquire a new customer than to retain a profitable customer. Churn prediction and management provide communications service providers with the means to more effectively retain their most valuable customers, segment their customer base, and focus marketing campaigns in terms of Customer Lifetime Value (CLV).

Understanding customers can have a positive impact on business. BI and BPM tools can help create transparency in your customer base by identifying potential churners, determining customer lifetime value, evaluating effective retention programs, and identifying the most responsive targets.

4.3.8.3 Integrate Cost and Revenue

The cost of customer acquisition and customer profitability is pushing companies to look more closely at profit margin analysis to reduce costs and increase revenue. The expansion into multiservice, multinetwork, and multipartner environments has made cost management complex, because many customers don't truly understand the accuracy of inventory and assets, customer profiles, and business users' needs.

While carriers struggle to compete by offering creative service bundles, they often fail to conduct detailed margin analysis to understand the costs associated with these bundles, making it difficult to implement profitable plans and eliminate unprofitable services and customers. For carriers to understand their profit margins, there must be a clear view of both costs and revenues on a per-product and per-customer basis. This requires that carriers change their mind set in treating Revenue Assurance (RA) and Cost Management (CM) as two totally different departments within their business.

Ultimately, cost management departments will have to communicate and share data with the revenue assurance department so that marketing can truly understand the impact of acquisition costs on profitability. This will require using fine-grained revenue measures that exceed the capability of existing general ledger and accounting systems. Service provider profitability will increase as a result of the RA group's focus on identifying revenue leaks, and the cost management group's managing and reducing costs. For synergies to be realized and control margins to improve, there need to be end-to-end revenue monitoring systems that start with provisioning of services and continue all the way through to invoice verification in cost management. In other words, if carriers look at only what they bill, without marrying information to what they know of costs (i.e., what wholesale providers charge), then they have only half the picture.

Because most companies are not at the point of sophistication where they can truly integrate cost and revenue management, some are managing to derive business intelligence directly from billing and OSSs via rapid-query tools, analyses, and reporting capabilities. As BI and BPM platforms become more prevalent in telecom, they will help carriers to establish integration among financial measures of profitability and operational key performance indicators.

4.3.8.4 Improved Procurement

Most companies can have only a rough idea of the productivity gains that can be derived from SCM (Supply Chain Management) as part of the entire setup. No manufacturing company can get a 360-degree

view of its operations if its SCM is a stand-alone system. BI has also made inroads in providing a consolidated view of the company including its supply chain.

This has helped reduce the traditional limited view of the supply side of operations. With this new approach of BPM, the BI system also acts as an intermediary between the organization's CRM and SCM systems. The consolidated view aside, this approach helps the company manage its supply to match the demands of its customers. It can also help the customer tailor product attributes to match changing customer needs.

4.3.8.5 Self-Service

Economic success and prosperity for businesses in the twenty-first century owe their origin to customer empowerment brought about by one overriding development: the introduction of self-service in almost every business area. The deployment of next-generation services, such as broadband, VoIP, IPTV, wireless TV, and Fiber to the Home (FTTH) will create customer pressure on telecommunications service providers to provide for customers' self-control of their own services and support. Self-service enables service providers to not only reduce the cost of interaction, but also collect more customer information to enable more personalized service. This, in turn, can increase customer retention and revenue. From the customer's perspective, self-service is valuable because it is convenient and flexible. A well-designed self-service system can help companies not only reduce customer-care costs, but also help retain customers and provide up-selling and cross-selling opportunities.

Although self-service technology is now widely taken for granted, it has transformed the way business operates, with business intelligence at the heart of any successful self-service initiative. Business intelligence tools improve overall customer satisfaction by making self-services more interactive and productive by making available and using information from multiple sources. It also makes such initiatives more relevant by applying experience and assumptions to develop an accurate understanding of customer usage dynamics.

Business intelligence tools enable service providers to achieve optimal effectiveness, while investing in self-service initiatives by assisting them in:

- Collecting sufficient, current, and accurate data
- Ensuring information and content are up to date and effective
- Integrating front-end and back-end systems
- Providing customers with features that enable control, customization, and ease of use
- Ensuring flexibility for service providers for offering new services and promotions

4.3.8.6 Risk Resolution and Management

Risk management is increasingly viewed as an integral element of the risk and decision making by organizations. Risk management may be undertaken prior to the decision itself (e.g., insuring against certain risks) or after the decision (i.e., effective management of relationships with customers to reduce the likely incidence of disputes).

These uncertainties and risks place new demands on the field of risk management, indicating that the conventional methods of managing uncertainty (e.g., buffer stocks, spare capacity, quoted lead times) are likely to be less effective in meeting the new demands and uncertainties.

With BPM and BI tools, managers are now equipped with improved knowledge, skills, and understanding to be able to identify, analyze, and manage these developments and to assess the consequences and risks arising from the more diverse range of markets contexts. BI provides the resources and solutions to facilitate this improvement in risk management.

4.3.8.7 Pre- and Postpaid Convergence

Traditional segmentation of wireless subscribers based on their payment modes has resulted in separate OSS/BSS systems. Fierce global competition is now challenging the basis of this segmentation. The

economies of maintaining two separate systems, the need to introduce advanced services, and pressure to reduce TCO (total cost of ownership) is compelling service providers to redraw OSS/BSS solutions.

Subscribers across the world markets have selected prepaid as a preferred payment method over postpaid. At the same time they demand the products and services offered to other postpaid customers. Existing prepaid systems heavily rely on network-based charging, which can deal with time-based charging, but are incapable of charging the combination of voice, data, and content services.

For service providers, the additional revenue streams that can be generated through offering advanced services to the postpaid subscribers are limited by the ability to charge for these services. With pressures to offer competitive services and increase ARPU, service providers find the ever-increasing need to adopt a converged BSS solution for their prepaid and postpaid customers.

Prepaid services, primarily introduced to develop new market segments, meant the service providers had to add stand-alone systems. These systems are limited in terms of scalability, rapid growth in the subscriber base, flexibility, and complex offerings of advanced services. Efforts to migrate prepaid subscribers to postpaid services to overcome these challenges have not been met with the desired results.

This has led service providers to think of alternate strategies and the emerging needs for a converged prepaid-postpaid solution to handle all the rating, billing, and customer relationship management demands. The converged solution provides service providers with operational efficiencies, increased customer satisfaction, and the ability to introduce products and services quickly and efficiently to subscribers regardless of their preference of payment method. With all the capabilities of their postpaid systems available to prepaid subscribers, the difference between these types of subscribers is reduced to a choice of payment method.

4.3.8.8 Business Challenges

- *Hybrid Prepaid and Postpaid Accounts:* Service providers cannot offer bundled prepaid and postpaid services and price plans.
- *Advanced Products and Services:* Many complex voice and data services cannot be offered to prepaid subscribers as the prepaid systems cannot charge them.
- *Integrated Customer View:* Individual subscriber needs cannot be addressed.
- *Customer Service:* Limited communication with prepaid subscribers restricts target promotion.

4.3.8.9 Operational Challenges

- *Vendor:* Different vendors for the prepaid and postpaid systems result in integration, upgrade, and maintenance issues.
- *IT and Operations:* Separate database and operations teams associated with prepaid and postpaid results in overhead and reduced bottom lines.
- *Customer and Product Data:* Prepaid and postpaid customers; product data is maintained separately, resulting in duplication and increased cost of maintenance.
- *OSS/BSS Infrastructure:* Diverse infrastructure resulting in higher maintenance costs and integration challenges.

The business technical and operational challenges are inefficiencies resulting from disparate prepaid and postpaid billing systems if addresses can effectively offer increased revenue streams and give a much required boost to the service providers' bottom line.

4.3.8.10 BI for Convergence

Currently, most service providers are maintaining separate systems for prepaid and postpaid. This situation has many drawbacks on both the marketing and operational levels, including duplication of effort and the lack of an integrated view of the customer. The end result is that service providers have higher operational costs and yet are not able to generate new revenue streams.

There are huge benefits of moving to a convergent BI environment that includes a single set of system modules, common infrastructure, and rapid introduction of new services. However, flexibility offered by the BI platform in transition from the current separate prepaid and postpaid systems to a convergent system is unprecedented. Most BI platforms allow transition in several ways, in accordance with the customer's requirements. The approach can be a rapid transfer to a single convergent billing system, or phased, in a manner that guarantees benefits and ROI at each stage. BI provides a smooth migration path from legacy systems.

Such platforms enable the service provider to nurture its prepaid customers through a unified customer view, sophisticated offerings for next-generation services, and enhanced customer services. Moreover, it can offer the benefits of maintaining a single billing system for prepaid and postpaid, with a single customer database, product catalog, and rating engine. The approach is effective across multiple functional areas and provides more efficient applications and services such as:

- Message acquisition and formatting
- Real-time charging environment
- Multidimensional rating
- Balance management
- Product catalog, development, and improvement
- Single customer data and billing

4.3.9 Summary and Trends

Today's CDI systems have evolved into highly sophisticated applications incorporating leading-edge research and development advances in fields such as information theory, natural language processing, artificial intelligence, and others. One major advance has been the recognition of users' needs to be able to fine-tune the matching and householding behavior to create a single customer view that more directly fits with the business needs.

CDI vendors no longer assume that they can dictate to businesses what the "correct" single customer view is. As businesses have become increasingly sophisticated with business intelligence, CRM, and one-to-one systems, they have demanded control of their customer definition. This is typically affected via business rules that control how the single customer view is resolved by the CDI system.

The BI and middleware vendors will collaborate with CDI vendors to provide tighter integration. In addition, given the importance of trustworthy data to both BI and CDI projects, trusted data sources will play a bigger role in the market. With most organizations investing big money in such initiatives and also opening up the system to Web self-service, salespeople entering data, field service entering data, and call centers entering data, the data can become corrupted pretty quickly. Rather than argue over which department's data is right, companies will look to trusted data sources to append or be the source of record for customer information. These data vendors are offering more services to the CDI market today, and in the future they may be more aggressive in this area. While companies grapple with CDI, many are also implementing business intelligence tools, and these technologies will meet in the middleware.

Competitive Intelligence (CI) is a specialized branch of business intelligence and has become an important initiative in the present competitive scenario. Next to knowing all about your own business, the best thing to know about is the other fellow's business.

Competitive intelligence is defined as a systematic and ethical program for gathering, analyzing, and managing external information that can affect your company's plans, decisions, and operations. In other words, CI is the process of ensuring competitiveness in the marketplace through a greater understanding of competitors and the overall competitive environment. CI is not as difficult as it sounds. Much of what is obtained comes from sources available to everyone, including government sources, online databases, interviews or surveys, special interest groups (such as academics, trade associations,

and consumer groups), private sector sources (such as competitors, suppliers, distributors, customers), or media (journals, wire services, newspapers, and financial reports).

The challenge with CI is not lack of information; it is the ability to differentiate useful CI from chatter or even disinformation. Of course, once you start practicing competitive intelligence, the next stage is to introduce countermeasures to make the CI task about you more difficult for other firms. The game of measure, countermeasure, and counter-countermeasure, and so on to counter to the nth measure is played in industry just as it is in politics and in international competition.

Acronyms

ALM	Application Life Management
CRM	Customer Relationship Management
DW	Data Warehouse
ERP	Enterprise Resource Planning
FTTH	Fiber to the Home
IRR	Internal Rate of Return

4.4 Service-Level Management

Christian Voigt and Kornel Terplan

4.4.1 Introduction

Service-Level Agreements determine the service and its security requirements in a legal sense. Using legal contracts, such as a Service-Level Agreement, disagreements between contracting parties can be avoided. Disagreements may include the type of service, limits of a service, quality of a service, and compensation for a service. In order to inform the customer about the actual quality of services, support interfaces are recommended between the service provider and the customer.

Service-Level Agreements can be signed between service providers and customers representing a usual legal contract between businesses or between various departments of a corporation. External Service-Level Agreements are always considered as legal business contracts by both contracting parties, but this is not always the case with internal contracts. The result is that for the internal service provider or operator, the limits of such contracts underestimate or do not deliver satisfactory service quality.

Service-Level Agreements have to be signed with all relevant suppliers of services and equipment; they must be provided in written form. Service-Level Agreements must describe and quantify the available service level, and all responsible persons must be identified in the contract.

4.4.2 Principal Terms and Metrics

Service-l evel Agreements are written agreements about the service and its quality of service (QoS) metrics between contracting parties.

Service-l evel Management is the process of determining QoS metrics, preparation and maintenance of Service-Level Agreements, continuous collection and distribution of information about QoS metrics, and supervising whether the conditions of Service-Level Agreements are met.

The Service-Level Agreement is more than just a list of service metrics; it lays out the ongoing monitoring, reporting, and response process. The Service-Level Agreement should clearly define the responsibilities of both parties. For each function that is defined, the person responsible for controlling that function needs to be identified by position.

A Service-Level Agreement is a formal negotiated agreement between two parties. It is a contract that exists between the service provider and the customer or among multiple service providers or between service providers and network operators. It is designed to create a common understanding about services, priorities, and responsibilities.

A Service-Level Agreement should also cover corrective actions; that is, the steps to be taken in the event that a service-level objective is not met. This section of the contract should define who resolves the problem of each service deficiency, as well as consequences for not resolving the problem. Consequences can appear in the form of penalty clauses or alternatively a bonus clause for meeting the objectives. The end result will be the same.

Service-Level Agreements can cover many aspects of the relationship between the customer and the service provider, such as quality and performance of services, customer care, billing, and provisioning. Performance reporting uses the Service-Level Agreement as a reference, but does not address the other parameters known to exist as part of the Service-Level Agreements.

Service-Level Agreements (SLAs) are an excellent tool for customer and service provider management. A well-crafted SLA sets and manages expectations for all elements of the service to which it refers. It assists the service provider in forcing operational change, improving internal measurement and reporting, assessing trends, improving customer relationships, and provides a vehicle for potential differentiation from its natural competitors.

Quality of service (QoS) parameters can be used to quantify the quality of service and are included in SLAs. In selecting quality of service metrics for SLAs, the following criteria should be considered:

- Importance of the metric for operations
- Ease of measurement
- Providing basis data for reporting
- Quantification of the service
- Importance of the metric for business applications

4.4.2.1 Service-Independent Metrics

Examples of service-independent metrics are outlined below.

Availability for:

- Service access points
- Applications
- Devices
- Transmission facilities

Mean time to service restoration for:

- Service access points
- Applications
- Devices
- Transmission facilities

Mean time between failures for:

- Service access points
- Applications
- Devices
- Transmission facilities

Mean time to repair for:

- Service access points

- Applications
- Devices
- Transmission facilities

Mean time of repair for:

- Service access points
- Applications
- Devices
- Transmission facilities

Relationship: proactive and reactive problem detection for:

- Service access points
- Applications
- Devices
- Transmission facilities

Relationship: preventive and reactive problem detection for:

- Service access points
- Applications
- Devices
- Transmission facilities

Relationship: Number of referred problems to all detected problems for:

- Service access points
- Applications
- Devices
- Transmission facilities

Meeting escalation guidelines for:

- Service access points
- Applications
- Devices
- Transmission facilities

Number of chronic problems to all detected problems for:

- Service access points
- Applications
- Devices
- Transmission facilities

Help desk performance:

- Receiving notification
- Solving trivial problems
- Performance during first, second, and third shifts
- Performance on weekends
- Performance on holidays

Number of outages for:

- Service access points

- Applications
- Devices
- Transmission facilities

Average number of outages for:

- Service access points
- Applications
- Devices
- Transmission facilities

4.4.2.2 Service-Dependent Metrics

Examples of service-dependent metrics are outlined below.

Resource utilization for:

- Service access points
- Applications
- Devices
- Transmission facilities

Leased lines:

- Bit error rate
- Nonavailable seconds
- Severely errored seconds
- Block error ratio

PDH (plesiochronous digital hierarchy) devices:

- Number of problems
- Average duration of problem restoration

SDH (synchronous digital hierarchy) devices:

- Number of problems
- Average duration of problem restoration

Frame Relay service:

- Network delay
- Committed Information Rate (CIR)
- Cyclic Redundancy Check (CRC) Errors
- Discarded frames
- Effective Permanent Virtual Circuit (PVC) throughput

ATM (asynchronous transfer mode) service

- Available bit rate (ABR)
- Constant bit rate (CBR)
- Unspecified bit rate (UBR)
- Variable bit rate (VBR)
- Network delay
- Errored cell rate
- Effective PVC throughput

XDSL service:

- Number of problems
- Average duration of problem restoration

Cable networks:

- Number of problems
- Average duration of problem restoration

Packet switching service:

- Throughput rate
- Relationship: discarded packets to all packets

IP services:

- Packet delay in one direction
- Packet delay in both directions
- Packet loss in one direction
- Packet loss in both directions
- Transmission throughput

4.4.2.3 Classification of Metrics

Surveys of market research companies clearly show where customer interests lie. This interest should be taken seriously by service providers and network operators. Customers are interested, first of all, in service-oriented metrics that are technology independent. This means that technology-dependent metrics need not to be included in SLAs. But they are important in computing technology-independent metrics.

The priority sequence of metrics from the perspective of customers is the following:

- Availability of all components (devices and transmission links) connected to the network
- Availability of application on the network (accessing applications)
- Availability of servers
- Network round-trip time or network delay
- Application response time during peak periods
- Availability of clients
- Server delay
- Mean application response time
- Client delay
- Median application response time
- Percentage of transactions completed within defined performance levels

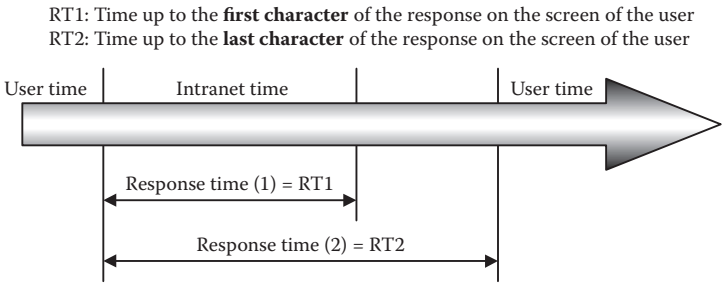
4.4.2.4 General Considerations

Service availability (SA) is a percentage (SA%) that indicates the time during which the contracted service at the respective service access point (SAP) is operational. *Operational* means that the customer has the ability to use the service as specified in the SLA.

An event affecting the service at the service access point can be defined as an *outage*. The duration of this specific event is the *outage interval*. Ordinarily this concept is used for the unavailability (UA%) and service availability percentage (SA%) calculations as follows:

$$SA\% = 100\% - UA\%$$

$$UA\% = (\text{Sum of outage intervals/Activity time}) \times 100\%$$



FIGu RE 4.4.1 Response-time definitions.

An additional issue for SLA management is to determine whether an event affecting the service at the SAP is causing a complete service outage (service fully unavailable) or a partial service outage (degraded service available).

Critical success factors for measuring and computing availability are:

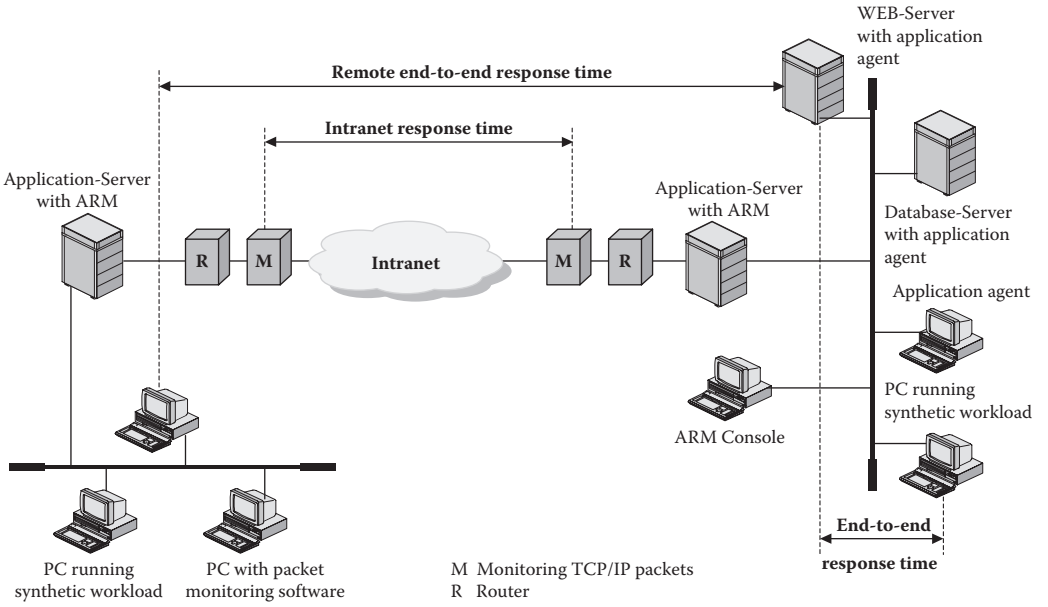
- Trouble tickets
- Proactive tools
- Workforce with dispatch capabilities
- Efficient help desk
- Skill and experience of subject matter experts

Response time is one of the key metrics in all SLAs. Its definition varies, but usually users consider the period between sending the inquiry and receiving the full answer as response time. Figure 4.4.1 displays the differences between two response-time (RT) alternatives:

- Time up to the first character of the response on the screen of the user
- Time up to the last character of the response on the screen of the user

The second definition is a better suited definition for the working cycle of users. The difference between RT2 and RT1 depends on many factors, such as the throughput of the backbone and access networks, servers in these networks, number of hops, and the hardware/software capabilities of the client's workstation or browser. Present measurement technology offers the following alternatives:

- Monitors and packet analyzers: These analyzers filter and interpret packets and draw inferences about application response times based on these results. These monitors are passively listening to the network traffic and calculate the time it takes specific packets to get from source to destination. They can read the content of packages revealing eventual application errors and inefficiency, but they cannot measure response time end to end.
- Synthetic workload tools: These tools issue live traffic to get a consistent measurement of response time on a particular connection in the intranet or for a given application. These tools are installed on servers, desktops, or both. They typically send Transmission Control Protocol (TCP) messages or Structured Query Language (SQL) queries to servers and measure the time of the reply. Results from multiple sources are correlated to give a more detailed view of intranet response times. They are very accurate with respect to the end-to-end response time.
- Application agents: These agents work within or alongside applications, using software that monitors keystrokes and commands to track down how long a specific transaction takes. They can run on both the client and server. They clock specific portions of the application at the server or at the workstation. The use of agents needs customization and the correlation of many measurements in order to give users a performance estimate about their intranet.
- Use of application response-time measurement (ARM) management information bases (MIBs): ARM defines application programming interfaces (APIs) that allow programmers to write agents



FIGu RE 4.4.2 Positioning response time measuring tools**.**

into an application so that network managers and Web masters can monitor it for a range of performance metrics, including response time. It is a complete offer to application management, but it requires rewriting of existing code, which many companies are unwilling to do.

Figure 4.4.2 shows the locations of these tools and agents.
When evaluating products, many components must be considered. These factors are:

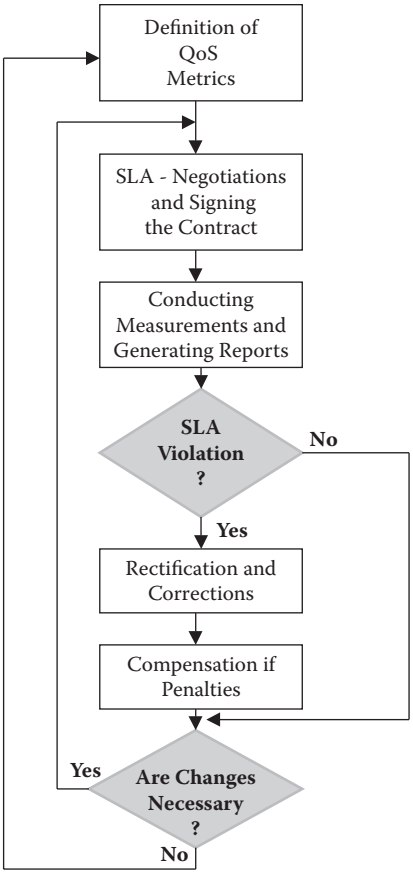
- Customization needs
- Maintenance requirements
- Deployment of code
- Overhead of transmitting measurement data
- Load increase due to synthetic workload
- Reporting capabilities
- Capabilities to solve complex performance problems
- Capabilities to conduct root-cause analysis
- Combination with modeling tools
- Price of the tools

Critical success factors for measuring and reporting of response time are:

- Measurement procedures
- Availability of application MIBs
- Proactive tools
- Workforce with dispatch capabilities
- Efficient help desk
- Skill and experience of subject matter experts

The pressure on service providers and network operators is increasing. Some guidance for metrics has been recently published by them. Examples are:

- Minimal network delay 70 to 100 ms



FIGu RE 4.4.3 Service-Level Management process.

- Minimal packet loss 1%
- Network availability 99 to 100%

Compensation policies have also been published. Examples include:

- One day free of charge services
- Three days free of charge services

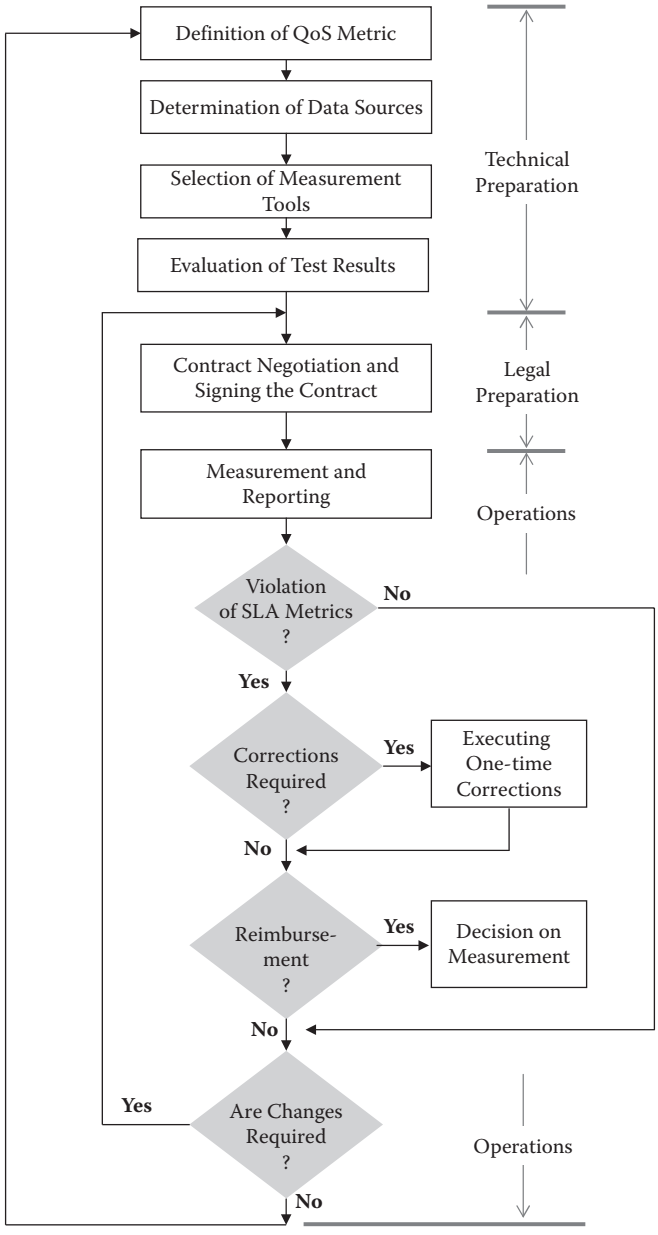
4.4.3 Process of Service-Level Management

Figure 4.4.3 shows a high-level overview of the principal activities of Service-Level Management.

Service-Level Management (SLM) requires that multiple metrics are continuously supervised and measured. Depending on the contract, several reports are generated and distributed. Data sources include:

- Trouble tickets
- Alarms (Simple Network Management Protocol [SNMP] traps or alarms from other sources)
- Logs
- Performance metrics that have been measured by various tools

All collected data must be formatted in order to guarantee a unified format (Figure 4.4.4). No complicated processing is expected in this phase. The results are summarized in a table (Table 4.4.1) including a number of events that are going to be used to evaluate service-level violations.



FIGu RE 4.4.4 Detailed Service-Level Management process.

TABl E 4.4.1

Events
Type of event
Identification of event
Time stamp
Severity
Importance

TABl E 4.4.2 Classification of Problems

Problem	Severity of Problem	Comments
Outage of nodes without backup	Critical problem	
Outage of nodes with backup	Major problem	
Generic node problems	Minor problem	
Node isolated	Critical problem	
Site isolated	Critical problem	Escalation step 2
Region isolated	Critical problem	Escalation step 2
Country isolated	Critical problem	
Change of circuit's media	Major problem	
NOC outage	Critical problem	Escalation step 2
Line problem without backup	Major problem	
Line problem with backup, but with performance problem	Minor problem	
Line problem without backup	Critical problem	
Performance degradation in the whole network	Critical problem	Escalation step 2
Performance degradation in network segments	Major problem	
Performance degradation for a limited number of sites	Minor problem	
Outage of the management system	Critical problem	Escalation step 2
Outage of network management applications	Major problem	
Routing problem in whole network	Critical problem	Escalation step 2
Routing problem in networking segments	Major problem	
Virus in the whole network	Critical problem	Escalation step 2

Table 4.4.1 is the basis to initiate escalation steps. First of all, problems should be classified by severity (Table 4.4.2). Severity has three classes:

- Critical problems
- Major problems
- Minor problems

For each severity level, escalation procedures should be prepared in advance. This preparation work includes the following steps:

- Definition of emergency: usually problems with a certain severity or a combination of problems with a certain severity. In the second case, not all problems must show the highest severity.
- Determination of escalation layers: it depends on the constellation of participating service providers and operators. Usually, two layers for the customer and one layer for each provider and operator should be considered.
- Identification of persons: for each layer, persons with location and reach numbers should be clearly identified.
- Description of processes: each severity triggers specific escalation steps that describe, in detail, manual and automated tasks. The time limit is also given for each severity.

TABl E 4.4.3 Customer Master File with Service-Level Agreement Entries

Customer ID
Service ID (A)
Service ID (B)
Service ID (N)

TABl E 4.4.4 Table of Services

Service ID (A)
Attributes (A)
Service ID (B)
Attributes (B)
Service ID (N)
Attributes (N)

TABl E 4.4.5 Service-Level Agreements

SLA ID (A)
Services in A
Metrics in A
SLA ID (B)
Services in B
Metrics in B
SLA ID N
Services in N
Metrics in N

Another table is prepared (Table 4.4.3) containing customer master data including the list of all services agreed upon with the service provider and network operator. The service portfolio is expected to be maintained in another table (Table 4.4.4), together with their typical service metrics.

The SLAs, which are standardized as much as possible, are maintained in a separate database. The number of different SLAs should be kept to a reasonable minimum in the best interests of service providers and network operators. Table 4.4.5 is used to identify when reports should be generated and their periodicity.

Table 4.4.6 shows a sample table to be used for documenting escalation procedures.

4.4.4 Sample SLA

In order to design and implement SLAs, the following activities are necessary.

- **Draft of the agreement**: A subject matter expert collects information about the functionality, application goals, and security requirements of the service. On the basis of this information, alternatives for service levels and their goals are defined. Depending on the specific requirements of the customer, SLAs may be specific as well.
- **Definition of metrics**: In the phase of defining and selecting service metrics, only metrics should be selected that can be met, measured, and controlled, along with those that cannot be interpreted differently.
- **Definition of security requirements**: The customer's internal security standards should be used when defining security requirements.
- **Agreement on the textual draft**: Prior to displaying the contract to the contracting parties for signing, the text of the contract should be coordinated with the legal department, the purchasing department, and with top management.

TABl E 4.4.6 Escalation Procedure for Service-Level Agreements

Severity	Escalation Steps 0	1	2	3
	Critical Problems			
Duration for the escalation step	0.5 H	1.5 H	2 H	24 H
Responsible for problem restoration	HD	Operator	Service Provider	Crisis Manager
Information is provided for	Customer Operators NO	Customer Operators NP	Customer Operators Service Provider NP	Customer Operators Service Provider NP
	Major Problems			
Duration for the escalation step	2 H	3 H	4 H	
Responsible for problem restoration	HD	Operator	Service Provider	
Information is provided for	Customer	Customer Operators	Customer Service Provider	
	Minor Problems			
Duration for the escalation step	4 H	8 H	18 H	
Responsible for problem restoration	HD	Operator	Service Provider	
Information is provided for	Customer	Customer Operators	Customer Service Provider	

Note: NCC = Network Control Center, HD = Help Desk, NP = Network Planning, NO = Network Operations.

- **Signing the contract and handling changes**: After signing the contract, it becomes legally valid. Due to rapid changes in communication technology, changes in the service contract cannot be avoided. Handling those changes should be included in the original contract.

The recommended standard SLA contract for SLA certification is detailed in the following section:

4.4.4.1 Template

The structure and content of SLAs should be unified, which should result in easier negotiations and more opportunities for control. The following items should be reviewed by both parties.

1. Parties to the agreement: All parties to the agreement should be listed, especially when there are multiple service providers and/or client groups. Each party defined as a contracting party should sign the contract. All parties must be documented; in particular, when multiple service providers, operators, and customers are signing the contract.

   It is unusual to include many parties in one single contract. The duration of negotiations and resolving disputes and misunderstandings would take too long. These facts represent arguments against complex contracts with many participants. Many bilateral contracts, however, require synchronization.
2. Terms of the agreement: The period of time that the agreement will be in place should be specified carefully. A typical length is from 3 to 5 years. There are multiple choices here. In case of strategic partnerships, the agreement may last more than 5 years. In the case of routine services, where back-out is easier, the duration can be less than 3 years.
3. Service included: Each service included in the agreement should be identified and described in detail. For each service, service-level metrics should be defined individually. The SLA should describe how the indicator is measured and who is responsible for performing the measurement. Further regulations should be included for emergencies, reaction time in case of outages, and escalation procedures to mutually inform relevant contracting parties.

   See contract item 18.

4. Optional services: All optional services that the service provider is willing to supply on request should be listed, in addition to those that are specifically part of the current agreement.

   In addition, the conditions, like provisioning and prices, must be detailed. Contracted and optional services may combined into bundles for pricing purposes.
5. Priorities of customers: In order to provide a SLA with multiple stages, clients are expected to set priorities for the following items:
   - Sites
   - Applications
   - Users
   - User groups

   These priorities may differ, particularly in the case of multinational enterprises.
6. Service offer: The service offer must be quantified. This contract item may be synchronized with contract item 5, where priorities are set for sites, applications, users, and user groups. The highest service level is $24 \times 7 \times 365$.
7. Metrics for services: All metrics to be considered are grouped in accordance with FCAPS (fault, configuration, accounting, performance, security). Examples are:
   - Fault metrics:
     - Number of outages by service
     - Availability by service
     - Network delay
     - Utilization of network and systems components
     - Mean Time between Failures (MTBF)
     - Mean Time to Incidents (MTTI)
     - Mean Time to Repair (MTTR)
     - Mean Time between Breakdowns (MTBB) and Mean Time of Diagnosis (MTOD)
     - Maximal duration of outages
     - Ratio: proactive and reactive problem detection
     - Ratio: referred problem and total number of problems
   - Performance metrics:
     - Utilization of managed objects
     - Throughput rates
     - Security metrics:
     - Number of security violations
8. Definition of outage: Include the exact definition of outages that do not impact the fulfillment of the SLA. In this category, include preventive maintenance and other client-ordered activities. The term *vis mayor* should be defined in writing.
9. Responsibilities of customers: Written confirmation that customers ensure physical access to managed objects and assist in outage detection and problem resolution. It is expected that security measures will be mutually agreed upon.
10. Reporting and reviews: Reports that are defined and created must be supported by the monitoring tools. The frequency of reporting, access to reports, and availability of real-time reporting, as well periodic reports should be defined. In many cases, Web access to these reports is required by customers. In order to become more dynamic, actual notifications, reports, and information are expected to be published on Web servers. Using "push" is a viable option.
11. Modifications: The process for changing the SLA, if necessary, should be defined along with the persons who are authorized to make changes. Rules are required for differentiating between corrections, rectifications, and changes.
12. Refinements: Technology may require refinement of the SLA and a redefinition of the commitment. For example, new equipment may be added, and the client may therefore have increased

expectations of performance. Extensions in writing are attached to valid SLAs; they are expected to be signed by all participating parties.

13. Tracking changes: Changes in the client organization—for example, an increase in size or acquisitions—can place unexpected traffic on the network, resulting in poorer response time. Introduction of new applications can also change QoS and the cost of delivering it. Changes must be documented, and parties must take into account the impact of these changes. If it is possible, a test should be run with the changed traffic profiles. Test results must be attached to the contract.
14. Eligibility: SLAs are legal documents. The following three questions should be answered up front:
    - Authorization for signatures
    - Representative persons
    - Authorization of changes
15. Nonperformance: An SLA also defines nonperformance, or what is to be done when the indicators do not meet the levels specified. However, some consideration has to be given to the amount of deviation. For example, instead of requesting a 2-second response time, it is more realistic to request a response time of 2 seconds for 90% of transactions, and 5 seconds for 99% of transactions.

    Actions against violating SLAs (IDG survey 2002):
    - Impose the built-in SLA penalty (41%)
    - Withheld payment for services (38%)
    - Changed providers (31%)
    - Renegotiated the SLA (26%)
    - Gave poor recommendations to others (21%)
    - Took no actions (12%)
    - Took legal actions (7%)
16. Agreement on tools: Determining what measurement techniques and tools are going to be used by contracting parties. Tools with standard interfaces for data collection, data processing, data storage, and for reporting are preferred. A mutual access (read only) to data is recommended.
17. Help Desk services: SLAs represent partnerships between contracting parties. Usually, service providers offer basic services, including:
    - What services are supported
    - Availability of the services

    The basic service is actually a hotline; subject matter experts handle client notification and inquiries. They are processed immediately or are referred to other experts. Availability is usually corresponds with contract item 6.
18. Escalation procedures: Contracting parties are expected to agree on the tasks, persons, priorities, and timeframes of the escalation. Severity tiers may help to streamline the escalation processes. Multiple tiers are recommended:

    Tier 1: No dial tone in voice networks, and no packet transfer are service incidents that require immediate attention.

    Tier 2: Latency spikes, which can affect user response time. Sustained latency of, for example, 250 ms or more for periods of longer than 1 minute may constitute another type of escalation procedure.

    Once tiers have been defined, customers should decide about the duration for which service providers are expected to trigger escalation steps and how quickly the customer may expect restored services.

    Names, job descriptions, and phone numbers for the individuals who will be directly responsible for repairs must be identified. Reach numbers of individuals who must be notified must be accurately identified.

    Each tier will have a different escalation schedule, depending on the terms of the SLA.

19. Communication between contracting parties: Selection of phone, fax, letter, e-mail, or a combination of all the above. In addition, authorization and authentication of the sender is required. The usual security agreements between contracting parties are in use.
20. Billing for services: This includes the price for services and discounts if objectives are not met. Service providers publish their price structures. Bundling opportunities in case of multiple services are strongly supported. Fixed prices or usage-based pricing are the alternatives. Conditions and prerequisites for discounts should be defined in writing.
21. Payment regulations: Collection alternatives are agreed to in this section. Basically there are two: flat rates for each month or usage-based payments. The manner of payment should be clarified. Alternatives are: paper bill, direct debit, bank card, credit card, or electronic payment.

4.4.4.2 Closing Comments

- Services outlined in the agreement must be realistic; in other words, the service provider can meet them.
- Measures for quality of service (QoS) should be accepted by both negotiating parties.
- The expected service level must measurable. If a service metric cannot be measured, it should be left out of the SLA.

4.4.4.3 Horizontal and Vertical SLAs

There are basically two alternatives for SLAs from the perspective of service providers and network operators:

4.4.4.3.1 Individual Service-Level Agreements

Customers (particularly large and important ones) may negotiate individually tailored SLAs with their providers. This is because they in turn have SLAs with their own customers (the end users), and there is a desire to make these SLAs line up. While this is certainly understandable, it is extremely difficult for service providers and network operators to keep track of the many SLA variants that might be developed or to respond effectively to problems when the terms of a service might vary widely from one customer to the next. Contract management and access to contracts in real time are extremely important.

4.4.4.3.2 Standardized Service-Level Agreements

Customers that are not large enough to demand tailored SLAs are subjected to multiple SLAs from multiple service providers and network operators. They see some measures, such as availability, that may sound the same but represent very different measures. They see multiple terms for the same thing. With no standard terms or definitions, they are left having to make the "translation" in order to produce some sort of measure of overall performance. Contract management and access to contracts in real time are extremely important. It is very beneficial when both parties are using the same tools to supervise SLAs.

In summary, SLAs are either

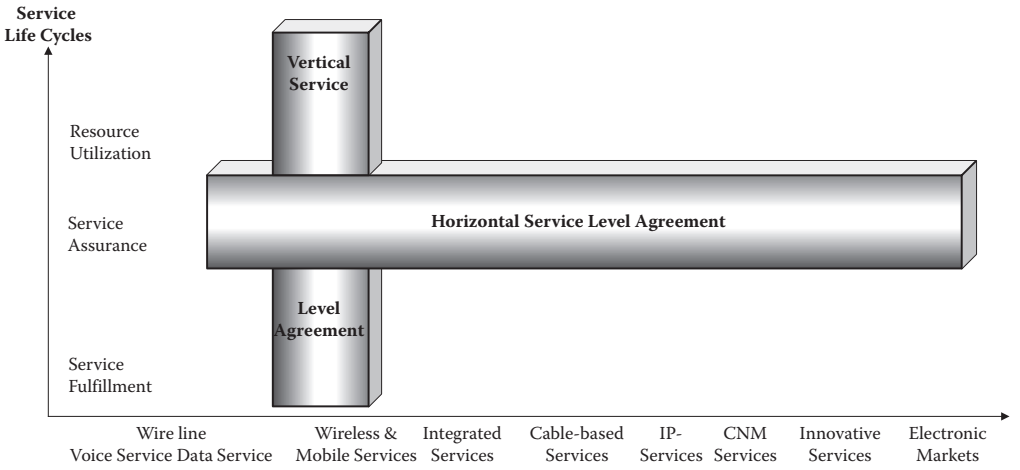
- Generic, robust, and simple (A)
- Specific, flexible, and complex (B)

There are different types of SLAs (see Table 4.4.7). In order to certify SLAs of service providers, both horizontal and vertical SLAs are under consideration. The axis of SLA dimensions are:

- Y-Axis: Service life cycles
  - Resource utilization and billing
  - Service assurance
  - Service fulfillment
- X-Axis: Service portfolios
  - Wireline voice services

TABl E 4.4.7 Types of Service-Level Agreements

Relationship	Type of Agreement
Service Provider–Operator	B
Service Provider–Service Provider	A or B
Service Provider–Customer	A or B
Operator–Customer	A
Operator–Operator	A



FIGu RE 4.4.5 Dimensions of Service-Level Agreements.

- Wireline data services
- Wireless and mobile services
- Integrated services
- Cable-based services
- IP services
- Customer Network Management services
- Innovative services
- Electronic market places

Figure 4.4.5 shows these dimensions.

4.4.4.3.3 Extensions of SLAs

Vertical SLAs are valid for individual services, usually including all components of life cycles. The result is that multiple metrics are being used. In addition to the typical metrics, such as availability, delay, error rates, other metrics, such as provisioning duration of services, security metrics, and billing parameters, must be considered.

Horizontal SLAs include multiple services, but usually just for one phase of the life cycle of the services. In most cases, service assurance is the target, with other areas coming later. In terms of metrics, service quality metrics are the first targets.

4.4.4.4 Traffic Classes

It is always important to unify QoS expectations in SLAs. SLAs may be signed for different traffic classes (also for application classes). Table 4.4.8 offers class types and their most important attributes.

TABl E 4.4.8 Traffic Classes and QoS Expectations

Traffic Classes	Bandwidth	Latency Sensitivity	Jitter Sensitivity	Packet Loss Sensitivity
Bulk data transfer	10 Mbps to 100 Mbps	Low	None	Low
Transaction data	Less than 1 Mbps	Moderate	None	None
Voice and fax	8 Kbps to 64 Kbps	High	High	Low
Multimedia (voice plus image)	Up to 384 Kbps for video	High	Moderate	Low
Video on demand and streaming	28.8 Kbps to 1.5 Mbps	Low	Low	Low

TABl E 4.4.9 QoS Techniques

Method	Benefits	Drawbacks
Bandwidth overprovisioning	Easiest to deploy Usually cheapest	Insufficient for voice
802.1p prioritization	Easy to deploy	Works only in LAN segments
Differentiated Services (DiffServ)	Works end to end in network	Requires supporting policy software
Add-on devices at LAN-WAN boundary	Easy to deploy	Limited granularity and flexibility
MPLS	Improves router efficiency	Complex; works only on routers
Resource Reservation Protocol (RSVP)	Guarantees priority bandwidth	Complex; mainly for backbone traffic

- Potential classes are:
  - Bulk data transfer
  - Transaction data
  - Voice and fax
  - Multimedia (voice and image)
  - Video on demand and streaming
- The attributes under consideration for the future are:
  - Bandwidth demand
  - Latency sensitivity
  - Jitter sensitivity
- Packet loss sensitivity

4.4.4.5 QoS Techniques

SLAs must be based on feasible and realistic techniques and technologies. This is the reason why existing QoS techniques play a significant role. This role must be emphasized for the process of the certification of service providers. Table 4.4.9 shows an overview of types of QoS control, including their benefits and disadvantages. These techniques include:

- Bandwidth overprovisioning
- 802.1p prioritization
- Differentiated services
- Add-on devices at the WAN/LAN boundary
- MPLS (Multiprotocol Label Switching)
- RSVP (Resource Reservation Protocol)

Any reference to the use of one of more of these techniques must be scored as positive.

SLAs are potentially a win-win situation. They include realism about service levels, relate price much more closely to service, and allow providers to differentiate and charge more by offering clear, guaranteed service that inspires confidence in users. Moreover, by looking at the world from the user's point of view, they force service providers to think creatively about how to achieve specific levels of service.

4.4.5 Certification of SLAs

4.4.5.1 Difficulties with Different Techniques and Tools

In a multiprovider environment, there are many procedures and tools for supervising the quality of Service-Level Management processes. The certification process helps to unify and simplify the implementation of processes and tools. Furthermore, the certification process helps determine whether processes and tools are capable of meeting expectations. The overall goal is to use as many standards as possible and for them to be as meaningful as possible to support operations. The certification process concentrates on the basics, but permits the implementation of multiple options. These options relate to the improvement of existing processes and tools through innovation.

4.4.5.2 Attributes of a Professional Solution

Open and standardized interfaces are required for the integration of measurement processes and tools into the existing OSS/BSS (Operations Support Systems/Business Support Systems) environment. Scalability is also extremely important in order to manage growth. A very flexible reporting system can help service providers to differentiate themselves from their competitors. These reporting systems are helpful in meeting the requirements of new technology, suppliers, services, and customer needs. Some of the reports address future needs as well. Change is constant, and SLAs are the key to success.

It is unfortunate when clients presume that they are not getting the right service for the right price. But focusing simply on revenue is a short-sighted view. Open-minded views of service quality are more important. These views allow the service provider to optimize network performance and meet customer needs based upon existing network infrastructures. Both short-range and long-range observations are important; modeling and traffic simulation help to predict the performance of future network configurations.

Service providers are expected to understand their networks very well. Only those who understand are able to utilize the infrastructure to offer high-quality services and achieve higher revenues. When weaknesses are known, resources may be optimized to fill gaps and avoid bottlenecks. This may be done without significant new investments in hardware and software. Return on investment (ROI) is elementary for each service provider in these days. Service-Level Management solutions are gaining momentum, and they help to achieve secure ROI for each service provider.

SLAs may vary greatly. Not everyone is interested in getting very detailed performance reports. Management needs overview-type reports for their decision-making processes for investments. Marketing and sales need reports about quality levels of new services or customer-level service quality. SLA metrics may also be used successfully in billing and for control purposes.

One of the key certification criteria for SLAs and performance management are client-specific content and a simple presentation format. Reports must be easily understandable for them to be used for service improvements. Customer satisfaction is the key to success. Overseer must certify SLAS on the basis of customer care and satisfaction.

New services will be accepted and implemented when quality requirements are met by using mutually accepted metrics.

4.4.5.3 Measurement Procedures

In all cases, a number of questions should by answered by the contracting parties at the very beginning:

- Art of information collection
- Time of measurements
- Overhead due to measurements
- Compression hierarchy of measurements
- Art of information presentation

- Art of information interpretation
- Archiving information
- Combinations with modeling tools
- Embedding measurement data into reporting
- Requirements for customization
- Types of customization
- Resource requirements of first installation
- Resource requirements of operations

In case of separate measurements, network domains (actually the geographical coverage of service providers) usually are separated from each other. Management stations are expected to exchange information with each other. Reports are generated and distributed on a provider basis. When Web-based solutions are in use, reports by different providers may be prepared separately and presented together. The authorized client can access information and reports on the Web server.

In case of collaborative measurements, management stations are eligible to mutually access measurement data. This access should be negotiated between the service providers whether on a permanent basis or only in the case of an outage. Management stations exchange information with each other and offer backup to each other. Reports are generated together and usually distributed via the Web. Maintenance of datamarts or data warehouses is also supported as a joint venture.

Measurements are grouped as follows:

- Active measurements against a measurement point
- Active measurements against the server
- Passive measurements over a TCP measurement point
- Passive measurement over a mirror port
- Passive measurements over an application measurement point

Additional subalternatives may be defined depending on the measurement tools in use.

The certification includes both QoS techniques and SLA control procedures. Both consist of policies and tools. The nature and goals of both are slightly different.

4.4.5.4 QoS Procedures

These procedures are policy determination and parameter setups that are valid for an existing infrastructure and for an existing service portfolio in order to reach a specific service quality.

The following alternatives are recommended for certification:

- Admission control: Determines whether a requested "connection" is allowed to be carried by the network. The main considerations behind this decision are existing traffic load, existing QoS, requested traffic profile, requested QoS, pricing, and other policy considerations. Admission control could be performed in the setup of RSVP flows or MPLS paths.
- Traffic shaping: It is necessary to specify the traffic profile for a connection to decide how to allocate various network resources. Traffic shaping ensures that traffic entering at an edge of a core node adheres to the profile specified. Typically, this mechanism is used to reduce the burstiness of a traffic stream. This involves a key trade-off between the benefits of shaping, such as loss in downstream network, and the shaping delay.
- Packet classification: Packets must be classified to allow for different QoS treatment using various fields in IP headers (source/destination addresses, protocol type) and higher-layer protocol headers (source/destination port numbers for TCP or User Datagram Protocol [UDP]). Efficient and consistent packet classification is a significant problem under active research.
- Priority and scheduling mechanisms: To satisfy the QoS needs of different connections, nodes need to be prioritized and scheduled. Priority provides different delay treatment, with higher-

priority packets always being served before lower-priority packets, both for packet processing and transmission on outbound links. Scheduling mechanisms ensure that different connections obtain their promised share of the resources, such as processing and link bandwidth. This mechanism also ensures that any spare capacity is distributed in a fair manner.

- Signaling protocols: To obtain the required QoS from a network, systems need to signal the network regarding the desired QoS as well as the anticipated offered traffic profile. Scalability and the corresponding capabilities to signal different QoS needs are issues under current examination.
- Queuing: Some network elements allow for fair queuing algorithms. This ensures that a misbehaving application will not punish other, better-behaved applications, and that the average of dropped packets is evenly distributed across flows and queues.

4.4.5.5 SLA Procedures

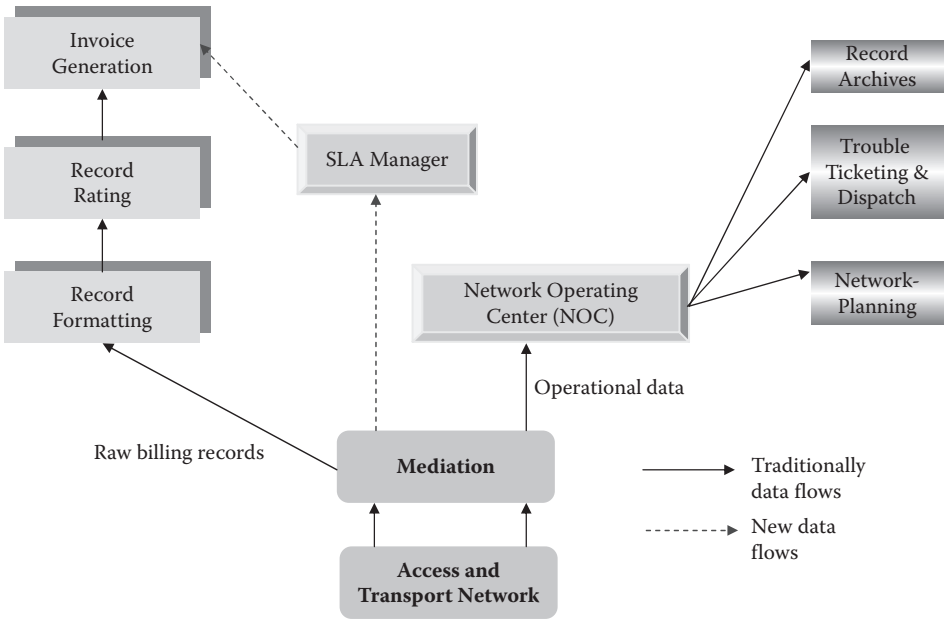
This area involves policy determinations for controlling service quality. This quality has been defined and agreed upon in the SLA between the contracting partners. The following procedure steps are recommended for certification.

Definition of data sources: The following options are considered as data sources: Managed Objects (MO), Element Management Systems, traps, and SLA probes. In certain cases, manual logs and notifications from the help desk and hotline may be considered.

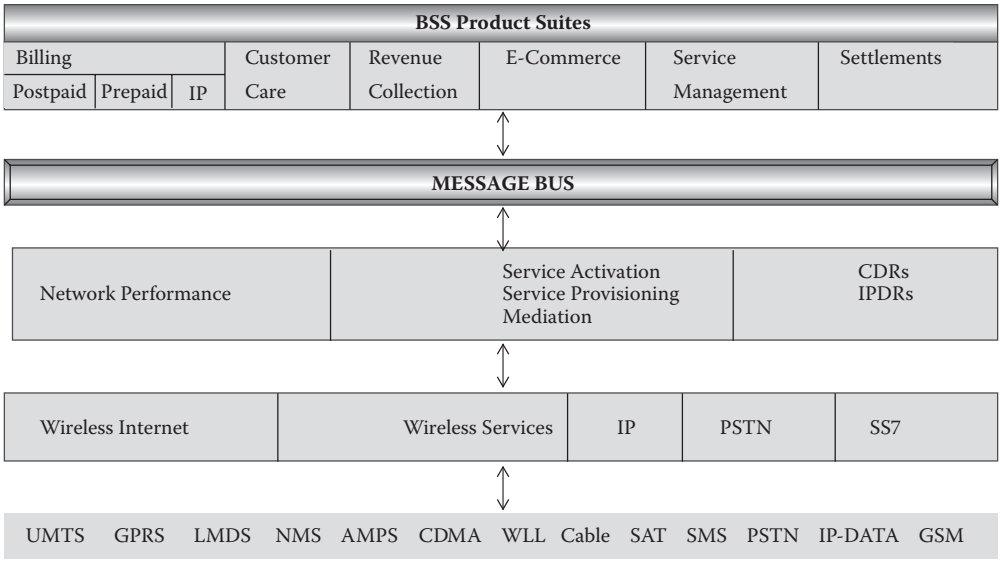
Determination of raw data to be collected: MIB-II entries using SNMPv1 and SNMPv3, and the remote monitoring (RMON)1 and RMON2 standards.

Determination of the mediation functions: The current focus of the telecommunications industry on SLA and QoS as a method of differentiating service providers creates a completely new mediation challenge: nonbilling transactions have the potential, when a series of conditions are met, to trigger a billing action.

Figure 4.4.6 shows an example of the impact of the new data flows generated by the addition of SLA agreements and an SLA manager that provides billing adjustment information to the billing system. The same is true with settlements among multiple providers. It becomes even more challenging when the information that triggers the billing action originates on a network element that belongs to another



FIGu RE 4.4.6 Impact of SLA on data flows.



FIGu RE 4.4.7 Multiservice mediation platform.

service provider. In order for this to occur, the mediation platform must contain rules that trigger these after-the-fact transactions. The concept of a mediation platform as opposed to mediation systems becomes critical in the evolution of OSSs for new service portfolios.

Individual mediation systems targeted on specific services or network elements can support a high volume of transactions and the associated data processing activities. However, they still cannot perform the correlation activities that are required to enable automated SLA and QoS programs without additional systems in the OSS.

The convergence of multiple services—voice, data, and video—is driving the need for new mediation tools as it increases the importance of mediation with Operations Support Systems. New mediation requirements are created with each new service or network element that is added to the product portfolio. Multiservice mediation highlights the complexities facing service providers when they evolve their networks and OSSs to meet the challenges of new infrastructures and technologies.

Information that drives transactions in the OSS can potentially originate anywhere in the network. In addition, these transactions can require multiple copies of a single record to be simultaneously transmitted for various types of billing transactions. These transactions include traditional billing, IP content, m-commerce, and other types of billing not even considered yet.

Figure 4.4.7 shows a new view of the architecture of mediation. Key points to note in this architecture are the increased importance in the mediation platform and the data bus, which enables the transfer of information among the various components. Due to the potentially dynamic nature of trading partners, the mediation layer becomes the logical point in the OSS to store and exercise business rules. It also becomes the logical point for accounting for all of the transactions that occur because all transactions will have had some form of treatment in this layer. This functionality is not possible without an integrated multiservice mediation platform.

Mediation systems can also drive downstream events, such as product and service delivery. They must be capable of recognizing and acting upon transactions that drive multiple events and provide records to each operation or trading partner that needs this information.

Determination of metrics that are expected to be supervised:

- Availability (multiple TMN layers)
- Delays and latencies (network and element layers)

- Packet loss (network and element layers)
- Throughput (network and element layers)

Determination of views: All views are expected to be prepared in accordance with TMN layers. The result is a hierarchy of the service layer, network layer, element management layer, and network element layer.

Determination of reports: All reports are expected to be prepared in accordance with TMN layers. The result is a hierarchy of the service layer, network layer, element management layer, and network element layer.

It is assumed that basic data are collected by unified collection techniques. Processing data should follow the same unified guidelines. The minimal requirement permits accessing all data on a Web server. Service providers and clients may access these data, and if required, they can generate reports on their own.

Just a minimal set of standard reports are targeted. These standard reports are in accordance with the reports agreed upon in the SLAs. These report on the following metrics:

- Availability (multiple TMN layers)
- Delays and latencies (network and element layers)
- Packet loss (network and element layers)
- Throughput (network and element layers)

4.4.6 Role of SLAs in Settlements between Service Providers

4.4.6.1 Present Difficulties

When customers are able to control the performance of service providers, their trust level toward the service providers will increase. The same is true with service providers when they conduct business with each other. Customer churn cannot be quantified easily, but the negative impact is obvious for business processes and references.

Some of the difficulties may be summarized as follows:

- How is availability defined? If it is the average downtime per month, the supplier could still meet the terms of the agreement for the month as a whole, despite one very critical day, by exceeding the terms on all other days.
- Complex SLAs may require very time-consuming negotiations between technicians and layers. SLAs should be as short as reasonable, and they should be off-the-shelf contracts, except in cases where very large, customized deals are involved.
- Guaranteed SLAs over several service providers are very difficult to achieve. So-called back-to-back SLAs, which create a kind of service level chain, would be the solution. But these are very hard to negotiate, especially across borders in Europe, where different jurisdictions prevail.
- SLAs must be lined up with special SLAs of the customers, since many companies offer agreed service levels to their own customers.

4.4.6.2 Carriers are Fighting for Stronger SLAs

Service providers are very much interested in their service portfolios. They use best-of-breed tools to differentiate themselves from their competitors. They occasionally publish performance metrics about the segments of the networks they support. Table 4.4.10 shows metrics that may also be included in SLAs. These metrics are very important for multinational enterprises because they represent performance guidelines in the global network. They help to set realistic goals. All values in this table may be accepted.

The lower-level infrastructure and the role of subcontractors are not further detailed.

TABl E 4.4.10 Performance Metrics of International Carriers

Carrier	Cable and Wireless	AT&T	UUNet
Intra-U.S.	Less than 55 ms	Less than 60 ms	Less than 65 ms
Intra-Europe	Less than 60 ms	Less than 65 ms	Less than 65 ms
Trans-Atlantic	Less than 100 ms	Less than 120 ms	Less than 120 ms
Trans-Pacific	Less than 130 ms	Less than 130 ms	Less than 150 ms

Note: Informal examples.

4.4.6.3 Assistance of SLAs in Settlements

SLAs can support a fair and mutually acceptable settlement between service providers by using accurate measurement data. Settlements are business agreements between service providers that may support the same clients with their service portfolios.

All revenues generated for certain clients should be subdivided between service providers following mutually agreed rules. Criteria for these rules are:

- Accuracy: Settlements should be based on measured data; the level of detail should be reasonable.
- Simplicity: Settlements must be simple and easy to implement and call for the use of very simple metrics.
- Periodicity: The timeframes of settlements should be agreed upon. They could be very different. Common sense should dictate the decision. Experiences show timeframes from one week to three months.

The first and second items are slightly contradictory, and compromises are required.

SLAs are signed to help clients. Clients are usually interested in signing end-to-end SLAs, but in such cases, multiple service providers may be included. Multiprovider models are addressed in other APs. Basically, we should consider:

- Bilateral SLAs between clients and service providers: In this case, SLAs are rarely used for settlements because they have been signed for individual use. Settlements are important, but they are not impacted by the SLAs.
- The client and the main service provider sign an SLA: There are multiple SLAs between the main service provider and other service providers (subcontractors). In this case, SLAs have a great impact on settlements because the business relationships are built hierarchically. An SLA violation may trigger a chain reaction of events.

Metrics, agreed upon in SLAs, may be used for the settlement (mutual billing) between service providers. The most important metrics are:

- Throughput: Transferred bytes, messages, and transactions in both directions
- Trouble statistics that are the basis for availability calculations
- Performance metrics that determine the service quality
- Security metrics that determine the level of protection in the networking segments of service providers

The supervision of SLAs is supported by measurements accomplished by standardized and certified measurement tools. They are able to provide the measurement results for throughput rates, error rates, response time, delays, latencies, and eventually security violations. When all or some of these metrics may be utilized for settlements between service providers, savings may be accomplished in the area of data collection tools and of generating special settlement-oriented reports.

Settlement between service providers are functioning as follows:

- Collection of data on traffic volumes at the peering point between service providers, considering both directions.
- The generated billing records, that have been derived from the traffic measurements, are presented to the partners using a paper-based or electronic invoice.
- These invoices originate from both parties; it makes sense to balance invoices and bill for the difference only. For this purpose, electronic bill presentment and payment (EBPP) is the ideal solution.
- When other metrics of SLAs are generated and distributed that require exceptional handling, they will be considered during settlements. Violating SLAs will have financial consequences.

In order to better support all the processes addressed above, Web-based services are recommended for implementation. As part of EBPP, it is recommended that the parties agree upon the access to views, results, and partial results.

In summary, we can conclude the following:

- SLAs can make the settlement process between service providers more objective.
- SLAs offer basic data that may be utilized during the settlement process.
- SLAs report on exceptions that might have an impact on settlements.
- Connecting SLAs and settlements are important, and must be addressed during the certification process.

There are very few examples; service providers are not willing to publish settlement rules. The overseer of the multinational enterprises will get insight into those settlement agreements as part of the certification process.

4.4.6.4 Tasks of the Overseer

In addition to the well-known tasks of the overseer, the following tasks are added due to the need for SLA certifications.

- Decision about the SLA contract template for SLAs: The overseer reviews the sample contract with its 21 contract items and executes changes if necessary. The overseer decides what contract items are mandatory for the certification process.
- Determination of the acceptable Service-Level Management processes: The overseer reviews the SLM process (or SLM processes) and executes changes if necessary. The overseer determines what changes are tolerable from the core process.
- Determination of active and passive measurement points: The overseer evaluates the recommendations for active and passive measurement points on the basis of the networking environment, existing and planned tools, and on the basis of the skill levels of subject matter experts.
- Determination of the administration boundaries, when multiple providers are serving the same client: The overseer evaluates the peering alternatives among multiple service providers. The overseer determines the administration boundaries on the basis of the configuration of peering points, existing and planned management tools, and on the basis of skill levels of subject matter experts.
- Evaluating additional certification criteria for suppliers of QoS and SLA tools: The overseer checks all certification criteria and occasionally executes changes.
- Certification of suppliers: The overseer certifies suppliers of management, administration, and measurement tools on the basis of technological capabilities and financial stability.
- Evaluation of the results of unifying service classes: The overseer evaluates all recommendations for the unification of service classes submitted by service providers. On the basis of these submissions, service providers will be certified. In cases where service providers are not willing to engage in unification, certificates may be revoked.
- Arbitration in case of settlement problems: The overseer executes his/her function as arbitrator when service providers—suppliers of the multinational enterprises—are unable to reach a settle-

ment among themselves. The overseer makes decisions based on available billing records that have been collected in or around the peering point.
- Determination of penalties, if required: The overseer determines the magnitude of penalties by selecting from the alternatives, agreed upon in SLAs. The overseer follows the best interests of the multinational enterprises by deciding about the right compensation measures.
- Arbitration in case of SLA problems: The overseer executes his/her function as arbitrator when clients and service providers are unable to reach an agreement about SLA problems. The overseer makes decisions based on the measurement data submitted by both parties.
- Clarification of mutual access rights to measurement data: The overseer determines who is entitled to access measurement data—read or write or both. These rights could be included in the SLAs and signed by all participating entities.

4.4.7 Summary and Trends

One of the big obstacles in implementing a Service-Level Agreement is the inability to create a solid baseline of data to quantify the historical performance of the service provider with various customers. Without a complete body of historical data, it is difficult to create a meaningful and realistic Service-Level Agreement, and expectations may be set unrealistically high. Not all tools will capture all of the components of the network, and the metrics will be incomplete. SLAs of the past have failed due to their lack of accurate measurement data. Data were manually recorded and were often unreliable. New monitoring tools have increased the quality and quantity of data available for SLA evaluation.

In order to reduce the load due to a large volume of collected data, sampling is gaining in importance. Accuracy is acceptable for both parties and overhead can be kept to a reasonable minimum. Most of the successful tools work with this principle. Due to mergers and acquisitions, the vendors who provide tools for SLAs are continuously changing. The big players, such as Hewlett-Packard, IBM, Computer Associates, and BMC remain serious solution providers. Also smaller vendors may play a role; even open source solutions could become significant.

High-tech measurements and documented performance of the service provider and operator are extremely important. The emphasis should be on:

- Efficiency of escalation procedures
- Offering reactive and proactive capabilities
- Speed of component repair and replacement
- Willingness of rectifying problems by corrective actions
- Jointly analyzing root causes of problems
- Unifying and simplifying communication paths
- Taking advantage of Web and intranet capabilities for information exchange

As competition is progressively introduced into all service provision markets and customers become increasingly discerning, service providers are realizing the need to differentiate their products through value-added services. Additionally, industry deregulation is transforming a traditionally monopolistic marketplace into an extremely competitive one. One prime mechanism to combat this is the provision of an off-the-shelf SLA document that clearly sets out the obligations of both the service providers and the customer.

In summary, critical success factors for successful SLAs are:

- Clear understanding of QoS metrics
- Powerful data collection and correlation capabilities
- Support of both real-time and historical reporting
- Low overhead of tools for the network, systems, and applications

Acronyms

ABR	Available Bit Rate
AIB	Archived Information Base
API	Application Programming Interface
ARM	Application Response-Time Measurement
CBR	Constant Bit Rate
CGI	Common Gateway Interface
CIR	Committed Information Rate
CRC	Cyclic Redundancy Check
DNS	Domain Name Service
DSL	Digital Subscriber Line
FTP	File Transfer Protocol
GUI	Graphical User Interface
HD	Help Desk
HIB	Historical Information Base
HTTP	Hypertext Markup Language
IP	Internet Protocol
ISP	Internet Service Provider
LAN	Local Area Network
MIB	Management Information Base
MTBB	Mean Time between Breakdowns
MTBF	Mean Time between Failures
MTOD	Mean Time of Diagnosis
MTOR	Mean Time of Repair
MTTI	Mean Time to Incidents
MTTR	Mean Time to Repair
NCC	Network Control Center
NNM	Network Node Manager
NO	Network Operator
NP	Network Planning
PVC	Permanent Virtual Circuit
QoS	Quality of Service
RMON	Remote MONitoring
SA	Service Availability
SAP	Service Access Point
SIB	Summarized Information Base
SLA	Service-Level Agreement
SLM	Service-Level Management
SMIB	Service Management Information Base
SQL	Structured Query Language
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
TNM	Telecommunication Network Management
UBR	Unspecified Bit Rate
URL	Universal Resource Locator
VBR	Variable Bit Rate
WAN	Wide Area Network
XML	Extensible Markup Language

4.5 Management Services and Outsourcing

Kornel Terplan and Christian Voigt

4.5.1 Introduction

As the marketplace around managed services continues to evolve and mature, managed service providers (MSP) will need a more open approach to delivering managed services.

Part of the disconnection problem between providers and customers today involves the definition of managed services. Providers think that if they provide statistics on circuit and device uptime and availability, along with the ability of troubleshoot and reconfigure malfunctioning components, they are fully managing the service. Customers have a broader definition: a managed service is one that ensures the critical applications are performing effectively. Moreover, they expect providers to detect issues that compromise business functionality and to take immediate action, and not wait to be informed. These requirements assume that providers are able to correlate application dependency from infrastructure components.

But another part of the problem also lies with IT executives that are unable to manage outsourcers. Using managed services is a form of outsourcing. IT departments of enterprises have become real business savvy during the last decades. They are now able to align business with IT infrastructures. In other words, they are dedicated to their business units. In the case of outsourcing, they expect the same level of dedication from the provider. But it does not happen. Outsourcers need to pay attention to their bottom lines with the result that the interests of outsourcers and IT organizations are not necessarily aligned.

This section will focus on governance first. It is important to define the structure of the enterprise that will significantly impact the relationship with providers. In most cases, multiple providers must collaborate in delivering the managed service to customers. Collaboration and settlements are key priorities between providers and between providers and customers. This section also gives criteria and guidelines for outsourcing decisions considering benefits and disadvantages of an outsourcing decision. Further details about near-shore and off-shore alternatives will not be addressed. In any case, contract management is the key. Components of contract management will be discussed in depth.

4.5.2 Policies and Tasks of Governance

4.5.2.1 Governance Models

IT governance can be defined as the decision rights and accountabilities that encourage desirable behavior in the use of IT.

Research continues to highlight the frequent coexistence of IT organization (ITO) subcultures and governance patterns that often diverge (or even contradict) the prescribed governance model of the merged (or trimmed) enterprise following Mergers and Acquisitions (M&A) transactions. This frequent occurrence of "parallel" or exceptional ITO governance patterns is usually explained (but of course not justified) by the historical evolution of the IT function as a special entity with knowledge and skills deemed arcane by (often technophobic) business people, singular labor market conditions (e.g., the gap between IT and non-IT salaries and retention schemes), and an enduring engineering culture that values deterministic planning and technology. Optimization is key and self-sufficient decision criteria. Furthermore, limited financial modeling and communication skills within ITOs usually hinder effective negotiation with other key enterprise functions. Analysts believe *Fortune* 2000 CIO (chief information officer) leadership will be further tested in carve-out merger acquisition (CMA) contexts on their capacity to analyze and diagnose governance divergence during the early phases of CMA negotiations and drive firm corrective action during the analysis, transition, and integration phases.

As a way of assisting CIOs in establishing a proactive diagnosis and assessment of the level of business/IT governance alignment, we recommend breaking down governance model practices into one or more of the following fundamental governance patterns:

- "Hard" management
- Centralizing models
- Centrifugal models
- Loosely coupled organizations
- Laissez-faire approaches

Although each enterprise governance model ultimately will be unique, analysts believe these models can be decomposed into a combination of the patterns listed above.

4.5.2.1.1 "Hard" Management Governance Pattern

Through 2010/2012, the global economy will follow a more irresolute and longer path to recovery than initially anticipated. During this period, *Fortune* 2000 enterprises will continue to embark on merger, acquisition, and divestiture operations as fundamental strategies for survival through economies of scale, risk distribution, and financial strength in a volatile global economy. Analysts believe CMA activity will be as strong in economic contraction cycles as it will be during growth periods. Indeed, while CMA action will tend to accelerate during economic growth periods as an instrument for swift corporate expansion, CMA activity in less favorable periods will center on one of the following:

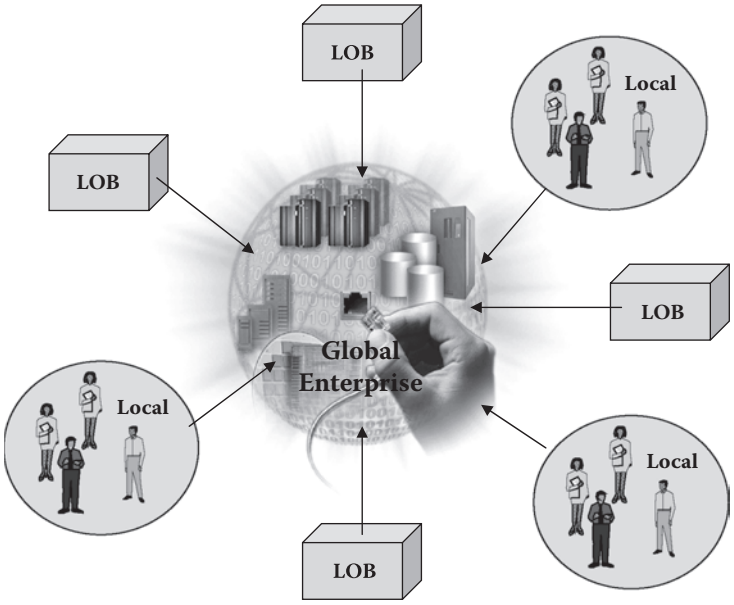
- **Protective CMA maneuvers:** In this context, struggling or challenged organizations join forces to improve economies of scale, market synergies, financial strength, and global reach, and to be in a better position to face stronger competitors (e.g., Telia's purchase of Sonera in a challenging European Telco market, Allianz-Dresdner, HP's acquisition of Compaq).
- **Consolidating CMA movements:** In these cases, stronger and bigger players in specific markets leverage their size and force to take over or outdistance weaker competitors. Usually the parent company (i.e., the acquirer) maintains an overwhelming (or at least preponderant) financial and operational weight in the resulting entity (e.g., IBM's acquisitions of PwC, Rational Software, CrossWorlds, and Informix's database business; Wal-Mart's relentless expansion in the retail sector).

Analysts expect protective and consolidating CMA maneuvers to call for extremely strong governance models, with highly centralized management processes as a prerequisite to attain the established CMA goals. The prevailing governance patterns in such settings will be characterized by an uncompromising drive to rapidly unify operations, establish a common enterprise process, and efface "local" or individual specificities.

In hard management (e.g., monarchic, colonial) governance patterns, there is little tolerance for local operational specificities or cultural idiosyncrasies across geographies (e.g., national and regional subsidiaries) or lines of business (LOBs). Such peculiarities are perceived as impediments to the ultimate CMA goals of consistency, economies of scale and time, global flexibility, and productivity. Therefore, in hard management governance models, local geography and LOB autonomy will be associated with inefficiency and ultimate failure.

The hard management governance pattern describes an organizational culture that will "instinctively" prefer and advocate common, strictly defined business processes and metrics across a large organization. In this pattern, the consensual general wisdom among enterprise executives is that centralization will always be the best approach, even when the context is fuzzy and the information available is incomplete. Therefore, this governance pattern describes not just a purely Cartesian, rational, metrics-based decision framework for CMA implementation, but also a model of sociopsychological management behavior in such circumstances.

From an IT perspective, this governance pattern will typically translate into a strong drive to adopt common corporate applications, e.g., an enterprise resource planning (ERP) backbone along with



FIGu RE 4.5.1 Hard management CMA governance pattern.

customer relationship management (CRM), commerce chain management (CCM), e-business, and business intelligence (BI) packages. Standard applications will be embraced as embodiments of best practice operations and business processes, and as tools to impart consistency and discipline across a multifaceted enterprise.

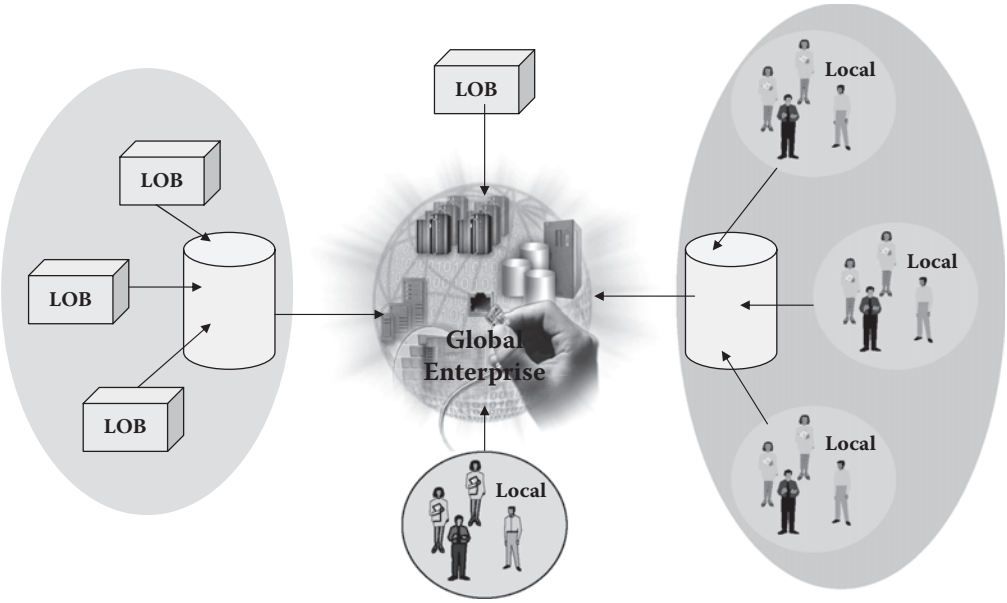
Although geographical, cultural, legislative, and business practice specificities across national and LOB boundaries will be recognized as important, management will share a strong conviction that these can be minimized through time. Persistent local/LOB specificities will be handled through individual configuration of corporate standard applications with strong central procurement, deployment, and (eventually) management and hosting. Furthermore, hard management governance patterns will favor aggressive infrastructure consolidation projects (e.g., data center, storage, and server consolidation) and global sourcing and service provider deals (see Figure 4.5.1).

In Figure 4.5.1, business perspectives are:

- Common global processes
- Rapid elimination of local/LOB specificities
- Consistency
- Economies of scale
- Discipline

Typical IT mandates are:

- Common enterprise application backbone (e.g., ERP, CRM, PRM (product resource management), SCM (supply chain management), BI)
- Common enterprise software infrastructure (e.g., portals, workgroup, application servers)
- Infrastructure consolidation (data center, storage, servers)
- Global sourcing and service provider deals
- Support for centrally controlled Web 2.0 solutions



FIGu RE 4.5.2 Centralizing CMA governance pattern.

4.5.2.1.2 Centralizing Governance Pattern

Analysts expect the hard management governance pattern described above to remain popular among enterprise executives, financial analysts, business consultants, and investors in formal communications concerning strategic directions and vision statements through 2008/2009. Nevertheless, and despite the most earnest and determined management intentions, analysts believe it will rarely be completely achieved in a short period of time (less than three years) following CMA operations. Indeed, certain cultural, operational, and technical differences will linger across global geographies and LOBs for longer periods of time (i.e., more than four to five years), or even permanently in post-CMA *Fortune* 2000 organizations.

The rapid execution of hard management patterns in CMA operations will often be hindered by the culture and structure of the individual historical entities, particularly in "defensive" mergers involving companies of comparable size but with significantly different operations and processes. In certain cases, geographical distinctiveness (e.g., national and regional laws, labor regulations, local supply/distribution networks) will present lingering impediments for a comprehensive consolidation of business processes and operations as prescribed in the hard management pattern. Furthermore, CMA operations will seldom be one-off occurrences. Indeed, such maneuvers are likely to take place several times in the life of *Fortune* 2000 Organization, requiring constant revisiting of otherwise successful implementations of hard management governance patterns.

Therefore, enterprises with an otherwise strong affinity for the hard management governance approach will find it difficult (or even impossible) to achieve it completely within the expected CMA execution time, forcing enterprise executives to compromise with enduring local and LOB autonomic processes. The management desire for process consistency and economies of scale and time will be limited by local specificities. Still, enterprise business executives retain a strong belief in the definitive benefits of the hard management approach and will therefore tolerate local/LOB process specificities only as temporary compromises, which will tend to diminish as CMA operations follow their due course and organizations mature accordingly. Thus, the centralizing governance pattern (see Figure 4.5.2) tends toward hard centralization of enterprise processes and operations, but with consent to certain levels of autonomy as a "necessary evil" and as a tactical approach with a limited planned life span (i.e., less than four to five years).

The general wisdom in centralized (e.g., federal, territorial) governance pattern is that anything that can be done centrally *should* be done centrally. So, with uncertain or incomplete information, enterprise executives will tend to back centralization, consolidation, and shared global business processes as a default position. Local/regional or Line of Business (LOB)-specific processes and systems will still be perceived as sources of inefficiencies and will be tolerated only as exceptions.

From an IT perspective, the centralizing governance pattern will usually focus on partial (from a geography or LOB perspective) consolidation of process, application, and infrastructures. Characteristic projects include the adoption of common enterprise application backbones and consolidation of data centers within various regional hubs (e.g., the Americas, Europe/Middle East/Africa, Asia Pacific) or specific LOBs (e.g., in a financial service enterprise, centralization is pursued within and not across private banking and retail banking divisions).

Despite the compromise with dissimilar regional hubs or LOB-centric consolidation initiatives, certain enterprise processes will be implemented on a cross-region/cross-LOB global scale (e.g., procurement, networking infrastructure, global sourcing and licensing contracts, identity and security infrastructure). Management will determine which specific processes/operations must be implemented globally, with some local exceptions being tolerated (e.g., regional service providers, local suppliers offering better and more competitive procurement possibilities). Characteristic of this pattern is that central and regional approaches are perceived as exceptions to the default central/global approach. Therefore, the "burden of proof" rests on the local/regional advocates, who must justify why such exceptions to the general corporate wisdom are necessary.

In Figure 4.5.2, business perspectives are:

- Common business processes within regions (e.g., the Americas, Europe, Middle East, Africa [EMEA], Asia Pacific)
- Tolerance of local/LOB specificities through justified exceptions
- Certain processes must be global

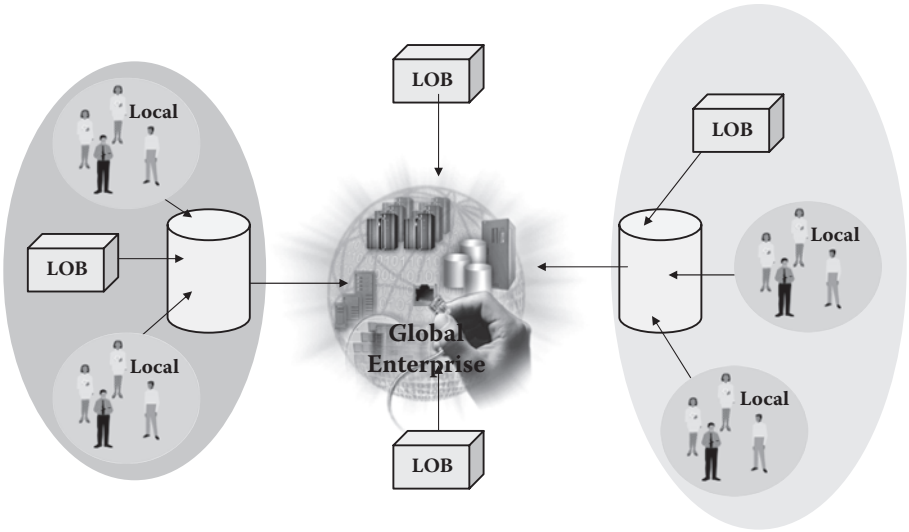
Typical IT mandates are:

- Regional enterprise application backbones (e.g., ERP, CRM, PRM, SCM, BI)
- Regional data center consolidation
- Global services (sourcing, licensing, procurement, communications, identity and security)

4.5.2.1.3 Centrifugal Governance Pattern

Despite the current popularity of hard management and centralizing governance patterns being perceived as more efficient in terms of cost reduction, analysts believe more decentralizing governance will regain interest in both the medium term (two to three years) and the long term (four to six years). Indeed, during the next six to seven years, leading *Fortune* 2000 executive management will seek competitive advantage through flexibility and will increasingly recognize the pragmatism and business value of more loosely coupled cross-enterprise integration models as a way to cope with constantly changing market conditions. This will translate into noninvasive CMA models, where integration of local or lower granularity processes is embraced as a strategic asset for the corporation and investment in enabling skills and technologies is a strategic corporate goal. Analysts identify these as "centrifugal" governance patterns (see Figure 4.5.3). By centrifugal, analysts mean that the center of gravity remains in the periphery (i.e., the local LOB or geography) level, while the global enterprise maintains a balance to avoid inefficiency and waste.

In the centrifugal governance pattern (e.g., decentralized, city-county), the individual entities involved in CMA operations will seek enhanced value as a whole without necessarily imposing a strict homogeneity of processes, applications, and infrastructures. The corporate CMA goals of achieving economies on a space scale (e.g., breadth of product/service offerings, geographical reach, shared/coordinated production and procurement, virtual inventories) as well as economies of time (e.g., just-in-time manufacturing, time-sensitive production/marketing/selling) and consistency (e.g., global branding,



FIGu RE 4.5.3 Centrifugal CMA governance pattern.

coherent service/product catalogs) will be achieved through harmonization (not homogenization) of local processes and supporting applications and infrastructures.

The general rule in centrifugal governance patterns will be that anything that *can* be done effectively on a local level *should* be done locally unless it can be established that a centralization or deep homogenization can benefit the corporation. Centralized processes will prove beneficial, but they will not be considered as superior by default to local (or lower granularity) processes (i.e., the burden of proof is on centralizing forces).

From an IT infrastructure perspective, achieving this goal in practice will require an earnest investment in cross-organizational integration infrastructures, e.g., enterprise application integration (EAI), interenterprise integration (IEI) with solid business process automation (BPA) capabilities. Requirements will extend beyond the mere acquisition of EAI/IEI/BPA toolkit packages to seeking the vendors' promise of a quick fix to existing application and business process silos among CMA enterprises.

CIOs should pay attention to the proactive development of planning and development skills in integration infrastructures. An obvious need for technical proficiency (e.g., networking, middleware, applications) must be complemented with nontechnical competencies in financial modeling and business analysis interpersonal communication. Indeed, the most important challenge in strategic integration infrastructures will be the planning phases, where IT organizations must be able to match business requirements with nonstandard service-level metrics (e.g., flexibility of processes in time, space, and scope; ease of integration of existing and future processes, applications and information, sourcing options) and matching those with appropriate infrastructure development plans.

In Figure 4.5.3, business perspectives are:

- Corporate goals of achieving economies of scale/time and consistency will be achieved through harmonization (not homogenization) of local processes
- Local autonomy means leanness and agility
- Processes are local (LOB or geographical) by default, unless corporate processes are proven necessary or more efficient

Typical IT mandates are:

- Toleration of local application backbones (e.g., CRM, SCM, BI, production)
- Support of certain processes by corporate apps (regional ERP)

- Significant investment in business process automation and application integration to establish enterprise processes
- Emphasis on process modeling
- Gradual reduction of unnecessary variety in applications and data centers
- Telecommunications infrastructure as corporate and global service

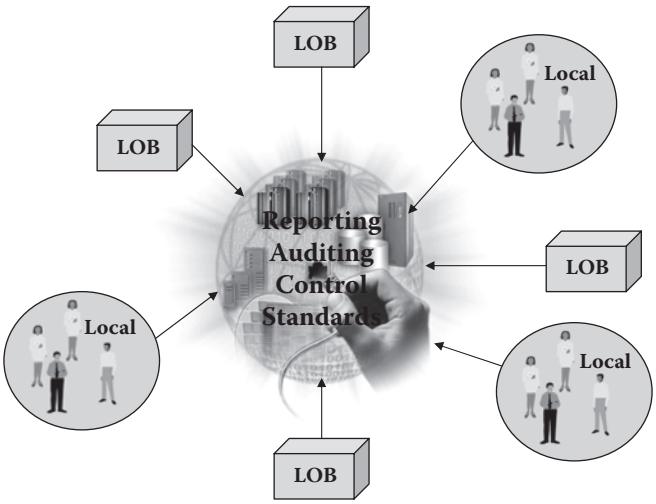
4.5.2.1.4 Loosely Coupled Governance Pattern

Despite the rational appeal of centralized processes in CMA settings, the complexity of the task and the time required to implement such projects will always leave room for the most common CMA governance pattern—loosely coupled organizations.

In loosely coupled (e.g., city-state, franchised) organizations, post CMA organizations are managed and governed as a conglomerate or holding structure of individually strong organizations (see Figure 4.5.4). In this governance model, central corporate instances play only audit and control functions, while local geographies and LOBs retain full execution and operational independence, as long they comply with established group performance metrics (e.g., profit and loss, productivity, growth, budgets). Local autonomy does not mean anarchy; corporate processes exist and are strong, but they are indirect and less invasive.

During the next five to six years, *Fortune* 2000 enterprises with a distributed or holding structure and culture will increasingly enforce stricter central control of local processes. Nevertheless, analysts expect this control to be established through soft (noninvasive) integration methods that preserve local responsibility and initiative. While noninvasive CMA integration models safeguard local/LOB entrepreneurship and operational independence, they enforce corporate efficiency, consistency, and economies of scale through a stronger emphasis on performance metrics, negotiated targets, incentives, and executive governance.

From an IT perspective, these organizations retain relatively autonomous (but not independent) IT units within LOB organizations and geographies, while central IT plays a supervisory and advisory role (setting standards, negotiating licensing contract frameworks, establishing common design principles, encouraging sharing of information and resources). In these integration models, global processes are not automated over local/global business applications, but rather through formalized person-to-person workflows and manual processes. Effective deployment of noninvasive CMA operational integrations



FIGu RE 4.5.4 Loosely coupled CMA governance pattern.

will require powerful, accurate, and timely analytical infrastructures supporting local/global decision and conflict resolution mechanisms.

Noninvasive CMA infrastructures will center on the exploitation of analytical infrastructures (e.g., hub-and-spoke data warehousing infrastructure patterns) and collaborative systems (e.g., person-to-person and person-to-system workflow tools provided in Lotus Notes and Exchange environments). Despite its falling out of favor during the last few years, analysts believe noninvasive integration models will regain standing among *Fortune* 2000 executives as experience exposes the gap between vision and implementation of various CMA integration approaches.

In Figure 4.5.4, business perspectives are:

- Enterprise managed as a "holding" of individually strong organizations
- Central instances audit, control, and guide local entities
- Local independence is retained as long as targets are met (profit and loss, productivity, growth, budgets)
- Corporate processes exist and are strong, but these are indirect and less invasive

Typical IT mandates are:

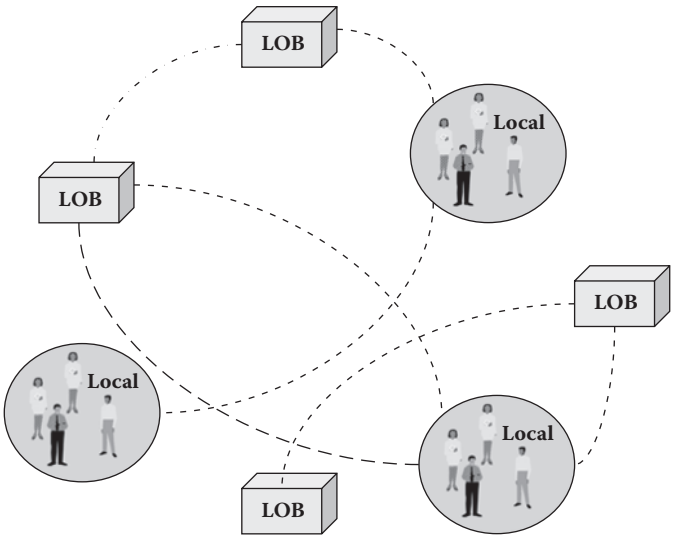
- Central IT steering groups
- Strong central architectural function
- Standards
- Design principles
- Solid reporting and analysis of cross-enterprise data
- Efficient analytical tools and infrastructure
- Corporate licensing and procurement

4.5.2.1.5 Laissez-Faire Governance Pattern

Very often, loosely coupled (e.g., holding company or auxiliary) organizations degenerate into anarchistic or silo organizational models, where the critical supervisory, advisory, and control function of the central organization fails to establish solid cross-organizational enterprise processes. In this case, a dysfunctional loosely coupled organization falls into laissez-faire governance patterns (see Figure 4.5.5). Although laissez-faire patterns have the advantage of encouraging local initiative (and therefore flexibility), weak central coordination will fail to eliminate or control redundant investments, incoherent developments and poor human and material resource use across the enterprise. Therefore, the laissez-faire pattern is the result of business units having agreed to not cooperate, collaborate, or communicate with each other.

The line between the loosely coupled and the laissez-faire governance pattern can be easily crossed in the absence of strong management or changing corporate cultures. Organizations that were once efficient but are now challenged by decentralized management schemes are good examples of such a transition. When business reviews were hailing these companies and their management in the 90s as the epitome of modern, networked, flat, dynamic organizations, they were running effectively under loosely coupled governance patterns. When central coordination and control weakened, their governance models fell into laissez-faire schemes. The question remains whether the way to manage a dysfunctional laissez-faire organization is by superimposing a hard management, centralizing governance culture (e.g., Carly Fiorina's remodeling of HP's structure), or reinforcing governance while keeping a strong measure of local optimization as prescribed in the loosely coupled governance model.

It is therefore critical that CIOs avoid the prevailing confusion between the highly decentralized, but potentially efficient and viable, loosely coupled governance pattern and its dysfunctional degeneration, the laissez-faire pattern. While both advocate a decentralized organization with local initiative and relative autonomy, in the loosely coupled pattern, central governance is indirect and noninvasive, but strong and effective. Clarity of metrics, reporting, and control structures enable properly governed loosely coupled patterns.



FIGu RE 4.5.5 Laissez-Faire CMA governance pattern.

In Figure 4.5.5 business perspectives are:

- Dysfunctional decentralized organization
- Weak central coordination
- Redundant investments
- Incoherent strategies
- Inefficient
- Costly

Typical IT mandates are:

- ITOs should follow corporate culture transition
- Transition can go various ways:
  - Shock treatment: Replace laissez-faire inefficiency with hard management or centralized governance models. This is risky, as this involves a radical culture change
  - Soft treatment: Fix dysfunctional decentralized anarchy with strong but noninvasive corporate governance. This is easier to transition, but risks extend cost and inefficiency state for a longer period of time.

In summary, despite (but sometimes because of) challenging economic conditions, *Fortune* 2000 organizations will continue to embark on merger, acquisition, and divestiture operations through 2010/2012, as they search for economies of scale and time as well as global resilience. Changing corporate topologies and structures will require strong management vision and leadership in selecting and enforcing appropriate governance models. This challenge will be particularly critical for CIOs, because the frequent misalignment of IT and business governance schemes is a major risk factor in CMA operations. The various fundamental IT governance patterns presented above cover a broad spectrum, from distributed organizations (holdings) to highly centralized enterprises. Users should be able to describe unique, complex governance cultures in their enterprise using one or more of these five basic governance patterns.

4.5.2.2 Tasks of Governance

The most important tasks can be summarized as follows:

- Process enforcement for carve-ins and carve-outs

- Security platform deployment and ensuring secure communications between partners
- Streamlining services for partners
- Streamlining management and administration for partners
- Unifying and simplifying metrics for all IT-related functions
- Supervising service-level management, including SLAs
- Professional partnership relationship management
- Professional contract management
- Professional collaboration management
- Ensuring the high quality of services to all partners
- Leading business process automation
- Arbitration if disputes
- Defining a collaborative data retention strategy

4.5.2.3 Security Frameworks of Governance

Service providers rely on a variety of security products including firewalls, intrusion detection systems, vulnerability assessment tools, antivirus applications, Web applications, operating systems, and networking devices, to monitor, investigate, and report on the many types of security issues that are experienced each day. Typically, these devices come from many vendors, as organizations seek best-of-breed products. But because each device type and vendor has its own message, log, and console format, as the security infrastructure is built out, it becomes increasingly difficult to understand, interpret and correlate the output of individual or even groups of devices and get a complete picture about threat profiles.

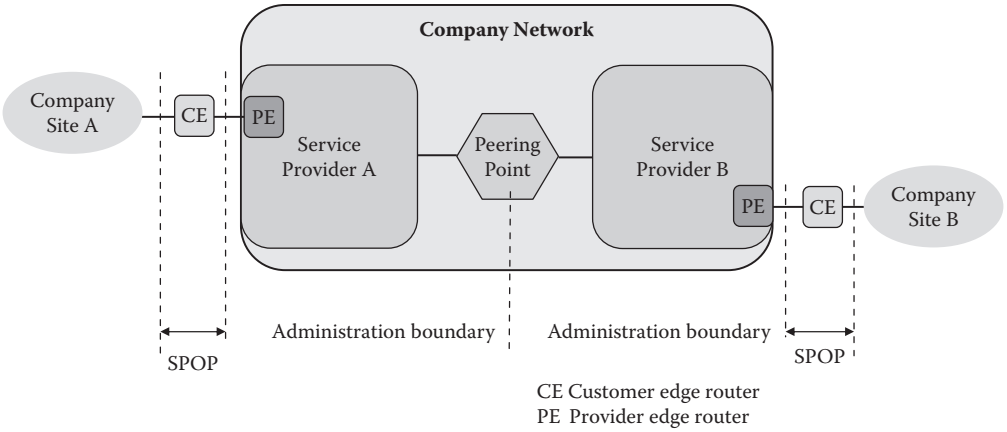
To obtain maximum value from these heterogeneous devices, they must be assembled into a framework that provides the necessary intelligence and tools to deal with a very large number of messages, events, alarms, and alerts per day. Security management frameworks provide a consolidated component set that collects security data from the network, puts it in a common format, stores it in a database, and executes a range of analysis, display, response, and reporting tasks.

A security management framework consists of software agents, server-based managers, and consoles. Agents can be deployed on the security devices, network devices, and applications that report security events at aggregation points or as listening posts for SNMP broadcast. The agents forward the data to server-based managers that consolidate, filter, and cross-correlate the events, using a rules engine and a central database. These managers report relevant information to consoles, where security experts monitor events, receive notifications, and perform incident investigation and response management. Consoles are available as applications for dedicated workstations or via a browser-based interface for remote access.

Real-time correlation is the key element in an effective security management framework, because it automatically examines and analyzes millions of events per day. It works by reading the original event, alarm, or alert, parsing it for its individual fields, and putting those fields into a common format, or schema. These messages, which are being forwarded by the collection component, are then assigned to a proper priority level; real-time correlation assigns them by combining the threats that the firewall intrusion detection system identifies with information about the targets, or assets. The correlation system contains a rule set that scores the threats according to the answers to the following questions:

- What else has occurred?
- Is the asset vulnerable?
- How valuable is the asset?

Because the point of all this is to take the right action at the right time, service providers can set up policies to govern automated responses and responses acted on by subject matter experts. These rules may separate lawful intercepts related to events, alarms, and alerts together and assign the highest priority to them. In other words, assets with lawful intercept functions will get the highest value.



FIGu RE 4.5.6 Peering point and administration boundaries.

4.5.3 Multiprovider Collaboration and Peering

It is likely that the needs of enterprises for managed services will be met by multiple service providers. These providers may be friendly or hostile to each other; but in every case, they will work together to meet customer expectations. Collaboration with fair peering and settlements are the prerequisites.

4.5.3.1 Peering Point and Administration Boundaries

Usually, the infrastructure of the enterprise is supported by multiple service providers domestically, and also internationally. These alternatives are:

- Leading provider (dependency, but just a few contracts)
- Bilateral relationships (many contracts, but decision freedom)
- Broker (everything is arranged by the broker)
- Combination (optimal solution; some of the contracts are arranged by the broker)

Figure 4.5.6 shows the peering point and the administration boundaries between two service providers that are in charge to provision services for different enterprise sites. This example assumes that the boundaries of administration between the two service providers are between the PE and CE routers. But the boundary may be shifted toward the CE router, depending on the service agreement between the providers and the customer.

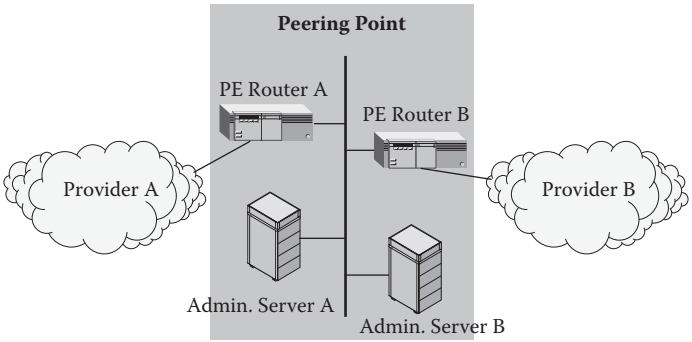
There are multiple alternatives for the peering point: LANs, VLANs, PE routers, or switches. The peering point is usually operated by both service providers. Operations include among others:

- Configuration of network elements
- Configuration of LANs and VLANs
- Problems management
- Load and utilization supervision
- Security control and
- Preventive maintenance

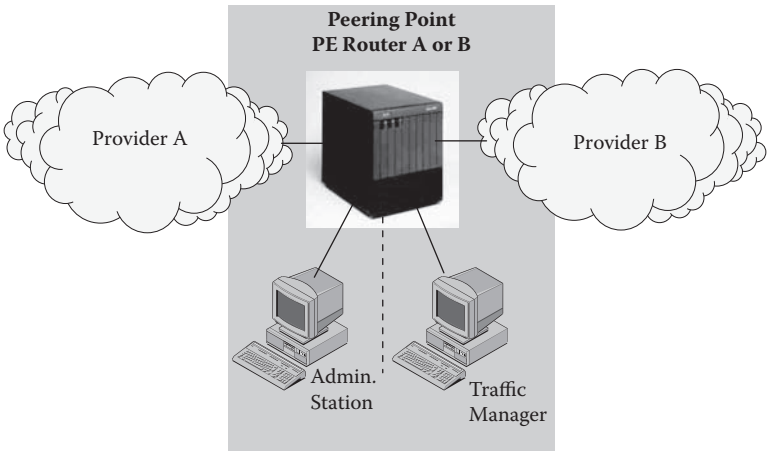
The peering point is usually hosted physically in the building of one of the service providers, but there are special cases with shared buildings.

A VLAN solution enhances the flexibility, e.g., for more rapid reconfiguration of the peering point. Figures 4.5.7 and 4.5.8 show how a specially configured router is in charge of the peering point.

This figure assumes that the peering point is operated jointly by both service providers. It is further assumed that the shared router is capable of providing all data for the settlements between the service providers.



FIGu RE 4.5.7 Peering point for two service providers using a LAN.



FIGu RE 4.5.8 Peering point—one or two PE routers.

Figure 4.5.9 shows a solution alternative with a switch that is well known for voice networks. The switch is in charge of the peering point. Using internal addressing schemes, traffic flows can be assigned and controlled. Further distribution and streamlining is taken care of in the own networks of the service providers.

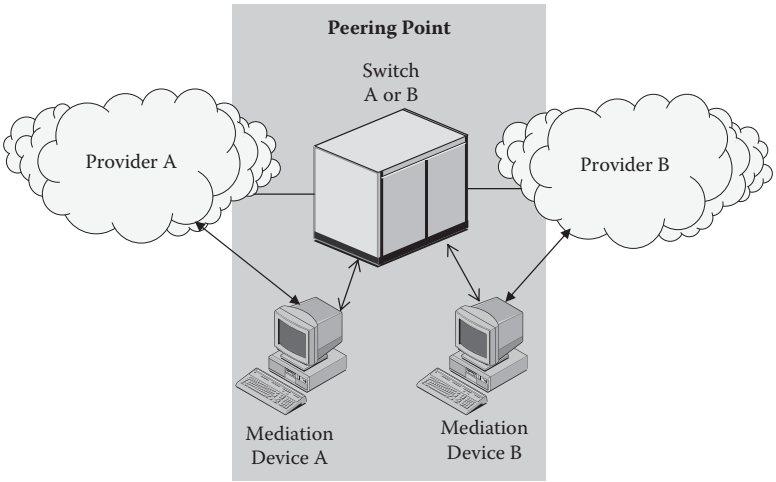
Using this figure, we assume that the peering point may be operated jointly by the service providers. It is further assumed that the shared switch is capable of providing all data for the settlements between the service providers. The CDRs (Call Detail Records) are used as a basis for settlements in the area of voice. In case of data and video, other metrics may be used as the basis for settlements. These metrics are provisioned and controlled by mediation devices. They are responsible for analyzing, formatting, processing, and distributing utilization profiles.

The administration boundary is usually known as the SAP (Service Access Point).

4.5.3.2 Process of Multiprovider Collaboration

This overall process is subdivided into the following process steps:

- Asset management for sites of clients and of service providers
- Defining administration boundaries
- Classification of SAPs
- Unification of service classes and SLAs
- Review of existing SLMs



FIGu RE 4.5.9 Peering point—one switch.

- Determination of peering points
- Establishing and unifying management concepts
- Defining the depth of reporting
- Adding criteria to certification criteria
- Adding tasks to overseer
- Detailing settlement rules

4.5.3.3 Detailed Process Steps of Multiprovider Collaboration

In this section, all of the process steps listed in Section 4.5.3.2 will be detailed.

4.5.3.3.1 Asset Management for Sites of Clients and of Service Providers

Assets must be known and their status and availability must be up to date. In this context, the ownership question must be addressed. Considering assets, the following object classes are relevant:

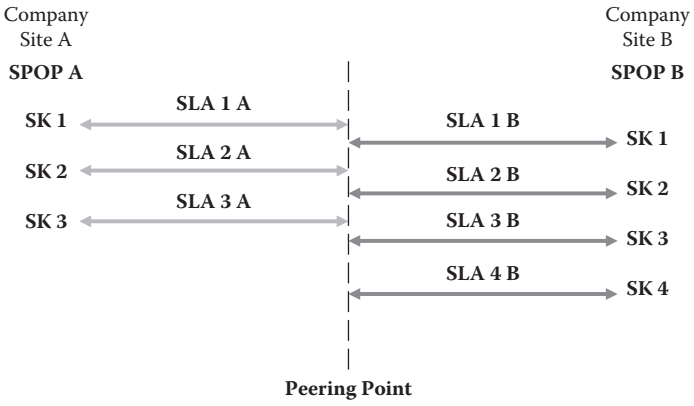
- Customer, client or subscriber of services
- Site
- Device or customer premises equipment
- Switch
- Router
- Distributor at multiple hierarchical layers

Attributes that must be maintained for each managed object, there are ITU and TeleManagement Forum recommendations. To accomplish a unified asset management is not possible due to the large number of isolated and different inventories. Usually, existing inventories include:

- Physical assets
- Business applications
- Customers
- Suppliers
- Parts (reserve, ordered, repaired parts)

4.5.3.3.2 Defining Administration Boundaries

At the boundary, the following managed objects may be identified (Figure 4.5.10 shows a schematic overview):



FIGu RE 4.5.10 Different service classes for service providers.

- PE router
- CE router
- Load balancer
- Firewall

This boundary determines who is in charge for the purposes of administration. All combinations are feasible depending on the agreement with the service providers. As an entry into these issues, we assume that, with the exception of PE routers, all managed objects are managed by the enterprise. Attributes and the actual configuration are maintained in a unified database.

The actual configuration is discovered and compared with the configuration database.

4.5.3.3.3 Classification of SAPs

Together with service providers, enterprises determine the necessary number of different service classes for SAPs. This depends greatly on the rating and accounting model. It is recommended to use maximal three service classes from the enterprise perspective.

- Best effort (Bronze)
- Basis (Silver)
- Premium (Gold)

It is recommended that the following Key Performance Indicators (KPI) are included:

- Availability
- Service acceptance
- Reaction time
- Periodicity of notifications
- Network delay
- Variance
- Packet loss
- Sampling rate
- Access bandwidth

These KPIs will be quantified for each service class. Also the evaluation period is going to be defined. Furthermore there is a correlation with different service classes that can be combined with network delay, variance, and packet loss. These service classes are:

- Real time
- Critical data

- High priority
- Best effort

Enterprises, usually represented by their overseer, are expected to negotiate the right service class for each of the SAPs, maintained by various service providers.

Each group of the service classes will be characterized by quantifiable service metrics. These metrics are included into Service-Level Agreements (SLAs).

4.5.3.3.4 Unification of Service Classes and SLAs

Customers need a transparent end-to-end (across multiple service providers) service agreement. Service providers do have their own ideas and experiences about realistic service classes from their perspective. To agree with deviation from these ideas is not easy. Many service providers show gaps and weaknesses in terms of their own services; practically, there are no service portfolios or service catalogs. Even worse, there are serious gaps in terms of supervising service KPIs for the following reasons:

- The KPIs have not yet been clearly defined
- No unification with metrics
- The tools to supervise metrics are different
- There are no concepts for measurement and supervision
- Reporting is obsolete and still batch oriented
- Status displays (e.g., service views) are not offered
- Accessing measurement data is not supported

Figure 4.5.10 shows a simple structure for one customer with two sites and with two service providers. There are three service classes for site A, and four for site B, respectively. At the peering point, there is no seamless transition between the different service classes. The customer is expected to negotiate each service class with each service provider separately. There is not end-to-end service guarantee. Due to the different number of service classes, the KPIs cannot be compared to each other.

All these shortcomings must be eliminated, or improved to some degree, at least. Otherwise, no good results can be accomplished with multiple service providers and multiple sites of the client. If disputes arise between customers and service providers, the overseer is in charge of arbitration.

The functions of this process step are:

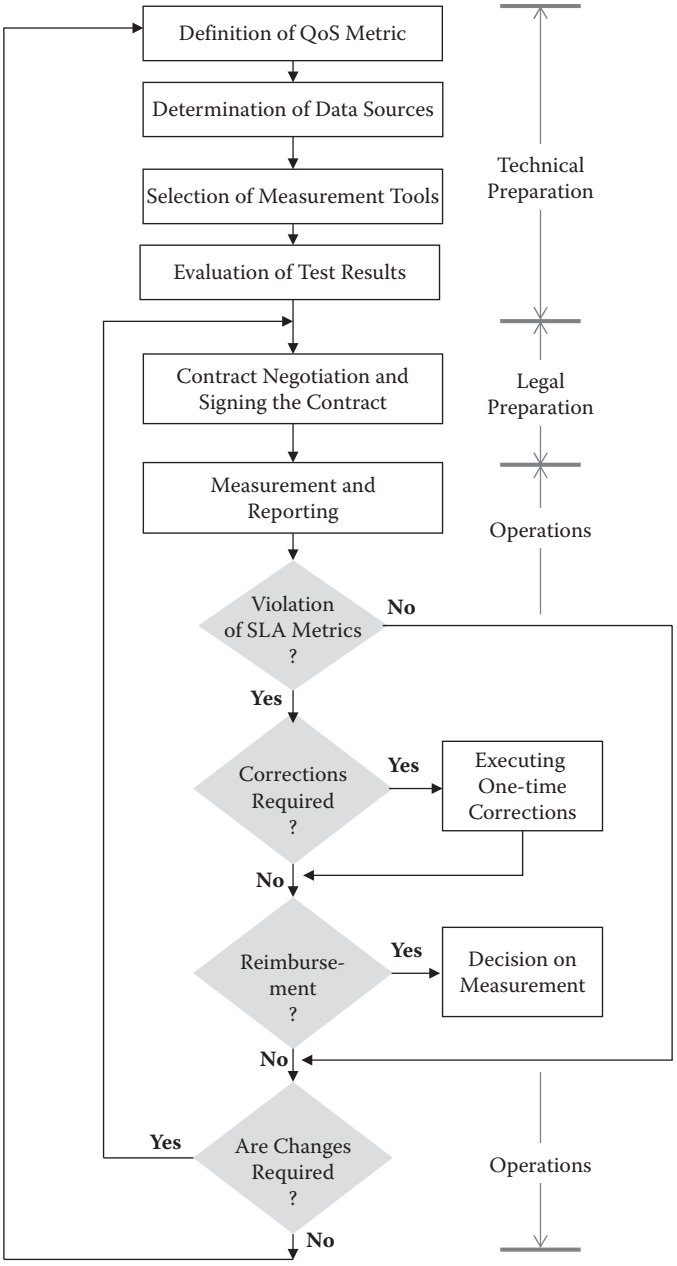
- Publishing service classes
- Publishing service metrics by each provider
- Publishing measurement tools by each provider
- Exchange of existing measurement results and sample reports
- Identification of deltas (deviations) between the desired state and actual state
- Agreements about compromises
- Agreements about the common denominator of all SLAs
- Comparing process steps of service-level management
- Agreements about joint process steps, including the periodicity of reporting

4.5.3.3.5 Review of Existing SLMs

Figure 4.5.11 shows an overview of the most important activities within the service-level management process, and identifies the following functional blocks:

Preparation of service-level agreements:

- Definition of QoS metrics
- Definition of data sources to compute QoS metrics
- Selection of measurement tools, which may be accepted by both parties
- Evaluation of test results with the measurement tool



FIGu RE 4.5.11 Detailed display of the service-level management process.

Contract negotiation and signing the contract: All items in Section 4.5 apply.

Measurements and reporting: The detailed process description includes the measurement data sources, the periodicity of measurements, data processing, combining different databases, and finally reporting. Feedback from this activity decides whether service-level agreements have been met or violated.

Corrections in SLAs: If corrections and rectification are required, they should be executed in quasi–real time. Measures include:

- Completing missing measurement data

- Data correction
- Correlation with manually collected data for some metrics

These corrections do not change the contract, but help to solve single and sporadic deviations. They represent a response for customer complaints over short periods of time.

Reimbursement for noncompliance with SLAs: If noncompliance happens, measures will be agreed upon periodically. Review periods should correspond to billing periods. This process step is heavily correlated with accounting.

Changes in SLAs: When the number of corrections and noncompliance cases is over a predefined threshold, SLAs are expected to be reviewed and if necessary then changed.

SLM requires that multiple QoS metrics are continuously supervised and measured. Depending on the agreements between clients and service providers, reports may be generated and distributed, or/and information on Web servers prepared.

Data sources for SLM include:

- Trouble tickets that are generated automatically or are prepared manually
- Events that are generated by network elements (managed objects), filtered, modified, and classified
- Alarms (SNMP traps or alarms from other sources) represent a specific class of events
- Logs of systems and network components
- Performance metrics, provided by various tools
- Manual logs based on observations

The middle part of Figure 4.5.11 can be further detailed. The collected data are unified and converted into a common denominator (Figure 4.5.12). No complex processing is expected here. The output is a special table (see Table 4.5.1) with a number of different events that are utilized to supervise SLAs. In the first step, separate tables are generated for each service provider.

Table 4.5.1 is the basis for triggering escalation steps. Alarms are derived from events. Basically, the following hierarchy is valid:

- Managed objects
- Status notification
- Filtering processes
- Generation of alarms
- Generation of notifications
- Distribution of alarms and notifications

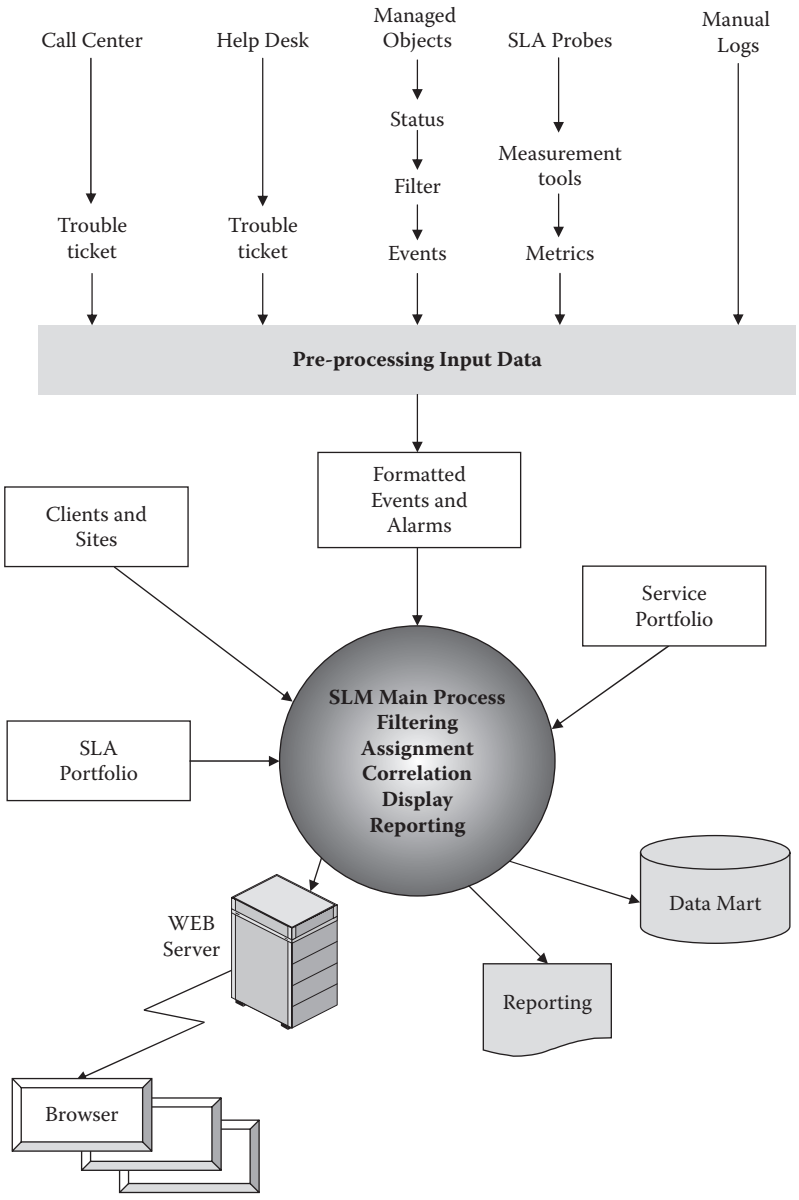
But problems must be classified first. There are usually three classes:

- Critical problems
- Principal problems
- Noncritical problems

The peering agreement includes clarification and unification of these problem classes and their actual content.

Escalation steps must be clearly defined for each problem class in advance. The preparation may look like the following:

- Definition of emergencies: Emergency is when critical problems occur or when multiple problems occur with a certain combination. In the second case, not all problems must be critical.
- Determining the escalation layer: The number of layers depends on the organization of the participating service providers and operators. Usually, two layers are defined for the customers and one for each service provider and operator.
- Identification of persons: For each layer, subject matter experts are named; also site and reach numbers are identified.



FIGu RE 4.5.12 Supervising SLAs as part of Service-Level Agreements.

TABl E 4.5.1 Events

Events	Example 1	Example 2	Example 3
Type	Outage	Brownout	Outage
Severity	High	High	Middle
ID	Router CE	Router PE	Load Balancer
Importance	Middle	High	Middle
Time stamp	31.1.2008	31.1.2008	31.1.2008
Class	Principal problem	Critical event	Noncritical
Comments	Customer Edge	Provider Edge	3 sites impacted

TABl E 4.5.2 Customer File with SLA Entries

Customer/Client	Site A	Site B	Site C
Service ID	SID1: FR2MEG	SID1: VOIP	SID1: FR2MEG
	SID2: ATM64MEG	SID2: Managed E1 Line	SID2: Managed E1 Line
	SID3: IP-VPN	SID3: IP-VPN	SID3: IP-VPN

- Description of processes: Each problem triggers specific escalation steps that may contain both automatic and manual steps. Also the time windows are determined for each step.

Another table (Table 4.5.2) is constructed for each customer (client). This table details the services of the client for each site. This table shows an example for two sites. Considering all the sites, each service must have at least one source and one destination site.

The list of offered services is summarized in another table (Table 4.5.3). The service ID describes the specific ID used by service providers. Also the metrics may be added onto this table.

This service portfolio table is generic. All service providers customize it to their own needs. This is the first step. In the second step, KPIs are defined and offered for each of the services.

The actual SLAs, standardized as far as possible, are maintained in a separate database. In best interest of clients and service provider, the total number of SLAs should be kept to a minimum. When reports are requested, this SLA database is going to be used (Table 4.5.4) for periodic and ad hoc reports.

In establishing such a table, different views may be represented:

- SLAs could be different for each client
- SLAs may be agreed upon for each SID depending on metrics, timestamps, and sites
- Converging all SLAs for clients by the service provider may reduce the administration requirements significantly

In every case, simplification, unification, and standardization of SLAs are the main goal.

4.5.3.3.6 Determination of Peering Points

In determining the peering points, the following tasks should be considered:

- Geographical identification of peering points
- Decision about the peering alternative
- Clarification of ownership at the peering point
- Testing data collection capabilities
- Testing administration software
- Testing processing capabilities of collected data for multiple purposes
- Testing back-up capabilities, when administration is separated by service provider
- Checking the completeness of collected data for service providers settlements
- Trial operations emulating problem cases
- Checking the efficiency of escalation procedures
- Checking the accuracy of collected data

The total path between end points can be frequently broken down into multiple subpaths (see Figure 4.5.6):

- LAN at site A
- WAN1
- WAN2
- LAN at site B

Each subpath is usually measured separately from each other. Between these so called disjunctions points, peering points are located. Basically, there are two cases:

TABl E 4.5.3 Portfolio of Services

Portfolio of Services	Service ID
Wire l ine Voice Services	
Basis services	
IN services	
AIN services	
Wire l ine Data Services	
Managed Lines	
Message Switching services	
Packet Switching services	
Frame Relay services	
Fiber/Copper Distributed Data Interface services	
High-Speed LAN-based services	
High-Speed Data Transfer services	
Audio Broadcasting services	
Wireless and Mobile Services	
Paging services	
Cordless services	
Cellular services	
Personal Communication Systems services	
Specialty voice services	
Wireless data services	
VSAT-based and direct broadcasting services	
Microwave services	
Point-to-multipoint services	
Distribution services	
Integrated Services	
ISDN	
ATM	
SDL	
Business Video and Multimedia services	
Cable-based Services	
IP-based Services	
VoIP and FoIP	
Internet access services	
IP VPNs	

- Measurement domains right and left of the peering points are supported by the same service provider or operators: There are technical problems, when different metrics are used right and left from the peering point or when the same metrics used, but the interpretation of measured values are different. Service providers and operators must pledge to unify and simplify their metrics and the interpretation of metrics for the same managed objects.
- Measurement domains right and left of the peering points are supported by different service providers or operators: Experiences show that there are always technical and organizational problems.

QoS beyond peering points can only be computed with the necessary accuracy when the same metrics and the same measurement tools and techniques are used for the same managed objects.

Figure 4.5.8 shows routers as a peering point. In this case, there are two networks terminated with the router, respectively. Routers are owned by the service provider and connected back-to-back. The rules and alternatives are as follows:

TABl E 4.5.4 Service-Level Agreements

SLA ID for Clients	Client 1 SLA ID = 01	Client 2 SLA ID = 01
	SID1	SID1
	SID2	
	SID3	
		Client 2 Sl A ID = 02
		SID2
		Client 2 Sl A ID = 03
		SID3

- Provider A router and provider B router must exchange routing information with each other. The alternatives are:
  - Policy Based Routing (bad scalability due to the fact of the need of many manual entries, very CPU intensive, acceptable performance can only be guaranteed by high-performance hardware)
  - Static routing
  - RIP (technically making sense)
  - OSPF is meaningful, when IP VPNs and MPLS are well supported
  - EBGP (technically meaningful)
- Measurement and testing points for QoS control are located on the router ports that are toward to the peering point; this makes it possible for routers to measure themselves.
- Due to the fact that the peering point represents the ingress point onto the "other" network, the ingoing packet will be repeatedly identified. This includes the following:
  - Traffic specification, e.g., policy selection
  - Class of Service decision, e.g., priority queuing
  - Call Admission Control
- Inserting firewalls into this path must be excluded; over firewalls, no quality guarantees may be given.

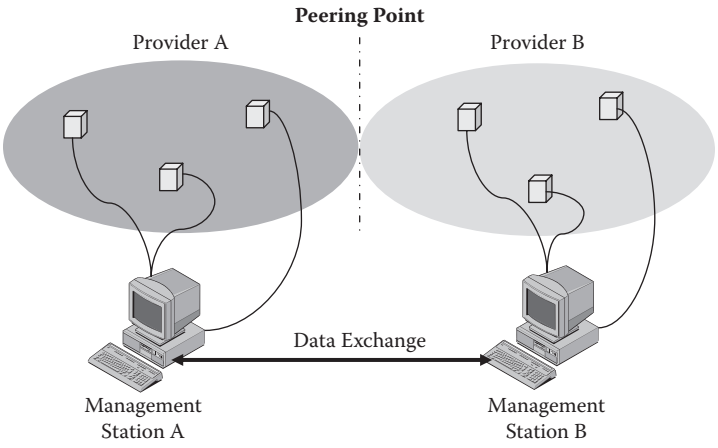
4.5.3.3.7 Establishing and Unifying Management Concepts

In case of multiple service providers, it is mandatory to administer and handle managed objects on the basis of unified guidelines. The following functional areas should be addressed by enterprises, represented by their overseers, and by all service providers, and their subcontractors.

The structure consists of a basis management platform and application areas.

The basis platform provides functions that are required by all management applications. It helps to coordinate all management tasks. The following functional grouping is recommended:

- Infrastructure (hardware, software, directory services, time services, and asset management)
- Basis services (internal communication, graphical user interface, Web server functions, database services)
- Support services for operations (installation, backup, workforce management, order processing, recovery, housekeeping, software distribution)



FIGu RE 4.5.13 Separated measurements to support SLAs.

- Integration services (events, monitoring and alarm management, reporting, policy determination, central security for management systems, information management)
- Multifunctional services (trouble ticketing, hotline, support desk)

The management applications are based on the traditional FCAPS (fault, configuration, accounting, performance, and security), extended by SLA management. These applications groups are:

- Fault Management
- Configuration Management
- Accounting Management
- Performance Management
- Security Management
- Management of Service-Level Agreements

It is assumed that general security measures have already been decided by the contracting partners. Accounting models will be addressed in a separate agreement. In this section, fault, performance, and configuration management should be addressed in greater detail.

First of all, it is very important that contracting partners agree on these management applications and use them as a basis for a checklist.

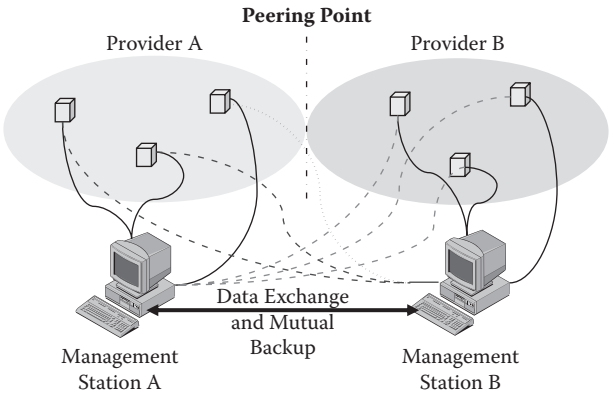
In collaborating with each other, trouble tickets may play a significant role. There are no standards yet, but the majority of service providers exchange trouble-related data using XML. Otherwise, CLI-based (Command Line Interface) solutions are still preferred. When XML-based solutions are selected, trouble tickets may be exchanged and distributed on the Web. Each partner summarizes the required information, completes the symptoms using entries from the inventory database, and forwards the trouble ticket to the next partner, eventually as part of a joint workflow process.

In the field of measurements, service providers may support SLAs with customers in the following two ways:

- Separate measurements (Figure 4.5.13) and
- Collaborative measurements (Figure 4.5.14)

In both cases, at the very beginning, a number of questions should be answered by contracting parties:

- Art of information collection
- Time of measurements
- Overhead due to measurements



FIGu RE 4.5.14 Collaborative measurements to support SLAs.

- Compression hierarchy of measurements
- Art of information presentation
- Art of information interpretation
- Archiving information
- Combinations with modeling tools
- Embedding measurement data into reporting
- Requirements for customization
- Alternatives of customizing
- Resource requirements of first installation
- Resource requirements of operations

In case of separated measurements, network domains usually stay separated from each other. Management stations are expected to exchange information with each other. Reports are generated and distributed on a provider-specific basis. When Web-based solutions are in use, reports by different providers may be prepared separately, and presented together. The authorized client may access information and reports on the Web server.

In case of collaborative measurements, management stations are eligible to mutually access measurement data. This should be negotiated between the service providers on either on a permanent basis or in cases of outages. Management stations exchange information with each other and offer mutual backup to each other. Reports are generated together and are usually distributed via the Web. The maintenance of datamarts or data warehouses is supported as a joint venture. Measurements are grouped as follows:

- Active measurements against measurement point
- Active measurements against server
- Passive measurements over TCP measurement point
- Passive measurement over mirror port
- Passive measurements over application measurement point

Additional measurement options can be defined depending on the tools in use. The network technology for the new infrastructure is very complex; it would be no surprise to select specific products to address specific needs. To go to best-of-breed is correct, but as a result, integration of many good, specific products must be addressed.

4.5.3.3.8 Defining the Depth of Reporting

It is assumed that basic data are collected and processed using unified guidelines. The minimal requirement is that eligible clients can access raw data, preprocessed data, or fully processed data on a Web

server. Service providers and clients can query these data; if necessary, they can generate their own ad hoc reports.

Only a minimal volume of paper-based reports are targeted. These reports are in accordance with supervising principal KPIs in SLAs. They report usually on the following metrics:

- Network availability
- Network delay
- Packet loss and
- Throughput

Each report is organized by service, by site, and by service provider.

4.5.3.3.9 Adding Criteria to Certification Criteria

In addition to the known certification criteria of service providers, the following criteria are recommended for consideration:

- Checking on the ability and willingness for collaborating with other service providers
- Providing a feasible solution for the peering point
- Exchange of fault, performance, and configuration management data with other service providers supporting a standardized format
- Protecting the peering point against noneligible access
- Supporting a clear and unified settlement strategy
- Publishing the own service classes
- Publishing all procedures for Service-Level Management
- Willingness to modify service classes
- Willingness to modify procedures for Service-Level Management
- Maintenance of management data (FM, CM, and PM) for a mutually agreed upon period of time
- Support of international standards for exchanging data
- Open mind in selecting measurement tools

4.5.3.3.10 Adding Tasks to the Overseer Duties

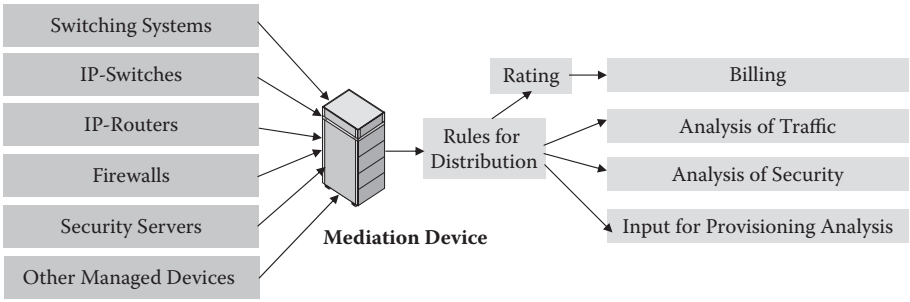
In addition to the known duties of the overseer, the following tasks are recommended for consideration:

- Determining the administration boundaries
- Review of additional certification criteria for service providers
- Review of results for unifying service classes
- Arbitration in case of settlement problems
- Arbitration in case of SLA disagreements
- Clarification of access rights to measurement data

4.5.3.3.11 Detailing Settlement Rules

The mutual beneficial collaboration of multiple service providers presumes usage-based accounting. In order to support usage-based settlements, data streams have to be measured into both directions. Each peering alternative, addressed previously, offers usage data. They must be collected, formatted, and processed. Mediation devices play a key role with settlement support. The main tasks of mediation devices include:

- Data collection
- Validation of collected data
- Corrections to data
- Filtering data
- Aggregating data
- Presentation and distribution



FIGu RE 4.5.15 Sample mediation architecture.

Figure 4.5.15 shows a simple mediation solution.

The following recommendations may be utilized to support settlements between service providers.

Settlement periods could be once in a month or once in a quarter. In case of a very close relationship, service providers may agree to a flat fee. But this agreement is still based on middle and long range traffic observations.

These settlement rules are independent from the services billed to clients of customers.

It is recommended that the following measurements be supported at peering points:

1. Availability: availability of the Inter-Provider-Access channel/peering point
   - Metrics: middle availability, total operating time, total outage time, number of outages, list of single outages
   - Measurement technique: Ping
   - Operational support: 24 x 365 in all 15 minutes
   - Reporting: overseer is the receiver; flexibility is required in compressing and archiving data and reports
2. Throughput: in the WAN between selected measurement points in Bits/s; utilization will be computed in dependence of the bandwidth
   - Measurement technique: permanent and optional measurements using generated traffic flows
   - Measurement frequency: daily once for permanent measurements
   - Reporting: overseer is the receiver; flexibility is required in compressing and archiving data and reports
3. Latency: signal delay < x ms between two selected measurement points
   - Measurement technique: 25 x 64 Byte-ICMP Ping; building averages over 25 measurements
   - Measurement frequency: 24 x 365 in all 15 minutes
   - Reporting: overseer is the receiver; flexibility is required in compressing and archiving data and reports
4. Variance: the high and low marks of latency measurements are considered as the basis
   - Measurement technique: permanent and optional measurements using generated traffic flows
   - Measurement frequency: daily once for permanent measurements
   - Reporting: overseer is the receiver; flexibility is required in compressing and archiving data and reports

The sites of a typical enterprise are going to be supported by multiple service providers. The depth of collaboration is different. The physical configuration of peering points may be different; it is expected that routers, switches, and network access control (NAC) devices will play a key role. For networking the peering points, LANs and VLANs are going to be used.

The administration boundaries decide about the breakdown of responsibilities between the service providers and multinational enterprise clients, and also between service providers, themselves. Both separated and joint measurements for supervising SLAs will be used.

Multiprovider collaboration requires that certification criteria for service providers are extended. The same is true for the tasks of the overseer of a company. All recommendations for these extensions are part of this work package.

4.5.4 Management Services and Outsourcing

Outsourcing has been on the agenda of many enterprises for the last thirty years. In many cases, the decision is highly political. There are heated debates, whether outsourcing is a threat to employment in certain highly developed industrial countries. Besides politics, financial considerations play an important role, but as surveys confirm, they are not the dominating factor at decision points. Time-to-deliver and gaining competitive advantages took over the highest priorities over the last two to three years.

Industry observations may be summarized as follows:

- Spending on outsourcing is picking up at large companies
- The benefits of outsourcing still outweigh the problems
- Outsourcing is not necessarily a money saver
- Offshoring is growing in popularity, but satisfaction with offshore firms lags behind domestic outsourcers
- Employees' fears of losing employment due to outsourcing is a growing problem
- Enterprises like the outsourcing idea. It is the outsourcers, they have got difficulties with
- Good governance improves outsourcing success and satisfaction

From today's perspective, the driving forces for outsourcing include:

a. In-house management has become too expensive—human resource costs are growing especially fast. Outsourcing can produce direct cost savings of over 10%, and there may be tax advantages to using external or managed service providers rather than in-house personnel.
b. There is a chronic shortage of skilled personnel that can keep up with service offers, innovative infrastructure components, and their management.
c. Enterprises need to speed up implementation of new technology, including new services, new equipment, and new facilities. When the managed services provider can share network management resources—platforms, tools, and people—among multiple customers, not only is implementation relatively rapid, but also the customer benefits from the MSPs economies of scale.
d. CIOs and CTOs may spend more time on strategic issues than on operative network management, including the very time-consuming need to deal with multiple suppliers of infrastructure components.
e. Enterprises need more accurate Service-Level Agreements. When outsourcing is working well, these service-level agreements become part of the service contract with the managed services provider.

Before deciding for or against outsourcing, enterprises should evaluate the following criteria very carefully:

1. The present asset value of networking components and network management.
2. The efficiency of existing processes, tools, and human resources. This review is absolutely necessary when deciding for which part a managed service is under consideration. In addition, CIOs and chief technology officers (CTOs) considering outsourcing have a good excuse to audit present operations and address areas that need improvement.

3. The dependence of application availability and whether the MSP can guarantee the targeted availability.
4. The grade of service and applications' security may dictate the type of provider the enterprise must use. In other words, enterprises must select providers that are not sharing resources among multiple customers.
5. The level of security that is required for the business. In some cases, it may be inadvisable to allow third parties to access the infrastructure components and eventually see the carried traffic.
6. Decide whether it is more cost efficient to concentrate on the core business or in addition to spend on building a sophisticated network management system and organization.
7. If the enterprise had to invest a substantial amount into network management, outsourcing would be a valid choice. If not, outsourcing might still be interesting, but it will be a lower-priority item.
8. The current and future needs for skilled personnel; the most sophisticated networks and their services would be useless if the enterprise cannot find human resources to operate it.
9. Potential acquisitions, mergers, and carve-out of business units, as well as changes in the application portfolio will affect agreements with providers and therefore need special and careful treatment in contracts.
10. Since outsourcing contracts are usually more than five years in length, the terms of the contract are of paramount importance.
11. CIOs and CTOs should evaluate the corporate cultures of all parties in general and the management philosophies in particular prior to the final outsourcing decision.

When outsourcing toward managed services and working with managed service providers, the following issues are high priority:

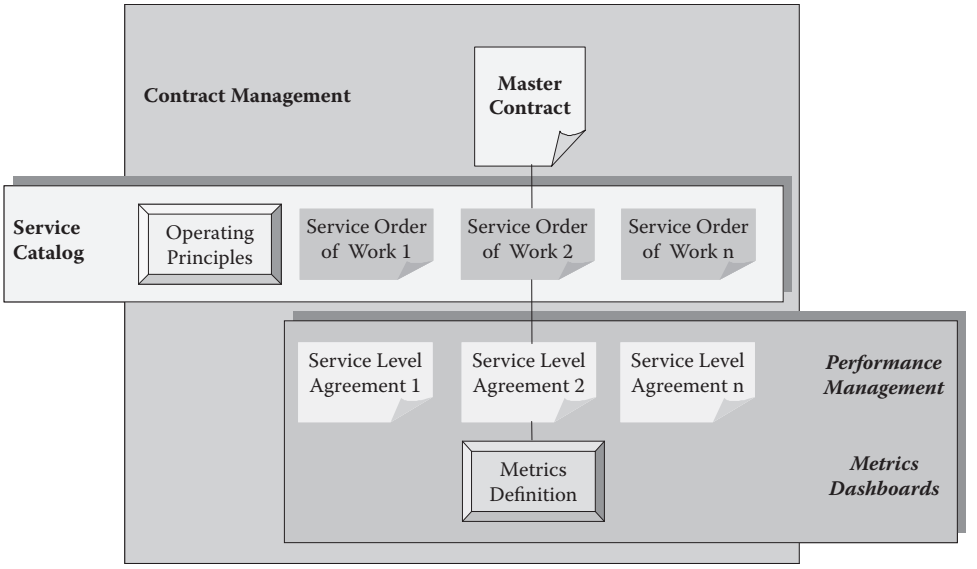
- **Customer support**: the MSP must dedicated high-quality support staff to the customers' account. A well-equipped network operations center is a prerequisite.
- **Integration**: with multiple best-of-breed services available, MSPs should answer the question, how they can integrate with other services customers might have in place in the future.
- **Reporting**: MSPs should be made accountable. They should provide extensive reporting as part of the managed service contract. Various reporting forms should be supported.
- **Scalability**: the MSP could be addressing an isolated IT task, but customers must be sure, the niche service can scale to the expanding networking needs of the customers.

The current market climate is right for customers, who have grown more confident of the reliability and security of managed services and for MSPs themselves. Specialty managed services such as security and storage, and software-as-a-service licensing models are popular among IT buyers. Some service providers offer their customers an on-demand management model with the result that they invoice customers just for the duration of actual usage.

4.5.5 Contract Management

A single sourcing arrangement of moderate complexity and scope may result in a contract that's several hundred pages long. Depending on the preferences and organizational capabilities of the negotiating parties, contract documents may be well structured and easy to follow, or they can be a convoluted mess. In earlier days, the client manager most likely would have locked the contract away in a drawer, taking it out only when a problem arose. Even if a manager desired greater day-to-day control and oversight of the terms, obligations, and promises made in the contract, virtually no commercial tools were available to help the manager in this effort.

Today's contract management tools, which are still evolving, work with and are driven by the various terms and conditions specified in the contract. The simpler the contract is to understand, and



FIGu RE 4.5.16 Documents composing a sourcing contract.

the simpler it is to find important terms and information, the easier it is to use an automated tool. Ideally, a contract will consist of several documents to improve understanding and ease administration. Figure 4.5.16 illustrates a typical (and recommended) contract structure, the documents involved, and the relationships between them. These include the following items.

- **Master contract.** The master contract is a legal document that specifies the rules by which the client and service provider will operate. Normally, one master contract governs the relationship, containing financial terms and standard legal protections, and covering such topics as issue resolution, work changes, and exit clauses—all from a legal perspective. The contract refrains from specifying the details of the work to be sourced but refers to attachments (described below)—structure that makes it easier to change service-related work without necessarily modifying the master contract.
- **SOWs (Statements of Work).** SOWs form the core of any sourcing arrangement and define the scope of the service to be provided or the work to be performed. A SOW specifies the assets and/or functions that the service provider will support, the types of work that will be performed, the inputs required, deliverables created, and each party's roles in the arrangement. Multiple SOWs tend to be easier to administer and adjust than one mammoth SOW, with the actual number depending on how many discrete projects, functions, or efforts will be transferred to the service provider. Depending on its construction, a SOW may relate directly to a service contained in the service catalogue. For example, a SOW may cover desktop PC maintenance services, which appears as an entry in the catalogue.
- **Sl As.** SLAs define the parameters by which the work identified in the SOWs will be performed and judged. The SLA defines performance criteria such as the volume of work that should be performed in a given time frame, system availability, and acceptable response times for requests, quality requirements, and efficiency measures. These commitments may run two ways, describing both service provider and client obligations. SLA performance criteria are described using metrics. Generally, a separate metric is used for each characteristic measured. For example, when outsourcing a help desk function, work volume may be specified by the number of calls handled per month, and responsiveness may be defined as average time to call back. SLAs are the measures that appear in dashboards to give managers a quick view of vendor performance.

- **Metrics definitions.** To avoid confusion or misinterpretation, each metric used in an SLA must have a definition. A definition will include a name, description, objective, measurement method, rules of initialization, and responsibility/roles for collection and interpretation.
- For most sourcing arrangements, a subset of metrics will suffice, and metrics generally are centralized in one definitions document. The rules by which a metric is calculated and the sources for data are items typically entered into a performance management tool.
- **Operating principles.** Operating principles define how the service provider and client will work together on a daily basis, paying special attention to logistics and handoffs between provider and client. Reporting relationships, governance, procedures, and processes for such activities as submitting work requests, turning completed work over to production, obtaining sign-offs, and raising problems are covered by the operating principles document. When creating the service catalog, this document is an important source of information.

4.5.5.1 Benefits of Contract Management Tools

When managing a single, complex contract, or many smaller contracts, tools offer significant benefits over manual, occasional oversight. These benefits include:

- **The ability to actively manage the contract throughout its life cycle.** By capturing all relevant contractual terms and obligations (i.e., SOWs, SLA metrics, payment), important dates and procedures (i.e., operating principles, notice provisions), and interested parties, tools automate contract administration, relieving managers of the burden of wading through pages of detail.
- **The ability to manage changes to contract documents.** Contracts will inevitably change as new services are added, existing services are terminated, service-levels are adjusted, or the parties want to modify terms and conditions. Because tools integrate all of the contract's components, they can assist in the change management process by allowing the parties to assess the impact of changes before they're rolled out. Once changes are approved, tools can help propagate them throughout all components and documents. This way, tools serve as a central place to route and track change requests and help formalize otherwise ad hoc change requests.
- **The ability to log contract changes and notify aȩected parties.** Once contract changes are approved, tools can keep a log of modifications so that all parties have a current view of the contract and associated terms, such as changed SLAs. Tools also can send alerts or notices to parties affected by selected modifications for legal compliance purposes and to enable them to prepare for no changes in services, price, responsiveness, hours of operation, or other terms.
- **The ability to automate the approval process.** In addition to changes in the contract, various other terms and services will inevitably change during the course of an engagement. The operating principles typically specify which approval processes to use and the parties involved in approving changes. Tools also can help automate the approval process by ensuring that the appropriate individuals are notified, that changes are captured and forwarded without falling through the cracks or languishing on someone's desk, and that sign-offs are obtained in the right order and from the right people.
- **The ability to perform initial setup and definition of Sl As and metrics.** Tools are ideal for defining metrics and service levels. Using predefined templates, wizards, lists of formulas, and calculations, tools allow companies to take the SLA and metrics definitions from the paper contract and translate them into a format that can be used to collect performance data automatically.
- **The ability to support issue elevation and resolution.** Operating principles usually describe a process for raising issues that affect the relationship (as opposed to project or application-type issues that are resolved as part of project management). When issues rise to a certain level—whether it's a noticeable trend, a failure, or a predicted failure—tools can automate the process described in the operating principles to alert stakeholders, forward relevant data, and monitor

the elevation process to completion. In case of a contract breach or application of a penalty, tools can integrate with other systems (such as billing and CRM) to ensure that the correct follow-up activities occur, such as adding credits to an invoice. Tools can even trigger a round of contract changes and associated approval process.

4.5.5.2 Selection and Setup Issues and Concerns

When automating contract management, a few important caveats apply. The ability to map the contract documents and terms into a tool depends greatly on the quality and structure of the contract documents and the degree of detail in the terms. If the contract is so large that it defies understanding, is confusing, or uses ambiguous terminology, then translating it into tool-friendly form will be challenging. Tools force definition, so it's easier if the contract is clear and precise from the outset. Even with a well-laid-out contract, taking the information from the contract and entering it into a tool can be a time-consuming effort, affected by the number of documents involved and the number of sourcing contracts administered. Tools can help ease the process based on the quality of their interfaces, but considerable setup time is the norm. On the bright side, this initial investment is quickly repaid with improvements in contract administration through the life of the outsourcing engagement. Furthermore, the data entered for contract management purposes—from SOWs, SLAs, and metrics—is also used for setting up service catalogs and performance measures, so once it's entered it can be reused for multiple purposes.

4.5.6 Summary and Trends

Two primary factors drive the need to a more open approach of delivering managed services to customers. The first is that not all solution providers can deliver all managed services to all customers. As a result, they will need to partner with other MSPs, and those partnerships cannot be limited to MSPs that happen to be running the same managed services platform. The second is the diversity of the MSP community, which will require integration with services provided by telecommunication companies, value-added resellers, and retail chains. MSPs are not going to be able to dictate terms to enterprises that are typically the multiple of their size. Standard organizations, such as OASIS (**Organization for the Advancement of Structured Information Standards)** or the WWW (World Wide Web) Consortium are helpful to integration, but most likely too slow. In the meantime, MSPs would be well advised to look for solution alternatives with a large amount of flexibility.

Given the growing demand of customers for services more tightly aligned with their internal business requirements, it was only a matter of time before customers began pressing managed services providers for services built with their needs in mind, rather than the one-size-fits-all approach. The challenge this presents to the providers is the requirement to build a service that appeals to one customer while still being applicable to other customers. If so, MSPs can derive additional business opportunities for the required technological investment.

Managed services in combination with software-as-a-service will guarantee great flexibility for enterprises. Besides flexibility and market agility, enterprises may save capital and operating expenses.

Acronyms

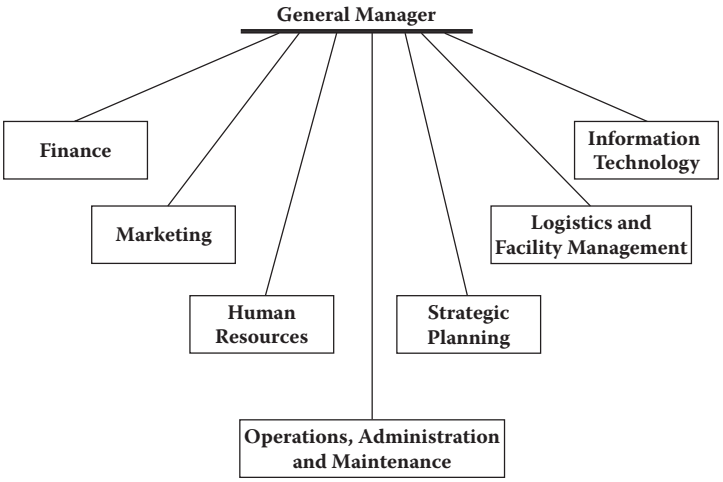
BI	Business Intelligence
CM	Configuration Management
CMA	Carve-Out Merger Acquisition
CIO	Chief Information Officer
CRM	Customer Relationship Management

CTO	Chief Technology Officer
EBGP	Enhanced Border Gateway Protocol
EMEA	Europe, Middle East, Africa
ERP	Enterprise Resource Planning
FCAPS	FCAPS (ISO model for network management)
FM	Facility Management
FoIP	Fax over Internet Protocol
ICMP	Internet Control Message Protocol
IP	Internet Protocol
ITO	IT Organization
KPI	Key Performance Indicators
LAN	Local Area Network
LOB	Lines of Business
M&A	Merger and Acquisitions
MPLS	Multiprotocol Label Switching
MSP	Managed Service Provider
OASIS	Organization for the Advancement of Structured Information Standards
OSPF	Open Shortest Path First (dynamic routing protocols)
PM	Performance Management
PRM	Product Resource Management
QoS	Quality of Service
RIP	Routing Information Protocol
SAP	Service Access Point
SCM	Supply Chain Management Software Configuration Management
SLA	Service-Level Agreement
SLM	Service-Level Management
SOW	Statement of Work
VLAN	Virtual Local Area Network
VoIP	Voice over IP
VPN	Virtual Private Network
WAN	Wide Area Network
XML	Extensible Markup Language

4.6 Network Management Organization

Kornel Terplan

Workforce is one of the critical success factors in building, maintaining, and operating communication networks. Motivated people are the differentiating factor between a well run and a badly run business in the service provider area. This chapter focuses on human resources and their management by introducing a sample organization. Human resources (HR) are assigned to principal business processes and support tools of the service provider. Sample job profiles help HR of the service provider to upgrade their own profiles, post jobs, and hire the right persons. Enabling technologies, such as document, knowledge, and workflow management help to increase the efficiency of the service providers' organization. The process and most likely results of benchmarks are shown. Benchmarks help to compare the performance of service providers with each other, with the industry average and with best practices.



FIGu RE 4.6.1 Organizational structure of an average telecommunications provider.

4.6.1 Organization Structure of the Average Provider

It is very difficult to create an organizational model that fits all providers. But there are certain attributes that can be identified with each provider. With a few exceptions, the following organizational units can be discovered with telecommunications providers: (Figure 4.6.1):

- Finance (enterprise resource management, revenue assurance, fraud management, credit analysis, etc.)
- Marketing (customer, care, customer relationship management, products/services portfolio management, sales, order processing, billing, etc.)
- Human resources (salary accounting, administration, recruiting, training, career management, etc.)
- Logistics and facility management (order management, physical inventory, asset management, logistics, transport, energy management, security management, investments, purchase, etc.)
- Strategic planning (new technology, design, development. Infrastructure optimization, etc.)
- Operations, administration and maintenance (provisioning, deployment, change management, service configuration, maintenance, tests, monitoring, quality and service assurance, support systems, documentation systems, management systems, etc.)
- Information Technology (application management, vendor management, software purchase, intranet, documents management, e-mail, groupware, knowledge management, Internet access, etc.)

The list of functions and responsibilities is not complete. At this stage, they are examples, only. Depending on the size of the service provider, both centralized and decentralized structures can be observed. Operations, Administration, and Maintenance are departments of a distributed organization with many regional managers. Central control should be combined with decentralized dispatch. Sales and marketing, and all customer-facing activities should be distributed, as well. Other organizational units may be organized centrally. Using state-of-the-art communication technologies, the physical presence of certain functions is not relevant anymore. To a certain limit, teleworking may be supported as well.

How an IT organization responds to the challenges of the competitive environment may become the key differentiator for the provider. Within this IT environment, network and IT managers continually struggle to find methods to cope with:

- shrinking headcounts
- tightly controlled operational budgets
- increasing needs for new applications and services

- deploying new network technologies into an already stressed environment
- integrating an automating complex support, documentation, and network management systems
- balance resources between developing new applications and customizing purchased applications
- upgrading skill sets of subject matter experts to capitalize on new information technologies
- manage and reimplement knowledge
- balance and optimize use of systems and networking infrastructures for internal use
- help to establish extranets with customers

4.6.2 Assigning Subject Matter Experts to Processes and Support Systems

Table 4.6.1 shows a high-level allocation matrix of principal support processes of service providers to organizational units. Service providers are expected to break down this matrix further and direct the allocation of each function to subject matter experts. Doing so, redundancies in responsibilities, including finger pointing in cases of emergencies, can be avoided.

Table 4.6.2 shows a high-level allocation matrix of principal support systems of service providers to organizational units. Service providers are expected to break down this matrix further and identify the individual ownership for support systems. Doing so, redundancies in responsibilities, e.g., procurements, customization, and integration, can be avoided.

4.6.3 Building the Management Teams

These factors complicate the hiring process by making it very difficult to write job descriptions and analyze candidates' background materials. The following is a list of recommended criteria when hiring network management staff.

1. Identify team members: The earlier sections of this chapter gave an overview of principal network management functions. Depending on the size of the network, the human resources demand may be computed. After subtracting the available staff from the total demand for each functional area, the demand on new hires can be quantified.
2. Recruit candidates: Advertisements, conferences, headhunters, and individual contacts to colleges, universities, and other companies help to find candidates to be interviewed.
3. Establish interview criteria: Guidelines and evaluation criteria have to be set prior to starting the interviews. In order to keep investment for both parties low, written applications must be filtered carefully. Occasional phone conversations may fill existing gaps. Invitations to personal interviews should be sent out to candidates whose applications are matching the expectations.
4. Hire properly qualified candidates: Hiring has to be for mutual benefits, and not just to fill the job. Future turnover can be avoided this way.
5. Assign and/or reassign responsibilities: Static job descriptions should serve as a guideline only. Within this framework, more dynamic descriptions with rotation in mind are necessary.
6. Institute performance evaluations: Periodic reviews are widely used. If possible, upward performance appraisals should be agreed upon as early as possible in the team-building phase.
7. Promote openness and handle complaints: In order to emphasize team spirit, opinions, even complaints must be encouraged on behalf of network management supervisors. The employees must have the feeling that their comments and suggestions are handled at the earliest convenience of managers.
8. Resolve personnel problems quickly: In order to avoid tensions within the network management organization, problems must be resolved for mutual benefits as quickly as possible. The reward system must provide opportunities to do so. Quite frequently, visibility of how the reward system works resolves problems almost automatically.
9. Institute systematic training and development programs: Systematic education should include training for network management functions, network management instruments, and for personal

TABl E 4.6.1 Allocation Matrix for Support Processes

Organizational Units	Finance	Sales and Marketing	Human Resources	Operations, Admin, and Maintenance	Strategic Planning	Logistics and Facility Management	Information Technology
				Processes			
Customer Care and Billing							
Customer interface management		X		X			
Sales		X					
Problem handling				X			
Customer QoS management				X			
Call rating and discounting	X						
Invoicing and collection	X						
Consulting and support		X	X	X	X		X
Provisioning and Order Processing							
Inventory management	X	X				X	
Service creation, planning and development		X			X		
Network planning and development					X		
Network provisioning				X			
Service ordering	X			X			
Service configuration				X			
Security management				X			
Network Operational Management							
Service problem resolution		X		X			
Service quality management		X		X			
Network maintenance and resolution				X			
Data collection and data management				X			X

TABl E 4.6.2 Allocation Matrix for Support Systems

Organizational Units	Finance	Sales and Marketing	Human Resources	Operations, Admin, and Maintenance	Strategic Planning	Logistics and Facility Management	Information Technology
				Support Tools			
Management Frameworks				X			X
Operation Support Systems	X	X		X			X
Business Support Systems	X	X				X	X
Marketing Support Systems		X					X
Documentation Systems	X	X	X	X	X	X	X
Management Systems				X			X
SLA Surveillance Instruments		X		X			
Customer Relationship Management Tools		X					
Billing Applications		X					
E-Care Tools		X		X			
Modeling Applications				X			X
Knowledge Tools				X	X		X
Geographical Information Systems				X			
Enterprise Resource Planning Tools	X		X			X	

skills. A curriculum in coordination with vendors and educational institutes would guarantee high quality and employee satisfaction.

10. Regularly interface network management staff with users: In order to promote mutual understanding of working conditions and problems, both parties should exchange views and opinions. The level of formality may vary from very informal to very formal; in the second case; written service-level agreements are evaluated.
11. Evaluate new technologies: As part of the motivation process, network management renovation opportunities must be evaluated continuously. This process includes new management platforms, new technologies of distribution, feasibility of new and existing solutions, new monitors, changes in de-facto and open standards, simplification of management processes, changes in the offerings of leading manufacturers, and monitoring the needs of users. Thus, enrichment of lower-level jobs may easily be accomplished.

4.6.4 Keeping the Management Teams

In order to keep the network management team together, expectations of employers and employees must match to a certain degree. Table 4.6.3 shows a sample of expectations on both sides. The individual–organization contract is implied because much of it often unwritten and unspoken.

TABl E 4.6.3 Expectations in the Employer/Employee Relationship

Expectations of the Individual	Expectations of the Employer
1. Compensation	1. An honest day's work
2. Personal development opportunities	2. Loyalty to organization
3. Recognition and approval for good work	3. Initiative
4. Security through fringe benefits	4. Conformity to organizational norms
5. Friendly, supportive environment	5. Job effectiveness
6. Fair treatment	6. Flexibility and willingness to learn and develop
7. Meaningful or purposeful job	7. No security violations

There are several reasons for this:

- Both parties may not be entirely clear about their expectations and how they wish them to be met. They may not want to define the contract until they have a better feel for what they want.
- Neither of the parties is aware of their expectations. For example, organizations are hardly able to define the term: loyalty.
- Some expectations may be perceived as so natural and basic that they are taken as granted, e.g., expectations of not stealing and an honest day's work for a day's pay.
- Cultural norms may inhibit verbalization, which is in particular very important with multinational companies hiring employees in various countries.

At a given time, there will be some relatively fulfilled and unfulfilled expectations; however, each party has to have a minimum acceptance level of fulfillment. If either party concludes that the fulfillment of its needs is below this minimum level, it will view the contract as having been violated.

Turnover in network management can be very disadvantageous for maintaining service levels to end users. Corporate and business units' management should try to avoid above-average turnover by implementing rewards to satisfy employees. Gaining satisfaction with the rewards given is not a simple matter. It is a function of several factors that organizations must learn to manage.

The individual's satisfaction with rewards is, in part, related to what's expected and how much is received. Feelings of satisfaction or dissatisfaction arise when individuals compare their input (knowledge, skills, experience) to output (mix of rewards) they receive.

Employee satisfaction is also affected by comparisons with other people in similar jobs and organizations. People vary considerably in how they weight various inputs and outputs in that comparison. They tend to weight their strong points more heavily, such as certain skills or a recent performance peak. Individuals also tend to correlate their own performance compared with the rating they receive from their supervisors. The problem of unrealistic self-ratings exists partly because supervisors in most organizations do not communicate a candid evaluation of their subordinates' performance to them.

Employees often misperceive the rewards of others; their misperception can cause the employees to become dissatisfied. Evidence shows that individuals tend to overestimate the pay of colleagues doing similar jobs and underestimate their colleagues' performance.

Finally, overall satisfaction results from a mix of rewards rather than from any single reward. Rewards fall into two principal categories: extrinsic and intrinsic. Extrinsic rewards come from the employer as compensation, benefits, job security, training, promotions, effective network management instruments, and recognition. Intrinsic rewards come from performing the task itself, and may include job satisfaction, sense of influence, quality of environment, and quality of assignment. The priority of extrinsic and intrinsic rewards depends on the individual person. The following list gives a frequently seen priority sequence to keep the network management team together.

1. Compensation: Payment is still the most important motivation factor. Organizations try to use a number of person-based or skill-based compensation techniques combined with the dependence on sales revenues of the larger organization, if applicable. Pay is a matter of perception and values that often generates conflict.
2. Benefits: Benefits take special forms, depending on the employer's business; e.g., company car, life insurance, lower interest rates, housing. The cost of benefits at companies can be as high as 35 to 45% of pay dollar.
3. Job security: Seniority with job assignments is a very valuable management practice, in particular, when the economy is stressed. Job security policies include retirement plans, options for early retirement, and agreements of non-layoff. Job security packages are more advanced in Europe and Japan than in the United States.
4. Recognition: Recognition may come from the organization or from fellow employees. The periodic form of recognition is the performance appraisal conducted by the supervisor. A relatively new form, the so-called upward appraisal is considered a form of subordinates recognition. It is difficult because most managers do not want to be evaluated by their subordinates. For the subordinates, it is the forum of communicating ideas for improvement.
5. Career path and creation of dual ladders: In order to keep motivation high, managerial and technical assignments must be compensated equally. Promoting technically interested persons into managerial positions may not have the desired results; these persons are usually high in affiliation motivation and low in power motivation. Helpful activities include career counseling and exploration, increased company career opportunity information, improving career feedback, enhancing linear career, slower early career advancement, and enrichment of lower-level jobs with more challenges.
6. Effective training: This type of motivation helps to keep the specific and generic knowledge of the employees at the most advanced level. Three to six weeks training and education annually is considered adequate in the dynamically changing network management environment.
7. Quality of assignments: Job descriptions are expected to give the framework for expectations. But dynamic job descriptions may help to avoid monotony and promote job rotation. The client contact point, systems administration and change control may be rotated periodically.
8. Use of adequate tools: Better instrumented networking environments facilitate the jobs of the network management staff, increase the service quality to users, and improve the image of the network management organization. At the same time, persons working with advanced tools are proud of their special knowledge, and of their employer. They are highly motivated to continue with the company.
9. Realistic performance goals: As part of dynamic job descriptions and job rotation, realistic performance expectations may help to stabilize the position of the network management team. Management must find the balance between quantifiable and non-quantifiable goals. Average time spent on trouble calls, response time to problems, time of repair, and end-user satisfaction or dissatisfaction are examples for both types of goals.
10. Quality of environment: This is more or less a generic term expressing the mix of network management related instruments, pleasant working atmosphere, comfortable furniture, adequate legroom, easy access to filing cabinets or to hypermedia, acceptance of opinions on shortcomings, and team spirit.
11. Employee control: Despite high team spirit, individuals need certain levels of control that can only be determined by managerial skills. Depending on the person, positive or negative motivation, or a combination of both, may work best.

The preceding list has tried to concentrate on key motivation alternatives only, but there are many more. In order to find the optimal combination for individual installations, a human resources management audit is recommended.

4.6.5 Job Profiles for Human Resources of Telecommunications Service Providers

Successful operations require a well-educated management team with adequate skill levels. In order to make hiring and cross-education easier, the most important profiles for the teams should be well prepared and well maintained. These job descriptions serve as a basis to evaluate the completeness of existing descriptions and of existing documents on both managers and subject matter experts of the service provider.

Job profiles should include the following items:

- Responsibilities
- External job contacts
- Qualifying experiences
- Required education
- Personal attributes
- Salary range

In order to help service provider to prepare job profiles, a few examples are shown:

4.6.5.1 Profile of a Network Operations Manager

Responsibilities:

1. Supervises and monitors the quality of network management
2. Estimates cost and resource requirements
3. Reviews and approves processes and instruments
4. Performs planning and scheduling of products' implementation
5. Develops, implements, and enforces procedural and security standards
6. Evaluates performance of processes, instruments, and people, and reports results to management
7. Plans and directs acquisitions, training, and development projects
8. Creates and supervises service-level agreements (SLAs)
9. Defines and selects quality of service (QoS) metrics

External job contacts:

1. Other managers within information systems
2. Some users
3. Some vendors
4. External consulting companies

Qualifying experience and attributes:

1. Prior experience in statistics, mathematics, accounting, computer science, telecommunications, or the equivalent
2. Training in advanced practices, skills and concepts, administrative management, supervisory techniques, resource management, budgeting, and planning
3. Excellent communication skills
4. Excellent negotiation skills
5. Excellent managerial skills

4.6.5.2 Profile of a Call Center Operator

Responsibilities:

1. Network supervision: Implements first-level problem determination procedures and maintains documentation to assist customer in terminal operation

2. Problem logging: Uses procedure guide for opening trouble tickets and reviews change activities log
3. Problem delegation: Determines problem area, assigns priorities, and distributes information
4. Additional duties when call center activity is low:
   - Data entry for configuration and inventory
   - Summary of active problems for problem coordinator
   - Entering change information for change coordinator
   - Monitoring of security
   - Generating management and technical reports
5. Recommends modification to procedures

External job contacts:

1. Customers
2. Vendor representatives
3. Problem and change managers
4. Network operation and technical support
5. Network administrator for trouble tickets
6. QoS and SLA managers

Qualifying experience and attributes:

1. Familiarity with functional applications and terminal equipment
2. Training in personal relationships
3. Clerical rather than technical
4. Data-entry skills
5. Problem determination know-how
6. Sensitivity to customers: Understands their business needs, has a pleasant telephone voice and language know-how

4.6.5.3 Profile of an Inventory Manager

Responsibilities

1. Manages the online configuration application, including establishment of requirements for this area
2. Maintains the network configuration
3. Maintains vendor information
4. Knows status of program and access methods used by the system
5. Maintains security of inventory control records
6. Tracks the delivery and installation of new equipment
7. Implements coordination

External job contacts:

1. Technical support
2. Network operation, change, and problem coordinators
3. Call center
4. Customers
5. Service and problem manager
6. Vendors

Qualifying experience and attributes:

1. Has knowledge of communications facilities and offerings
2. Has some knowledge of systems programming and database structure

3. Has inventory-control skills
4. Is familiar with conversion procedures and general project management

4.6.5.4 Profile of a Service and Problem Manager

Responsibilities

1. Ensures that problems are routed to proper person or function for resolution
2. Monitors status of outstanding problems via open trouble tickets
3. Enforces priorities and schedules of problem resolution
4. Maintains up-to-date problem records, which contain problem descriptions, priority, and status
5. Schedules critical situation meetings with appropriate parties
6. Fulfills administrative reporting requirements
7. Cross-organizes resources if required
8. Assumes responsibility for total communication network
9. Provides input to experience files
10. Provides input to what-if catalogs
11. Evaluates security logs
12. Evaluates SLA violations

External job contacts:

1. Vendor representatives
2. Technical support
3. Call center
4. Change manager
5. System and application programmers
6. Network operations manager

Qualifying experience and attributes:

1. Has broad knowledge of QoS metrics
2. Has broad information system and teleprocessing background
3. Has good aptitude in communication and coordination

4.6.5.5 Profile of a Service Technician

Responsibilities;

1. Provides in-depth problem determination (third-level), as necessary
2. Provides technical interface with vendors, as necessary
3. Designs and maintains up-to-date problem determination, bypass, and recovery procedures
4. Provides technical interface, as necessary, with application and system programmers
5. Uses inventory control data base
6. Ensures valid run procedures
7. Assists with network configuration/reconfiguration
8. Reads dumps from network equipment
9. Starts and evaluates special purpose diagnostics
10. Evaluates QoS metrics

External job contacts:

1. Change manager
2. Problem manager
3. Call center operator
4. Vendor technical personnel

5. Application and system programmers
6. Network infrastructure operator
7. Inventory manager

Qualifying experience and attributes:

1. Several years experience with a broad range of communication equipment and the tools, and aids necessary to maintain the network, including:
   - Network operation
   - Network-control programs
   - Access and transport networks
   - Network equipment
   - Configuration/reconfiguration procedures
2. Good aptitude for communicating with people
3. Can use diagnostic tools
4. Understands vendor standards and procedures
5. Has patience in pursuing problems

4.6.5.6 Profile of a Service-Level Manager

Responsibilities:

1. Assumes responsibility for all access and transport networks
2. Negotiates service levels
3. Evaluates service-oriented parameters
4. Provides feedback to capacity planning
5. Designs and generates service reports
6. Costs service levels and negotiates chargeback
7. Assumes responsibility for QoS management
8. Execute base lining
9. Measures QoS metrics

External job contacts:

1. Capacity planning and design
2. Operations support
3. Performance monitoring and tuning
4. Customers

Qualifying experience and attributes:

1. Has excellent communication skills
2. Has overview on communication networks
3. Understands the relationship between service level, costs, and resources utilization
4. Has knowledge of TMN

4.6.5.7 Profile of a Business Planner

Responsibilities:

1. Pursues business plans of the service provider
2. Creates portfolios
3. Characterizes and represents present networking workload
4. Projects networking workload
5. Develops installation and migration plans

6. Assigns due dates for installations
7. Supervises pilot installations
8. Prepares network design alternatives that will meet future needs
9. Produces fallback plan
10. Selects modeling tools
11. Executes benchmarks

External job contacts:

1. Business planners of the larger organization
2. Network performance analysts
3. Vendors of projection tools
4. Modeling coordinator
5. Application systems developers

Qualifying experience and attributes:

1. Communication skills
2. Knowledge of the business of the larger organization
3. Some experience in networking technologies
4. In-depth knowledge of projection techniques
5. Political skills
6. Planning and scheduling skills
7. Communication skills in working with vendors
8. Detailed knowledge of fallback procedures

4.6.5.8 Profile of Technology Analyst

Responsibilities:

1. Prepares feasibility studies for network planning and for network changes
2. Evaluates technologies, architectures and protocols
3. Makes and evaluates reference visits
4. Evaluates utilization and service level
5. Accomplishes comparative lab measurements
6. Performs hardware and software selection

External job contacts:

1. Application systems developers
2. Service-level manager
3. Communication providers
4. Vendors
5. Inventory and assets coordinator
6. Other users

Qualifying experience and attributes:

1. Communication skills
2. Overview on communication forms, services, and networks
3. In-depth knowledge of service level
4. In-depth knowledge of networking facilities and equipment

4.6.5.9 Profile of the Security Analyst

Responsibilities:

1. Defines monitoring and surveillance functions
2. Evaluates and selects security management services
3. Evaluates the performance impacts of security techniques
4. Constructs threat matrix
5. Recommends instruments to be selected
6. Supervises installation of instruments
7. Customizes passwords and access authorization
8. Programs instruments
9. Establishes procedures for securing the network management systems

External job contacts:

1. Security officer
2. Vendors
3. Security auditor
4. Other internal users

Qualifying experience and attributes:

1. Has superior personal record
2. In-depth knowledge of security management services and tools
3. Has technical skills to customize products
4. Has some communication skills toward vendors

4.6.5.10 Profile of the Operations Manager for Lawful Intercepts

Responsibilities:

1. Supervises and monitors the quality of lawful intercepts
2. Estimates cost and resource requirements
3. Reviews and approves processes and tools
4. Performs planning and scheduling of products' implementation
5. Develops, implements, and enforces procedural and security standards
6. Evaluates performance of processes, instruments and people, and reports results to management
7. Plans and directs acquisitions, training, and development projects
8. Creates and supervises SLAs together with law enforcement agencies (LEAs)
9. Defines and selects QoS metrics
10. Manages outsourcers

External job contacts:

1. LEAs at management level
2. Other managers within information systems
3. Some vendors of surveillance tools
4. External consulting companies

Qualifying experience and attributes:

1. Prior experience in statistics, surveillance technologies, accounting, legal procedures, telecommunications, or equivalent

2. Training in advanced practices, skills and concepts, administrative management, supervisory techniques, resource management, budgeting and planning
3. Excellent communication skills
4. Excellent negotiation skills
5. Excellent managerial skills

4.6.5.11 Profile of a Network Infrastructure Operator

Responsibilities:

1. Observes ongoing operations and performance to identify problems
2. Initiates corrective action where required, within the scope of knowledge and authority
3. Interprets console messages from network software or applications programs and perform required actions
4. Assists with network oriented problem determination
5. Implements backup procedures
6. Implements bypass and recovery procedures for system/network problems
7. Fulfills administrative-reporting requirements on network problems
8. Maintains communications with systems control
9. Understands monitoring technologies supporting lawful intercepts
10. Monitors all network activities
11. Uses and invokes network diagnostic aids and tools
12. Uses and provides input to data base for problem and inventory control
13. Does second-level problem determination
14. Does network start up and shutdown
15. Schedules of network activities, such as testing and maintenance
16. Understands lawful intercepts related standards
17. Identifies Intercept Access Points in the networking infrastructure

External job contacts:

1. Technical support staff
2. Configuration and inventory function
3. Problem and change coordinators of network administration
4. Customer education and customer support desk
5. Vendor representatives
6. Network administration

Qualifying experience and attributes:

1. Training in concepts of network infrastructure operations
2. At least one year of network experience with access and transport networks; lines, clusters, and terminal types; and service levels
3. Alert, intelligent, strives for efficiency
4. Can execute bypass/recovery procedures
5. Can perform authorized network alterations
6. Understands escalation procedures, problem and change management, and reporting requirements and procedures
7. Has communication skills
8. Can use various tools, depending on the availability of such tools

4.6.5.12 Profile of a Database Administrator

Responsibilities:

1. Defines data basing and data retention functions and procedures
2. Estimates data volumes
3. Determines the hierarchical layers for data archiving
4. Supervises all related processes
5. Selects, deploys, and maintains tools
6. Keeps in touch with regulatory bodies
7. Audits compliance with regulations
8. Supervises format translations
9. Prepares evidences
10. Uses intelligent analysis tools
11. Writes escalation procedures
12. Determines criteria for business continuity, backup, and restoration

External job contacts:

1. LEAs
2. Security analyst
3. Legal counsel
4. Regulatory bodies

Qualifying experience and attributes:

1. Has deep database and data warehousing knowhow
2. Has over average product knowhow
3. Has technical skills to customize products
4. Some communication skills towards LEAs and vendors
5. Has superior security record

4.6.5.13 Profile of a Legal Counsel

Responsibilities:

1. Investigates the lawfulness of LEA requests
2. Approves surveillance activities from the legal perspective
3. Maintains contacts to peers at LEAs
4. Evaluates and follows precedence cases
5. Differentiates what is going to be intercepted: call-related data or content or both
6. Differentiates between telecommunications and information services
7. Interprets the law for telecommunications service providers
8. Prepares evidences
9. Uses intelligent analysis tools
10. Protects privacy of customers of telecommunications service providers

External job contacts:

1. Legal counsels of LEAs
2. Security analyst and officer
3. Supervisor for lawful intercepts
4. Regulatory bodies
5. Courts, judges, police departments

Qualifying experience and attributes:

1. Has a superior security record
2. Legal background

3. Deep knowledge of lawful intercepts and their authorization
4. Good communication skills towards peers

4.6.5.14 Profile of a Contact Administrator

Responsibilities:

1. Understands legal basis of lawful intercepts
2. Checks compliance with regulations
3. Differentiates between intercepts-related and content-related warrants
4. Differentiates between communications and information services
5. Supervises handover of raw data or intelligence
6. Initiates intelligent analysis
7. Helps to prepare court evidence
8. Physical presence (occasionally) at surveillance actions
9. Selects tool for intelligence analysis

External job contacts:

1. Other LEAs.
2. Telecommunications service providers
3. Access providers
4. Network operators
5. Outsourcers
6. Legal counsel

Qualifying experience and attributes:

1. Superior security clearance
2. Legal background
3. Understands warrants
4. Well informed about regulations
5. Basic networking background
6. Some knowhow of lawful intercept standards
7. Managerial skills

4.6.5.15 Profile of a Manager of LEMF

Responsibilities:

1. Supervises all processes in LEMF
2. Supervises handover of raw data and intelligence
3. Maintains data
4. Distributes data to other LEAs
5. Supervises security measures
6. Decides about priorities when multiple requests from other LEAs

External job contacts:

1. Other LEAs
2. Service providers
3. Access providers
4. Network operators
5. Outsourcers
6. Legal counsel

Qualifying experience and attributes:

1. Superior security clearance
2. Legal background
3. Understands warrants
4. Basic networking background
5. Some knowhow of lawful intercept standards
6. Knowhow on data retention technologies
7. Experience of data center operations

4.6.6 Summary and Trends

Critical success factors for successful network management are processes, support tools and people. This section has addressed the importance of building and retaining the network management teams. Typical job profiles help HR departments to hire more pointedly for the function; middle management to compare existing skill levels with expectations; and top management to review the scorecard about the effectivity of subject matter experts. Outsourcing might change the network management organization by shifting certain responsibilities to overseas. India, China, and Latin America are the primary candidates for outsourced jobs. In such cases, new job profiles are going to be created for the telecommunications service providers, such as contract manager, customer experience officer, chief innovator, and chief influencer. Strategic and architectural responsibilities, however, are not going to be outsourced.

4.7 Best Practices Benchmarks for Service Providers

Kornel Terplan

The orientation towards best practices is always a good guidance. Service providers can differentiate themselves by products, services, customer support, operational efficiency, and prices. In each case, they need a continuous comparison with the industry average and with best practices. Benchmarking is well accepted discipline for that purpose.

This section starts with benchmarking, considered as a tool of best practices. Benefits, risks, myths, and key performance indicators are addressed in depth. But the main emphasis is on the process of benchmarking, consisting of data collection, comparison with best practices and industry average, gap analysis, and elaborating of recommendations for improvement.

Due to standardization, the barriers between the typical service provider and other enterprises are continuously disappearing. In both cases, IT plays a basic supporting role. Industry analysts observe that successful best practice standards, such as the **Information Technology Infrastructure l ibrary** (ITIL), Control Objectives for Information and related Technology (CoBIT), and International Organization for Standardization (ISO) 17799, known as accepted techniques by enterprises outside the telco world, are penetrating the IT and operations departments of service providers. CoBIT tells management what to monitor and control. ITIL describes how to go about implementing the processes for doing that. ISO 17799 lays out a process for securing these services and addressing legal requirements.

Scorecards and dashboards are recommended for use. They give a multiple level view of how the corporation and its division are performing. This section gives a short overview with implementation examples and preparation guidelines.

4.7.1 Benchmarking

4.7.1.1 Description of Benchmarking

Benchmarking is the comparison of a company's performance against itself and against other organizations for the purpose of organizational improvement. Although obtaining key performance

measures is important in benchmarking, the main focus of benchmarking is to understand and compare processes.

Benchmarking can be used in a variety of ways. The four basic types of benchmarking are:

- Benchmarking against internal operations, called *internal benchmarking*: In most large companies there are similar functions in different business units. The objective of internal benchmarking is to compare these internal operations and identify the internal performance standards of an organization.
- Benchmarking against external direct product competitors, called *competitive benchmarking*: Benchmarking can be done externally against competitors. Direct competitors are the most obvious to benchmark against. The objective is to compare companies in the same markets that have competing products or services or work processes, e.g., Coca-Cola vs. Pepsi.
- Benchmarking against external functional best operations or industry leaders, called *industry* or *functional benchmarking*: You can benchmark against customers, suppliers, or other companies that are in the same industry who may have the same products or services but are not competitors in the same market. Industry benchmarking tends to involve comparisons between firms that share some common technological and market characteristics and to concentrate on specific functions. For example, Telecom Australia might benchmark its billing process against the billing process of British Telecom.
- Benchmarking a process in one or several unlike organizations, called *generic* or *process benchmarking*: Some business functions or processes are the same regardless of the dissimilarities of the industries. This type of benchmarking focuses on identifying excellent work processes rather than examining the business practices of a particular organization or industry. Although generic benchmarking requires a lot of creativity, it is most effective in identifying practices that can lead to new developments and breakthroughs.

In general, benchmarking requires a large amount of effort. Many clients have the expectation that we have all the data at hand, when in the majority of cases we actually don't have the information. In reality, benchmarking is not an activity that Accenture routinely performs. For these reasons, one must be quite cautious about starting a benchmarking activity.

Budget pressures force service providers to improve the efficiency and effectiveness of the use of their networking infrastructures. The key to improvements lies in processes, tools, standards, and in human resources of operations, administration, and maintenance. In most cases, however, service providers are not clearly aware of the value of their operation: are they stronger or weaker than the industry average. Audits and benchmarks help to answer this question.

An additional target is the preparation of outsourcing decisions. There is a mutual interest for both parties, the service provider and the outsourcer, that they quantify the value of the network, its infrastructure, management, and its human resources in charge of operations, administration, and maintenance. Benchmarks can easily be incorporated into the *diligence* segment of the outsourcing process.

Thanks to the growing acceptance of audits and benchmarking, it is becoming increasingly common for service providers to get together and compare notes on how they solve problems and what the results are. Theoretically, these service providers are coming away with the stronger sense of their operations position relative to others in and out of their industry. The mediator role is usually supported by consulting companies.

Auditing and benchmarking help determine what accomplishments really exist or can be achieved and give companies the chance to match or exceed the best in the business. Now, these activities are an integral part of the total quality management program, based in most cases on ISO 9000 or TQM (Total Quality Management). ISO 9000 offers practical guidelines for how to improve the quality of conducting businesses. It became a condition of business from the European point of view. TQM has similar goals, but from the American point of view.

4.7.1.2 Advantages of Benchmarking

Benchmarking is an improvement technique that has the following main advantages:

- Comparisons inside the company allow managers to understand how their performance compares with other functional areas.
- External comparisons provide an objective basis to compare specific business functions and processes.
- It can be applied to virtually any or all areas of an organization.

In addition, each type of benchmarking has its own advantages.

- Internal benchmarking:
  - Allows functional areas to share information
  - Information is easily accessible
  - Allows organizations to obtain immediate gains by identifying their best internal practices and transferring them to other parts of the organization
  - The knowledge about internal best practices can become the baseline for later investigation and measurement involving external benchmarking partners
- Competitive benchmarking:
  - Most useful in assessing key performance measures, as it allows organizations to see their related performance.
  - Information about competitors can be used to get a burning platform and therefore apply pressure to change.
- Industry or functional benchmarking:
  - Easier to identify willing partners, since the information is not going to a direct competitor.
- Generic or process benchmarking:
- Has the potential of revealing the best of best practices.

4.7.1.3 Disadvantages of Benchmarking

Benchmarking only helps in shooting for the best practice, it does not guarantee (although it may help) leaping ahead of the competition. Benchmarking can be quite time consuming and may not produce useful information.

The disadvantages specific to each type of benchmarking are outlined below.

Internal benchmarking:

- Fosters an introverted view and it is all to easy to ignore the fact that other firms have the competitive advantage over you if you are concentrating on outperforming internal rivals.
- There is a risk of political conflict.
- It generally has limited value, as the possibility of finding large improvement opportunities is not very strong.

Competitive benchmarking:

- The main disadvantage is that information, beyond that in the public domain, is difficult to find.

Industry or functional benchmarking:

- The disadvantages are cost and the fact that the most renowned companies are beginning to feel overwhelmed with benchmarking visits and some are even charging a fee for access.

This type of benchmarking is most useful in benchmarking process designs as opposed to performance measures and strategies.

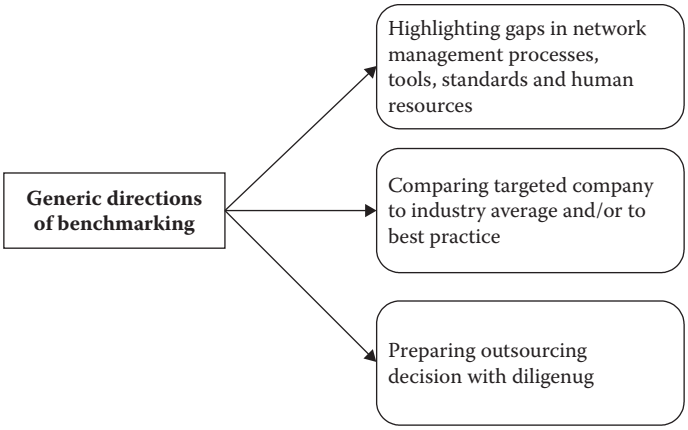
Generic or process benchmarking:

- This is generally the most difficult type of benchmarking.
- This type of benchmarking is not effective when benchmarking performance measures and strategies.

4.7.1.4 Myths and Mistakes

When doing benchmarking, one must be careful to avoid the following common mistakes:

- *Mistake 1:* Confusing benchmarking with participating in a survey. A survey of organizations in a similar industry to yours is not really benchmarking, whatever it may be called. Such a survey will give you some interesting numbers, but benchmarking is the process of finding out what is behind the numbers. In other words, a benchmarking survey may tell you where you rank, but it won't help you improve your position.
- *Mistake 2:* Thinking there are preexisting benchmarks to be found. Just because some survey or study says that a cost of $2.35 is the "benchmark" cost of a particular transaction, does not mean that you must perform that transaction for that price. The so-called benchmark may simply not be applicable to your markets, customers, or resource levels. Insist on identifying your own benchmarking partners and finding out from them what is achievable, and then whether you can achieve a similar level of performance.
- *Mistake 3:* Forgetting about service delivery and customer satisfaction. Benchmarking stories abound of organizations that have become so fixated on the cost of providing their product or service that they have failed to take the customer into account. Paring down the costs often rebounds in lesser service delivery, so customers go elsewhere and ultimately you don't have a business. Take a "balanced scorecard" approach when developing your benchmarking metrics.
- *Mistake 4:* The process is too large and complex to be manageable. A process is a group of tasks. A system is a group of processes. Avoid trying to benchmark a total system—it will be extremely costly, take ages, and be difficult to remain focused. It is better to select one or several processes that form a part of the total system, work with it initially, and then move on to the next part of the system.
- *Mistake 5:* Confusing benchmarking with research. Benchmarking presupposes that you are working on an existing process that has been in operation long enough to have some data about its effectiveness and its resource costs. Commencing a new process, such as developing a new employee handbook by collecting other people's handbooks and taking ideas from them, is research, not benchmarking.
- *Mistake 6:* Misalignment. Choosing a benchmarking topic that is not aligned with the overall strategy and goals of the business is misalignment. A Lead Team at the strategic level needs to oversee the benchmarking project and make sure that it is in line with what is happening in the business as a whole.
- *Mistake 7:* Picking a topic that is too intangible and difficult to measure. Employee communication is probably the most slippery concept that exists in an organization, but it is often cited as one of the worst problems, so many organizations try to benchmark it. Encourage your benchmarking team to select instead a part of the topic that can be observed and measured; for instance, the process of distributing memos around the organization.
- *Mistake 8:* Not establishing the baseline. When participating in a benchmarking partnership, you must analyze your own process and its level of performance thoroughly. After all, that information is what you have to offer to your benchmarking partners in exchange for the information you are seeking from them. Make sure your benchmarking team is very clear about what it wants to learn, before you approach potential benchmarking partners.
- *Mistake 9:* Not researching benchmarking partners thoroughly. This is essential in selecting the right benchmarking partners, so you don't waste their time or yours. There is a rule of



FIGu RE 4.7.1 Benchmarking directions.

benchmarking etiquette that says you should never ask a benchmarking partner a question that you should have been able to answer for yourself through researching the literature in the public domain.

- *Mistake 10:* Not having a code of ethics and contract agreed to with partners. Your partners should be clear about what you are seeking to learn from them, how that information will be treated, who will have access to it, and for what purposes it will be used. Ideally, this should be formally agreed. The benchmarking code of practice offered by the American Productivity and Quality Center provides a useful model.

4.7.1.5 Network Management Benchmarks

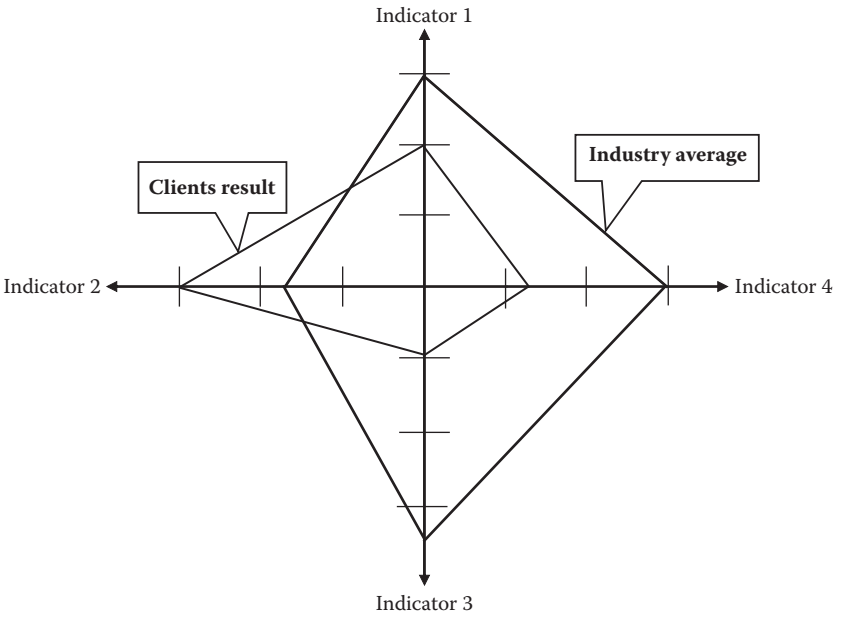
Budget pressures force network managers to improve the efficiency and effectiveness of the use of their communication networks. The key to improvements lies in processes, instruments, standards, and in human resources of network management. In most cases, however, network managers are not fully aware of the value of their operation: are they stronger or weaker than the industry average. Network management audits and benchmarks help to answer this question.

Network management means deploying and coordinating resources in order to plan, operate, administer, analyze, evaluate, design, and expand networks to meet service level objectives at all times at a reasonable cost, and with reasonable capacity.

Basically, network management benchmarks can be utilized for the following three purposes (Figure 4.7.1):

- Highlighting gaps in network management processes, tools, standards, and human resources
- Comparison with industry average or with best practice
- Preparation of additional outsourcing decisions to service providers

In the first two cases, indicators are needed to quantify performance. Benchmarking requires the use of various data collection techniques, such as forms, to evaluate what network management processes are supported, what instruments are used, what management protocols have been implemented, how human resources are assigned to processes and instruments, and what are the skill levels of the management team. In order to get feedback about the networking environment, investments made, and process details, three different questionnaires are used. They can be filled in prior to or during the benchmarking process. In order to quantify and compare the quality of network management, a number of benchmarking metrics are used, grouped around generic, organizational, process specific, and costs indicators. The results may be used by the overseer to identify areas of improvements or to compare performance with the industry average or with the best practice. Usually, benchmarking clients are interested to do both.



FIGu RE 4.7.2 Typical benchmarking star.

The third target is the preparation of outsourcing decisions. There is a mutual interest for both parties—the client and the outsourcer; they quantify the value of the network or network segment, its management, and its human resources who are in charge of management. Network management benchmarks can easily be incorporated into the diligence phase of the outsourcing process.

Thanks to the growing acceptance of audits and benchmarking, it is becoming increasingly common for companies to get together and compare notes on how they solve problems and what the results are. Theoretically, these companies are coming away with the stronger sense of their operations position relative to others in and out of their industry. The mediator role is usually played by consulting companies.

Figure 4.7.2 shows an example of benchmarking results, using four metrics to compare the clients' performance with the industry average.

The metrics are:

- *Indicator 1:* Number of proactive fault detection in comparison to all faults (proactive means the use of intelligent, continuous monitoring in order to quickly recognize deteriorating equipment or facility status)
  - Industry average: 60%
  - Own company: 40%
- *Indicator 2:* Percentage of multiple responsibilities for network management functions (multiple responsibilities risk efficiency by finger pointing if serious or subtle faults are identified)
  - Industry average: 35%
  - Own company 60%
- *Indicator 3:* Percentage of single referrals in comparison to all faults (usually, multiple referrals or with other words, multiple persons are involved in fault resolution process—the more referrals, the longer the resolution process)
  - Industry average: 80%
  - Own company: 20%

- *Indicator 4:* Percentage of correctly completed change request forms in comparison to all change requests (incomplete forms delay the change execution process and risk successful execution)
  - Industry average: 60%
  - Own company: 20%

The indicated shaded area in Figure 4.7.2 is a clear sign for the company of where and how to improve.

The basic idea of auditing and benchmarking is to identify several organizations that represent best practices in the same functions and services where the interested company needs improvement. After the identification of comparative companies, several meetings are arranged, research done, questions answered, and potential improvements are highlighted, in most cases, to the mutual benefit of the participants.

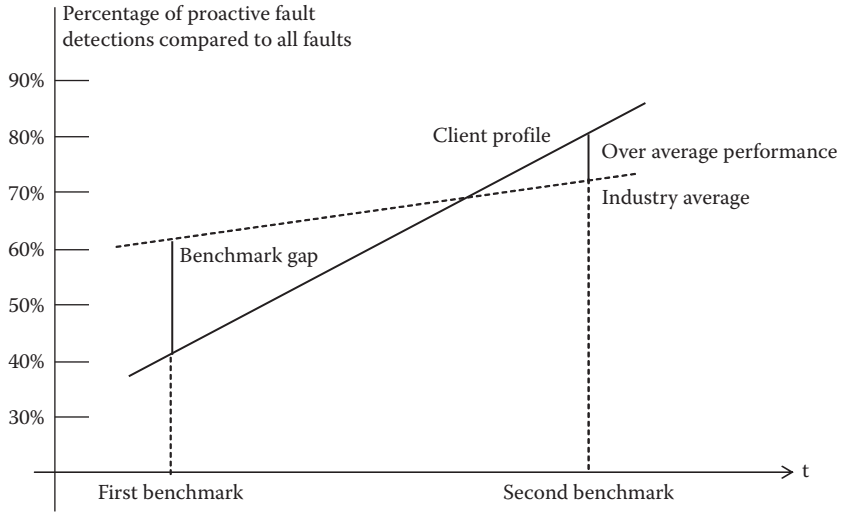
An additional goal is to prepare outsourcing decisions. Network management audits and benchmarks give a good estimate on the value of existing network management processes, instruments, and personnel. Besides the actual value, top management receives detailed reports on the following items:

- Inventory of network components
- Inventory of network management instruments
- Organization structure of network management
- Skill levels, ages, and salary ranges for network management personnel
- Allocation of functions and instruments to network management personnel
- Statement of mission or list of objectives

Benchmarking helps to observe dynamic changes in the performance of certain network-related indicators. Figure 4.7.3 shows the dynamic of one special indicator (percentage of proactive fault detection in comparison to all faults); starting at lower than the industry average at the initial benchmarking study and overtaking the industry average at the second benchmarking study. This performance change can be influenced by the level of investments into network management processes, instruments, and education.

Benchmarking is of course not without problems. A number of items may cause problems and impact the results of the benchmarks. The most important in this respect are:

- The industry average does not contain enough samples and thus the comparison with the specific results of the company is not representative.



FIGu RE 4.7.3 Dynamic changes between two benchmarks.

- Errors in measurements due to not using the right tools or not using the right settings of parameters of the tool.
- Dishonest or biased interview partners can skew practically all answers in the questionnaire and in interviews making filtering out the true answers on behalf of the benchmarking team almost impossible. All interviews to be conducted in the Network Operating Centers (NOC) could prove just the opposite. Each person is willing to help and is motivated for improvements.
- Lack of documentation on processes, functions, tools, and protocols delays the data collection phase significantly. It may substantially delay the execution of the benchmarking process.
- Obsolete documentation may lead the benchmarking work in a false direction. Usually, the team is warned too late about the out-of-date status of the documentation.
- False interpretation of certain metrics may be caused by the specific use of certain metrics within the company. If the team is informed about it, ad-hoc redefinitions may solve this problem easily.
- The nonapplicability or nonavailability of monitors delay the data collection phase considerably. Manually collected data are no replacement for accurate monitored data.
- The overseer and his/her organization may not permit observations of operations in end-to-end mode; in particular, results may be significantly impacted when shifts takeovers, performance of the client-contact point, and remote diagnostics cannot be analyzed end to end.

Most likely, all of these negative impacts will not happen during the same benchmark, but some of them might. In order to avoid them, each should be addressed during the very first meeting between the company and the benchmarking team.

4.7.1.6 Benchmarking Phases

The principal phases of network management audits and benchmarks are the following:

4.7.1.6.1 Data Collection

- Documentation (analysis of topologies, managed objects, performance reports, network management instruments, closed trouble tickets)
- Interviews with various persons in the operations, administration, and maintenance organization (what are the major problems, what are feasible solutions, job assignments, education, skill levels)
- Observation of operational efficiency (reaction to trouble calls, how are support, documentation and management tools used, support quality in various shifts, shifts takeover process and documentation)

4.7.1.6.2 Comparison with Best Practice and Industry Average

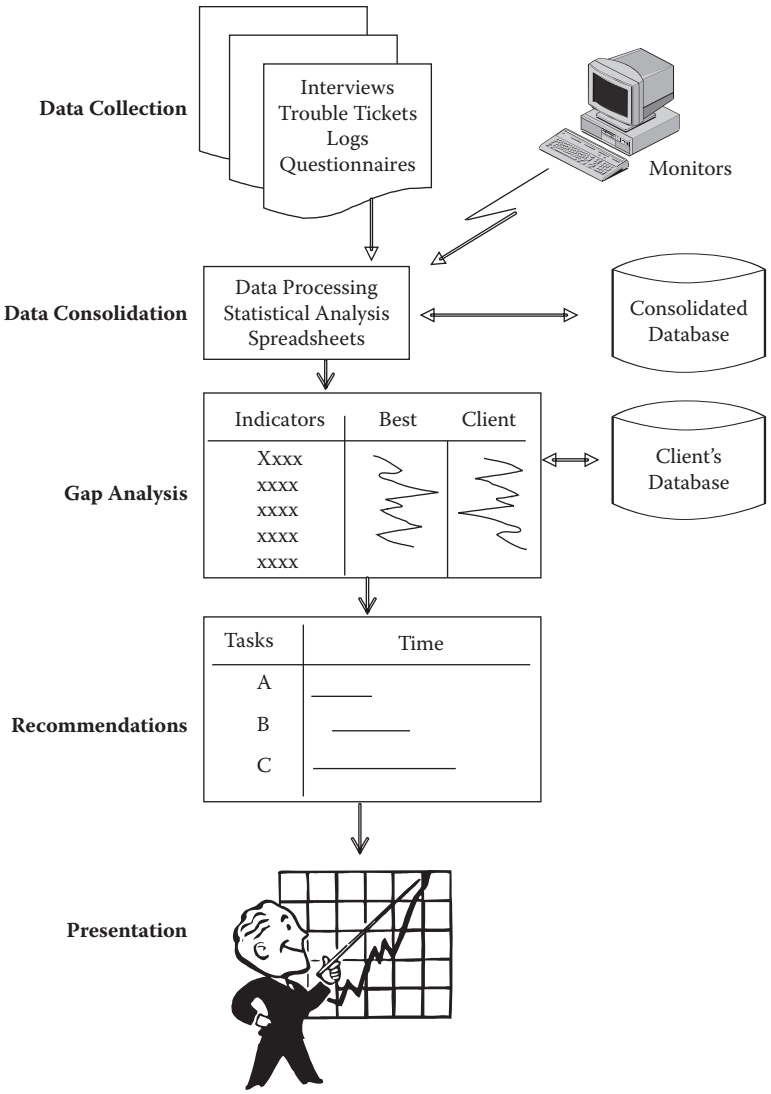
This phase includes just the specific branch or cross-branch evaluation. Usually, cross branch indicators are used, first. Results are usually visualized and may indicate considerable differences in certain areas between industry and client results.

4.7.1.6.3 Gap Analysis

This phase provides real details and addresses the functions of all business processes of service providers for customer care, billing, order processing, provisioning, and network operations management. Also tools, such as support, management and documentation systems, monitors, analyzers, element management systems, integrators, and modeling tools are addressed in depth. Also the assignment of human resources to processes and tools are investigated in depth. Recommendations to avoid redundancies are also included.

4.7.1.6.4 Elaborating the Recommendations

The gaps identified in the previous phase are the basis for improvements or for supporting the decision for outsourcing certain business processes. If insourcing, the tasks are: setting priorities and timeframes,



FIGu RE 4.7.4 Process of benchmarks.

and estimating the demand for reengineering business processes, acquiring tools, and hiring or cross-educating human resources.

Benchmarking operations of service providers is not a trivial exercise. Usually, small teams of 2–5 persons are assigned to the benchmarking team. There are two different types of persons who conduct benchmarks: senior consultants and analysts. Skill levels overlap in certain areas, but usually both types of persons do have specific areas of experiences. Benchmarking definitively requires teamwork. Figure 4.7.4 shows the process of benchmarks.

Benchmarking can be used in the following circumstances:

- When establishing effective goals and objectives to become competitive with the best in the industry
- When assessing the relative cost and performance of key business functions
- When identifying company strengths and weaknesses and pinpointing the areas that need improvement

Benchmarking can be applied throughout the Business Integration Methodology. Typically, the most useful comparisons are made during the planning phase, and the early stages of delivering (capability analysis). As mentioned before, benchmarking can be a very time consuming task and whether it is appropriate to do a rigorous benchmarking during the above phases must be determined based on time availability.

Inputs to benchmarking include:

- Industry-specific background information:
  - Literature searches
  - Annual reports
  - Government reports
- Originator company:
  - Internal interviews
  - Internal data
- Benchmark participants:
  - Interviews, questionnaires
  - Site visits
- Other benchmarking studies:
  - General industry studies, environmental scans (e.g., accounting firms/other consulting companies)
  - Other benchmarking/industry ranking studies
  - Industry case studies
- Information in wikis and blogs:
  - Previous presentations
  - In-house interviews
- Case studies

Outputs from benchmarking include:

- Presentation:
  - Diagnostic; where the originator stands; correlation with industry
  - Conclusions that can be drawn from best practices:
    - Best processes
    - Best customer service
    - Best quality
    - Empower teams, etc.
  - Next steps—areas for improvement, changes to future strategy
- Competitor results:
  - Release generic (no names) diagnostic, highlighting participants positions for comparison to industry's "best"

4.7.1.7 Applicability of Benchmarks

The search for best practices can be mapped against two dimensions—process and industry. This provides a framework for determining the focus of benchmarking and the expected level of improvement (see information below).

4.7.1.7.1 Same Process, Same Industry

This provides a basis for direct comparison, especially with direct competitors. Therefore, the results of such a benchmarking exercise can be very valuable and directly applicable. However, in order to maintain competitive advantage, competitors may be reluctant to participate in such exercises and therefore it may be very difficult to get accurate and relevant data.

4.7.1.7.2 Same Process, Different Industry

This is probably the most widely used approach for identifying best practices. This is due to the difficulty in getting competitor data discussed above, and because it fosters innovation. The objective is to ensure processes are optimized by adopting practices used in different industries. For example, the stock control process of an automotive manufacturer may be directly improved in terms of speed by adopting stock control practices from the food retail industry.

4.7.1.7.3 Different Process, Same Industry

The emphasis of this approach is to learn and adapt practices from other processes. In addition, it fosters innovation and integration by bringing different stakeholders together. For example, the use of new technology in distribution management by a supplier may be reused or adapted to improve lead time and hence improve the order-to-cash process.

4.7.1.7.4 Different Process, Different Industry

This can result in substantial competitive advantage for the first to adopt it. This approach provides maximum room for innovation and it usually is the result of an innovation workshop. For example, library management systems using bar code technology pioneered in the food retail industry.

It is important to emphasize that benchmarking does not have to be a comparison of metrics. From a best practice view point, it is more important to understand how others achieve results.

4.7.1.8 Support Tools for Benchmarking

The following generic tools are available to collect, store, process, and report data and information:

- Manual logs—handwritten notes
- Event logs (e.g., active components, such as servers, switches, routers register processes, inquiries, transfers, etc.—they store these events, which may be pulled periodically)
- Workforce manager (e.g., time stamps of process records may be utilized to evaluate the timeliness of processes)
- Trouble tickets (e.g., special products for opening, progress control, and closing of incidents, problems, and other specific events in service areas)
- Case-based-reasoning records (e.g., consolidated trouble tickets, sorted by symptoms including solutions by resource)
- Manual reports—observations in written form, including individual opinions on certain system and network events
- Observations
- Automated Call Distributors (ACD) (e.g., at the help desk, incoming calls and inquiries are distributed to the next available agent; in addition, very simple statistical data are collected and processed)
- Interactive Voice Response (IVR) (e.g., during answering inquiries, simple questions are answered by using synthetic voice out of customer files)
- Management systems (e.g., management framework with basic services and many management applications)
- Element Managers (e.g., switches, routers usually have a dedicated manager with limited scope and functionality; this manager has been implemented by the supplier of the systems and/or networking component)
- Simple Network Management Protocol (SNMP) agents (e.g., they can be implemented in all possible managed objects; they collect Protocol Data Units [PDUs] for Management Information Base [MIB] that are polled and processed)
- Remote Network Monitoring (RMON) agents (e.g., they can be implemented in all possible networking objects; they collect data for MIBs that are polled and processed)

- Probes (e.g., hardware or software probes, implemented in managed objects or around managed objects; they collect data for MIBs that are polled and processed)
- Server monitors (e.g., special tasks for servers are executed)
- Service monitors (e.g., special tasks for services are executed)
- Application server monitors (e.g., special tasks for the application layers are executed)
- Security server monitors (e.g., special tasks of authorization and authentication are executed)
- Intrusion detection tools (e.g., special tasks of detecting intrusions are executed)
- Intrusion prevention tools (e.g., special tasks preventing intrusions are executed)
- Virus detection tools (e.g., special tasks of detecting and eliminating virus attacks)

4.7.2 ITIL (Information Technology Infrastructure Library)

This public library represents a Best Practice Framework consisting of 60 books. The ITIL is supplier independent; it exists for the selection, implementation, and operation of IT solutions, and defines this kind of support of IT services.

Processes, functions, roles, responsibilities, and implementation elements are defined; they are the basis and prerequisites for efficient and effective IT operations.

The central elements of ITIL are subdivided into the principal areas:

- Customer orientation
- Process orientation
- Quality orientation

ITIL provides alternatives for the "what" questions; in other words, content, process, and goals for the IT organization. Answers for the "how" questions depend on the concrete IT environment of enterprises. ITIL is a guideline and blueprint at the same time; it leaves a lot of room for customization and individual implementation of its modules.

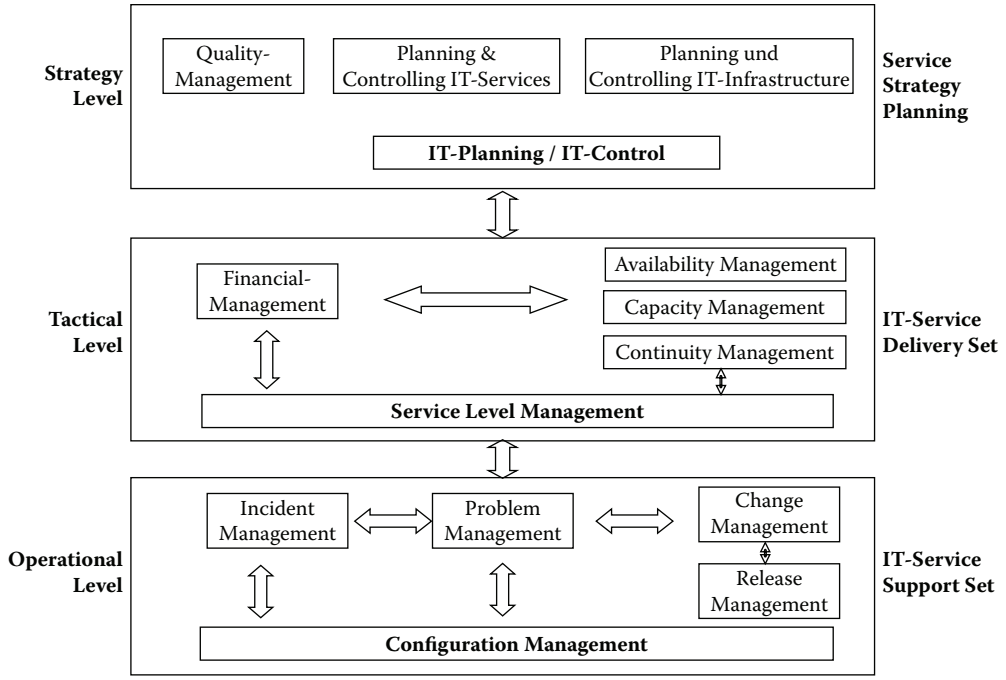
The ITIL compendium consists of a large number of SETs that address various topics and customers. There are three sets with 10 principal processes (Figure 4.7.5):

- Service strategy planning set
- IT service delivery set
- IT service support set

The service support set concentrates on permanent tasks of IT organizations. They are basic processes for IT operations. The processes are heavily interrelated.

On the operational layer, incidents, orders, and inquiries are handled by Incident Management at a central site (User Help Desk). They are accepted, classified, and described in more depth. Using an identification process, incidents without clear understanding of their causes are classified as problems. Problem Management deals with finding solutions for these problems by addressing multiple service layers. Solutions of problems may require the modification of the IT infrastructure. Change procedures are supported by Change Management. Attributes of all objects of the IT infrastructure and their changes are maintained by Configuration Management. Checking feasibility, testing, and rollout of hardware and software components are supported by Release Management. All these processes are interrelated with each other.

- *Incident Management with Service Desk:* This process is the central interface to customers with the result that is also responsible for the good communication between the IT organization and customers. The Service Desk accepts messages about outages, breakdowns, service impacts, and complaints, and forwards them to Incident Management. Incident Management diagnoses and classifies. The overall goal is to be back to normal operation conditions. Incident Management is not responsible to find the causes of service interruptions. It is supporting the communication



FIGu RE 4.7.5 The ITIL model**.**

with customers and as such it is part of an "Image Building" function of the IT organization. First of all, it assures the continuity of business processes by restoring services after service interruptions.

- *Problem Management:* Central goal is a systematic problem analysis and restoration by identifying the real causes of problems. Problems are identified, registered, and classified. They will be analyzed, diagnosed, and finally eliminated with the overall goal of reducing the total number of incidents middle and long range. Also additional functions are supported, such as providing information to management on solution rates and impacts of existing problems, recognizing trends of events and problems, and finally providing advice for management about proactive corrective and preventive measures.
- *Change Management:* This area is dealing with all kinds of changes in the IT environment. These changes must be implemented efficiently, on time, with reasonable costs and with minimal risks for the IT infrastructure and for the IT service scenario. A missing change management is the most frequent source of problems for IT. The majority of problems within IT can be allocated to insufficient (lack of planning and tests) change management functions. In order to avoid interruptions and deviations in service quality due to change management inefficiencies, change requests should be classified and approved first. Their impacts should be checked carefully before being planned and executed. The most important tasks of change management are: investigations, analysis, and decision making. The right organizational structure and subject matter experts are key prerequisites for successful change management. The Change Manager coordinates and estimates risks, costs, and impacts. A so-called Change Advisory Board clears eventual conflicts. Due to the central role and position of Change Management, this process maintains direct communication relationships with all other ITIL modules. The Service Desk should get status reports on changes; the desk can then inform the customers.
- *Release Management:* This process is responsible for checking readiness of hardware and software components for a seamless and on time rollout. Releases are expected to be planned, and rollouts

to be coordinated. Impacted customers must be informed and trained. If completed, components can be distributed, and hardware and/or software properly installed. When the IT organization does everything right, it is assured that just tested, and released versions of software are being installed for customers.

- *Configuration Management:* This process collects, supervises, and maintains all IT configuration elements. This is the basis for secured and actual information delivery for all IT projects. This process can forward these IT elements to other processes. These IT objects are maintained in the Configuration Management Data Base (CMDB). Thus, configuration items can be identified and grouped into classes with common attributes. Status is continuously updated; responsibilities are clear, as well. There is also information available how objects may be used. On the basis of this CMDB, all other ITIL processes can be improved and their efficiency increased.

The service delivery set concentrates on tactical issues, such as Service Level Management, Financial Management, Capacity Management, Availability Management, and Continuity Management.

- *Service Level Management:* This process describes the qualitative and quantitative management of IT services being provisioned by the IT organization for customers. This process controls service agreements and service quality by comparing actual and agreed upon services. Service Level Management is positioned central within ITIL. It is connected to all other modules of ITIL. It is, as a result of this central position, one of the most important processes. Service-level agreements (SLAs) determine objectives and metrics for IT services and are responsible for making IT more measurable. SLAs bring customer requirements and cost of services closer to each other. It is important to notice that Service Level Management is a continuous process; measurable service goals are determined and their fulfillment is controlled. Formal negotiations of SLAs are not enough. It is more about a service culture and customer orientation. It is recommended to define and distribute roles and responsibilities. The Service Level manager supports the service orientation of the service organization.
- *Financial Management:* This process is responsible for controlling the costs, for charging of fulfilled IT services, and for the financially healthy IT organization. Prerequisites are the clear structure of costs and the resource demand of the IT service. On this basis, customers may be charged on the usage of services and resources. The benefits of such a cost management are obvious: visible costs for IT services, profitable operations within the IT organization, and better predictions of impacts of strategic decisions about new IT services.
- *Capacity Management:* This process is responsible for the provisioning the right level of capacity to meet present and future IT resource requirements. It determines the right capacity that can be realized with reasonable costs to meet requirements of existing and future SLAs. One of the tasks is to estimate and predict future demand on IT resources. Demand Management represents this function. Workload Management analyzes the resource demand by applications. Resource Management elaborates a resource plan that serves as the basis for the capacity plan. All information is put together into a Capacity Management database. This ensures an objective planning for all capacities, their supervision, and for generating analysis reports.
- *Availability Management:* Enterprises depend today to a large extent on IT. The overall goal is to ensure reasonable availability to customers. Central planning functions identify and analyze availability requirements, determine availability, define security concepts, and elaborate an availability plan. All actual availability related metrics of IT components are collected, evaluated, and used for checking on the compliance to existing SLAs.
- *Continuity Management:* This process is a special subset of Availability Management. It concentrates on restoring all relevant IT services as soon as possible in case of unexpected exceptional situations. Reactions and escalation procedures are defined for unexpected IT service interruptions. As part of Risk Management, potential threats are analyzed to help with the right reactions in case of serious emergencies.

The importance of efficiency and customer orientation of IT organizations is continuously increasing. These organizations support business processes of enterprises. ITIL is a guideline and a tool the same time for a process-oriented structure of IT Service Management. It is a blueprint for IT organizations to offer services with measurable quantity, quality, and timeliness to their customers.

4.7.3 ISO/IEC 17799

ISO/IEC 17799 is a framework for information security management published by the International Organization for Standardization and the International Electrotechnical Commission. It specifies best practices for security in 10 areas and offers guidance on such topics as protecting personal data, internal information, and intellectual property.

The guiding principles are the initial point when implementing information security. They rely on either legal requirements or generally accepted best practices.

Measures based on legal requirements are (among others):

- Protection and nondisclosure of personal data
- Protection of internal information
- Protection of intellectual property rights

Best practices mentioned are:

- Information security policy
- Assignment of responsibility for information security
- Problem escalation
- Business continuity management

When implementing a system for information security management, several critical success factors should be considered:

- The security policy objectives and activities reflect the business objectives.
- The implementation considers cultural aspects of the organization.
- Open support and engagement of senior management are required.
- Thorough knowledge of security requirements, risk assessment, and risk management is required.
- Security targets all personnel, including management.
- Security policy and security measures are communicated to contracted third parties.
- Users are trained in an adequate manner.
- A comprehensive and balanced system for performance measurement is available and supports continuous improvement by giving feedback.

The guidance is structured into 10 sections. Information security should, following the standard, consider at least the following parts:

4.7.3.1 Security Policy

- An information security policy should define the direction and contain the commitment and the support of management.
- The policy should be communicated throughout the organization.

4.7.3.2 Organizational Security

The definition of adequate organization structures for the management of information security within the organization should include:

- A management information security forum
- A forum for coordination

- Assignment of responsibility for information security to individuals
- Definition of responsibility areas for managers
- Definition of an authorization process for IT facilities
- Definition of responsibility for investigation of security-relevant know-how
- Defined range for cooperation with third parties as well as independent security reviews
- Comprehensive measures should exist for management of third-party services (definition of risks and security requirements).
- Risks caused by outsourcing contracts should be managed

4.7.3.3 Asset Classification and Control

The inventory of assets and the assignment of the responsibility should be seen as a prerequisite to sound accountability for assets.

Information should be classified following a generally accepted system, thus ensuring an appropriate level of protection.

4.7.3.4 Personnel Security

- Security responsibilities, confidentiality agreements, and the contract of employment should be part of the job responsibility.
- Adequate controls for personnel screening should be in place.
- Information security education and training should increase users' security awareness.
- The process of reporting security incidents, weaknesses, and software malfunctions should be defined. This should include assessing the adequacy of the controls, implemented by a permanent process of learning from incidents.

4.7.3.5 Physical and Environmental Security

- Central equipment should be installed only within a secure area, where adequate access controls and damage prevention are implemented. These areas include offices, rooms, and facilities. There is also a need for special attention to delivery and loading areas.
- Equipment should be protected against loss, damage, or compromise by being sited and protected in an appropriate manner. Power supplies, an adequate level of cabling security, and correct maintenance of the equipment should be in place.
- Equipment installed off-premises and disposal or reuse of information should be considered.
- General controls (such as a clear desk and clear screen policy) to protect information processing facilities or to prevent damage caused by unauthorized offsite usage of equipment should be in place.

4.7.3.6 Communications and Operations Management

- Operations should follow documented procedures.
- All changes to equipment should be documented.
- Procedures for sound incident management should be defined.
- Duties should be segregated, ensuring that no individual can initiate and authorize an event.
- Development and operational facilities should be separated.
- Risks caused by contracted external facilities organizations should be covered.
- Capacity demands should be observed, and future demands should be projected.
- Acceptance criteria for new systems should be defined.
- Damage caused by malicious software should be prevented, using preventive and detective controls, formal policies, and defined recovery procedures.
- Information should be backed up and the backup files tested regularly.
- Activities performed by operational staff and errors should be logged.
- Networks should be set up and managed with a view to ensuring the necessary level of security.

- Removable media should be handled with special care.
- Media with sensitive information should be disposed of in a secure manner.
- Adequate controls in information handling procedures (e.g., labeling of media, ensuring completeness of inputs, and storage of media) should be considered.
- System documentation should be protected, as it may contain sensitive information.
- Agreements for the exchange of information and software should be established, including media in transit, electronic commerce transactions, electronic mail, electronic office systems, publicly available systems, and other forms of information interchange.

4.7.3.7 Access Control

- Access to information should be granted in accordance with business and security requirements.
- A formal access control policy should be in place.
- Access control rules should be specified.
- User access management (registration, privilege management, password management, review of user access rights) should follow a formal process.
- Responsibilities of users should be clearly defined.
- Networked services, operating systems, and applications should be protected appropriately.
- System access and use should be monitored constantly.
- Mobile computing and teleworking should be performed in a secure manner.

4.7.3.8 Systems Development and Maintenance

- Security issues should be considered when implementing systems.
- Security in application systems should take into account the validation of input data, adequate controls of internal processing, message authentication, and output data validation.
- Use of cryptographic systems should follow a defined policy.
- Access to system files (including test data and source libraries) should be controlled.
- Project and support environments should allow for security by being rigorously controlled (e.g., change management procedures, arrangements for outsourced development).

4.7.3.9 Business Continuity Management

- A comprehensive business continuity management process should permit prevention of interruptions to business processes.
- The business continuity management process should not be restricted to IT-related areas and activities.
- An impact analysis should be executed, resulting in a strategy plan.
- Business continuity plans should be developed, following a single framework.
- Business continuity plans should be tested, maintained and reassessed continuously.

4.7.3.10 Compliance

- Any unlawful act (e.g., data protection acts) should be avoided.
- Compliance with the security policy should be ensured by periodic reviews.

4.7.3.11 Business Drivers That Call for Implementation of the Standard

- Definition of an information security management system, applying best practice in security management based on a systematic approach
- Identification of critical assets via the business risk assessment
- Enhancement of the knowledge and importance of security-related issues at the management level
- Definition of responsibility and organizational structures for information security
- Need for a basis for certification of the information security management system
- Need for contractual relationships

4.7.4 CoBIT (Control Objectives for Information and Related Technology)

While many organizations recognize the potential benefits that technology can yield, successful ones also understand and manage the risks associated with implementing new technologies. Among the challenges service providers and enterprises face, the following ones are the most important:

- Aligning IT strategy with the business strategy
- Cascading strategy and goals down into the enterprise
- Providing organizational structures that facilitate the implementation of strategy and goals
- Insisting that an IT control framework be adopted and implemented
- Measuring IT's performance

Effective and timely measures aimed at addressing these top management concerns need to be promoted by the governance layer of the enterprise.

Organizations must satisfy the quality, fiduciary, and security requirements for their information, as they do for all other assets. Management must also optimize the use of available resources, including data, applications systems, technology, facilities, and people. To discharge these responsibilities, as well as to achieve objectives, management must understand the status of its own IT systems and decide what security and control should be provided.

CoBIT is IT governance, a control framework, and a maturity model. Its purpose is to ensure that IT resources are aligned with an enterprise's business objectives so that services and information, when delivered, meet quality, fiduciary, and security needs. It is also intended to provide a mechanism to balance IT risks and returns. CoBIT defines 34 significant processes, links 318 tasks and activities to them, and defines an internal control framework for them all.

CoBIT can be used by business or IT management, but its origins are in auditing. It was developed by the Information Systems Audit and Control Association, which is an international organization based in the United States. More recently, the IT Governance Institute has made some contributions. CoBIT is often introduced in an enterprise via the audit route. As a result, IT managers often view CoBIT as a threat to their positions rather than a useful and powerful framework for communicating effectiveness and value for their companies.

CoBIT processes and control objectives are segmented into four domains (GARD02):

- Planning and Organization (PO)
  - PO1: Defines strategic IT plan
  - PO2: Define the information architecture
  - PO3: Determine the technological direction
  - PO4: Define the IT organization and relationships
  - PO5: Manage the IT investment
  - PO6: Communicate management aims and directions
  - PO7: Manage human resources
  - PO8: Ensure compliance with external requirements
  - PO9: Assess risks
  - PO10: Manage projects
  - PO11: Manage quality
- Acquisition and Implementation (AI)
  - AI1: Identify automated solutions
  - AI2: Acquire and maintain application software
  - AI3: Acquire and maintain technology infrastructure
  - AI4: Develop and maintain IT procedures
  - AI5: Install and accredit systems
  - AI6: Manage changes

- Delivery and Support (DS)
  - DS1: Define and manage service levels
  - DS2: Manage third-party services
  - DS3: Manage performance and capacity
  - DS4: Ensure continuous service
  - DS5: Ensure system security
  - DS6: Identify and allocate costs
  - DS7: Educate and train users
  - DS8: Assist and advice customers
  - DS9: Manage the configuration
  - DS10: Manage problems and incidents
  - DS11: Manage data
  - DS12: Manage facilities
  - DS13: Manage operations
- Monitoring (M)
  - M1: Monitor the processes
  - M2: Assess internal control adequacy
  - M3: Obtain independent assurance
  - M4: Provide for independent audit

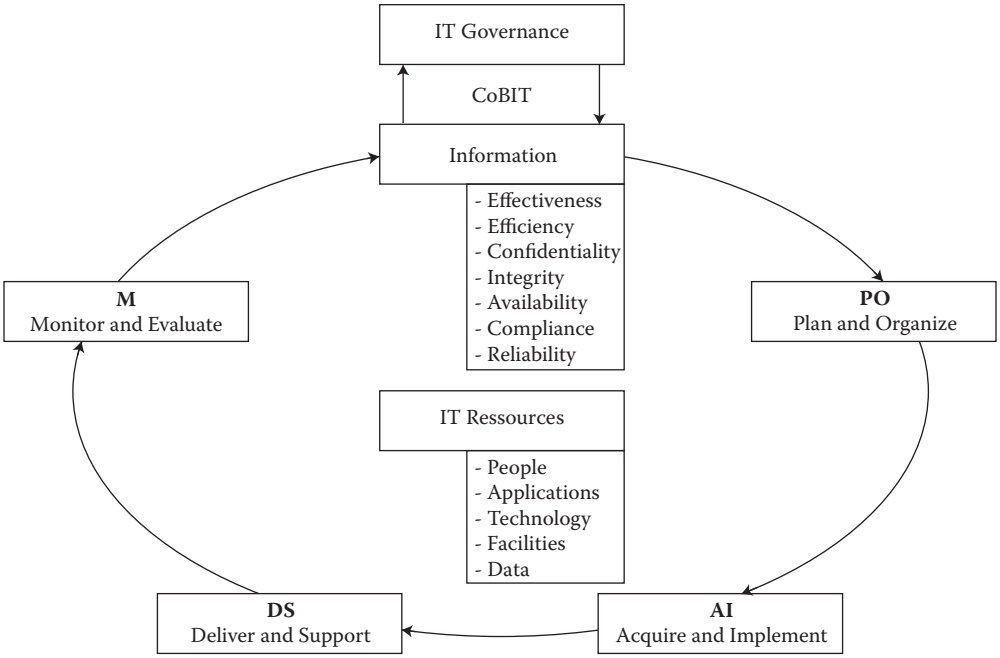
CoBIT is based on established frameworks, such as the Software Engineering Institute's Capability Maturity Model, ISO 900, and IT Infrastructure Library (ITIL). CoBIT does not include control guidelines or practices, which are the next level of detail. Unlike ITIL, CoBIT does not include process steps and tasks because it is a control framework rather than a process framework. CoBIT focuses on what an enterprise needs to do, not how to do it, and the target audience is auditors, senior business management, and senior IT management.

Figure 4.7.6 shows CoBIT IT processes defined within the four domains (ITGI06), which are characterized as:

- Quality requirements
  - Effectiveness: Deals with the relevance and pertinence of information to the business process as well as the timely, correct, consistent, and usable delivery
  - Efficiency: Concerns the provision of information through the optimal (most productive and economical) use of resources
- Security requirements
  - Confidentiality: Concerns the protection of sensitive information from unauthorized disclose
  - Integrity: Relates to the accuracy and completeness of information, as well as to its validity, in accordance with business values and expectations
  - Availability: Relates to information being available when required by the business process now and in the future. It also concerns the safeguarding of necessary resources and associated capabilities
- Fiduciary requirements
  - Compliance: Deals with complying with those laws, regulations, and contractual arrangements to which the business process is subject, i.e., externally imposed business criteria
  - Reliability: Relates to the provision of appropriate information for management to operate the entity and for management to exercise its financial and compliance reporting responsibilities

Following the CoBIT definition, the resources used by IT are identified as:

- Data: Objects in their widest sense (e.g., external and internal), structured and nonstructured, graphics, sound, file, record, etc.



FIGu RE 4.7.6 CoBIT IT processes defined within the four domains.

- Applications: Understood to be the sum of manual and programmed procedures
- Technology: Hardware, operating system, database, networking, switches, routers, cabling, multimedia, etc.
- Facilities: All resources to house and support information systems
- People: Staff skills, awareness, and productivity to plan, organize, acquire, deliver, support, monitor, and evaluate information systems and services

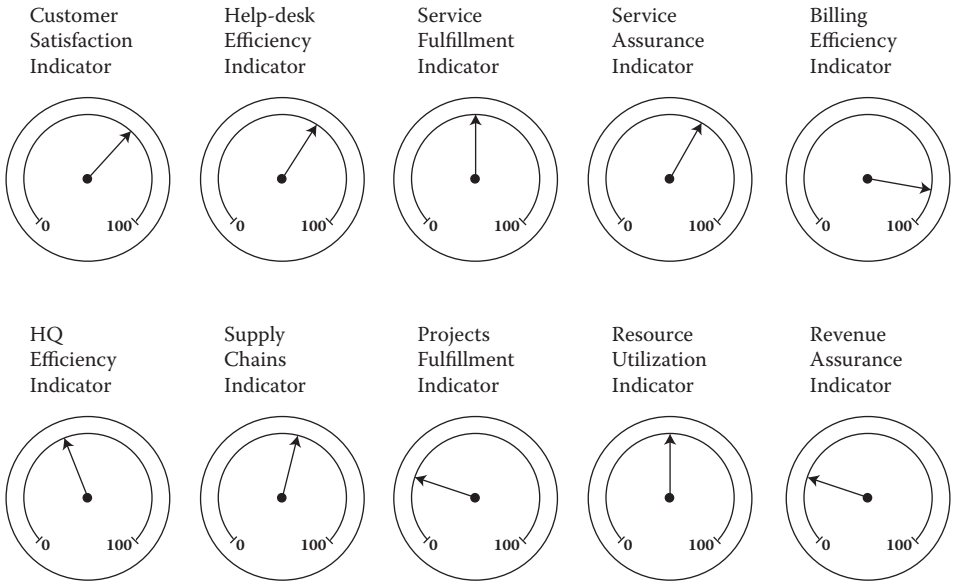
4.7.5 Dashboards and Balanced Scorecards

When management of operations cannot get a clear picture of the real-time data that underlines critical applications, infrastructure, and projects, management ends up reacting to issues, after users and customers are experiencing problems.

The answer is a dashboard that gathers key performance indicators, or KPIs, into a central repository that in a single window identifies the performance of critical systems, applications, and projects in real time. Leading vendors, such as IBM, HP, BMC, and CA are melting business service management, business intelligence, project, and portfolio management tools into dashboards. But this is a very complex task. The technical challenge of providing hooks into several vendors' reporting tools is remarkable, requiring standard interfaces.

a. Data sources
b. Key performance indicators

First, key performance indicators should be defined that drive the overall organization or parts of it. Instead of trying to grab as much data as possible, focus should be on the critical aspects for the business. This approach will make implementing the dashboard more straightforward. Next appropriate information sources must identified. The dashboard will need direct access to underlying data, typically requiring an API, or integration bus, or Web services to allow for loose coupling among information sources.



FIGu RE 4.7.7 Dashboard example for service providers.

Figure 4.7.7 shows an example for a service provider dashboard including areas, such as customer satisfaction, help desk, service fulfillment, service assurance, billing, human resources, supply chain, project, resource utilization, and revenue assurance.

Actually, there are eight steps in preparing a powerful dashboard. These steps are (BIDD08):

Step 1: Define the key performance indicators (KPIs) that need to be measured and displayed in the dashboard.

Step 2: Map KPIs to specific data requirements. Determine if the data exists in systems or needs to be collected.

Step 3: If data collection gaps exist, explore improvements to fill the gaps. Develop a plan and timeline to implement those systems.

Step 4: Investigate business service management, project, portfolio management tools, and BI tools based on your KPI requirements. Pay attention to how tools integrate with the existing infrastructure.

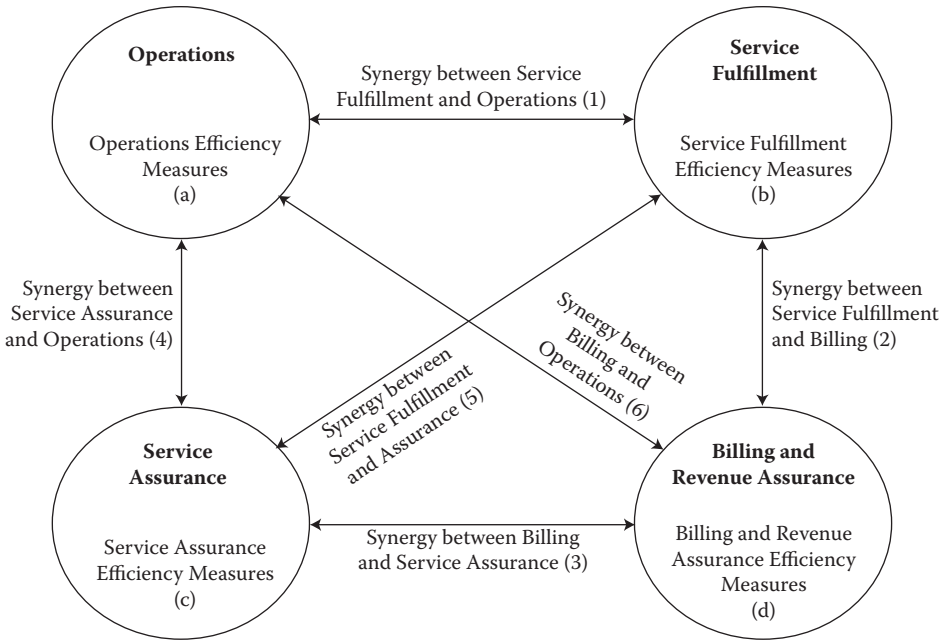
Step 5: Budget for the initial cost of the dashboard, annual maintenance, and fees to implement the system. Take into account the complexity and cost of changes and updates.

Step 6: Develop an implementation plan that provides dashboard visibility into key systems one at a time.

Step 7: After systems are integrated, focus on correlating data across those systems to provide meaningful visual information and alerting capabilities should a metric violate a threshold.

Step 8: When new components are considered, evaluate how they will be integrated into the dashboard.

Measuring the performance is a complex issue because it both drives and constrains behavior. A well-designed measurement platform can motivate both customers, whether internal or external, and staff to perform better and build more effective relationships. Scorecards may be used as part of benchmarks. They may also be used to highlight gaps, and to justify investments. Figure 4.7.8 shows a possible basic structure of interrelated scorecards for service providers, focusing on four key areas: operational effectiveness, service fulfillment, service assurance, and billing and revenue assurance. Recommended measures are listed below:



FIGu RE 4.7.8 Basic structure of interrelated scorecards.

- Operational effectivity: To what extend does operations and its processes improve:
  - Infrastructure life-cycle costs and TCO of equipment and facilities
  - Applications life-cycle costs and TCO of applications
  - Return of investment of operations-related projects
  - Staff productivity, capability, and skills developed
  - Managerial effectiveness
  - Availability of equipment, facilities, and applications
  - Business continuity capability
  - Level of convergence of products and services
  - Alignment with IT processes
- Service fulfillment effectivity: To what extend does service fulfillment and its processes improve:
  - How accurate and how up to date product and service catalogs are
  - Effectiveness of customer-facing procedures and CRM
  - State-of-the-art ordering process
  - State-of-the-art provisioning process
  - State-of-the-art activation process
  - Unique ownership of customer data
  - Support of customer-initiated and control fulfillment processes
  - Staff productivity, capability, and skills developed
  - Managerial effectiveness
  - Alignment with IT processes
  - Hit rate of fulfillment (number of activations/number to orders)
  - Return of investment of fulfillment-related projects
- Service assurance effectivity: To what extend does service assurance and its processes improve:
  - Definition of the right KPIs
  - Implementation of the right service assurance tools
  - Implement and supervise service level agreements (SLA)

  - Effectiveness of network-facing procedures
  - Staff productivity, capability, and skills developed
  - Managerial effectiveness
  - What level of integration between data, tools, and reports
  - Alignment with IT processes
- Billing and revenue assurance effectivity: To what extend does billing and its processes improve:
  - Level of convergence for products and services
  - Level of integration between pre- and postpaid products and services
  - Level of integration for rating, charging, and mediation
  - Support of electronic bill presentment and payment
  - Staff productivity, capability, and skills developed
  - Managerial effectiveness
  - Maturity for reducing the number of billing systems
  - Accuracy of billing systems
  - Connections to revenue assurance
  - Alignment with IT processes

The mutual interrelationships from this figure are the following:

1. Connecting and comparing each of these sets of measures, management can asses whether service fulfillment is aligned with operations. Fulfillment parameters can help setting the right and realistic thresholds in operating.
2. Connecting and comparing each of these sets of measures, management can asses whether all provisioned services are billed, and revenue assured. It ensures the synchronization between physical and logical inventory items.
3. Connecting and comparing each of these sets of measures, management can asses whether service levels are accordingly set and controlled by billing procedures, and whether in cases of SLA violations, reimbursements are granted.
4. Connecting and comparing each of these sets of measures, management can asses whether service assurance is aligned with operations. Continuous performance measurements help operations to set realistic values for operating thresholds.
5. Connecting and comparing each of these sets of measures, management can asses whether the right service assurance tools are in use for provisioned products and services.
6. Connecting and comparing each of these sets of measures, management can asses whether billing and revenue assurance are aligned with operations. It ensures the on-time delivery of invoices to customers.

Table 4.7.1 shows the customizing of scorecards for different audiences of service providers, including management of functional areas, executives of operations, and business unit and corporate executives. Besides the objectives and attributes, scorecard report examples are shown.

4.7.8 Summary and Trends

Telco executives find benchmarking a useful way to periodically check the overall performance of their organization. But the comparison with industry average and best practices should be based on very carefully selected key performance indicators. The old saying: "you got what you measure, so measure what you want to get" must be avoided in any case at any price. Periodic benchmarking should be conducted by an independent knowledgeable consulting company with industry experience. The combination of ITL, CoBIT and ISO 17799 could lead to multiple benefits, such as compliance with regulations, higher performance, and guaranteed security. But entry investments might be considerable. The use of

TABl E 4.7.1 Customizing Scorecards for Different Audiences

Audience	Management in Each Functional Area	Executives of Operations	Business Unit and Corporate Executives
Objectives	To provide operations managers with basic operating measures, so they can quickly identify unfavorable trends and take actions to resolve them.	To provide operations managers and direct reports with key indicators of overall performance across each functional area. Enables executives to quickly pinpoint risks and identify dependencies and issues; includes customer input as well.	To demonstrate where operations investments are going and how well they are supporting services OpEx expenses Where the expenses are
Attributes	Contains detailed operational data Is distributed weekly Guides managers in day-to-day operations Viewing capability	Contains higher-level information on key performance indicators Is distributed biweekly/monthly Should guide operations status meetings Provides customer perspectives on service delivery Viewing capabilities	OpEx details SLA fulfillment and violations Is distributed quarterly Viewing capability
Scorecard Reports	Detailed report 1 Detailed report 2 Detailed report 3	Summarized info 1 Summarized info 2 Summarized info 3	BU specific info 1, 2, 3 Activity specific info 1, 2, 3 Function specific 1, 2, 3

dashboards, scorecards, and Six Sigma quality control might help to meld BSM, BI, and PPM tools into a high level control mechanism of the efficiency of the overall performance of the business.

Acronyms

ACD Automated Call Distributor
API Application Programming Interface
BI Business Intelligence
BSM Business Service Management
CoBIT Control Objectives for Information and Related Technology
CRM Customer Relationship Management
ITIL Information Technology Infrastructure Library
IVR Interactive Voice Response
KPI Key Performance Indicator
PPM Project and Portfolio Management
SLA Service Level Agreement
TCO Total Cost of Ownership
TQM Total Quality Management

References

BIDD08: Biddick, M. 2008. Hunting the elusive CIO dashboard, *InformationWeek*, March 8, 47–52.
GARD02: Gardner Group, 2002. *Combine CoBIT and ITIL for powerful IT governance*, TG-16-1849, Research Note, June 10.

ITGI06: IT Governance Institute, 2006. *CoBIT mapping of ISO/IEC 17799*, Research Note.
IEC: International Electrotechnical Commission, http://www.iec.org.
ISO: International Organization for Standardization, http://www.iso.org.

Summary and Trends

Kornel Terplan

Governance is key to ensuring the effectivity and efficiency of a lean service provider. The basic infrastructure to offer products and services has been implemented by now in most countries. Most tier 1, tier 2, and tier 3 service providers are capable of supporting triple-play, or even quad-play for both business customers and consumers in households. Convergence of fixed and wireless is on the way.

Competitive advantages can be achieved by effectively using and combining existing infrastructure resources. Time-to-market, real-time enterprise features, agility, and meaningful use of information are a few examples where service providers may realize significant improvements. Service providers will be able to better control their infrastructures by continuously measuring KPIs. The result is that the relationship between customers and service providers will be based on mutually measurable metrics. Fewer disputes and penalties are expected in the future with Service Level Agreements.

Business intelligence is a business strategy, aimed at understanding and anticipating the needs of an enterprise's current requirements. It is knowledge about the enterprise's customers, competitors, business partners, competitive environment, and its own internal operations that gives management the ability to make effective, important, and often strategic business decisions.

CoBIT, ITIL, and ISO 17799 are not mutually exclusive. They can be combined to provide a powerful IT governance, control, and best-practice framework in service management of the telecommunication service provider. Enterprises that want to put their ITIL program into the context of wider control and governance framework should use CoBIT. Enterprises that want to secure all processes and want to meet compliance requirements should use ISO 17799.

This section gave a few innovative state-of-the-art ideas in areas, where service providers need a visible improvement. IT alignment with business might redefine the role of IT from cost center to innovation driver. Properly dealing with data and information helps recognize trends and helps to meet regulatory compliance. The simple structure of the management organization guarantees redundancy-free lean operations and finally, benchmarking helps to position the enterprise.

5

Future Telecommunications Services

Introduction **5**-1

5.1 User Needs **5**-2
Types of Users • Different Users Have Different Needs • Unified Messaging • Instant Messaging • Location Services • End-User Requirements Summary

5.2 Application Trends **5**-13
Application Functionality • Functionality Implementation • Network Neutrality • Semantic Web • Social Networking

5.3 Systems and Service Integration for Management **5**-22
Introduction • Drivers for Integration • Integration for Service Providers • Integration for Business Users • Integration for Mobile Professionals • Integration for SOHO Users • Integration for Residential Users • Unified Threat Management (UTM) • Network Behavior Analysis (NBA) • Mobile Device Management (MDM)

5.4 New Product and Service Creation **5**-33
Introduction • Drivers and Constraints • New Service Creation • Increasing Bandwidth

5.5 Telecommunications Tariffing **5**-38
Introduction • Regulatory Trends • Service Pricing Trends • Impact of New Technologies

5.6 Telecommunications Strategies **5**-42
Introduction • Service Providers • Goals • Green Computing • Software as a Service

Summary **5**-48

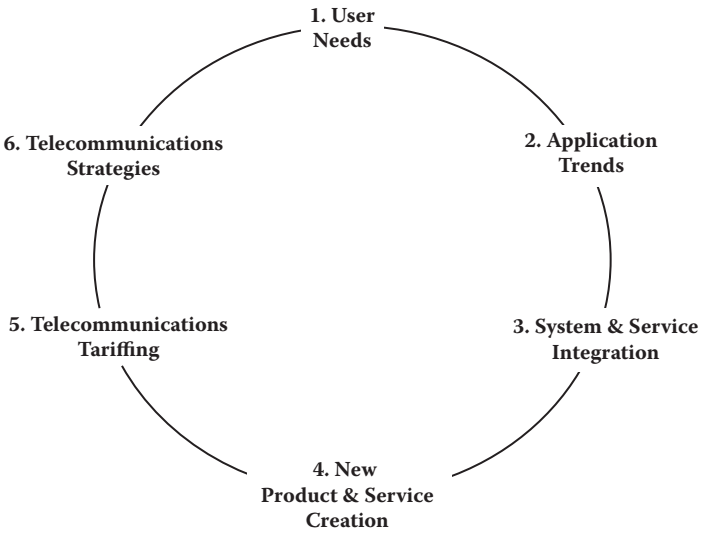
James Anderson
Verizon

Patricia Morreale
Kean University

Kornel Terplan
Industry Consultant and Professor

Introduction

Imagine for a moment how daily life would be affected if the telecommunications services and applications that we take for granted were to be removed. The daily paper would contain mainly local news and any international stories would be describing events that were weeks or months old. We would spend much of our time during the week traveling from house to house and town to town as we tried to keep in touch with our friends and business associates. We would tend to live close to where we were born and raised otherwise we would risk losing contact with friends and family. Finally, the amount of envelopes, paper, and stamps sold would be constantly increasing as people wrote letters in order to have their presence felt in far-off locations without having to travel. The contrast between our everyday life and



FIGu RE 5.1 Telecommunications trend lifecycle model.

this example clearly shows just how significant the impact of today's telecommunications services has been on how we communicate. As hard as it is to imagine a day without the communications systems and services that have become such an integral part of our lives, so too will it be impossible for future generations to imagine living in our times with our "primitive" telecommunications infrastructures and applications.

In this chapter we will be looking at where the evolution of the field of telecommunications will lead to. This type of prediction is not without a great deal of risk: a similar analysis done as recently as 1990 could not have hoped to accurately identify the impact that the Internet now has on the way we communicate today. However, the basic building blocks that will control the evolution of the field of telecommunications, the telecommunications DNA if you will, are reflected in the state-of-the-art services, applications, and equipment available today. We will look at the current trends along with the end-user requirements and competitive market forces that will shape the future of telecommunications.

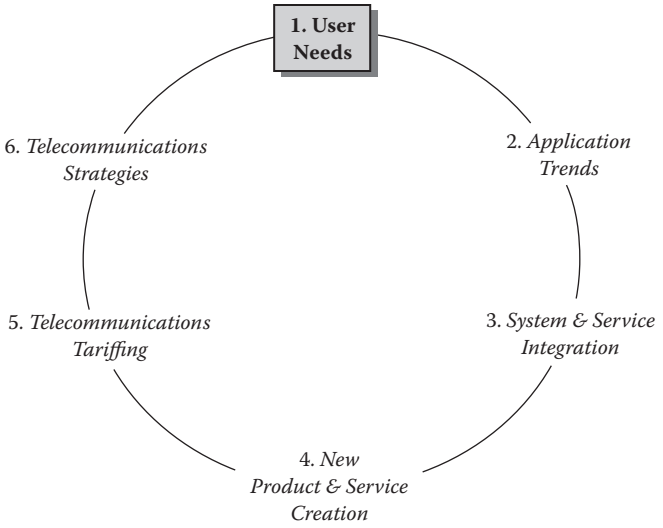
To help focus the consideration of such a large topic as the future of telecommunications, it is helpful to have a model to frame the discussion. The model that we will use in this chapter to identify future trends in telecommunications is shown in Figure 5.1.

This telecommunications trend lifecycle model that we will be using is intended to provide a high-level view of how the effects of changes "ripple" throughout the telecommunications field. We will be discussing the model in a sequential manner, starting with an analysis of the changing needs of end users. It is important to keep in mind that innovation and change in real life is often chaotic and seems to resist following orderly models. Therefore, as long as we understand that a new telecommunications trend can potentially start at any step of the trend lifecycle model (i.e., a new equipment technology is invented in a research lab and only later is it understood well enough to be used to address end-user needs), then we will be able to correlate this chapter's analysis and the real world.

5.1 User Needs

James Anderson and Patricia Morreale

The modern world is currently undergoing its third major communications transformation. It took 38 years for radio to garner 50 million listeners; likewise, it took 13 years for television to achieve a



FIGu RE 5.2 Trend analysis—user needs.

similar number of viewers. Incredibly, the worldwide computer communications network known as the Internet has required far less time to reach that milestone. In the United States, as of this writing, there are more than 220 million Internet users and more joining every day. As is to be expected, when more people make use of the Internet, more information needs to be processed by the networks and computers that make up the Internet. The needs of these users and others like them will form the drivers of telecommunication trends in the future.

In this section we will examine the user needs that form the basis—and demands of—tomorrow's telecommunications systems and applications (Figure 5.2). We will start by determining exactly what types of users' needs we have to understand. Next, we'll explore the specific problems and challenges that each group of users is currently trying to solve. Finally, we'll identify several general trends in user needs that will have the greatest impact on future telecommunications services.

5.1.1 Types of Users

It can be argued that almost everyone in industrialized countries could be considered to be a potential end user of telecommunications services and applications. A recent study by the International Telecommunications Union (ITU) standards body reported that in high-income countries (per capita GDP of more than U.S. $8955) there exists a "teledensity" of more than 50 phone lines for every 100 people. This would lead one to conclude that in these countries, telecommunications services and technologies will evolve to meet the needs of the general public. However, in order to identify specific future trends in telecommunications, we need to limit our focus to only those users who either have the financial resources or sheer numbers to generate and sustain a trend in telecommunications. We will also avoid focusing on narrower vertical application segments such as healthcare and banking in order to identify trends because their influences on future applications and services can be safely generalized into broader end-user groups without losing their contribution. In this chapter, we will segment end users into four primary groups for further study. These groups can be characterized in the following ways:

- **Businesses:** This segment of telecommunications end users is defined to be a group working toward a common goal at one or more locations. As a rule, businesses need to interconnect each of their workers on a frequent basis. Depending on the size and type of business, this interconnection requirement can result in the need for large amounts of bandwidth. The business segment is

also characterized by its growing need for 7 days per week × 24 hours per day × 365 days per year connectivity in order to support globally distributed operations. Businesses are fairly price resistant—they are willing to pay more for access to applications that they feel will provide enough of a competitive advantage to recover their costs.

- **Mobile Professionals:** These end users generally interact with business segment end users. The difference between these segments is that mobile professionals generally operate either by themselves or as part of small focused teams. Mobile professionals don't have a fixed location connected to telecommunications services; rather, they need to have services find them or permit them to access the services from a wide variety of remote locations. Once again, the mobile professional segment is fairly insensitive to the price of telecommunications services that have a direct correlation to a competitive advantage.
- **SOHO:** The small office/home office (SOHO) segment is a rapidly growing portion of the market, as larger businesses discover it is more economical to outsource many of the tasks they used to perform internally. Tax incentives from many local and federal governments designed to decrease commuting congestion and pollution have also added to the economic incentive for this segment to experience explosive growth. Telecommunications applications have been crucial to fueling the growth of this segment. Existing applications have permitted home office workers to have access to similar communications resources that centralized workers also enjoy. The SOHO segment is price sensitive; however, their large numbers can often be used to create attractive business cases for both the end users and the service providers.
- **Residential:** This segment of end users wants to have telecommunications services delivered to their homes. The telecommunications applications desired by this segment often are used to communicate with other residential end users, businesses, or for entertainment. This segment is very price sensitive; in order to pay for a telecommunications application or service, something else will have to be given up. Each application is subjected to a trade-off evaluation by the end user.

5.1.2 Different Users Have Different Needs

Each of the different user groups we have identified is facing a different set of challenges that can be addressed in a variety of ways by telecommunications services. In this section we will explore the environmental and social drivers that have created these end-user needs. In the final section of this chapter we will identify the common drivers that apply to each segment of end users. As you read this section, it is important to keep in mind that although the specific details of how end user problems will be addressed may change over time, the core set of conditions that have created the needs will not change.

5.1.2.1 Business End-User Needs

Businesses exist to earn a profit and they do this by offering some combination of better products, lower prices, or by meeting the specific needs of a particular customer better than any other firm. For the purposes of this discussion, we group together businesses of all sizes from the very small to the very large. Although the specifics of the problems they are trying to solve may differ, all businesses face the same basic set of challenges.

The communications needs of business end users can be divided into two basic groups: internal needs and external needs. A business's internal needs relate to how it communicates the way that it wants to do business to its employees and how those employees communicate status and learned information throughout the firm. The external communication needs of a firm relate to how it exchanges information with members of its business environment. These members include other businesses (trading partners) and customers alike. We will now examine the drivers in each of these different groups of needs in detail.

In the last decade, firms have come to realize that one of their primary sources of competitive advantage can come from how well they exchange information internally. Having used the recent explosion

of networking and computer storage technology to collect, store, and distribute large amounts of information, firms are now looking to refine their operations. What businesses have realized is that they are facing a major challenge in order to provide everyone in their organization with access to the specific types of information that they require in order to perform their jobs better. A key challenge is that each employee in a firm performs a different task (or performs the same task in a different business context) and therefore needs to have access to different types of information at different times. How to provide such connectivity presents a significant challenge to businesses of all sizes.

One of a business's most valuable resources is its internal knowledge of how problems were identified and solved in the past. A key communications objective for a firm is to find a way to manage and share problem-solving experiences throughout the firm. Meeting this challenge is critical for the firm, otherwise it will face the expense of solving the same problem for the first time over and over again. The solution involves communication solutions that not only provide access to detailed records of past projects, but also include identification and access to the employees who were involved in solving the problems. Only by finding a way to meet this challenge can firms refine their problem-solving processes and become more competitive.

The cost of producing products or services has received a great deal of attention in recent years. Businesses have implemented a wide range of control and monitoring systems that are able to evaluate the operations of different internal processes. Such systems include enterprise resource planning (ERP) systems that can control the supply chain of a product's production process, quality improvement tracking systems, and just-in-time manufacturing systems. One of the primary purposes of each system is to permit a firm to more effectively use its resources and raw materials—in other words, they help a firm run a "lean operation" in which all of its assets are fully utilized. Such tightly run operations require a business to establish and maintain a wide variety of communications between its internal divisions no matter where they may be located. Additionally, there is a direct correlation between how fully the firm's assets are utilized and how rapid communications between the different parts of the firm are executed. These processes and systems force a business to walk a tightrope between operating at peak efficiency and not having to correct materials to operate at all. Firms must identify what communication is required to support such mission-critical systems and then implement and use their telecommunications solutions to gain a competitive advantage.

Finally, businesses are often thought of as a collection of employees who come together at company-owned locations to perform work. Businesses are now starting to realize that the arrival of relatively inexpensive computing resources, coupled with the availability of numerous communications services, call for rethinking about how they conduct their daily operations. Firms have already realized that many of the noncritical or nonstrategic processes they perform can be effectively outsourced to other firms that are able to perform these processes more efficiently and at a lower cost. Firms are now starting to reexamine how and where their remaining employees work and interact. The popularity of telecommuting and rotating "work from home" days shows how firms are starting to explore these uncharted waters. One of the primary keys to making a widely distributed workforce successful is to identify communications solutions that permit the firm's employees to interact as though they were together in an office, without the actual expense of the office.

Advances in transportation and communication have permitted businesses of all sizes to compete on a global scale. New businesses are able to offer their products to almost any international market, starting on their first day of operation. Existing businesses that have saturated their traditional domestic markets are able to seek new revenue streams in unexplored global markets. One side effect of operating and competing on a global scale is that all of the telecommunication systems that a business established to facilitate internal communications for its domestic operations must now be extended to become both location and distance insensitive. This requirement affects all forms of communication including voice, video, and data. Traditionally, such services have been very distance sensitive, thus making telecommunications expenses a significant expenditure for a globally distributed business. As the number of firms that operate internationally has increased, so too has the number of telecommunications service

providers. This increase in service providers has provided businesses with an opportunity to seek out and use those providers who are able to help them minimize their telecommunications costs. Once again, the FCC's statistics show that the composite cost of an international phone call has dropped from U.S. $0.53 in 2000 down to U.S. $0.14 in 2004. While cost-per-call was declining, call volume from 1990 to 2004 increased from 1.9 billion to 10.9 billion.

As businesses study how they can maximize their profits, they have realized they can reduce their costs by streamlining interactions with their suppliers. This new understanding has led to the sharing of information, such as current sales results and stocking data between retailers and their many suppliers. The high volume and near real-time characteristics of this information have created a growing need for more sophisticated telecommunications services. Once again, since retailers and suppliers may be located in different areas, the telecommunications systems must be distance insensitive.

Finally, the most important interaction that a business has is communication with its customers. Customers are demanding that it become easier and quicker to interact with a firm. They want to see updated product lists and information; in some cases they want to be able to custom-design their own solution from a firm's product lines; and they want to be able to review and perhaps pay their bills electronically. This increased level of interaction with customers who are not physically located in a firm's place of business demands an entirely new set of sophisticated telecommunications services.

5.1.2.2 Mobile Professional End-User Needs

As business becomes more decentralized and at the same time more customer-focused, the ranks of the mobile professionals are swelling. This new breed of employee can no longer be thought of as being only a salesperson; rather, the mobile employee may be part of any one of a number of project teams that have been brought together to solve a specific problem. As more and more employees start to operate away from the firm's offices for longer periods of time, the ability to use communications systems and services to provide information, obtain status updates, and share learned knowledge becomes even more important. Let's take a look at some of the specific needs of this group of end users.

Arguably, the most critical need of a mobile professional end user is his need for up-to-date information. Since a mobile user is operating away from a centralized office environment, his ability to learn about changes in products or company strategy is limited to what information is sent to him—the critical real-world "water cooler" information exchange system is no longer available to him. New means of identifying important information need to be created along with an effective two-way system for distributing that information and getting end user responses and feedback.

Since the mobile end user is often away from the office and in fact may be spending much of the time with a customer, it is impractical to carry all of the product and service reference material that may be required to perform the tasks. Therefore, it's important that the mobile end user be able to quickly access all of the material that may be required to support the current task. Note that the information required may take many forms including text, pictures, animation, and video. Many firms that sell large, complex software systems have changed the way that they now perform product demonstrations. Instead of taking complex computer systems to the customer's site, they use a standard laptop and establish a communications link back to their office, where the application is running on the more complex hardware system. This is one way for the firm to better utilize its expensive resources and better support its mobile users at a lower cost. Such services are only the start of what will be required to support the growing mobile user community.

The type of data that can be accessed by mobile users is another critical issue. Current wireless links (ex. EV-DO Revision A) limit mobile users to average download speeds of 450 to 800 kilobits per second (kbps) and average upload speeds of 300 to 400 kbps, which is acceptable for accessing e-mail and small to medium-sized text documents. As more and more information is stored in richer formats such as video and integrated multimedia documents, new telecommunications services will have to be created to support mobile users. The need for access to multimedia information

is especially critical for mobile users whose firms design, manufacture, or sell complex products. The multimedia information for these products can help the mobile user to shorten the selling cycle by permitting such complex products to be clearly and simply communicated. New telecommunications solutions are required to ensure that mobile users are able to access *all* of the information they require in order to perform their jobs.

A unique requirement of mobile end users is that, unlike stationary users, information must "find" its way to the mobile user. The mobile user is expected to change locations quite often and can't be expected to be reached via an addressing scheme that requires the user to always be at a given geographic location. This applies not only to voice communication but also to all forms of electronic information interchange. This issue has been partially addressed by some of today's current telecommunications solutions; however, such solutions generally work only within a limited geographical area (country or artificially determined service provider territory) and completely different solutions have been designed for voice and data services. Mobile users require solutions that provide seamless integrated voice and data solutions of ubiquitous coverage.

Although many of the needs of a mobile end user relate to ensuring reachability at all times, the opposite is also a concern. One of a mobile user's more valuable resources is time. Giving others the ability to communicate with the mobile user also gives them the ability to appropriate time. The mobile user needs to be able to limit who has what level of access. Additionally, the mobile user needs to be able to decide if and how to respond to each request for valuable time.

A mobile user's toolkit consists of several groups of information to help do the job on a daily basis. These groups of information consist of a variety of phone lists, customer names and addresses, customer lead lists, internal corporate directories, etc. As this collection of data grows in size, so too does it grow in value to both the mobile user and the company as a whole. The telecommunications challenge is how this information can be shared among the wide variety of communication devices used by the mobile user without having to retype the information each time.

The demanding lifestyle that being a mobile user requires often results in the lines between a worker's personal and professional lives being blurred. Since the mobile user may be away from home for long periods of time, it is critical that personal messages from various sources and in varying formats must be able to find their way to where the mobile professional is. Additionally, personal communications must be clearly identified as such and must be easily differentiated from work-related communications. Both mobile workers and the firms that employ them appear to be drivers for this type of requirement—both parties realize that good communications can help a mobile worker strike the correct balance between different roles and responsibilities.

Change and movement are key components of a mobile user's typical day. Because of this, there is no single best way for messages and information to reach the mobile user. Therefore, the mobile user needs to be able to access a message in any one of several different ways: e-mail via the phone, and voice mail via the laptop. It is critical that the information is able to reach the mobile end user as quickly as possible without restricting how the user chooses to retrieve the information.

The era in which groups of the same people worked together for years or even entire careers is quickly coming to a close. Mobile users are at the forefront of this change and represent the new breed of worker: they are part of dynamic teams quickly created to solve specific problems. Once the problem has been solved and a solution implemented, the team is then dissolved and its members go on to join other dynamic teams. From a communications perspective, the mobile user needs to be able to easily exchange and work on the same information with other members of the dynamic team during the time that the team exists. The security associated with such communication is a critical factor. In today's customer-focused markets, employees of the customer may be part of the same dynamic team as the mobile user. In such cases, the ability to filter and restrict a team member's access to sensitive data is required in order to ensure that the internal and external team members are able to work together smoothly.

5.1.2.3 SOHO End User Needs

In contrast to large established firms, employees of small firms have different communications needs. We include in this group those workers, who may work for firms of almost any size, operating out of their homes. Corporate outsourcing and the increasing number of new businesses have caused this small office/home office (SOHO) group of end users to increase in size on a yearly basis. As the telecommunications service marketplace becomes more and more competitive, the SOHO segment of end users has started to receive the attention of telecommunications service providers. The key to a provider being able to successfully serve this market will be an ability to correctly identify the needs that will motivate the SOHO end users to purchase telecommunications services.

Unlike either the business or the mobile user, the SOHO end user is extremely price conscious. Smaller organizations naturally tend to have smaller budgets and therefore will have less to spend on telecommunications services of any kind. However, SOHO end users are generally involved in very competitive market niches and so they feel that it's necessary to their continued survival that they arm themselves with any tools that provide a competitive advantage. The end result of these two conflicting conditions is that the SOHO end user will purchase or subscribe only to those telecommunications services that are priced within budget and which can be clearly demonstrated to give them a competitive advantage.

SOHO end users do share some of the same basic needs that mobile end users have. Specifically, those SOHO end users who operate out of their homes will have the need to be able to separate personal messages from business messages. This issue is a little more complex than it was for mobile end users because all of the messages are delivered to a single location—the user's home. An extension to this need is that the at-home SOHO end user, just like the mobile end user, needs to be able to control who can communicate with them and when. Since all requests for time (phone calls, e-mail, etc.) will come to home, the SOHO end user needs to be supported by telecommunications services that can be told which role the end user is currently playing—homeowner or worker.

Most SOHO establishments share a desire to one day be bigger then they are now. As a move in that direction, SOHO end users want to be able to start projecting a "big company" facade at all times when dealing with customers. This requirement manifests itself in several different ways: addresses and staffing levels. In the days prior to electronic addresses, small firms could use postal boxes to obscure their less impressive residential or strip mall addresses. As we move into the future of electronically linked businesses and electronic commerce, the importance of an impressive electronic address will take the place of the postal box. Additionally, since SOHO operations are generally staffed at very lean levels (i.e., perhaps a single employee), SOHO end users are always on the lookout for telecommunications services that can take the place of additional nonexistent staff members and which can be used to provide superior customer contact. An example of such an application would be the "automated attendant" feature on many small business phone systems, which automatically provides company information and basic directory services.

For the SOHO end user, the previous requirement can be further extended. It is once again the limited amount of staff available in the SOHO environment that generates the need for additional telecommunications services. These services are needed to permit potential customers to easily show themselves the SOHO firm's products, prequalify themselves, and then get in touch with actual employees. This use of telecommunications services to handle initial customer interest and then using valuable human resources only when the customer has demonstrated that they are a viable potential customer may be one of the most important drivers for SOHO telecommunications requirements. It certainly is one of the easiest ways to justify spending money.

Like the mobile end user, a SOHO end user must often work with others in order to secure large business orders, due to a SOHO's small size. This can often result in a SOHO establishment being required to ad-hoc partner with another business on a per-project basis. The telecommunications requirements that would be driven by this opportunistic type of limited partnering would be to support the exchange among the temporary partners of such information as schedules and project information. Once again,

security would be critical; just because partnering is occurring on this project does not exclude the possibility that these partners may be competing against each other in the future.

Unlike the mobile end user, the SOHO end user has a base of operations—an office. It will be used to store almost all of the information related to the SOHO operation. This organizational structure produces a telecommunications need to permit the SOHO end user to access the information while away from home. Such access requirements include the ability to retrieve voice messages, electronic data, and any other information or formats that may be required. There is also the need for notifying SOHO end users that new information has arrived at the office in their absence. Note that once again, the information can arrive in a multitude of different formats.

Finally, since a SOHO end user faces the dual dilemma of operating under a tight budget today but believing that the operation will grow larger tomorrow, whatever telecommunications decisions are made today must be able to grow and change with the business. Solutions that must be removed and replaced are unacceptable both in terms of costs and time lost.

5.1.2.4 Residential End-User Needs

Our final segment of end users is also arguably the largest. In the United States there are currently over 120 million homes; it is these residential end users to whom a wide variety of service providers hope to sell additional telecommunications services. The marketing success of standard telephone service and the mixed success of various cable and Internet-related services clearly shows that the residential end-user community is a complex and multifaceted group. The service providers hoping to capture a significant share of this diverse group must be willing to spend the time to understand what shared needs are currently unsatisfied.

Perhaps the most important factor that must be considered when attempting to understand the needs of the residential end user is that, unlike the other end user segments that we've studied, the residential end user has a relatively fixed budget from year to year. The result of this is the simple fact that every purchase is a trade-off: if a new telecommunications service is to be purchased, then something else must be passed over. In most cases, this means that any service that does not provide a clear return for the residential end user's investment is certain to fail. A good example of this occurred when the next generation of phone services based on the Integrated Services Digital Network (ISDN) technology were introduced. Despite the technology being sound, one reason that they failed was because residential end users judged them to not provide enough of a benefit to justify their cost.

As communication systems have improved our lives, they have also permitted us to move faster throughout the day and get more done. The result of this has been that the residential end user views the ability to manage time as a critical need. Any product or service that can provide more control over how limited time resources are spent seems attractive. However, as we have previously discussed, other factors such as price and availability will still play a very significant role in determining the residential end user's final acceptance.

As more and more information arrives at a residence, a striking advantage of postal mail over telephone service starts to emerge: information that is delivered via the postal system clearly identifies its intended recipient. On the other hand, a phone call arrives with no attached address and so whoever is first to answer the phone is required to perform a crude routing function in order to ensure proper delivery. This problem will only continue to grow as Internet access requires separate e-mail addresses and cable services permit channel and scheduling selections to be customized on a per-viewer basis. Any services that seek to address these needs of the residential end user must make sure that they are able to handle information that arrives in a variety of formats and that both end user addresses and information processing preferences are handled by the service.

People are tribal by their very nature—we accomplish our daily activities by interacting with a wide variety of other people in our community, neighborhood, and extended family. Residential end users have a need to stay in touch with their contact group, which resides locally as well as their extended families who may not live locally. The specific relationship defines the frequency of this contact and the

format in which it needs to occur. Today, such contact is mainly limited to text (letters or e-mail) and voice (via the phone). However, the arrival of the Internet and its support for a diverse set of multimedia communication formats has started to acquaint residential users with new options for communicating.

A very important constraint on any new telecommunications service is that it must be easy for the residential end user to use. Since the educational background and technical sophistication of residential users can vary widely, the majority of residential end users require that systems they purchase be easy and intuitive to use. One of the reasons that basic telephone service has been such a success is that the service is intuitive and simple to use. Note that the amount of end-user training time that it takes to learn to use a phone is very short! A key point for service providers to remember when introducing new services is that, in the mind of the residential end user, ease of use is a more important factor than additional bells and whistles.

Residential end users are always on the lookout for bargains whenever they are preparing to make a purchase. This mentality can be seen in the types of retail establishments that dominate the U.S. landscape: Walmart, Kmart, and an almost infinite variety of strip malls. One of the greatest advantages of the Internet and e-commerce is support for comparison shopping prior to purchase, whether via the Internet or at a bricks-and-mortar site. In the future, telecommunications services that standardize such comparisons and permit product offerings to be compared on multiple criteria including price, features, and availability would meet a need of the residential end user.

As we move into a new millennium, it is becoming evident that the skills required to survive and thrive in the modern world are changing. An example is found in automobile repair. The number of residential end users who service and maintain their car themselves has dropped substantially due to increased complexity in automobile design (antilock brakes, turbocharged engines, etc.) and a decrease in the amount of time available to perform such basic tasks. Interestingly enough, when a car is taken to a repair shop to be worked on, one of the first steps that the mechanics perform is to attach computer input cables to various parts of the car in order to diagnose its operational health. Residential end users understand that this change in required life skills is occurring and they are eager to not be left behind. Therefore, they see access to education and information resources as a critical need and they desire telecommunications products and services that can improve, supplement, or provide greater access to such educational resources.

One of the greatest benefits of modern communications services is allowing people to interact with others who share a common interest. Without such services, perhaps these people would otherwise never know about each other. Residential end users desire services that will permit them to interact with other (potentially) remote end users who share a common interest. Examples would be collectors, fantasy-league sports players, online auction houses, and support groups. New telecommunications services offer the possibility of permitting such interactions to occur on a global scale.

In the past, if a residential end user wished to gain access to valuable resources such as technical help, a stockbroker, etc., they had few options: schedule an appointment and then travel to meet with the resource provider face-to-face or phone them and either wait on hold or wait for them to call back. Telecommunications services that can streamline access to such valuable and limited resources are desired by all residential end users.

Access also plays a key role when it comes to a residential end user's finances. Better access to financial resources such as loan information, checking/savings account information, and stock portfolios has always been desired but not widely available. Key barriers to such services in the past have been concerns regarding both the security of transactions and the inability to validate the identity of the user, and the lack of appropriate equipment at the end user's residence to support such services. Both of these issues are being dealt with and will not continue to be barriers.

Residential end users seek ways to supplement other activities and thereby produce a richer experience for themselves. Users desire a way to gain more information or to follow up on something else that they have read about or seen. An example would be PBS's *Nova* programs, which display different Web links that point to supplemental material about the portion of the show that is currently being

viewed. Additionally, residential end users would like to be able to follow up and obtain more information on advertised products that they see in different media—note that this accounts for the fact that Web addresses have become a standard part of any auto advertisement!

In a fashion similar to both mobile and SOHO end users, residential end users are very concerned about both their privacy and how they spend their valuable time. Residential end users want to be able to control who is able to get access to them and when such access is permitted. Therefore, they are interested in finding solutions that permit them to control who is able to send them information and how they are notified when that information arrives.

Finally, the ultimate benefit of technology is that it permits residential end users to plan events around their schedule rather than the other way around. Residential end users would like to be able to pick what time they want to be entertained instead of having to arrange their lives around external entertainment schedules.

5.1.3 Unified Messaging

Unified messaging (UM) refers to the union of all messaging media, which might include voice messaging, text messaging, e-mail and any other messaging services in a single communications experience. At a minimum, unified messaging takes the form of a single mailbox or alert service, which would permit the end user to have a single source for message delivery, storage, access, and notification. Further, UM permits the use of a single interface for sending and receiving messages. This permits a client to use a mobile phone, computer, or other PDA to retrieve and send messages with ease.

Unified messaging is an advanced messaging capability that enables the delivery of value-added services to end users. UM is designed to make mobile communications simpler, easier to use, and more productive. Example unified messaging features and services include:

- Single Delivery—a service in which any users may send any message to a specific end user by addressing the message to a single alias.
- Single Repository—all messages are stored in one place for ease of access and retrieval, as well as integration with other offered services, such as conversion of speech to text and vice versa.
- Single Access—using a unified addressing approach, users can access all messages through simple mailbox interface commands.
- Single Notification—using SMS (short message service) or other forms of notification, the user can be alerted to missed messages, incoming messages, and communication attempts.

Unified messaging has the ability to integrate with other applications. For example, consider a UM system that could detect a received e-mail, locate the intended recipient on his mobile phone, and send an e-mail notification to a short message service center (SMSC). This, in turn, would result in the SMSC sending an SMS to the user. Once the user received notification, the subsequent actions by the user could range from ignoring the new e-mail, choosing to receive the e-mail immediately on the mobile phone via SMS, initiate a voice call to the e-mail sender, or listen to the e-mail using a text-to-speech conversion.

Future services actively being integrated into UM include location-based services and mobile IP technologies, which are expected to further personalize the user messaging services. Additionally, presence and availability technologies are anticipated additions.

An example of the utility of presence and availability systems for UM services is the detection of the users in a specific location, and the delivery of location-specific details, such as restaurants, cinemas, parking, and other geographic location-specific services that might be of interest to the user.

5.1.4 Instant Messaging

Instant messaging (IM) is a form of real-time communication exchange with text being carried over a corporate intranet or the Internet. This synchronous exchange, which requires all communications parties to be present to be effective, is comparable to a traditional voice telephone call, but IM is a text-based

exchange rather than an audio exchange. IM has grown in popularity due to the availability of Internet computer access and the perception that IM exchange is just an extension of electronic mail exchanges, but with a more immediate response available.

One of the highlights of IM is that multiple parties can participate in an IM exchange, with each party receiving the exact same text on their screen at approximately the same time as all other participants. This is comparable to having a conversation with many overseers and participants. However, with IM, in contrast to an audio conversation where all participants must be co-located, IM exchanges can take place regardless of geographical or time constraints. This permits friends or colleagues in far-reaching places to exchange ideas and comments as though they were all in the same room at the same time.

Security remains a concern with IM, as the ability to spoof or masquerade as an IM alias user is possible. During a voice phone call exchange, at least the timbre of a voice is perhaps recognizable. However, IM does not have such a distinctive personal signature, so once an IM connection is established, unless the sending or receiving party disconnects or disengages in some manner, the communications path remains open. This can encourage misrepresentation or duplicity, unless a timeout or authorization is implemented to establish user validity before sending messages.

5.1.5 Location Services

Location-based services, or location services, are becoming more and more widespread. Using GPS information provided by mobile devices, location-specific information regarding nearest office, restaurants, and services can be provided to the mobile user. This is also referred to as customized service delivery, as the features and services delivered to the user are dependent on their geographical location. Future expectations are that in addition to customized service delivery, costs for delivering information will be variable depending on the specific location, time of day, and the amount of information available.

In addition to the (outbound) delivery of services to the user, the (inbound) request for services can be received by the network from a user. While this is an evolving application, currently inbound location service requests are handled by the nearest service provider node. It is anticipated that as contention for this mobile client increases, service requests received from the user could be directed to the application server of the user's choice or the service provider could auction off the opportunity to reply to the user to the highest bidder, as the user is potentially a valuable client for future service delivery.

5.1.6 End-User Requirements Summary

As we conclude this section, it is important that we review the needs that are facing the four main segments of end users who will be driving the evolution of telecommunications into the future: business, mobile, SOHO, and residential. It is important to note that each of these segments is attempting to accomplish a different set of goals with different sets of available resources. This simple fact becomes quite evident when one looks at the differences in how much each of the different segments is going to be willing to spend on new telecommunications applications and services.

Although there are significant differences between each of the major end-user segments, several common themes have emerged. One of the most fundamental needs that each segment is trying to address is the ability to better control how its time is spent. Telecommunications services have the unique ability to eliminate distances and to permit time to be *shifted*—that is, to allow interaction between different parties to occur when it is most convenient for all of the involved parties. This need is further supported by each segment's desire to be in control of when and how they communicate with someone. The curse of modern technology is that it severely limits our ability to make ourselves unreachable when we so desire. The ability to regain this ability is a need that has been expressed by end users in all segments. Finally, the realization that end users are working harder at their jobs and the fear that this will cause

their professional and personal lives to blur into an undistinguishable mass has generated a common set of needs. Users are seeking a way to be able to clearly distinguish communication and information that is associated with each of the different roles that they play.

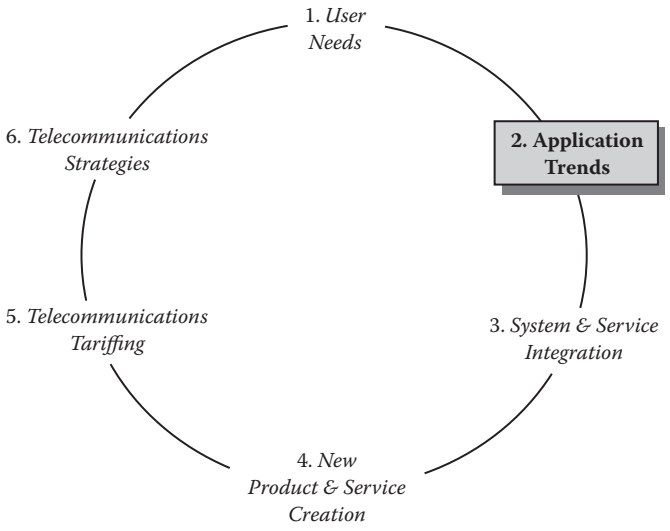
The recognition of these common basic end-user needs provides a clear prioritization for the development and deployment of future telecommunications applications and services. At its core, telecommunications is a field that exists to improve lives and solve problems. Advances in telecommunications often appear to be based on the latest "gee-whiz" technologies; however, for a new service or application to be successful, it must address one or more of the basic end-user needs that we have identified.

5.2 Application Trends

James Anderson and Patricia Morreale

Telecommunications applications provide solutions to the problems faced by people who wish to exchange information (end users). We define telecommunications applications as the software that provides end users with access to the functions that permit information to be exchanged. A wide variety of telecommunications applications are in use today: the software in telephone switches that provides such services as emergency 911, caller ID, three-way calling, etc; e-mail and Internet Web browsers; distributed synchronized databases such as Lotus Notes™, etc. Each type of application was developed to solve a specific set of end-user problems. Future telecommunications applications will also be developed to meet the needs of end users.

The types of future telecommunications applications will be directly related to the end-user needs discussed in the previous section. In order to focus our investigation of telecommunications applications, we will use the same segmentation of end users from the previous section. Figure 5.3 shows the stage of the Telecommunications Trend Analysis Model that is covered by this section. Our investigation will consist of two main parts: application functionality and functionality implementation. Looking at the application functionality trends that are occurring will help us to understand how application developers and service providers are working to address end-user communication needs. We will explore how this new functionality will be deployed in the real world when we go one step further and look at how vendors and service providers are planning to implement the new application functionality.



FIGu RE 5.3 Trend analysis—application trends.

5.2.1 Application Functionality

All four major categories of end users will require more functionality from their telecommunications applications. Because of the large purchasing power of each segment, competition among service providers has started to increase in the past few years. This trend is most noticeable in the United States and United Kingdom; however, the arrival of the European Union (EU) and a unified currency (the Euro) in Western Europe is also helping to make those telecommunications markets competitive.

A result of multiple competing service providers means that, at the very least, all segments of end users will shortly be presented with multiple sources for all existing services. Additionally, the number of services offered to end users will increase more rapidly than in the past due to the need for providers to distinguish their offerings from each other.

The eventual result will be that the telecommunications applications offered to all end-user segments will become more customized in order to meet the specific needs of a particular segment. Since end users are best suited to determining their exact needs, the process of subscribing to a telecommunications application will change from the selection of "all-or-nothing" applications, in which the end user had little or no choice, to participating in a "build-your-own" functionality selection in order to create a customized application.

This ability for end users to design their own applications will be the arrival of true multimedia applications that combine voice, video, and data features into a single customized application. This customization will cause the functionality provided by applications to increase over what is available in today's applications. New functionality will be apparent in the following five areas: Internet services, e-mail, videoconferencing, wireless services, and enhancements to traditional services. We will now look at application functionality improvements we can expect in each of these areas.

5.2.1.1 Internet Applications

The recent explosion in the popularity of the Internet (an unmanaged collection of interconnected computer networks that are all able to "speak" the same communications protocols) has forever changed what telecommunications applications will be expected to do. Studies of Internet usage are difficult to do because of its rapid growth; however, in the early 1990s the Internet was used by a handful of researchers and scientists. Current industry discussions center on the near-saturation of Internet access, with many users having more than one Internet-capable device, when computer and cellphone applications are considered together. With this kind of growth, it is very conceivable that the usage of the Internet will catch up to, and perhaps surpass, the use of the telephone in the not so distant future.

Today's Internet applications have secure and reliable functionality required for end users to perform e-commerce transactions efficiently. Electronic commerce (e-commerce) is the use of the Internet to facilitate the buying and selling of goods. The Internet offers sellers of goods the ultimate virtual storefront: without having to rent physical space, they can display and demonstrate their products for potential buyers.

Users are well aware of the fact that as they exchange information with a retailer's Internet application, it is possible for a malicious user to monitor and record their transaction. This could result in the malicious user obtaining credit card or bank account identification information that could then be used to steal funds from the unsuspecting user. Enhancements to functionality are being made to both the retailer's and the end user's applications. Basic encryption is now available that can be used to secure the transaction information before it is transmitted in order to negate the effect of any interception of the transmission. As this type of functionality is added to the end user's browsers and Internet-aware applications, user confidence in secure Internet transactions will increase and e-commerce can be expected to grow at an explosive rate. In the short term, some service providers are offering guarantees to make good on any losses incurred while using their networks in order to jumpstart e-commerce activities.

E-commerce is currently complicated by the lack of an agreed upon form of "digital cash." Despite gains in the past decade regarding the increasing use of credit and debit cards, the majority of retail transactions still occur using either paper money or checks. Neither of these two popular forms of exchange translate well to being used in the Internet's all-electronic environment. Once again, several different approaches to this problem are currently being investigated. Recent agreements among many of the major credit card companies have identified the required functional and exchange procedures that will be required to support electronic forms of currency for existing and new Internet applications.

Finally, as more and more of everyday life becomes computerized, consumers are starting to become concerned about how much information retailers are able to obtain regarding personal habits and buying patterns. As the use of the Internet to purchase goods increases, a retailer's ability to track the user's entire buying experience will also be increased. Such information could include a history of goods that the consumer looked at but did not purchase, how often and at what times of day the user visited a specific electronic store, and all of the products that the customer has ever purchased. Consumers have become alarmed that retailer's applications will be able to mine their purchasing history to target other goods for advertising purposes or that retailers will sell their information to other retailers for their use in trying to sell goods to the consumer. As Internet e-commerce applications mature, consumers are going to insist that retailers clearly identify what consumer-related information is being tracked and post their polices regarding use or sale of that data. Internet applications will have their functionality enhanced to support and enforce such privacy policies.

Although the Internet is a worldwide phenomenon, the majority of its content has been created in the English language. The reasons for this are varied; however, the origination of the Internet in the United States and the high availability of both computers and Internet access in English-speaking countries has definitely played a major role. Future Internet applications will be required to be able to deal with multiple languages. The tools to make this possible are slowly starting to emerge. Internet-based language translation products are now available that offer translation services for several languages. Whereas the amazing translation devices seen in some popular science-fiction movies may still be a long way off, the ability to translate text found on the Internet into another language or the ability to select a language for the purchasing process are just around the corner.

- Some service providers who are deploying high-speed digital access services are also establishing online communities built around high-speed access. These communities provide an opportunity for businesses to set up online shops, as well as a place for both residential and business customers to receive e-mail, purchase goods, access applications, and find out current event information for their local areas.
- Service providers are starting to explore the opportunities presented by integrated bills, accepting payment and providing customer care electronically over the Internet. Voice services and Internet services can be consolidated onto a single bill. Additional applications can electronically present the bill to customers and accept payments over the Internet. This type of application can be used with all types of telecommunications services including paging, IP voice, and long distance. An additional benefit of this approach is that it permits targeted marketing of specific customers and offers a better chance of capturing an impulse buying opportunity.
- Many vendors are looking for ways to replace today's ubiquitous fax machines. Some of the more innovative solutions are coming from companies that are trying to reduce their product support costs. One approach to directly provide a user with only specifically requested information uses Internet based "push" technology. This information delivery technique requires a user to log on to the company's server via an Internet connection. Then the company is able to push or force the display of specific information. The true power of this approach becomes clear when the user is able to talk with the company at the same time by using a separate line. These hybrid solutions are a cross between e-mail and fax services. Companies have found that this type of solution works

best when the company has a great deal of information that the user would otherwise have to work through in order to find what is needed.
- Firms are discovering that an estimated 10% of customers sometimes need assistance when using the firm's Web site. So-called "chat" applications are being added to Web sites to provide customers with the ability to receive real-time one-on-one guidance from employees of the firm.

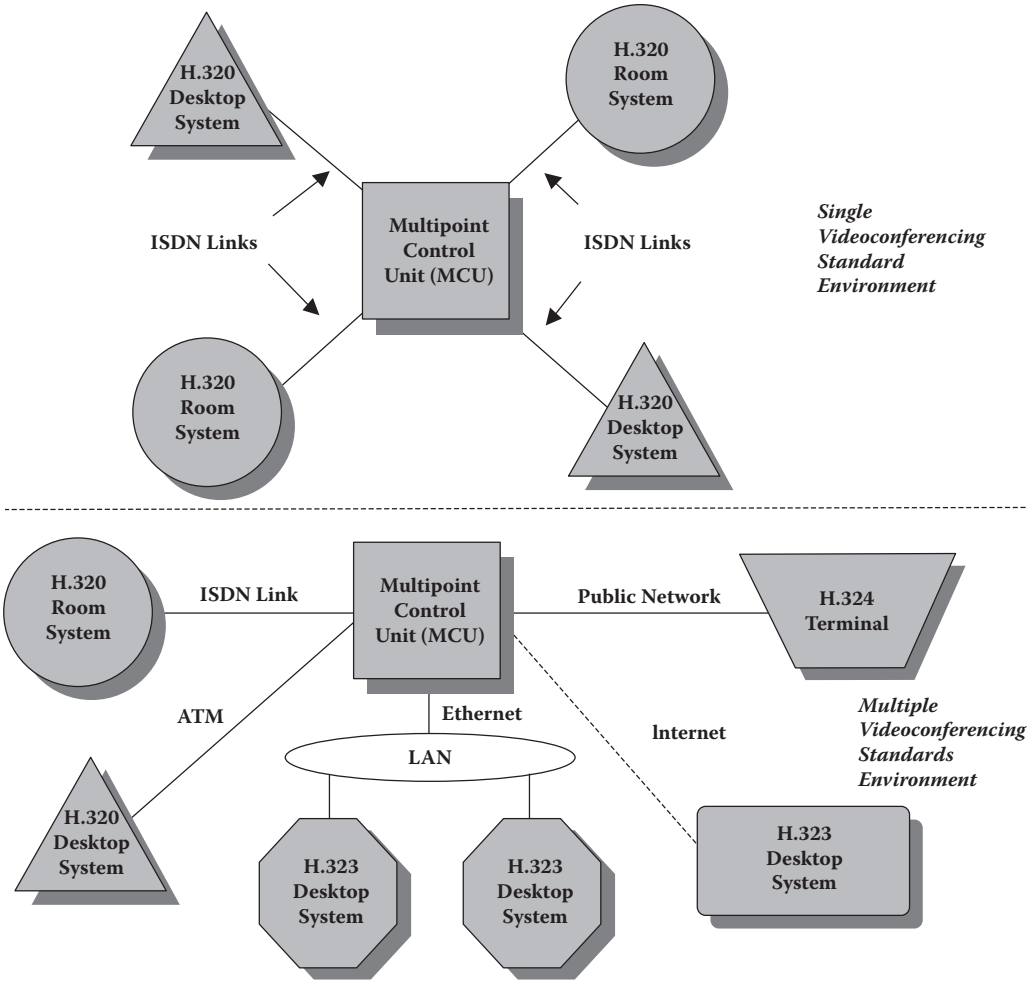
5.2.1.2 E-Mail Applications

E-mail has become such a critical part of how so many people communicate that we choose to treat its functionality separately from that of Internet applications. E-mail is widely understood to be the most popular online activity. Surfing the Web is regarded as the second most popular.

- Adding voice and video to e-mail represents the next step in e-mail's evolution. Service providers are now able to deliver e-mail that contains embedded links to additional voice and video components of the message. The additional e-mail components are then sent to the user through streaming technology that uses a service provider's computers to do the majority of the required processing, and then ships only the resulting images to be displayed on the end user's Web browser application. The challenge is to avoid disappointing the end users with poor application performance that causes them to revert to standard text-only messages.
- Estimates show that more than 40% of users' time on the Internet is spent on e-mail.
- E-mail is pervasive, fast, and relatively free. One of the next logical steps is to make it secure. The problem with existing e-mail is that it can be easily faked. Internet security has five key requirements: access control, authentication, privacy, integrity, and nonrepudiation.
- Current secure e-mail solutions use a public key infrastructure (PKI). PKI is a set of security services that can be used to provide security over public networks. PKI services consist of encryption, digital signatures, and digital certificates. PKI services require the use of a two-part key: a public key and a private key. Information is sent to a user after having been encrypted using their publicly advertised *public key*, and can only be decrypted using the user's secret *private key*. Every PKI exchange is monitored and authenticated by a company that provides digital security services.

5.2.1.3 Video Conferencing Applications

- Videoconferencing (Figure 5.4) offers many benefits, including savings in corporate travel and savings in employees' time. Important pieces must be in place for videoconferencing to happen: rising demand from multinational corporations, improvements in technology, solidification of key standards, and proliferation of standards-compliant video-enabled products from heavy hitters such as Microsoft and Intel. Key issues for service providers are reliability, quality, and ease of use. Current standards include those outlined in Table 5.1
- A major videoconferencing issue is service complexity: call-establishing can be delayed while all endpoints are configured to the same line speed, audio rate, frame speed, and resolution rate. Vendors and service providers' interpretation standards can also affect the service: a mismatch in interpretations can result in dropped calls. Videoconferencing systems that are able to talk to different standards-compliant endpoints are now becoming available (e.g., H.323/H.320 gateways).
- IP muliticasting will be able to provide multipoint H.323 videoconferencing. IP multicasting will save users' bandwidth on packet networks because the information needs to be transmitted only once over a given link, with routers replicating information as required. One challenge associated with multicasting is that it imposes a significant communications load on the processor at each endpoint since each endpoint must send information to every other endpoint. This means that IP multicasting is not currently scalable for large videoconferences.



FIGu RE 5.4 Videoconferencing environments.

TABl E 5.1 Video Conferencing Standards

Video Conferencing Standard	Purpose
H.320	Videoconferencing over ISDN
H.323	Videoconferencing over LANs, WANs, intranets, and the Internet
H.324	Videoconferencing over regular dial-up telephone lines

5.2.1.4 Wireless Applications

- Wireless data service providers are starting to shift their focus from vertical to horizontal applications. In the past, wireless data applications have been traditionally targeted at the public safety and utility markets. Newer applications target members of the financial community, such as bankers, analysts, and traders, by providing real-time access to stock information. One of the key success factors to entering horizontal business markets will depend on the service provider's ability to create appealing service bundles.
- In the United States, the future of the mobile data market is based on the cellular digital packet data (CDPD) technology. CDPD is TCP/IP implemented over cellular networks. CDPD is well

suited for certain types of transmission, especially short file transfers. CDPD was first specified in 1992; however, it has been slow to be adopted and there are currently fewer than 500,000 data customers on all U.S. cellular networks. Although CDPD may be well suited to supporting Internet-related applications, it is currently limited by two factors. The first is the fact that CDPD-based services are only available in selected markets. The second is that CDPD's bandwidth is currently limited to 19.2 kbps and actual connection throughput can drop as low as 2.4 kbps when network voice traffic is high. CDPD transmission rates as high as 56 kbps have been discussed; however, support for such rates is not currently provided.

5.2.1.5 Enhancements to Traditional Services

- Vendors are starting to work with service providers to create service solutions that meet end user needs. Unified messaging products are the first examples of such services. The service alerts users that they have e-mail via their service provider's Web site and their own voice/fax mailbox. This will be provided as a first step for customers who only want basic service. To be added: products that use text-to-speech technology. Good approach: everybody doesn't need everything. Future services include integrated e-mail, voice, and fax mailbox; nonsubscriber voice connect—allows e-mail users to send voice messages to anyone; and consolidated wireless/wireline mailbox with improved phones that contain text display screens.
- The Universal International Freephone Number (UIFN) system allows a single toll-free number to be used around the world. Users apply to the International Telecommunications Union (ITU) for an eight-digit number that is accessible by dialing the appropriate international access code, 800, and then the new number.
- Many new telecommunications applications are being developed for call centers. These applications are being designed to help companies gather information about their customers and make sure that the products and services that the company offers are meeting the needs of their customers. This type of application uses computer telephony integration (CTI), automatic call distributors (ACDs), and interactive voice response (IVR) systems.
- One of the primary motivations for firms to use virtual private networks (VPNs) is to avoid the costs of expensive dedicated leased lines. Vendors are now making VPN products that contain combinations of functions, including serving as IP routers, corporate firewalls, and certificate authorities, along with the required VPN functions of encryption and authentication. A key drawback to today's VPN products is that the processing power required to perform VPN functions such as encryption severely limit the throughput of the devices.

5.2.2 Functionality Implementation

The enhanced telecommunications application functionality described in the previous section requires that the way applications are designed must be radically altered. As the computing equipment available to end users continues to improve, the intelligence required to support the application is migrating from within the network to the endpoints. In new and emerging applications, much of an application's functionality may reside in the end user's equipment. This is dramatically changing how networks are designed. We will investigate these types of changes later in this chapter.

The competitive environment that service providers are starting to operate in will no longer permit deployment of new applications at the current somewhat leisurely rate. End users will demand new applications as soon as they identify problems they need to solve. The service provider who is the first to be able to offer a solution to such end users stands the best chance of capturing the largest share of the market. Past history has clearly shown providers that it is better to be first to market and bring additional functionality later rather than wait until a new application is perfect.

The arrival of networking equipment that is able to provide exponentially larger amounts of bandwidth will aid developers of new telecommunications applications. Table 5.2 identifies several of the

TABl E 5.2 Standard Transport Bandwidths

Transport Type	Bandwidth
OC-3	156 Mbps
OC-12	622 Mbps
OC-48	2.5 Gbps
OC-192	10 Gbps

network bandwidths now available for use with new applications. The result of greater bandwidth availability is that less development time will have to be spent attempting to minimize the amount of data that telecommunications applications exchange. This reduced development time will result in applications that are richer in functionality being made available to end users more rapidly.

Recent increases in the amount of bandwidth provided by data networking equipment, coupled with the initial availability of products that can provide voice services over a data network, have fueled a focus on Internet Protocol (IP)-based networks. As competitive service providers build new networks to provide services, they are selecting networking equipment that permits them to build IP-based networks rather than the traditional Class 5 voice switches. These new service providers believe that in the very near future all information (voice, video, and data) transported by a provider will be viewed as data and can be encapsulated in the IP data network protocol.

If current application trends continue as expected, almost all future telecommunications applications will be "Web-Aware." Simply put, this means that such applications will have the ability to obtain information from and provide information to other applications via World Wide Web (WWW) Internet protocols. The Java programming language has popularized a highly distributed programming model that will influence the design of such future applications. In this model, the network has the responsibility for advertising what applications it supports and storing the logic required to provide the application. The end user's customer premises equipment (CPE) will then download the needed functionality and execute it locally, thus distributing application processing from the network's limited resources.

The use of data networks for telecommunications application interconnection will have the interesting side effect of causing what has been called the "death of distance." Because end users are currently charged for the size of the connection that they use to access the Internet, it no longer matters how far the data travels once it is transmitted. This will result in a greater use of more widely distributed applications. A more detailed discussion of the effects of changes in telecommunications application pricing is provided later in this chapter.

One of the greatest bottlenecks in deploying new telecommunications applications resides in the back-office operations of the service providers themselves. After an end user selects a service and negotiates any customizations, there is often a delay (sometimes a significant delay) as the provider processes the order and reconfigures its network to deliver the requested service. The telecommunications applications introduced in a competitive environment must be deployed with minimal support costs and must start to generate revenue as quickly as possible. One of the most promising means of accomplishing both of these goals simultaneously is to automate the telecommunications application service ordering process. Assuming that the obvious security issues can be solved, interfacing the application directly to the provider's operation support systems (OSS) will reduce the support costs for the application while at the same time decreasing the delay between when the service is ordered and when the application is available to the end user.

Finally, the near-panic caused by the so-called Year 2000 (Y2K) software bug, which caused some applications to be unable to distinguish between 1900 and 2000 due to historical efforts by software developers to minimize the amount of memory required to execute an application, will forever change how telecommunications applications are developed. Immediately after having experienced the expense and turmoil caused by the hunt for potential Y2K errors in hundreds if not thousands of hardware platforms, operating systems, and applications, end users can be expected to demand protection from

future errors. Although complete protection from software errors can never be guaranteed, new telecommunications applications will most certainly contain enhanced testing capabilities that will permit the end user to simulate program execution in an off-line environment in order to determine how it will react to a given set of inputs.

Let's now take a close look at some of the issues surrounding how some of this enhanced application functionality will be implemented in two important segments of telecommunications applications: Internet applications and wireless applications.

5.2.2.1 Internet Functionality Implementation

- A carrier-grade IP telephony gatekeeper that complies with the emerging H.323 standard is now available in evaluation versions. This product can be used to tie together IP and public network gateway systems from other vendors. This product is significant because it represents the first phase of multivendor interoperability. Ericsson plans on using applications to differentiate its gatekeeper product—specifically for applications that are better suited to reside inside the carrier network.
- Hammer Technologies has introduced an IP test system that monitors the quality of voice on IP networks. The system automatically tests voice quality, measures audio quality, and includes a Voice over Internet Protocol (VoIP) analysis tool as an IP traffic generator.
- Microsoft, Sun, and others are competing to supply commercial Web server and application platforms to public network service providers. Each of these companies has a different vision of what the next generation of telecommunications applications will look like. Microsoft sees applications being built on top of low-cost PC-based Microsoft operating systems. Sun sees applications running on open, fault-tolerant systems that use the Java language.
- Some service providers are aggressively deploying advanced high-speed digital subscriber line (DSL) services and fiber-optical service (FiOS). Many of these service providers own and operate switch-based networks and feel that a switched network infrastructure routes packets faster and more reliably than a routed one. Such providers are offering Internet access and LAN-like services.
- In order to provide access to popular Internet content to users in other countries, creative applications are being developed to distribute the information. Using a combination of satellite links, multicasting software, and local caching, service providers are using public Internet kiosks to permit users to view the most popular Web pages. This eliminates long waits for dial tones and conflicts over access to what precious bandwidth exists. This new approach "pushes" content to where the user is instead of requiring the user to pull content off of North American servers.
- Dynamic HTML will allow designers to make richer, multilayered pages. Dynamic HTML will allow designers to create Web pages more efficiently so each link of information doesn't have to be downloaded from the server.

5.2.2.2 Wireless Functionality Implementation

- Microsoft has announced that it is developing a nonstandard microbrowser as a part of its goal to enter the wireless data marketplace. The microbrowser will permit wireless users to browse the Web, provision services, and access billing information. The Wireless Application Protocol Forum released the open wireless mark-up language (WML) microbrowser specification.
- Researchers have been able to crack the messaging encryption algorithm used in U.S.-based code division multiple access (CDMA) and time division multiple access (TDMA) digital cellular networks. The researchers have broken the Cellular Message Encryption Algorithm (CEMA) code. The CEMA code has been designed to safeguard dialed digits that are sent over the airwaves. Different encryption algorithms are employed for user authentication and voice privacy. The reason that the researchers were able to crack the CEMA code was, in part, due to the fact that the wireless industry has watered down its security algorithms in order to appease the U.S. federal government.

5.2.3 Network Neutrality

Net neutrality has been assumed ever since the Internet was nothing more than a diagram on a napkin. However, as the number of users grew, on a global magnitude, and finite resources were shared and maintained by all, governs over the assumed neutrality of the net have grown. This is true both from the perspective of government authorities, not always democratic and wanting to manage the ideas and images their citizens viewed, to commercial entities, wanting to ensure that all customers have equal access to bandwidth and services. The contrast between e-commerce companies that wish to make money and ISP networks that provide the pathway to revenue (WEIT08) has come into conflict.

Network neutrality, frequently referred to as net neutrality, refers to a commonly held principle applied originally to the earliest ARPANet and extended to today's Internet, including all residential and commercial broadband networks, and is defined as follows: a neutral broadband network is one that is free of restrictions on the kinds of equipment that may be attached, on the modes of communication allowed, without restriction of content, sites or platforms, and where communication is not unreasonably degraded by other communication streams.

Previous discussions of net neutrality centered on whether on not it was a problem. However, in recent years, as global use of the Internet has increased and the time-critical functions that both individuals and governments place on the network has reshaped the earlier question from "Is it a problem?" to "Should actions be taken now to assure network neutrality in the future?" Within the United States, this is considered a policy decision. Other countries may address the problem in a stronger or lighter fashion, depending on the nature of their own government and the expectations of their citizens. Currently, the debate continues and it is likely that whatever approach the United States adopts will serve as a comparative standard for the rest of the world.

5.2.4 Semantic Web

The semantic web refers to an evolving extension of the Web in which the semantics, or meaning, of information and services is defined in such a way that Web content can be used to provide information to people. The elements of the semantic web will most likely include features and other characteristics of content definition not yet found on the Web today. The semantic web is to today's Web as structured programming was to object-oriented programming in years past.

The growth of the semantic web is motivated by the increasing amount content placed on the Web and the need for a navigation process, transparent to the user, to collect the greatest amount of material with correspondence using a method independent of users.

5.2.5 Social Networking

Social networking refers to the informal establishment of online communities, exemplified by MySpace, Facebook, or the professional counterpart, LinkedIn. These online communities are composed of people who share interests and activities and are organically grown, that is, there is no master plan or compulsory joining expectation. In contrast to years past when social networks were established in a face-to-face manner, most of today's social networks are Web based and provide many ways for users to interact, ranging from instant messaging to e-mail. Many individuals who are connected by a social or professional networking site may have, in fact, never met. Their online linkage through Web page connectivity is predicated on knowing mutual friends or having been coworkers.

While the benefits of social networking include keeping in touch with friends and colleagues in a manner never before possible, the persistent nature of Web communication is such that materials an individual may not wish to display, such as photographs or videos, may be linked to his or her site by acquaintances who attended an event with him. Never before has the expression "your friends are an

extension of you" been so true. Much more information is available via social networking sites than an individual might care to reveal to possible employers or teachers. Therefore, the advantages of social networking are almost as great at the potential disadvantages.

Reference

WEIT08: Weitzner, D. 2008. Net Neutrality … Seriously This Time, *IEEE Internet Computing* 12, No. 3 (May/June) 86–89.

5.3 Systems and Service Integration for Management

James Anderson and Kornel Terplan

5.3.1 Introduction

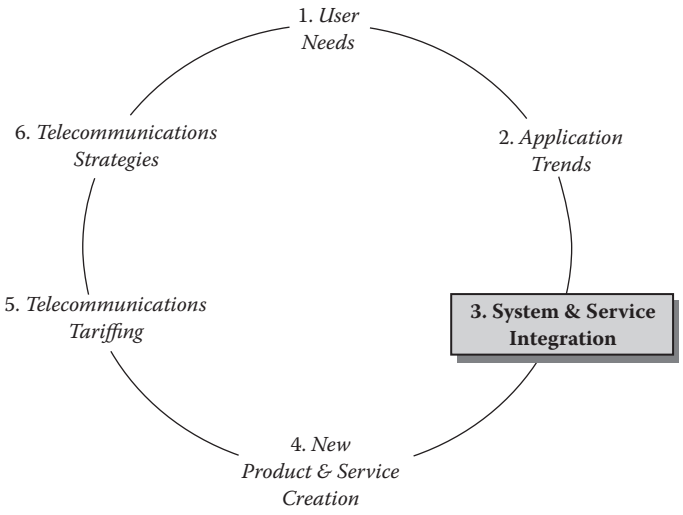
As telecommunications technologies are able to provide end users with more and more complex services with an increasing number of interrelated features, end users have started to complain. Just as when you go to purchase a car, you don't want to be required to make decisions regarding issues that are relatively unimportant to you. An example of this would be when taking an airline flight, you do care about flight times and where you sit; however, you don't care what altitude you fly at or what movie is shown. End users simply want to be able to use telecommunications services to make their lives easier and to make themselves more productive—they don't want to be telecommunications experts in order to select and use such services.

Service providers and network equipment vendors are responding to these needs by integrating what are currently separate service offerings into new feature-rich services and by consolidating technology-specific networks into single networks that are simultaneously able to handle voice, video, and data information exchanges. In this section, we will explore some of the drivers for service and system integration and identify how these are going to affect the telecommunications services available to end users in the future. We will then take a look at specific trends in service and system integration as they relate to each one of the four classes of end users identified earlier in this chapter (see Figure 5.5).

5.3.2 Drivers for Integration

Integration of services and systems requires both considerable effort and expense. In order to make such an investment worthwhile, there needs to be a future payoff for service providers and application developers who make the integrated solutions. In fact, there are several distinctly different motivations that are in the process of creating integrated solutions. The drivers for integrated services are as follows:

- **Competitive differentiation:** As the number of service providers is increasing, the number of end users in each segment is remaining relatively constant. This means that service providers will only be able to grow by wooing end users away from their current providers. In order to accomplish this, a provider will have to be able offer the end user a compelling reason to switch. Integrated services can be such an enticement, and such tactics are starting to appear in the form of "follow me" offerings where voice, paging, and mobile services are linked to a single service. With such a service, someone trying to contact the end user dials a single number, which then attempts to establish a connection with the called party via each different communication method. If the desired end user is not reachable, then a message can be left on a voice mail system that the end user can check via any of the available technologies.
- **Single provider:** Recent surveys of end users have revealed that, all other things (such as price) being equal, users desire to receive all of their services from a single provider. The reasons are



FIGu RE 5.5 Trend analysis—system and service integration.

simple: a single provider means a single bill and one number to call in the event of any problems with the service.

- **Technology advancements:** The integration of multiple services into a single offering to the end user has its own potential risks. An integrated application requires a significant amount of end-user customization in order to provide the maximum benefit. An example of an integrated service that has suffered from low end-user acceptance due in part to its complex configuration is the Integrated Digital Services Network (IDSN). Advances in network intelligence and equipment processing allow the configuration of new services to be simplified and have allowed much of the configuration process to be performed automatically by the network equipment itself. Additionally, improved network element processing has permitted multiple elements in different technology domains to exchange the required information to support integrated services.
- **Improved billing systems:** Amazingly enough, one of the greatest limitations on integrating services has been the billing system used by service providers. Such large and complex billing applications were originally designed to support a specific set of services offered via a single technology network.

Likewise, there are several identifiable drivers working together to motivate service providers to create integrated systems. The drivers for integrated systems are:

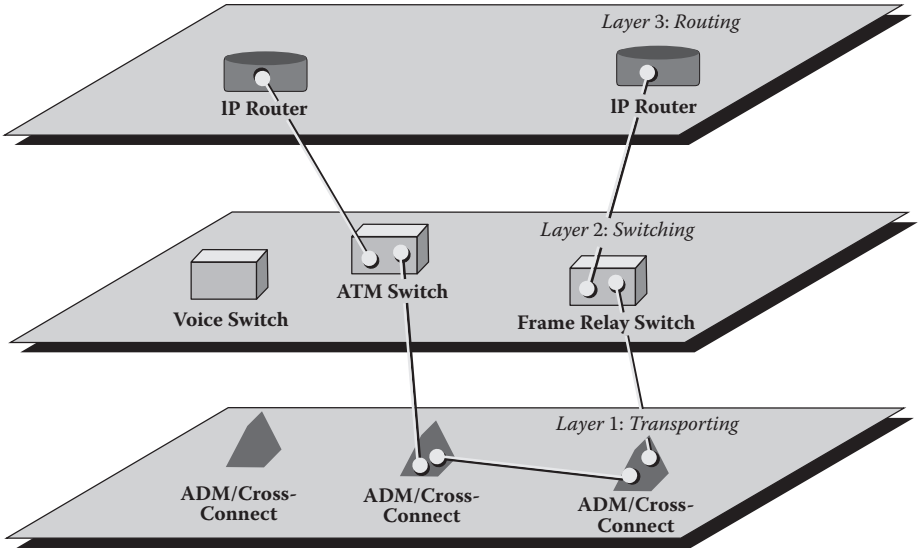
- **Reduced network deployment costs:** End users are starting to demand services that have voice, video, and data components. Service providers will have the choice of building separate redundant networks to provide such services or of building a single high-speed network to handle all three forms of communication. As you may well imagine, the decision to build a single network becomes very straightforward once the economics of building a single network to deliver all services is considered.
- **Reduced operations costs:** A significant cost of delivering a service to an end user can be attributed to the operational expenses required to keep the network working correctly. The use of an integrated network reduces the number of network elements required to deliver services, simplifying operational requirements, and thereby lowering the ongoing cost of offering the service.
- **Bandwidth breakthroughs:** The possibility of using an integrated system to offer services to end users could not be realized until improvements in network equipment and transport technologies

occurred. Recent increases in the bandwidth that can be provided by a single network have made it possible to build a single network that can support multiple services.

- **Tariffing:** Existing tariffing of telecommunications systems was designed years ago when the primary offerings to end users were voice services. Data networks are currently free of many of the limitations that restrict what and where services can be offered via traditional voice networks. We will discuss specific tariffing-based motivations later in this chapter.

5.3.3 Integration for Service Providers

- The arrival of wavelength division multiplexing (WDM) systems has caused service providers to reevaluate their existing time division multiplexing (TDM) systems. Network planners currently believe that the two different approaches can be used together to create networks that provide the lowest bandwidth costs.
- As service providers prepare to reshape their circuit-switched networks into IP packet-routed networks, the issue arises about what type of operation support system (OSS) will be needed. Service providers have a range of functions to support: provisioning and integrating of new services, service assurance, and network management. Existing service providers will most probably address this problem by reusing part or all of their existing billing or customer care systems. The greatest challenges will come in the areas of network management and provisioning.
- Traditional circuit-switched system service providers are watching the success of facilities-based Internet service providers (ISPs) and their packet-switching and routing forwarding networks. In addition, the large circuit-switched network equipment vendors are also modifying their equipment in order to transition them to work in a packet-switched environment. A key challenge is that the existing circuit-switched, connection-oriented public network infrastructure has been built up over decades and includes layers of resiliency and fault tolerance built in, which are not currently part of packet-based data networks. Another key point is that time-slotting information in hardware allows for guaranteed latency and delay parameters that simply cannot be guaranteed in many packet-switched systems. One possible future evolution is that the circuit-switched network will serve as a mission-critical backup system for the public packet-based system and will only be used for those cases where the "call must go through." A possible causality of the move to a packet-based public network could be the current computer-telephony integration (CTI) market.
- ISPs are now at the front line of telecommunications equipment design. Today's ISPs are building their own facilities, laying their own optical fiber, and installing their own carrier-class switches in points-of-presence (POPs). Traditional circuit-switched service providers have been taking more data and even voice traffic off traditional circuit switches and putting the traffic on packet-switched networks that were formerly considered to be data-overlay networks.
- The current public network consists of voice switches interconnected via transport systems. New user demands are causing this network architecture to be reshaped to now support voice, video, and data services. This new network architecture (Figure 5.6) uses a transport infrastructure that supplies the required interconnectivity to create its foundation. The architecture's switching layer provides the call setup and teardown functions throughout the network that are required to deliver services using a variety of protocols. Finally, a routing layer is used to provide the final step in the process of delivering data services to end users.
- As service providers start to offer services that use multiple technologies, equipment vendors are modifying their existing equipment to support the providers' new needs. Traditional voice switch vendors are enhancing their wireline switching products to also support wireless services. Some switch architectures are so flexible that providers can mix and match wireline and wireless modules to permit subscribers to connect to a cell site or the public network using the same switch.



FIGu RE 5.6 A new network architecture.

- U.S. West reports that the average voice call is approximately 5 minutes. The average data call is about 32 minutes—this is causing congestion in the central office.
- U.S. West is conducting trials with two ISPs to weed out Internet traffic from voice calls. Using distributed SS7 technology, ISP-bound data calls are identified at the user's ingress switch and immediately routed to the ISP over a parallel data network. This differs from the way traditional end-user data calls are set up. Data calls are normally routed to the ISP through the phone network via ISDN or digital switched services such as channelized T1. The data call rides the public network the whole way.
- Service providers are showing a renewed interest in video services. Vendors are demonstrating products that can push 26 Mbps over existing twisted pair wiring for up to 4000 feet. Broadband wireless equipment vendors are mainly focused on data applications; however, some have demonstrated videoconferencing and distance learning applications.

5.3.4 Integration for Business Users

- Some observers suggest that more than 60% of the costs associated with modern data networking lie in the cost of ownership.
- Standard Ethernet LANs typically offer 10 Mbps of shared bandwidth (see Table 5.3). As the volume of network traffic increases, however, this amount of bandwidth quickly becomes inadequate to maintain acceptable performance to support demanding applications. These traffic jams are fueling the need for higher-speed networks. Fast Ethernet, or 100BASE-T, has become the leading choice of high-speed LAN technologies. Building on the near-universal acceptance of 10BASE-T Ethernet, Fast Ethernet technology provides a smooth, nondisruptive evolution to 100 Mbps performance. The growing use of 100BASE-T connections to servers and desktops, however, is creating a clear need for an even higher-speed network technology at the backbone and server level. Ideally, this technology should also provide a smooth upgrade path, be cost effective, and not require retraining. The most appropriate solution now in development is Gigabit Ethernet. Gigabit Ethernet will provide 1 Gbps bandwidth for campus networks with the simplicity of Ethernet at a lower cost than other technologies of comparable speed. Gigabit Ethernet will be an ideal backbone interconnect technology for use between 10/100BASE-T switches, as a connection to high-

TABl E 5.3 Applications Driving Network Growth[a]

Application	Data Types/Sizes	Network Traffic Implication	Network Need
Scientific modeling, engineering	Data files 100s of megabytes to gigabytes	Large files increase bandwidth required	Higher bandwidth for desktops, servers, and backbones
Publications, medical data transfer	Data files 100s of megabytes to gigabytes	Large files increase bandwidth required Low transmission latency High volume of data streams	Higher bandwidth for desktops, servers, and backbones
Internet/Intranet	Data files now Audio now Video is emerging High transaction rate Large files, 1 MB to 100 MB	Large files increase bandwidth required Low transmission latency High volume of data streams	Higher bandwidth for desktops, servers and backbones Low latency
Data warehousing, network backup	Data files Gigabytes to terabytes	Large files increase bandwidth required Transmitted during fixed time period	Higher bandwidth for desktops, servers, and backbones Low latency
Desktop video conferencing, interactive whiteboarding	Constant data stream 1.5 to 3.5 Mbps at the desktop	Class of service reservation High volume of data streams	Higher bandwidth for desktops, servers, and backbones Low latency Predictable latency

[a] Source: Gigabit Ethernet Alliance, www.ethernetalliance.org, 1998.

performance servers, and as an upgrade path for future high-end desktop computers requiring more bandwidth than 100BASE-T can offer.

- Although Gigabit Ethernet is primarily an enterprise LAN technology, several service providers (most of them ISPs) have begun evaluating it for use in local and metropolitan area sections of their networks. Gigabit Ethernet can connect network equipment such as the server, routers, and switches within a service provider's POP, both inexpensively and at high speeds. One of Gigabit Ethernet's biggest selling points is that it is cheaper and faster than asynchronous transfer mode (ATM) or Synchronous Optical Network (SONET), which many service providers now use to link gear in their POPs. Gigabit Ethernet's heavy data orientation and distance limitations are red flags, however, for established telcos looking for technologies that can support voice, video, and data.
- Inverse Multiplexing for ATM (IMA) is a specification for provisioning multiple ATM circuits in T1 increments. IMA was created to bridge the bandwidth gap between T1 (1.544 Mbps) and T3 (45 Mbps). Using IMA, several low-cost T1 lines can be used to aggregate the bandwidth and distribute ATM traffic across multiple physical circuits.
- Frame relay speed and capacity improvements are being designed in order to keep pace with the needs of the new public network for data services. The two major changes to frame relay are the emerging frame relay over SONET (FROSONET) and multilink frame relay (MLFR) standards. FROSONET provides specifications for frame relay to run at OC3/STM-1 or OC12/STM-4 speeds. MLFR adds scalability to frame relay networks, thus helping service providers keep pace with growing traffic demands while providing an incremental capacity jump for users who are outgrowing T1 capacity but are not ready for the speed or expense of DS3/E3 lines. The FROSONET specification uses the same high-level data link control (HDLC) over SONET mapping that is being used for Point-to-Point Protocol (PPP) over SONET (PoS). This saves costs by allowing the same hardware to be used for both PPP and frame relay interfaces. MLFR trunks combine

multiple physical links between switches in the public network into a single higher-capacity logical facility. Additionally, frame relay's existing quality of service (QoS) functionality permits it to be used by service providers to offer such capabilities as service-level agreements (SLAs) and customer network management (CNM) functionality.

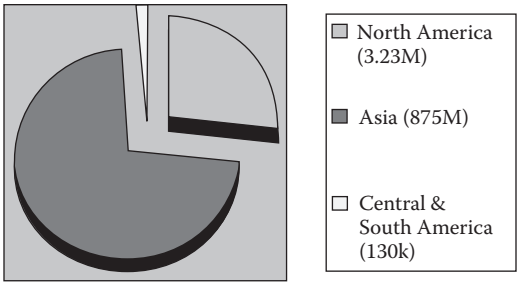
- Time division multiplexing (TDM) is used to combine individual connections in order to traverse longer distances. Switched circuits, such as those used in telephone networks, provide dedicated connections between two points. Switched-packet protocols, such as Ethernet, provide good utilization of the backbone but have no provisions for providing the equivalent of a switched circuit over a network. Switched-cell protocols such as ATM provide good utilization of the backbone and have provisions for providing CBR and UBR virtual circuits, but are expensive when compared to the newer switched packet systems such as Gigabit Ethernet.
- Wide area networks (WANs) tend to be rings like fiber distributed data interface (FDDI) or various star configurations. Generally, the number of entry points into the network tends to be very limited. WANs are designed to transmit data over long distances, tend to be focused on isochronous data, and lean toward circuit switching because they were often devised by telephone companies to carry voice.
- Five leading car and truck manufacturers have banded together to lead the Automotive Network Exchange (ANX) project. This is designed to create a specialized high-end, Internet-like VPN to link North American automakers and their suppliers. This very reliable and secure network may act as the beginnings of a parallel "Business Internet."

5.3.5 Integration for Mobile Professionals

- The CDMA Development Group (CDG) is coordinating location technology trials among member carriers and vendors. Trial focus will be on the three types of technology: global-positioning system-based, network-based, and a combination of the two. The Federal Communications Commission (FCC) has mandated that carriers are required to be able to locate callers within 125 meters. Some carriers believe that they will ultimately implement multiple location technologies. Network-based solutions are less precise but may meet the FCC mandate. A handset-based solution may locate users more accurately and, if used with a network-based solution, may allow a carrier to offer enhanced services such as location-sensitive billing and concierge services. See Figure 5.7 for information on worldwide CMDA subscribers.
- What standard will be used for the next generation of wireless services is still undecided. Possibilities include W-TDMA, which suffers from limitations in growth, and W-CDMA, which suffers from limitations in power and processing.

5.3.6 Integration for SOHO Users

- In September 1998, the ITU ratified a single standard (V.90) for 56-kbps access over the Public Switched Telephone Network (PSTN). V.90 data transmission technology overcomes the theoretical limitations imposed on standard analog modems by using the digital server connections that most Internet and online service providers use at their end connection to the PSTN.
- Community area networks (CANs), as represented by cable modems, have a unique topology that is not served well by existing LAN or WAN topologies. They have a large connection count of shared wire-like LANs, but have distances like those of WANs. Current CAN implementations generally utilize a single downstream CATV channel that is shared by all network participants. A separate upstream CATV channel is also shared for transmitting from the home to the cable head end.
- Low and medium earth orbit satellite systems (LEO and MEO) provide yet another possible way to connect SOHO users who are in remote locations. The challenge with this type of solution is ensuring that there are enough LEO/MEO satellites in orbit to provide constant coverage for users.



FIGu RE 5.7 Worldwide CDMA subscribers (CDMA Development Group).

- Geostationary (GEO) satellite systems are being used to deliver data broadcasts. There are two primary types of GEO services: very small aperture terminals (VSAT) and direct broadcast satellite (DBS). Two DBS services have been announced: an entertainment service and a data service. The data service will serve corporate customers with occasional and regular broadcasts, along with residential customer service. One possible use of the data service is for software distribution. Another company is offering three versions of its DBS Internet access service: direct delivery of text files at 12 Mbps, multimedia at 3 Mbps, and Internet access at 400 kbps. This service is asymmetrical: customers send information requests to the service provider via telephone lines and receive data via the customer's 24-inch antenna. VSATs are used for corporate broadcasts of data including price updates. VSATs operate at speeds between 14 kbps and 64 kbps, with high-speed bidirectional communication
- Home-based LANs are becoming more and more popular as homes start to have multiple computers, which means that users want to share limited resources such as printers and storage systems as well as exchange files. The arrival of Wi-Fi networking for residential use has greatly simplified the effort required to set up such a LAN.
- According to U.S. West, DSL services can be up to 250 times faster than a 28.8 kbps modem. An additional advantage is that while the DSL service is being used, the phone line can still be used to both make and receive voice calls.
- Virtual private networks (VPNs) provide an alternative to leased-line connections. VPNs provide an inexpensive and secure way to extend the corporate network to telecommuters, home workers, day extenders, remote offices, and business partners.
- VPNs are implemented through tunneling, in which the data to be transmitted and header information are encapsulated inside standard IP packets, usually after encryption and sometimes after compression as well.
- Three VPN tunneling protocols are currently in line to become industry standards: PPTP (Point-to-Point Tunneling Protocol), Layer 2 Tunneling Protocol (L2TP), and IPSec (IP Security).
- Security is a critical component of a VPN implementation, especially for those implemented over the public Internet. Encryption delivers the "private" in virtual private networking, but it is very process intensive. Because of this, hardware-based VPN products deliver the best performance.
- Despite some severe quality limitations, users have already started using the Internet to deliver videoconferencing. Existing corporate videoconferencing systems deliver full-screen images at 30 frames per second—a much higher-quality image. Internet-based videoconferencing services won't be real until new QoS standards are in place such as the Resource Reservation Protocol (RSVP) and IP version 6 (IP6).

5.3.7 Integration for Residential Users

- Researchers are issuing cautions regarding unsolved problems with digital subscriber line (DSL) and cable modem services. Complexities have been identified regarding the mixing of plain old

telephone service (POTS) and DSL services without using a splitter. One of the biggest issues concerns what happens when a user picks up a telephone handset in a splitterless service—the result is an immediate change in load on the local loop, resulting in a loss of one to two orders of magnitude in signal amplitude. Additionally, crosstalk can occur when several POTS twisted pairs in the same bundle are used to provide DSL service.

- A new service that is being considered combines Internet and television services so that end users can simultaneously surf the Internet and watch enhanced broadcast television at home. One approach uses cable television systems to deliver downstream data to advanced set-top boxes. Other approaches are more software oriented and don't necessarily need set-top boxes.
- The FCC has mandated that broadcasters must have started offering some high-definition television (HDTV) digital programming in 1999 and complete their transition to digital by 2009.
- MSNBC, the cable broadcaster owned by Microsoft and NBC, has experimented with technology that allows broadcasters to send digital signals embedded within television signals to PCs. MTV, along with Intel, launched Intercast Jam, which broadcasts videos to PCs alongside rock artist information via a Web browser in the broadcast signal.
- For years, TV stations have been beaming out data in small doses in the form of closed captioning, test signals, ghost canceling, and messages to affiliates. That data is carried mainly through the vertical blanking interval (VBI). This offers a total of between 150 and 200 kbps of available bandwidth. This bandwidth is now being used by the Intercast consortium to transmit ancillary data streams to PCs via the VBI.
- Vendors are starting to create products that support videophone services. One such product puts a video camera in a set-top box and displays its image on any cable-ready TV. A touch-tone phone provides audio, dialing, and navigation of the system's on-screen controls. The system includes a built-in 10BaseT Ethernet interface to link directly to a cable modem, digital subscriber line modem, or corporate network.

5.3.8 Unified Threat Management (UTM)

Many of the advantages of integrating security functions into a single device are obvious: reduced cost, consolidated reporting, a consistent interface, simplified network architecture, and ease of management. In fact, these trends are also driving consolidation of the endpoint security suite market. In addition, there are less obvious benefits, too. One of them is green computing. Consolidation does not only mean virtualization. Reducing the number of security devices by the way of UTM is another way to save on power bills. For more mature products in the space, UTM provides synergy. For example, an antivirus module, an antispam module, and a content-filtering module might all share the same database of known bad URLs that each would apply appropriately. Some vendors maximize this integrated approach; their entry platforms are designed around UTM from the beginning. These vendors usually maintain their own modules and signature databases what is rare among UTM vendors.

But there are disadvantages, too. It is unlikely that any single UTM product offers the best-of-breed in each of its parts. Actual performance, when all modules are turned on, might be completely different from estimated KPIs. Many products feature hardware acceleration for performance reasons.

From today's UTM solutions, the following features are expected to be met:

- Support of firewall functions: firewalls have fallen from favor lately. However, in a true default-deny deployment, a firewall is still a very solid security move, and the firewall origin of UTM devices illustrates their appropriate deployment model. Anywhere the network is well enough understood to deploy a firewall likely a good place to apply other security features. Moreover, firewall vendors have not stood still. The latest iteration of the firewall philosophy sees concepts once applied to Layers 1 to 3 of the Open Systems Interconnection (OSI) network model targeted across the entire stack, allowing rules to be applied to data and applications regardless of port or protocol.

- Support of IDS and IPS functions: While the philosophies behind IDS and IPS are different, their value over traditional firewalls centered on the ability to peer inside packets, as opposed to just basing security decisions on packet headers. Unfortunately, an IPS represents a default-allow. Rules are created to block attacks, vulnerabilities, or other suspicious behavior, but otherwise everything gets through. While much malicious traffic can be mitigated with a properly tuned IPS, in general IDS/IPS vendors have had a difficult time keeping up with the shift from server attacks to client-side exploits. Network gateway intrusion-prevention systems suffer from a perspective problem. Attacks that don't simply target the network socket are difficult to detect. For example, a browser plug-in exploit can be sent inside encrypted JavaScript that is compressed inside HTTP content encoding that is encrypted inside of an SSL transaction. Without software running on the endpoint or the ability to perfectly emulate an endpoint—something no network IPS is capable of—there is no reliable way an IPS to stop these types of attacks.
- Support of anvirus: The original UTM definition included gateway antivirus, which typically meant SMTP and HTTP scanning. Some products extended their protections into peer-to-peer protocols, file transfer protocols, or chat clients. There is no pure antivirus in recent UTM products. Instead, antispyware, antispam, and antimalware features are represented. The latest technologies use behavioral scanning to implement checks on files being transferred to identify potential threats without relying on a static fingerprint database. Of course, whether detection is behavioral or signature-based, it is still an example of a default-allow policy.
- Network infrastructure capabilities: Many UTM products are fully functional network infrastructure-in-a-box, including VPNs. They may include such features as Natural Address Translation (NAT) quality of service. The VPN functionality in UTM products includes site-to-site as well as SSL or other client-server VPN technologies that let remote employees access internal resources.
- Content filtering: Usually associated with Web-content filtering via a URL blacklist service, content filtering is often touted as a productivity increaser. Of course, savvy users almost find ways to get access to the content they want to view and occasionally download.
- Data leakage prevention: DLP products have been all the rage these past few years, and they must become part of UTM. A number of products now have some simple data-leakage mechanisms, like e-mail keyword filtering or blocking of attachments in e-mails.
- Network access control: While NAC or admission control, depending on who is selling or using it, can be hard enough to define even when not mixed in with many other capabilities, it makes sense for products that are doing everything else security related on the network to control access, enforce endpoint compliance, or integrate with other NAC systems.
- Identity-based access control: This is a relatively new security technology for network gateway products. Operating systems, applications, and other devices have used the idea of authentication and authorization since the beginning of computing. Applied to network gateways, however, it is a relatively new approach.

In summary, UTM solutions should seamlessly collaborate with OSS and BSS at service providers, and with network management platforms in enterprise environments.

Acronyms

DLP	Data Leakage Protection
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
KPI	Key Performance Indicator
NAC	Network Admission Control
NAT	Natural Address Translation

SSL Secure Socket Layer
UTM Unified Threat Management
VPN Virtual Private Network

5.3.9 Network Behavior Analysis (NBA)

Security breaches are very difficult to predict. The next attacker could be anyone from a script hacker to a crime syndicate to malicious insider. Yet the intrusion-detection and intrusion-prevention systems many enterprises employ in response to all this uncertainty suffer from the same weakness that has plagued antivirus product for many years—a reliance on signatures. Antivirus vendors realized early on that to stay competitive, they had to develop techniques to enable their products to identify suspicious traffic, even if they had not seen that particular activity before. The answer was heuristics and behavioral analysis methods that detect files and processes that behave in ways deemed threatening. Network Behavior Analysis (NBA) is addressing this need.

Most enterprises can benefit from NBA, since most are missing security events of interest because of overwhelming bandwidth or a lack of pervasive visibility. But, as with any product that interacts closely with networks and impacts security, a proper fit is absolutely necessary.

NBA solutions should integrate with current IDS, IPS, vulnerability scanners, and security incident and event managers while handling the volume needs of the enterprise.

Because NBA requires an intimate understanding of an enterprise's unique traffic patterns, it is a natural fit for vendors to add network performance monitoring features ranging from simple functions, like identifying top communicators, to more advanced reporting to assist with network optimization and planning. Essentially, this feature set is why NBA vendors promise both network and security teams visibility that they have previously not possessed, including alerts when new hosts appear on the network and the ability to find where bottlenecks exist and tie users directly to their network traffic flows.

NBA products need access to network traffic, either through flow data collection or via direct packet capture. Network flow data can best be described as metadata about a unidirectional sequence of packets that includes such information as time stamps for the start and finish of the flow number of bytes, and packets in the flow, source and destination IP addresses, source and destination ports, TCP flags if applicable, and IP information. There are several formats of network flow data. The three mainstream implementations—NetFlow, SFlow, and IPFIX, which is based on Cisco's NetFlow version 9—all are supported by leading NBA vendors.

NBA products serve as collectors receiving network flow data from switches and routers that they in turn process into meaningful performance and traffic information. With direct packet capture, the NBA system acquires network traffic directly from a switch or a router using a SPAN port or network tap, and exports it into the equivalent of what would be received if the NBA product had simply gathered network flow data. Going a step further, NBA systems also can leverage deep packet inspection through direct packet capture to flag attacks that could not be detected by monitoring only network flow data. This method also provides awareness of applications that may be piggybacking on other normal application ports.

A baseline of normal behavior is the core of NBA, but these systems also spot pattern-matching signatures to detect network scans, anomalous application behavior, and worms. NBA vendors recognize that customers like to have immediate feedback from security products when they flip the "on" switch, so pattern matching is available out of the box. Of course, the most value comes once a solid baseline is in place, but this take some time to accomplish.

In summary, NBA is a combination of network, security, and performance management. It helps to adding policies about what is allowed within the corporate network, such as instant messaging, P2P, locations services, and others. NBA also helps to meet regulatory requirements that require monitor-

ing and tracking of all network activity back to the user. As a result, NBA products must interface with network equipment, network facilities, and various databases.

Acronyms

IDS Intrusion Detection System
IPS Intrusion Prevention System
NBA Network Behavior Analysis
P2P Peer-to-Peer

5.3.10 Mobile Device Management (MDM)

Mobile devices are entering organizations at a fast pace. Unfortunately, many businesses have not kept up in terms of their ability to manage these products. In unmanaged environments, end users cannot be sure of secure and reliable access to their data, and IT departments struggle to support systems into which they have little visibility. IT leaders need to get ahead of this issue. Mobile device management will become a must-have capability for most IT departments over the next few years. As mobile devices and applications take over many of the functions previously carried out on desktop PCs, the ability to secure, track, update, and provision such devices will present a significant competitive advantage in many industries.

The process starts with IT assessing its organization's current state of mobility, surveying end users and departments to determine what devices are in use, how they are being used, and with what carriers. It is also important to gain an understanding of how IT staff has been supporting the devices. This information will provide insight into what the organization needs to support now, which is likely a heterogeneous environment, even if it plans to introduce more uniformity into the infrastructure in the future. This information will also reveal the existing IT support procedures so their effectiveness can be examined and will lay the foundation for incorporating into formal management practices those procedures that provide the greatest value. While management plans need to be practical about the existing environment—ensuring the policies, processes, and tools for mobile device management are infrastructure-agnostic—the plan should build in a timeline to achieve as much hardware standardization as is possible.

The core of a mobile device management plan is its policies and processes. Much of the work in setting policies will involve segmenting users and ensuring that their rights and privileges are appropriate to their jobs. To support flexible policy assignments of access and other rights, IT will want to consider a mobile device management platform that lets it assign users to groups and set policies systemwide, at the group or at the individual level.

Security policies and processes must be the focus of the mobile device management plan as well. Many organizations rightly are concerned that widespread mobile device usage potentially leaves corporate data more vulnerable to loss or theft; that wireless users may accidentally violate compliance policies concerning data usage; and that the chances mobile users will introduce viruses or otherwise compromise the corporate network are increased as a result of these deployments. Permissions granted through policy assignments can play a key role in addressing some of these concerns, as will developing a plan that outlines how to secure mobile users' connections to the network.

Creating the right policies and choosing the right mobile device management platforms to meet the organizations' requirements will have many positive results. Users will gain a reliable support platform that can be adapted to their job requirements. IT staff gets a road map to guide their interactions with mobile device users—and by and large, a much needed helping hand. Organizations themselves will benefit, as improved management enables them to save costs and ensure security.

Existing problems may be summarized as follows:

- Order and bill reconciliation for mobile devices are out of synch
- Endpoint security management is incomplete
- Software updates are not managed properly
- Lack of integration with enterprise management platforms
- Recovering assets from departing employees is unsolved

Features that matter most are:

- Managing the physical devices
- Password protection
- Policy enforcement
- Remote site wipes
- Compliance and policy setting

In summary, solutions are expected from both device suppliers and from independent management vendors. IT management should decide whether they prefer best-of-suite or best-of-breed solutions.

Reference

WIEN08: Wiens, J. 2008. With Security, More Is Better, *Information Week*, March 10, 43–48.

5.4 New Product and Service Creation

James Anderson

5.4.1 Introduction

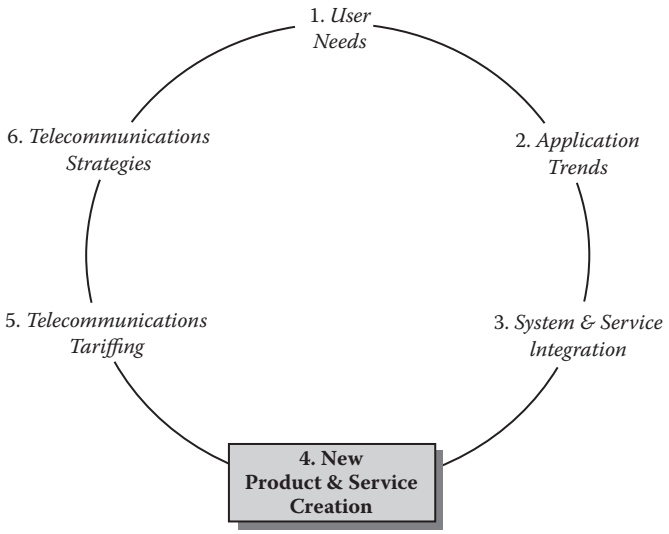
The increasingly competitive telecommunications environment will require service providers to create and deploy new products and services faster than ever before. Providers were able to create new services at a much more leisurely rate in the past. A provider could wait until the next generation of technology had been deployed into the network before introducing the services that used the new technology.

As the number of service providers increases, it is generally agreed that the ones who will succeed are those who are able to best understand their end user's needs and deploy services that meet those needs. In order to accomplish this, a provider will need a new approach to designing and deploying future services (see Figure 5.8).

Changes in network equipment and in the types of networks that are being deployed are equipping providers with the essential tools. In this section we will first look at some of the drivers and constraints that providers are facing as they struggle to change how they create services. Next, we'll focus on how services will be created in tomorrow's network. Finally, we'll investigate how network bandwidth affects the types of services that can be created and what is being done to provide more bandwidth for new services.

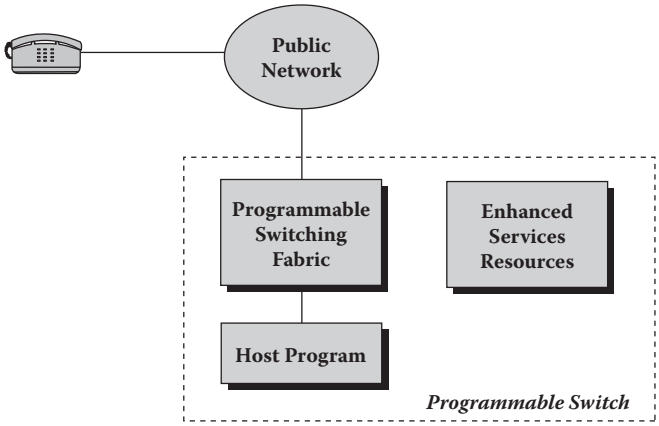
5.4.2 Drivers and Constraints

- U.S. West is now offering a PCS service that includes mobile dial tone and advanced messaging and routing capabilities. The dial tone service combines a handset-generated and network-generated dial tone. The handset portion allows users to hear a dial tone while dialing, and the network-generated portion allows users to hear a dial tone while they are initiating features. Customers have said that they associate dial tone with reliability and quality. The service also includes a same-number feature that routes calls made to a home, office, or PCS number to a PCS phone. It can also route all messages to a single mailbox, notifying users of messages via a light on the handset.



FIGu RE 5.8 Trend analysis—new product and service creation.

- Two different standards are being considered for using ATM to switch IP traffic: (1) Multiprotocol over ATM (MPOA), developed by the ATM Forum, is based on LAN emulation, and is seen as a campus backbone solution; and (2) Multiprotocol Label Switching (MPLS), developed by the Internet Engineering Task Force (IETF), which was designed with the large-scale WAN in mind.
- Service providers are looking for new ways to rapidly introduce new services to meet growing customer demands. Existing circuit switches can generally only be modified by their vendors, a process that takes too long and costs too much. Programmable switches (Figure 5.9) consist of three main parts: a programmable switching fabric, controlling software (host program), and external media used to provide enhanced services functions. New functions can be added to programmable switches by simply adding services and features to the host program. Open interfaces and APIs permit third-party developers to create vast libraries of available services.
- In order to help ISPs that are not ready to make a full-fledged investment in electronic commerce with an option to start a little smaller, e-commerce software vendors are getting creative. One



FIGu RE 5.9 Programmable switch architecture.

vendor permits ISPs to operate small online stores (less than ten items to be sold) for free. As they increase the size of the store, they then start paying the vendor for the use of the software.

- Both the competing standards for digital television, one promoted by the Advanced Television Systems Committee in the United States and the other promoted by Europe's Digital Video Broadcasting Group, offer an almost unlimited potential to broadcast data to end users. In tests, broadcasters have been able to transmit 60 Mbytes of data during a 51-second commercial.
- Some cable operators are now able to offer traditional switched voice services using their cable networks. These services are proving to be more popular than cable modem services.
- Two sets of networking protocols for permitting Internet traffic to be prioritized are working their way through the IETF. The first set is called Differentiated Services (DiffServ). It provides routing mechanisms designed to manage various QoS profiles or performance parameters. The other set of protocols is the multiprotocol label switching protocol. MPLS is a routing mechanism designed to group all packets within an IP session into a single "flow" at the networking layer (Layer 3) and "tag" each session as such for expedited passage through router hops.
- Wireless service providers who use TDM are looking for ways to differentiate their services. Their latest attempt is called Wireless Office Services. These allow users to access PBX features from their wireless phones while they are in the office and when they leave the office. This permits them to use such features such as as four-digit dialing and call forwarding in all environments.
- Advanced Intelligent Network (AIN) platforms and capabilities are ways that new services can be introduced into the public network. However, the change to a packet-based network puts the future of AIN services in some doubt.
- Vendors' research labs are starting to produce products that implement some of the latest advances in speech recognition technology. This type of interface is seen as a major step toward the convergence of telephony and Internet applications. Call center applications are expected to be among the first to benefit from these types of products.
- Vendors are offering service development products for Internet protocol-based voice service providers. Service providers will be able to use these products to add key features such as universal messaging, follow-me services, and paging to their IP/public network gateways. All incoming messages are stored in a single mailbox and can be converted to a variety of formats that the user can then access via Web browser, e-mail, or telephone.
- The Voice-over-IP (VoIP) Forum recently ratified an implementation agreement that defined an interoperability profile based on the H.323 standard from the ITU. H.323 was designed to be a technology for multipoint-multimedia communications over packet-based networks, which include IP-based networks, such as the Internet. It can be applied in a variety of ways—audio only (IP telephony or VoIP); audio and video (video telephony); audio and data; and audio, video, and data.
- One debate in the communications community is how to successfully deliver QoS and implement service-level agreements (SLAs). QoS, a networkwide performance characteristic, refers to the network's ability to fulfill a traffic contract—the SLA—between the WAN network provider and the subscriber for the minimum service provided by the network.
- ISPs are replicating content across multiple services in order to balance user demand loads. Vendors are now starting to offer products that allow service providers to automatically route end-user requests to the replicated server that has a low enough load to facilitate the request.
- User demand for access to multimedia Internet content has resulted in novel solutions being created by vendors. One approach uses satellite links to bypass the Internet and deliver multimedia content to local ISPs where it can be cached for access by local users. This approach can be further extended to caching of popular Web sites in order to speed up local access speeds.

5.4.3 New Service Creation

- The emerging consumer vehicle tracking service is called *telematics*. Telematics systems combine GPS and cellular networks to offer safety and concierge services to consumers in automobiles. Most U.S. telematics operate on advanced mobile phone system (AMPS) due to its near-ubiquitous coverage.
- Smaller ISPs are using audio and video conferencing capabilities to distinguish themselves from competitors. These service providers are starting to investigate using client and server software solutions that permit videoconferencing over the Internet. Initial users include schools that have a need to provide a one-on-one tutoring experience but don't need an elaborate room-based videoconferencing system. The supporting software systems are all H.323 compliant.
- Business travelers want to be able to access the Internet even when they are traveling internationally. This is currently not easy to do; such travelers must reach their ISP's POP in order to access the Internet. Some service providers are attempting to build international POPs to meet this need. Other smaller ISPs are banding together to create consortia to offer Internet access to their collective customers. An additional service that is being investigated would offer roaming users access to their corporate intranets via secure tunneling.
- Verizon Wireless is offering utilities the ability to read customer's meters automatically via wireless data transmission using the cellular digital packet data (CDPD) network. The service would allow utilities to automatically read meters and monitor energy flows, among other services, from a central location, skipping the need to send personnel to customer locations. This service offers utility companies an advantage in a deregulated market because they can offer their customers a better picture of their usage patterns and then offer them a special deal to keep them from going to other utility providers.
- Service providers are starting to offer enhanced fax services. These services include a mailbox, which provides a secure fax mailbox accessible from any location; a never-busy fax transparent service stores faxes for later delivery; fax-on-demand lets businesses create a library of faxable documents that customers can access; and fax broadcast delivers a document to as many as 10,000 locations with just one transmission.
- ISPs are starting to roll out Internet protocol voice services to corporate users. Initially, business users can connect their PBXs to the ISPs IP network, thereby cutting costs on internal long-distance calling. The next step is to combine IP voice with extranets. Businesses would then be able to call other businesses at remote locations using five-digit dialing.
- Many PC games now come with multiplayer Internet options. Users first connect to the Internet, then select a specific server that hosts a gaming session. Then as the end user plays the game in multiplayer mode, the server allows them to exchange information with other players in real time.
- Consumers and businesses will soon have the ability to both view and pay bills via the Internet thanks to various forms of electronic bill presentation and payment (EBPP). This new service will allow billers to cut paper processing costs and garner customer loyalty and Web site hits. Financial institutions, bill consolidators, Internet portals, and makers of personal financial manager (PFM) software products look forward to capturing market share.
- Electronic commerce is struggling with the issue of how to reach customers who are not connected. Companies that have to use both the telephone and the Internet to reach customers are looking for a way to tie the two systems together—*v-commerce*. These firms want to develop new applications that will link voice and data, telephone, and PC to let Internet vendors reach customers who can't reach their Web pages. These new applications will use Motorola's VoxML markup language that simplifies embedding speech into Web pages.
- The Web provides an opportunity for delivering a new type of picture called *immersive photography*. This technology allows you to use your PC to navigate around a digitized 360-degree photo. This technology is targeted toward Internet retailers who want to give their customers a

wraparound view of their goods, including high-end real estate agents, travel agents, cruise lines, and destination marketers.

- Visual communication services are poised for proliferation as new advances eliminate the final technological and market obstacles. The ideal solution for multimedia services combines the organization and simplicity of the telephone system with the multimedia and open nature of the Internet.
- Telemedicine is a broad term for several facets of medical care. Collaborative videoconferences between sites; online access to patient records, medical libraries, and databases; and continuing medical education all fall under this term. Most telemedicine programs today are either simple store-and-forward systems or videoconferencing systems adapted for use in a health care setting.
- Automobiles are being equipped with more and more electronics and telecommunications devices. Many cars now have Global Positioning System (GPS) receivers and computers to keep the driver from becoming lost. The U.S. government, state governments, and a variety of industries are considering spending U.S. $200 billion on the Intelligent Transportation System (ITS) initiative. ITS will provide automated cross-border fleet services for North America, enhanced driver navigation, automated accident reporting, and toll collection. Futurists foresee a day in which a car monitors its "health" and can then use wireless communications to identify repair stations in the event that a potential part failure is detected.
- Prepaid wireless services have become a big business in the United States. Customers generally must pay to have their wireless service activated, then they must purchase a prepaid denomination, often in the form of a card from a retail distributor. The next step is to initialize the prepaid service via an interactive voice response (IVR) service.

5.4.4 Increasing Bandwidth

- Wireless cable operators are starting to offer high-speed Internet access services using multichannel multipoint distribution service (MMDS). Without converting to digital, the most wireless cable operators could offer in video is 33 channels, which can't compete with average landline cable or satellite providers.
- Wavelength Division Multiplexing (WDM) technology is being added to the network in order to increase backbone capacity to handle new high-speed access technologies. Initial WDM systems were only two to four channels. Recently, 32 channels appeared in dense WDM. Now hyperdense or ultradense WDM (UDWDM) systems with channel densities of 40 and up and capacities of 400 Gbps are becoming available.
- Cable operators that want to start offering high-speed Internet access services to their subscribers without having to perform expensive upgrades to make their cable network two-way are getting creative. They are using their existing one-way cable networks to deliver content to end users while the end users use their telephone to send information requests. Although this solution may be well suited for rural cable providers who will never have the funds to make their systems two-way, this one-way approach may not provide the bandwidth required by the growing SOHO market.
- Cable operators are able to offer residential subscribers Internet access at 1.5-Mbps rates using a cable modem. FiOS is quickly surpassing these cable modem offerings, with speeds of 100 Mbps available in major metropolitan regions.
- U.S. West markets its DSL services to three types of residential users: consumer/Web browsers (want "always on"), gamers ("entertainment"), and work-at-home users ("looking for bandwidth and the user experience").
- The cable company MediaOne has found, through internal studies, that nearly all cable modem owners use their Internet connections seven to nine times more often than when they had a dial-up connection.
- MediaOne marketing cites a recent study that claimed the average Internet user wastes a total of 50 hours a year waiting to connect to the Internet and waiting for pages to download.

- Telecos, ISPs, and competitive local exchange carriers (CLECs) that are rolling out ADSL services are finding that the earliest adopters of the services are in the small business market. Telephone companies will stress the security of ADSL over cable modem's shared media to small business owners.
- In the U.S., the FCC has auctioned off 1.3 GHz of spectrum in the 28 and 31 GHz ranges for use in local multipoint distribution service (LMDS) two-way services.
- Broadband wireless networks have many benefits: they are fast and easy to deploy; they have minimal infrastructure and real estate requirements; they feature grow-as-you-go network buildout; and they can deliver voice, video, and data services from 64 kbps to 155 Mbps.
- LMDS can be used to offer many services. Business-oriented services include wire-speed LAN interconnect and fractional and full T-1. Teleworking at 10 Mbps is virtually as fast as being at the office. Megabit-per-second Internet access is geared to residential users. Other services include 100 broadcast video channels in competition with cable, and second and third phone lines at home or the office.
- LMDS services compete with DSL and hybrid fiber/coax (HFC) services. LMDS is better than both DSL and HFC at offering high-speed symmetrical services.
- Wireless cable operators have spectrum in the 2.5 GHz range (MMDS).
- The H.323 protocol, used to provide VoIP services, defines ways in which multimedia formats such as phone calls, computer data, pictures, or video can be exchanged and managed seamlessly across packet-switched networks.
- A variety of broadband wireless providers have already introduced services that use multichannel multipoint distribution service (MMDS) and local multipoint distribution service (LMDS). MMDS service providers have been around for awhile, whereas LMDS providers have only recently bought their licenses. MMDS offers a broader coverage reach while LMDS offers greater capacity. Current service offerings use either the public network or a cable modem for the return path.

5.5 Telecommunications Tariffing

James Anderson

5.5.1 Introduction

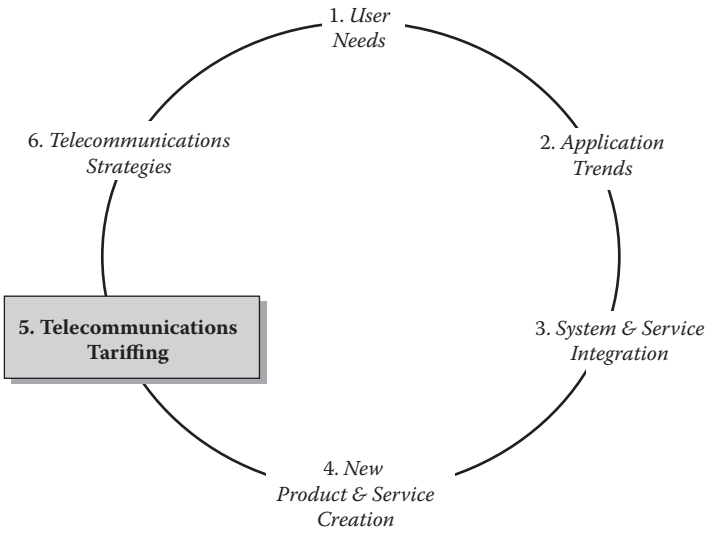
Perhaps no aspect of telecommunications is as overlooked as how services are priced. All segments of end users have differing amounts of funds available to spend on telecommunications services. Pricing a service too high will cause end users to seek lower-priced alternatives. Pricing a service too low will result in the service provider missing out on revenues that could have been used to fund the next service (see Figure 5.10).

In the past, service providers have enjoyed monopoly status in both North America and Western Europe. Under this system, prices for services were closely regulated by governments. This is in the process of changing, and in the future, service prices will be driven by market factors. This change will require existing service providers to change the metrics used to measure service and the pricing philosophies that have been used to create service rates in the past.

In this section we will explore trends in tariffing in the leading markets of North America and Western Europe. The effect of competition on service pricing will also be examined. Finally, we'll discuss the impact that new technologies will have on the pricing of future services.

5.5.2 Regulatory Trends

- The FCC is proposing that the Bell companies be permitted to create separate subsidiaries to offer data communication services. These subsidiaries would be less regulated and could set interstate



FIGu RE 5.10 Trend analysis—telecommunications tariffing.

service prices without filing with the FCC. The Bells' regulated units would still be required to sell capacity to competitors but the separate subsidiaries would not.

5.5.3 Service Pricing Trends

- The average long-distance call in the United States costs about 13 cents per minute, but the average international price is 89 cents per minute.
- Cable & Wireless USA hopes to use pricing and inexpensive long distance to draw residential customers to its Internet service. CWIX will offer customer 150 hours of online service, e-mail, and a free Web page for a monthly fee of U.S. $14.95. Some analysts doubt if bundling long distance with Internet access will attract new customers. They point out that the intersection of households that are online and use long distance heavily is not large—perhaps 15% of the total.
- In most U.S. telco service areas today, termination fees of up to U.S. $36,000 to break a tariffed service contract are still alive and kicking, despite efforts by competition to eliminate them. These contracts can prevent a customer from purchasing the services offered by a competitive provider because they still have a year or two to go on their current contracts.
- The paging industry grew by 14% in 1997 to a total of 50 million subscribers. However, in 1997 four of the top firms, which together control almost 40% of the market, reported almost a half-billion dollars in losses on combined record revenues of more than $2 billion. Many paging companies are suffering from expensive network buildouts. Paging companies seem to hold a high number of customers who refuse to upgrade beyond basic plans, according to analysts. Price wars and new technologies such as low-cost cell phones have driven down the costs of average basic local service from $20 a month a decade ago to less than $10 a month. In some markets, the price has shrunk to less than $5 a month.
- Although extending wireless service to the high percentage of credit-challenged users was a chief driver in the development of prepaid service, wireless carriers are discovering that prepaid strategies may be almost as critical to future growth of their overall customer bases as traditional post-paid service. AT&T Wireless intends to have prepaid accounts for 30% of its new sign-ups.

- Traditional methods of buying and selling bandwidth are not adequate in today's competitive market. A new Internet-based service permits providers with bandwidth to sell their available bandwidth for bidding purposes. Buyers are then able to see the available bandwidth along with information regarding destination country, size (T1, OC-3, etc.), and the length of the contract. If a qualified registrant posts a bid, then the service puts the bidder in touch with the service provider to see if they can work out a deal.
- An interexchange carrier has entered into a partnership with one of Florida's tourism groups. The carrier will share its profits with hotel property owners when hotel guests make calls from their hotel rooms using the carrier's service.
- Cable operators that provide Internet access services via their cable networks are already dropping the price of their service in order to capture more of the Internet access market. Some cable operators see this as the only way to push their Internet access service beyond the early adopters. These cable providers hope to use their lower prices to attract lighter users and cut into the market share of ISPs.
- The majority of U.S. cell phone users pay for all incoming and outgoing calls that use their phone. About 80 to 85% of all cellular calls originate from a wireless phone—this means that cellular subscribers are either not giving out their phone numbers or they are turning off their phones. One way to balance traffic is to upgrade equipment to accommodate calling party pays (CPP) billing. The caller typically pays $0.35 to $0.45 cents a minute, an average rate for an outbound call from a cell phone.
- One reason that domestic long-distance services have not switched to an IP network is because circuit-switched voice is already cheap: rates are below $0.05 per minute for corporate customers and below $0.09 for residential customers. The bottom line is that to make the numbers work domestically requires 10,000 minutes a month to a single location to justify the cost of a private IP telephony network.
- The cost to complete an international voice call is much higher. Carriers charge as much as U.S. $4.00 per minute to complete a call to North Korea and other countries where it is hard to find a good termination.
- To make greater wireless penetration and increased billable minutes a reality, carriers must embrace "calling party pays" (CPP) as the prevalent billing model, rather than "wireless party pays" (WPP).
- Juan Fernandez of Frost & Sullivan reports that when CPP was implemented in Argentina, the market grew from 700,000 subscribers to 2.1 million subscribers in 11 months.
- Giving a customer the first incoming minute of a call for free is an interim way that service providers are trying to increase the number of billable minutes.
- ISPs jumped *en masse* onto the flat-rate bandwagon in 1996, only to find that "all-you-can-eat" pricing has a way of eating away at the bottom line. Some service providers have found that the flat-rate strategy delivers something that they wanted to get all along: lots and lots of customers.
- Flat-rate pricing can be a nightmare for providers, especially if their costs are largely dependent on usage and that usage is difficult to predict. Frame relay and Internet services fall into this category.
- Providers gain from using flat-rate pricing because they don't have to cover the cost of administering usage-base pricing. That can be a significant gain considering that these expenses can run as high as 18% of the total cost of the service.
- Usage-based pricing becomes just as attractive as flat-rate pricing if the cost to deliver a service increases substantially as service usage grows.
- Wireless service providers are starting to offer prepaid services in order to address the 20 to 40% of the market that didn't qualify for service because of bad or nonexistent credit histories.
- Prepaid systems have become more attractive in recent years due to several improvements: they lacked a real-time billing engine (couldn't cut off calls in mid-conversation), and they didn't accommodate incoming calls.

- Wireless service providers can use either a switch-based or a handset-based approach to implementing prepaid services. Most providers have selected the switch-based approach because it works with any handset and it is less prone to tampering.
- The initial investment in prepaid infrastructure can be heavy, but payback periods can be quick. Along with expanding the potential customer base, prepaid wireless lowers the cost of acquiring a customer, since it eliminates the need to do a credit check.
- Despite the up-front charges (for a phone), prepaid services aren't necessarily a tough sell to credit-challenged customers. The per-minute charges are comparable with those levied under low- and mid-tier pricing plans, and they include taxes and interconnection charges. Prepaid customers aren't charged monthly access fees.
- Carrier consolidation and interconnection, increased competition, service bundling, and new technology introductions all are contributing to the need for more intelligent and flexible customer care and billing systems.
- Convergent billing means using a single billing system to create all bills—it does not necessarily mean sending a customer a single bill!
- New and existing service providers competing against each other are selling telephone services that are roughly the same. Their goal is to avoid a commodity war of attrition.
- In the United Kingdom, there are 150 licensed telecommunications providers contending to supply the country's 30-odd million adults with fixed, wireless, data, voice, and video communications.
- Although pricing is becoming increasingly important in telecommunications (especially voice telephony service), customer service, branding, billing, and value-added services are all keys to success.
- Types of carefully constructed rates and calling plans include bundling, demographic profiling, loss leaders, incentive schemes, flattened prices, calling circles, postalized rates, and special rates.
- Service providers seek to bundle multiple telecommunications services in order to provide one-stop shopping for their customers.
- There are concerns that bundling may reduce churn for a company as a whole, but not necessarily for individual lines of business.
- When customers are asked which company they would use for bundled services, customers overwhelmingly prefer local and long-distance carriers.

5.5.4 Impact of New Technologies

- Networkwide QoS is needed to deliver priority service to higher-paying customers. Service providers want to use QoS as a basis for offering various classes of services to different segments of end users. By doing this, they can create different pricing tiers that correspond to QoS levels. That might be one of the best ways to offer new revenue-generating services in public data networks.
- Smaller ISPs are using centralized functionality to improve their competitive situation. Most ISPs store subscriber information on up to five different servers, thus preventing them from using data-mining tools that are essential to customizing services. This opens the door to content-based billing. Software can be used to create something similar to the call detail records used with voice calls, but it will consider a subscriber's profile.
- PCS services have reduced many of the advantages of paging through longer battery life, first minute free, free/bundle voice mail, free caller ID, prepaid plans for less creditworthy customers, and competitive pricing.
- From a connectivity perspective, the Internet is well suited for telephony because of its global reach. From an engineering perspective, it is efficient—a dedicated T1 can support as many as 130 IP voice calls vs. 24 simultaneous calls as in today's carrier networks.
- When talking about billing for IP services, the two key issues are metering and settlements. Metering is relatively straightforward. Settlements introduce trouble because the number of

billing arrangements between carriers grows exponentially with the number of Internet telephony service providers.
- Finland has the greatest wireless penetration of all markets: 42% at the end of 1997.
- Some service providers are now betting millions of dollars that Web-based electronic billing systems are essential for hooking lucrative but finicky business customers—and eventually even some residential ones—who are interested in fast, responsive billing.
- Online billing, however, has its challenges. It not only requires Internet access but is also costly and complicated to set up, especially for big service providers with massive billing systems already in place.
- One of the major benefits of electronic billing is that it saves the service provider money. The more customers opt to pay their bills through a Web site, the lower the cost of running a paper-based billing system. By some estimates, the entire paper trail from stuffing an envelope, mailing the bill, and processing the payment costs a service provider 75 cents to $1.50 per account every billing cycle.
- Initially, IP services did a poor job of tracking and generating the appropriate data to accurately measure usage for customer billing. Changes are being investigated because of the interest in using IP telephony for voice and fax.
- Many different usage-based services are currently being planned: least-cost routing, time-of-day routing, dynamic bandwidth allocation, volume discount rates, callback, security enhancements, Web hosting, e-mail, chat lines, whiteboards, videoconferencing, work group collaboration and multimedia sessions, software applications distribution, applications rental, and classes of service quality.
- Many technical challenges of IP-based services must be tackled. Foremost are extrapolating and scrubbing down traffic information from routers and switches and matching that against customer account data for bills. This allows the invoicing and tracking of packet volumes, counting bits or bytes, and logging origination or destination IP addresses.

5.6 Telecommunications Strategies

James Anderson and Patricia Morreale

5.6.1 Introduction

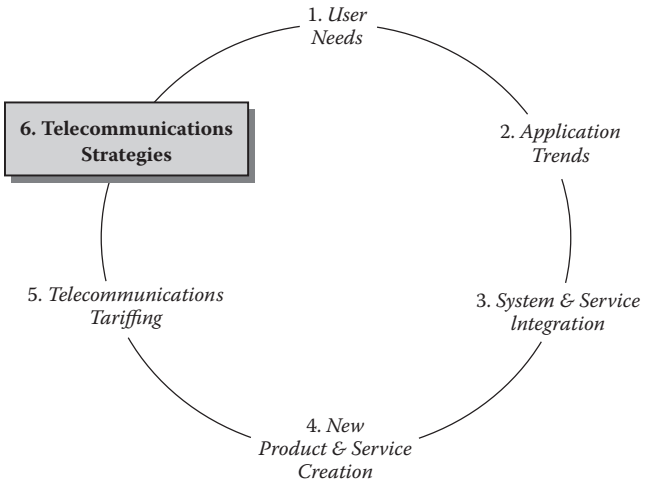
The brave new world that represents the future of telecommunications will consist of a group of aggressive global service providers who are competing for the same segments of end users. How each of the service providers hopes to succeed at the expense of its competitors is a fundamental part of its long-term strategy (see Figure 5.11).

A provider's strategy for increasing its market share must be in part based on its current situation. In this section we will look at the current situations that describe many of today's up-and-coming service providers as well as some of the well-established players. We will examine their business goals and how they may go about achieving them. Finally, we'll identify some of the possible events that could dramatically change existing strategies.

5.6.2 Service Providers

The value chain of products and services will dictate the positioning of telecommunications service providers. The positioning process usually starts with answering a number of questions, such as:

- What is the perceived quality of my network?
- Is the network keeping pace with the growth of subscribers?



FIGu RE 5.11 Trend analysis—telecommunications strategies.

- How much should be invested?
- Where do I need to invest?
- How can I get more revenue out of existing services?
- How can I reduce operating costs?
- How do I know if a problem is just a solitary abnormality or a building problem?
- How can I reduce customer churn?
- How can I predict future capital expenditures?
- How can I get system usage information to improve marketing and sales?

The traditional value chain was very simple. The equipment suppliers—a closed market of monolithic suppliers—have provided hardware with hard- or soft-wired integrated services. This equipment was key for network and service providers who have based their service offers to their customers on the capabilities of this equipment. Change cycles and service creation were extremely long, hardly meeting the customer's expectations.

The actual value chain includes the following principal components (TERP01):

1. Infrastructure
   - IT component suppliers
   - OSS application suppliers
   - Network element suppliers
   - System integrators
2. Network Products and Services
   - Network operators
   - Service providers
3. Hosting and Processing
   - Hosting services providers
   - Processing services providers
4. Applications and Media
   - Applications services providers
   - Context, content packaging, and management
   - Content services providers
5. Customer

The players are not yet evenly distributed. Most of them are still emerging from the traditional service providers, and can be allocated to Network Products and Services. Examples are:

- ILEC (Incumbent Local Exchange Carrier): Strong provider who owns a considerable amount of telecommunications facilities and doesn't want to give away this position easily. Most likely has a number of legacy support systems with little interoperability and integration in use. The result is high operating costs.
- CLEC (Competitive Local Exchange Carrier): Smaller, flexible provider who owns little or no telecommunications facilities (facility-less). By offering excellent customer care and new services, they try to build the support structure step by step. Their support systems are state-of-the-art, lightweight, and less expensive to operate. In certain cases, they use service bureaus for billing and provisioning.
- IEX (Inter Exchange Carriers): Primarily responsible for long-distance services with stepwise penetration of the local exchange area. They can be both incumbent and competitive providers with the result of the need for very heterogeneous support systems.
- PTT (Post, Telegraphy, and Telephone): Strong provider who owns a considerable amount of telecommunications facilities and doesn't want to give away this position easily. Most likely has a number of legacy support systems with little interoperability and integration in use. The result is high operating costs.
- CAP (Competitive Access Provider): Facilities-based or non-facilities-based; similar to the ILEC, but have carefully selected local loops for high-profit commercial customers.
- NSP (Network Service Provider): Responsible for providing a highly reliable networking infrastructure consisting of equipment and facilities. Its responsibilities are usually limited to the physical network only, but element management systems are usually included into their offers.

However, integration is important for many customers. Thus, ISPs and ICPs will play an important role as well. The short definitions are:

- ISP (Internet Services Provider): Its main goal is to provide Internet access to business and attract customers. Major challenges include peering to each other and to other carriers, managing quality, and offering acceptable performance.
- ICP (Integration Communications Provider): Emerging provider with integrated services offer, concentrating on next generation, high-speed data and wireless services, in particular for profitable business users. Its acceptance in the marketplace is expected to be high. In terms of support systems, they buy instead of build; occasionally, they use service bureaus for billing and provisioning. They take advantage of the fact that intranet, extranet, virtual private networks, e-commerce, and multimedia applications require more bandwidth than is available over traditional circuit-switched voice networks.

Hosting and processing will be most likely dominated by traditional mainframe and server manufacturers that are flexible enough to make the necessary facelifts to their equipment to meet requirements of load distribution, load balancing, storage management, and security. IBM and Compaq may be mentioned here as examples of providing reasonable services, using server farms with high availability features.

Application and media need many new competitive players. At the beginning, ASPs and ESPs will dominate this market. The short definitions are:

- ASP (Application Services Provider): Emerging service provider that must combine application, systems, and network management. Service-level expectations are extremely high; the whole business of customers may rely on this provider.
- ESP (Enterprise Services Provider): Emerging service provider from the enterprise environment. It offers services for a limited user community with similar attributes to the provider. It uses and customizes its existing support systems, which may not scale well.

Hosting and processing enable Web presence and interactivity on the Internet. These are typically provided by ISPs and NextGen service providers that are active in IP services. They mainly include hosting of Web server infrastructures and content and Web-enabled transaction software and hardware that allow the execution of online transactions. It is an infrastructure type of activity, although it is characterized by added value and a significant amount of additional services. Service offer alternatives are:

- Web hosting: Keeping content on Web server farms and offering access with good performance
- Value-added Web hosting: In addition, content, and database maintenance and Web master services
- Data hosting by offering Storage Area Networks
- Data management services, including search engines
- Public Key Infrastructure services, including trust-center functions
- Centralized Web transaction services
- Web community and Internet account management
- Transaction authentication services

The typical customers of these services are businesses. While large businesses previously deployed their own Web infrastructure in-house, they now also realize the efficiencies of lower complexity and economics of scale given by professional service providers. This customer base can easily extend to the future. The key differentiators will be the service level and complexity of services offered to customers. Hosting and value-added hosting emerge as a volume business. Large data centers with server farms, load balancers, and traffic shapers combined with high availability and excellent performance will take business away from smaller service providers with lower availability and limited Internet access capability. Most service providers are inexperienced in this area.

The continuation of the value chain is dominated by innovative services that are to a certain extent IP based. It means that the traditional circuit-switched architectures are replaced by packet-switched architectures. For the underlying physical architecture, there are many choices, including:

1. IP + ATM + SDH/Sonet = B-ISDN	Traditional approach, which has the most supporting network elements and their element managers
2. ATM transport	Includes both SDH/Sonet-less ATM transport and ATM/SDH/Sonet hybrids
3. Switched routing	ATM/IP hybrids
4. IP over SDH/Sonet	PPP- or HDLC-framed IP mapped to SDH/Sonet
5. Optical IP	Transport of PPP- or HDLC-framed IP over WDM with fast photonic restoration
6. Use of enhanced frame relay	Substitution of ATM by frame relay in any of the approaches 1, 2, or 3 above

There is no doubt that the new area of competition is content.

Content delivery management is taking off and service providers are well positioned to earn revenues there. Content delivery management helps content owners to provide seamless and fast Web site access for customers through the following:

- Large scale caching
- Distribution of Web server farms
- Complex Internet routing services on managed network segments

All these aspects help to deliver reasonable performance. Processing is an increasingly important revenue generation opportunity as traditional infrastructure business shrinks. Transaction, and therefore processing, which is the infrastructure and software enabler of transactions, is believed to grow to become the single most important revenue input of the Internet value chain. Processing is by no means related to the core business of service providers, but it is important for e-commerce ser-

vice offers. Services can be created by the IT organization of service providers in collaboration with systems integrators.

The Application and Media elements of the value chain create and translate traditional and digital content into Web-ready format and create the actual interface between the digital product and the customer. This service is targeting an end-to-end process that covers creation, manufacturing, delivery, and presentation of content to customers. This is believed to be the most promising business opportunity of the Internet. It carries the highest growth potential, but at the same time the highest risks, too. Telecommunications service providers, IT companies, media enterprises, retail chains, and several other industries are competing for revenues.

Service offers are:

- Application services to be provided by service providers, integrating IT, software, system integration, telecommunication, and consulting skills.
- Content authoring, auditing, deployment, and maintenance combined with bandwidth management, server-load management and traffic management, supporting generic, corporate, and specialized niche portals, and B2B and B2C operations.
- Content creation targeting videos, movies, audio, photo archives, encyclopedic articles, analyst reports, financial evaluation, and many others. Music and written material combine with broadband access to support multimedia to be delivered over the Internet.

These innovative service areas must be seriously investigated by service providers. In other traditional areas, the profit margins are narrowing; in the IP area they have to face other competition. To be successful, innovative minds are required. It means more collaboration with customers, mergers, acquisitions, investment ino smaller companies that may be acquired later, and flexibility in service creation, fulfillment, and quality assurance.

Another Internet-based service is *immersive photography*, allowing customers to use PCs to navigate around a digitized 360-degree photo. This technology is targeted toward Internet retailers who want to give customers a wraparound view of their goods, such as high-end real estate agents, travel agents, cruise lines, and destination marketers.

Whether retailers are between service providers and customers depends on the marketplace. No general guidelines can be given in this respect.

5.6.3 Goals

The goals are different for each cluster of service providers. Table 5.4 summarizes the most obvious goals and future targets for each cluster of service providers, referenced in Section 5.6.2 (TERP01).

5.6.4 Green Computing

Green computing and other sustainable networking initiatives refer to the study and practice of using computing resources efficiently. The primary objective of these environmentally sound programs is responsible stewardship of scarce and expensive planetary resources. Elements of a green computing initiative include a wide range of people and services, from the building housing the computer and network systems to the types of energy used to power the electrical delivery system underlying the network, disposal of electronic waste, pros and cons of telecommuting opportunities, server resources, and end-user satisfaction. Assessment of green computing impact, both initial and longer term, is in an embryonic stage.

5.6.5 Software as a Service

Software as a service (SaaS, typically pronounced “sass”) refers to a software deployment such that an application is hosted as a service to all Internet customers. The traditional customer’s burden to maintain

TABl E 5.4 Goals and Future Business Targets for Service Provider Clusters

Service Provider Clusters	Goals	Business Targets
	Infrastructure	
IT component suppliers	Sell more software Sell professional services	Replace legacy solutions Acquire OSS application suppliers Integrate legacy and innovative systems
OSS application suppliers	Sell more software Sell more professional services	Full-line of offerings of support systems Target ILEC legacy replacement Acquire other vendors of support systems Compete with system integrators
Network element suppliers	Sell more equipment via best-of-breed and best-of-suite offers	Outsource element management systems to vendors of support systems Use of open interfaces Develop solutions for e-commerce
System integrators	Sell custom design, development, and deployment Sell custom integration Sell consulting	Acquire vendors of support systems Conduct many projects Consolidate products Compete with OSS suppliers
	Network Products and Services	
Network operators (ILECs, PTTs, IEXs, NSPs, CAPs, and global carriers)	Rapid introduction of new services Cost reduction Customer retention Multivendor management Convergent ordering Up-to-date asset management	Less internal software development More use of systems integrators More packaged software of support systems Pervasive interconnection of support systems Customer relationship management Self-care with support systems for customers
Service providers (CLECs, ISPs, ICPs)	Build network capacity Customer acquisition Improve service quality Add facilities More carrier interconnection Support of micropayment and prepaid services	Minimal internal development Automated processes More packaged software for support systems Less service bureaus Integration of support systems Customer relationship management Self-care with support systems for customers
	Hosting and Processing	
Hosting (mainframe manufacturers, server manufacturers)	Use existing storage resources Reengineer business processes Use load balancers	Penetrate the Web market Support of e-commerce Advanced asset management Support of Storage Area Networks (SAN)
Processing (mainframe manufacturers, server manufacturers)	Use existing processing resources Reengineer business processes Use caching	Penetrate the Web market Support of e-commerce Advanced asset management

Continued

TABl E 5.4 Goals and Future Business Targets for Service Provider Clusters (*Continued*)

Service Provider Clusters	Goals	Business Targets
	Applications and Media	
Applications (ASPs, ESPs)	Sell service Customer acquisition Early profitability	Resource integration Good management of the infrastructure Advanced asset management Excellent service levels Use of packaged software
Context, content packaging, and management (ISPs, ESPs)	Real-time rating Service creation on-the-fly Mid-range profitability	Usage-based billing Multimedia support Multicasting for distribution
Content providers (ISPs, ASPs)	Real-time rating Service creation on-the-fly Mid-range profitability	Billing for content value Web switching technology
Customer	Increase service reliability Lower transport costs Faster service provider responsiveness Customer network management	Self provisioning via Web Custom quality of service reporting Flexible billing formats Electronic bill presentment and payment Usage-based accounting

and run the application on the customer's own computer is eliminated. This can be done for a number of reasons, with customer affinity being a primary motivator. SaaS alleviates the customer's burden of software maintenance, ongoing operation, and support.

Reference

TERP01: Terplan, K. 2001. *OSS Essentials: Support System Solutions for Service Providers*, John Wiley & Sons, New York.

Summary

Patricia Morreale

The rapid growth of services and demand has outpaced the traditional industrial approach to pricing and service delivery. With the user needs discussed here, we can see application trends emerging. Future service offerings must be anticipated by service providers, but in a much shorter timeframe than previously used by service providers. The classic model of service providers "offering" new network services and features has been radically transposed into a new model of customers asking for new services and features, frequently in advance of service provider offering or pricing. This customer-demand driven model is not limited to services, but also addresses an interest in green computing, and an awareness of the nonmonetary price that may be paid by unchecked user demand. Future service models must take into account the prudent, eco-wise delivery of customer services.

Index

A

Amdocs, taxonomy by, **3**-154
APM, *see* Application performance management
Application performance management (APM), **2**-81–**2**-99
  acronyms, **2**-99
  need for APM in communication services, **2**-81–**2**-90
    acceleration of business, **2**-83–**2**-84
    areas of applicability, **2**-86
    assuring high quality of customer self-service, **2**-87
    benefit of homogeny, **2**-90
    business APM, **2**-89–**2**-90
    doing more with APM data, **2**-90
    enhancing contact center operations, **2**-87
    ensuring availability of information hubs and busses, **2**-85
    increasing profitability and efficiency of service delivery, **2**-86–**2**-87
    leading performance management issues today, **2**-85
    meeting service-level agreements, **2**-85
    need for new type of service assurance solutions, **2**-85–**2**-86
    operational APM, **2**-87–**2**-89
    optimizing OSS/BSS, **2**-85
    possibilities and challenges, **2**-81–**2**-83
    rise of software, **2**-83
    use cases, **2**-89, **2**-90
  overview, **2**-91–**2**-93
    APM tool requirements, **2**-92–**2**-93
    evolution of APM, **2**-91–**2**-92
    interested parties for APM, **2**-91
    reactive and proactive APM, **2**-92
  performance management for next generation OSS/BSS and SDP, **2**-96–**2**-98
  performance management for next generation Service Delivery Platforms, **2**-93–**2**-96
    challenges with traditional network management tools, **2**-94–**2**-96
    SDP overview, **2**-93–**2**-94
    strategic investment, **2**-94
  trends, **2**-98

B

Benchmarking, **4**-123–**4**-134
  advantages, **4**-125
  applicability, **4**-132–**4**-133
  description, **4**-123–**4**-124
  disadvantages, **4**-125–**4**-126
  myths and mistakes, **4**-126–**4**-127
  network management benchmarks, **4**-127–**4**-130
  phases, **4**-130–**4**-132
  support tools, **4**-133–**4**-134
Businesses, information technology alignment with, **4**-17–**4**-26
  baseline IT plan, **4**-18–**4**-19
    affordability, **4**-19
    assignment of resources, **4**-19
    business strategies supported, **4**-19
    manageability of investments, **4**-19
    plan elements depending on IT organization, **4**-19
    scope and scale of investments, **4**-19
  directions for service orientation, **4**-20–**4**-22
    business service management, **4**-21–**4**-22
    service-oriented architecture, **4**-21
  importance of IT, **4**-17–**4**-18
  real-time enterprise, **4**-19–**4**-20
  role of IT infrastructure, **4**-22–**4**-25
    applications and services, **4**-22
    computers, **4**-22
    networks and network equipment, **4**-23
    real-time infrastructure solutions, **4**-23–**4**-25
  support of business by IT, **4**-18
Business intelligence and analytics, **4**-27–**4**-51
  customer intelligence, **4**-42–**4**-46
    customer data in telecommunications industry, **4**-43
    customer relationship management, **4**-43–**4**-45

using client data for intelligence, **4**-45–**4**-46
customer is king, **4**-29–**4**-31
disappointments from past and barriers to business intelligence, **4**-39–**4**-41
closing the loop, **4**-40–**4**-41
effort to build BI solution, **4**-41
information gathering, **4**-40
risk data, **4**-40
trust, **4**-40
four Cs of telecommunications industry, **4**-28–**4**-29
impact of BI and BPM on telecommunications industry
average revenue per user boosting, **4**-47
BI for convergence, **4**-49–**4**-50
business challenges, **4**-49
churn prediction and management, **4**-47
improved procurement, **4**-47–**4**-48
integrate cost and revenue, **4**-47
operational challenges, **4**-49
pre- and postpaid convergence, **4**-48–**4**-49
risk resolution and management, **4**-48
self-service, **4**-48
strategy at work, **4**-37–**4**-39
campaign assignment and management, **4**-38–**4**-39
customer retention, **4**-39
scoring and segmentation, **4**-38
strategic decision support, **4**-37–**4**-38
telecommunications industry, **4**-32–**4**-37
BI application areas, **4**-34–**4**-37
BI requirements, **4**-34

C

CA Wily Technology, performance management solution from, **2**-96–**2**-98
CGI scripting, **2**-24–**2**-25
CoBIT, *see* Control Objectives for Information and Related Technology
Cognitive radio networks, multimedia applications for, **1**-58–**1**-67
cognitive radio devices, **1**-62–**1**-63
cognitive radios and cognitive radio networks, **1**-59–**1**-60
dynamic spectrum access, **1**-60–**1**-62
spectrum management, **1**-61
spectrum mobility, **1**-62
spectrum sensing, **1**-60–**1**-61
spectrum sharing, **1**-61–**1**-62
policies for cognitive radio operation, **1**-63–**1**-64
pricing schemes for multimedia applications, **1**-65–**1**-67
quality of service, **1**-65
Computer Telephone Integrated (CTI), **1**-2–**1**-12
applications and trends, **1**-11–**1**-12
basic definitions, **1**-2–**1**-3
brief history, **1**-3–**1**-4
call control, **1**-7–**1**-8
components and models, **1**-5–**1**-11
first-party and third-party CTI, **1**-8–**1**-11
media processing, **1**-5–**1**-7
Control Objectives for Information and Related Technology (CoBIT), **4**-140–**4**-142
CRM, *see* Customer relationship management
CTI, *see* Computer Telephone Integrated
Customer relationship management (CRM), **3**-138–**3**-143
customer contact management, retention, and loyalty, **3**-138
customer information management, **3**-138–**3**-139
customer interface management, **3**-139
customer problem handling, **3**-141
customer quality of service and service-level agreement, **3**-141–**3**-142
customer self-service, customer self-management, **3**-142–**3**-143
selling/sales process, **3**-139
service order management, order handling, **3**-140–**3**-141

D

Database administrator, **4**-120–**4**-121
Data communications, *see* Voice and data communications
Data marts, **2**-40
Data warehousing, *see* Web-enabled data warehousing
Denial of service (DoS) attacks, **2**-29
Digital signatures, **2**-25
Digital subscriber line technologies, **3**-29–**3**-31
Distributed management systems, **3**-12–**3**-4
distributed management systems architectures, **3**-12–**3**-14
distributed network and computing systems, **3**-12
in-band and out-of-band management systems, **3**-14
Document encryption, **2**-26
Document life cycle, **4**-4–**4**-7
archive, **4**-6–**4**-7
create, **4**-5–**4**-6
distribute, **4**-6
manage, **4**-7
Domain name system, **2**-107
DoS attacks, *see* Denial of service attacks
Dual-homed host firewalls, **2**-24

E

eBusiness, **3**-154–**3**-164
implications for service providers, **3**-156–**3**-162
reference model, **3**-163–**3**-164
service provider migration toward, **3**-162–**3**-163
E-commerce technologies, **2**-100–**2**-103
back-end systems, **2**-102–**2**-103

client systems, **2**-100–**2**-101
front-end systems, **2**-101–**2**-102
Internet, **2**-101
supply chain systems, **2**-103
Electronic technologies (intranets), **2**-99–**2**-105
e-commerce technologies, **2**-100–**2**-103
back-end systems, **2**-102–**2**-103
client systems, **2**-100–-**2**101
front-end systems, **2**-101–**2**-102
Internet, **2**-101
supply chain systems, **2**-103
management of Web services, **2**-105
Web service delivery challenges, **2**-103–**2**-105
building relationships, **2**-104
high cost of small errors, **2**-103–**2**-104
security, **2**-104
speed, **2**-104
technology evolution, **2**-105
E-mail management, business case for, **4**-11–**4**-12
Encryption, **2**-25–**2**-26
digital signatures, **2**-25
document encryption, **2**-26
link encryption, **2**-25
pretty good privacy, **2**-26
public-key encryption, **2**-25
secure HTTP, **2**-26
Secure Socket Layer, **2**-26
Enterprise management, **3**-134–**3**-137
assurance, **3**-136–**3**-137
billing and revenue assurance, **3**-137
end-to-end vertical process groupings, **3**-135
fulfillment, **3**-136
fulfillment, assurance, billing processes, **3**-135–**3**-136
infrastructure lifecycle management processes, **3**-135
operations support and readiness, **3**-137
product lifecycle management processes, **3**-135
strategy and commit processes, **3**-135
strategy, infrastructure, and product process, **3**-135

F

File transfer protocol (FTP), **2**-12
Firewalls, **2**-2–**2**-24
application-level firewall, **2**-23
circuit-level firewalls, **2**-24
duel-homed host firewalls, **2**-24
firewall architectures, **2**-24
load-balancing, **2**-61
network-level firewall, **2**-23
screened host firewalls, **2**-24
screened subnet firewalls, **2**-24
Free clip art and images, **2**-17
FTP, *see* File transfer protocol
Funding growth, **2**-6

G

Gen-2 RFID protocol, **1**-43–**1**-49
adaptive Q algorithm, **1**-48–**1**-49
basic operations of identifying tags, **1**-43–**1**-47
Gigabit Ethernet, **1**-34, **1**-35
Global telecommunications market, growth of, **3**-111
Green computing, 5-46–5-48

H

HTML, **2**-11
Human resources of telecommunications service providers, job profiles for, **4**-114–**4**-123
business planner, **4**-117–**4**-118
call center operator, **4**-114–**4**-115
contact administrator, **4**-122
database administrator, **4**-120–**4**-121
inventory manager, **4**-115–**4**-116
legal counsel, **4**-121–**4**-122
manager of LEMF, **4**122–**4**-123
network infrastructure operator, **4**-120
network operations manager, **4**-114
operations manager for lawful intercepts, **4**-119–**4**-120
security analyst, **4**-119
service-level manager, **4**-117
service and problem manager, **4**-116
service technician, **4**-116–**4**-117
technology analyst, **4**-118
Hypertext, **2**-12

I

IBM, **3**-183–**3**-189
IBM, Tivoli Network Manager IP Edition, **3**-186
Netcool/Portal, **3**-185–**3**-186
Netcool/Proviso, **3**-186
Netcool/Realtime Active Dashboards, **3**-187
service-oriented architecture, **3**-184, **3**-189–**3**-190
Tivoli Composite Application Manager for Internet Service Monitoring, **3**-187
Tivoli Monitoring, **3**-187
Tivoli Netcool Customer Experience Management Solution, **3**-188
Tivoli Netcool/Impact, **3**-186
Tivoli Netcool/OMNIbus, **3**-185
Tivoli Netcool Performance Manager for Wireless Technology Packs, **3**-188
Tivoli Netcool/Reporter, **3**-186
Tivoli Netcool Service Quality Manager, **3**-188
Tivoli Netcool/Webtop, **3**-185
Tivoli Network Manager IP Edition, **3**-186
Tivoli Network Manager Transmission Edition, **3**-186
Tivoli product portfolio, **3**-184–**3**-185

Tivoli Security Operations Manager, **3**-187
WebSphere, **3**-189
Information life cycle management, **4**-3–**4**-16
critical success factors of document management, **4**-14–**4**-15
document life cycle, **4**-4–**4**-7
archive, **4**-6–**4**-7
create, **4**-5–**4**-6
distribute, **4**-6
manage, **4**-7
hot topics, **4**-7–**4**-14
data leak prevention, **4**-8–**4**-10
deleting and retaining documents, **4**-10–**4**-11
documents for optimal communications, **4**-7–**4**-8
e-mail management, **4**-11–**4**-12
impact of compliance, **4**-12–**4**-13
links to other applications, **4**-14
outsourcing, **4**-14
terms, standards, and statistics, **4**-3–**4**-4
document standards, **4**-4
statistics, **4**-4
terms, **4**-3–**4**-4
Information Technology Infrastructure Library (ITIL), **3**-164–**3**-166, **4**-134–**4**-137
matching ITIL to eTOM, **3**-166
value of, **3**-165–**3**-166
Insight Research, taxonomy by, **3**-154
Intelligence Support Systems (ISS), **3**-200–**3**-226
basic requirements for lawful intercepts, **3**-206–**3**-207
basics and application areas, **3**-201–**3**-204
positioning among other support and security systems, **3**-204–**3**-206
positioning lawful intercepts and surveillance, **3**-200–**3**-201
principal functions of interception, **3**-207–**3**-208
principles of monitoring and intercepts (hardware and software probes), **3**-214–**3**-220
access function implementation approaches, **3**-216
intelligence transmission, **3**-219–**3**-220
internal and external lawful interception, **3**-215–**3**-216
use of probes, **3**-216–**3**-219
reference models for lawful intercepts, **3**-208–**3**-214
Internet basics, **2**-8–**2**-12
client programs and browsers, **2**-11
connection, **2**-8–**2**-9
FTP, **2**-12
home pages, **2**-10–**2**-11
HTML, **2**-11
hypertext, **2**-12
Java, **2**-12
packet switching, **2**-8
TCP/IP, **2**-8
terminology, **2**-9–**2**-12
Uniform Resource Locator, **2**-10
Web browsers, **2**-9–**2**-10
World Wide Web, **2**-9
Internet management concepts, **2**-2–**2**-30
benefits of intranets, **2**-3
content development and design checklist, **2**-5–**2**-6
cross-platform compatibility, **2**-4
funding growth, **2**-6
gaining support, **2**-4
improved corporate culture, **2**-4
improved customer service, **2**-3
improved help desks, **2**-3
intranet planning and management, **2**-4–**2**-8
management overview, **2**-2–**2**-4
management summary, **2**-7–**2**-8
management support skills checklist, **2**-6
organizational challenges, **2**-7
paper-less office, **2**-3
planning intranet strategy, **2**-4–**2**-5
selecting implementation team, **2**-5–**2**-6
technical support skills checklist, **2**-5
Total Quality Management, **2**-6–**2**-7
training employees, **2**-7
Internet protocols, **2**-105–**2**-112
addressing for Internet, **2**-106–**2**-107
communication protocols in Internet, **2**-107–**2**-108
IP, **2**-107
TCP, **2**-108
UDP, **2**-108
domain name system, **2**-107
file transfer in Internet, **2**-110–**2**-111
information search in Internet, **2**-112
category indexes, **2**-112
Web crawlers, **2**-112
information transfer in Internet, **2**-108
Internet e-mail, **2**-109–**2**-110
IMAP4, **2**-110
POP, **2**-110
SMTP, **2**-109–**2**-110
mailing lists in Internet, **2**-111–**2**-112
Netscape and Microsoft, **2**-112
news and usenet, **2**-111
telnet in Internet, **2**-110
types of Internet access, **2**-108–**2**-109
Intranets, **2**-1–**2**-126
acronyms, **2**-113–**2**-115
application performance management, **2**-81–**2**-99
acceleration of business, **2**-83–**2**-84
acronyms, **2**-99
APM tool requirements, **2**-92–**2**-93
areas of applicability, **2**-86
assuring high quality of customer self-service, **2**-87
benefit of homogeny, **2**-90
business APM, **2**-89–**2**-90
doing more with APM data, **2**-90
enhancing contact center operations, **2**-87

ensuring availability of information hubs and busses, **2**-85
evolution of APM, **2**-91–**2**-92
increasing profitability and efficiency of service delivery, **2**-86–**2**-87
interested parties for APM, **2**-91
leading performance management issues today, **2**-85
meeting service-level agreements, **2**-85
need for APM in communication services, **2**-81–**2**-90
need for new type of service assurance solutions, **2**-85–**2**-86
operational APM, **2**-87–**2**-89
optimizing OSS/BSS, **2**-85
overview, **2**-91–**2**-93
performance management for next generation OSS/BSS and SDP, **2**-96–**2**-98
performance management for next generation Service Delivery Platforms, **2**-93–**2**-96
possibilities and challenges, **2**-81–**2**-83
reactive and proactive APM, **2**-92
rise of software, **2**-83
trends, **2**-98
use cases, **2**-89, **2**-90
electronic technologies, **2**-99–**2**-105
e-commerce technologies, **2**-100–**2**-103
management of Web services, **2**-105
Web service delivery challenges, **2**-103–**2**-105
Internet and intranet management concepts, **2**-2–**2**-30
application-level firewall, **2**-23
automating HTML/XML authoring, **2**-21
benefits of intranets, **2**-3
centralized vs. distributed control, **2**-21
CGI scripting, **2**-24–**2**-25
circuit-level firewalls, **2**-24
content development and design checklist, **2**-5–**2**-6
content management, **2**-18
content structure, **2**-16
conversion of paper documents into electronic form, **2**-19–**2**-20
cross-platform compatibility, **2**-4
data security, **2**-28
denial of service, **2**-29
desktop clients running TCP/IP, **2**-15
digital signatures, **2**-25
document encryption, **2**-26
dual-homed host firewalls, **2**-24
encryption, **2**-25–**2**-26
firewall architectures, **2**-24
firewalls, **2**-2–**2**-24
free clip art and images, **2**-17
funding growth, **2**-6
gaining support, **2**-4
improved corporate culture, **2**-4
improved customer service, **2**-3
improved help desks, **2**-3
information organization, **2**-15–**2**-16
interface design, **2**-17
interface to legacy database(s), **2**-20
Internet basics, **2**-8–**2**-12
Internet security, **2**-27–**2**-29
intranet components, **2**-12–**2**-15
intranet deployment, **2**-19–**2**-21
intranet document tracking information, **2**-20
intranet functionality, **2**-17–**2**-18
intranet implementation, **2**-15–**2**-19
intranet planning and management, **2**-4–**2**-8
intranet security issues, **2**-22–**2**-27
intranet security threats, **2**-26–**2**-27
IP spoofing, **2**-28
link encryption, **2**-25
maintaining information content on intranet, **2**-21
management overview, **2**-2–**2**-4
management summary, **2**-7–**2**-8
management support skills checklist, **2**-6
managing document links, **2**-21
modems, **2**-27–**2**-28
network-level firewall, **2**-23
network operating systems, **2**-13–**2**-14
network requirements, **2**-13
organizational challenges, **2**-7
paper-less office, **2**-3
passwords, **2**-28
physical security, **2**-27
planning intranet strategy, **2**-4–**2**-5
pretty good privacy, **2**-26
protecting against ICMP redirects, **2**-27
public-key encryption, **2**-25
screened host firewalls, **2**-24
screened subnet firewalls, **2**-24
secure HTTP, **2**-26
Secure Socket Layer, **2**-26
selecting implementation team, **2**-5–**2**-6
server hardware, **2**-14
source-routed traffic, 2-26–2-27
spoofing, **2**-27
standardizing hardware and software, **2**-21
TCP/IP security, **2**-28–**2**-29
technical support skills checklist, **2**-5
technological considerations, **2**-19–**2**-21
Total Quality Management, **2**-6–**2**-7
training employees, **2**-7
training and support, **2**-18–**2**-19
UNIX, **2**-13–**2**-14
UNIX Web servers, **2**-15
use of multiple servers, **2**-20–**2**-21
Web browsers, **2**-15
Web server software, **2**-14–**2**-15
Windows NT, **2**-14
workstation security, **2**-28

Internet protocols, 2-105–2-112
addressing for Internet, 2-106–2-107
communication protocols in Internet, 2-107–2-108
domain name system, 2-107
file transfer in Internet, 2-110–2-111
information search in Internet, 2-112
information transfer in Internet, 2-108
Internet e-mail, 2-109–2-110
mailing lists in Internet, 2-111–2-112
Netscape and Microsoft, 2-112
news and usenet, 2-111
telnet in Internet, 2-110
types of Internet access, 2-108–2-109
open source software, role of, 2-115–2-125
application servers, 2-122–2-123
browsers, 2-121–2-122
business process management, 2-124
commercial and noncommercial applications and tools, 2-116–2-117
content management systems, 2-123
cost considerations, 2-119–2-121
CRM and ERP, 2-124
database management, 2-122
e-mail servers, 2-123
examples, 2-121–2-124
languages and development environments, 2-122
licensing, 2-119
nature of open source, 2-115–2-116
office suites, 2-123
opportunities and vulnerabilities, 2-117–2-119
public domain, 2-119
security software, 2-122
trends, 2-124–2-125
virtualization, 2-123–2-124
Web servers, 2-121
wikis, 2-123
virtual private networking solutions, 2-30–2-38
frame relay, 2-37
Layer 2 or Layer 3 comparison, 2-38
Layer 2 protocols, 2-30–2-35
Layer 3 tunneling protocols, 2-35–2-37
Web-enabled data warehousing, 2-38–2-46
advantages, 2-39
benefits, 2-41
concepts, 2-39
data marts, 2-40
future growth, 2-40
future trends, 2-45–2-46
making it happen, 2-41–2-43
obstacles and limitations, 2-43
overview, 2-39–2-40
vendors, 2-43–2-45
Web farming, 2-45
Web performance management, 2-46–2-81
accounting management challenges, 2-53–2-54
application performance management, 2-75–2-79
changes in networking infrastructures, 2-69–2-70
configuration management challenges, 2-54
content-smart flow admission control, 2-58–259
content-smart link management, 2-59–2-60
content-smart load balancing, 2-60–2-61
content-smart quality of service and resource management, 2-58
in data collection, 2-70–2-72
fault management challenges, 2-54–2-55
generic intranet management challenges, 2-49–2-55
Internet, intranets, and extranets, 2-47–2-48
intranet performance management, specific challenges, 2-55–2-62
load balancing, 2-72–2-75
load-balancing firewall, 2-61
load-balancing switches, 2-61
load-balancing traffic shapers, 2-61–2-62
load distribution and balancing, 2-59
log file analysis, 2-62–2-68
managing content, 2-56–2-57
performance management challenges, 2-49–2-51
security management challenges, 2-51–2-53
technologies of access networks, 2-62
traffic monitoring tools, 2-72
Web performance management trends, 2-79–2-80
Web server management, 2-57–2-59
wire monitors and network analyzers, 2-68–2-72
ISO/IEC 17799, 4-137–4-139
access control, 4-139
asset classification and control, 4-138
business continuity management, 4-139
business drivers calling for implementation of standard, 4-139
communications and operations management, 4-138–4-139
compliance, 4-139
organizational security, 4-137–4-138
personal security, 4-138
physical and environmental security, 4-138
security policy, 4-137
systems development and maintenance, 4-139
ISS, *see* Intelligence Support Systems
ITIL, *see* Information Technology Infrastructure Library

J

Java, 2-12
Job profiles, human resources of telecommunications service providers, 4-114–4-123

business planner, **4**-117–**4**-118
call center operator, **4**-114–**4**-115
contact administrator, **4**-122
database administrator, **4**-120–**4**-121
inventory manager, **4**-115–**4**-116
legal counsel, **4**-121–**4**-122
manager of LEMF, **4**122–**4**-123
network infrastructure operator, **4**-120
network operations manager, **4**-114
operations manager for lawful intercepts, **4**-119–**4**-120
security analyst, **4**-119
service-level manager, **4**-117
service and problem manager, **4**-116
service technician, **4**-116–**4**-117
technology analyst, **4**-118

L

LANs, *see* Local area networks
LDAP, *see* Lightweight Directory Access Protocol
Lightweight Directory Access Protocol (LDAP), **3**-80–**3**-82
attributes, **3**-80–**3**-82
directory centralization with, **3**-81
limitations, **3**-82
Link encryption, **2**-25
Local area networks (LANs), **1**-21–**1**-38
IEEE 802.3 (CSMA/CD) specifics, **1**-23–**1**-36
adding collision detection, **1**-26–**1**-29
building cabling specifications, **1**-36–**1**-38
carrier sense multiple access, **1**-25–**1**-26
collision backoff scheme, **1**-27–**1**-29
example implementations, **1**-30–**1**-32
Fast Ethernet, **1**-33–**1**-34
fiber optic media, **1**-35
frame structure, **1**-23–**1**-24
full-duplex operation, **1**-36
Gigabit Ethernet, **1**-34, **1**-35
higher speed extensions, **1**-33–**1**-35
Logical Link Control layer, **1**-36
multisegment guidelines, **1**-33
repeaters, **1**-32
sample frame transmission, **1**-24–**1**-25
shielded copper cable, **1**-35
switched hubs, **1**-32–**1**-33
system components, **1**-29–**1**-30
topology extensions, **1**-32–**1**-33
major MAC standards **1**-22–**1**-23
overview, **1**-21–**1**-23
reference model, **1**-21–**1**-22
standards, **1**-21

M

Mailing lists in Internet, **2**-111–**2**-112
Management, *see* Network management and administration; Web performance management
MDM, *see* Mobile device management
Mobile device management (MDM), **5**-32–**5**-33
Modems, **2**-27–**2**-28
Multimedia applications, *see* Cognitive radio networks, multimedia applications for

N

NBA, *see* Network behavior analysis
Network behavior analysis (NBA), **5**–31–**5**-32
Network management and administration, **3**-1–**3**-256
Intelligence Support Systems, **3**-200–**3**-226
access function implementation approaches, **3**-216
acronyms, **3**-225
basic requirements for lawful intercepts, **3**-206–**3**-207
basics and application areas, **3**-201–**3**-204
intelligence transmission, **3**-219–**3**-220
internal and external lawful interception, **3**-215–**3**-216
positioning among other support and security systems, **3**-204–**3**-206
positioning lawful intercepts and surveillance, **3**-200–**3**-201
principal functions of interception, **3**-207–**3**-208
principles of monitoring and intercepts (hardware and software probes), **3**-214–**3**-220
receiver applications, **3**-220–**3**-222
reference models for lawful intercepts, **3**-208–**3**-214
trends, **3**-222–**3**-224
use of probes, **3**-216–**3**-219
management concepts, **3**-4–**3**-19
communications general model, **3**-4–**3**-5
distributed management systems, **3**-12–**3**-14
distributed management systems architectures, **3**-12–**3**-14
distributed network and computing systems, **3**-12
driving forces behind management technologies, **3**-6–**3**-7
high-level management functions, **3**-6
high-level users management requirements, **3**-6
in-band and out-of-band management systems, **3**-14
justifying network management investment, **3**-7
management base model, **3**-7–**3**-8
management and data communications, **3**-4–**3**-6
management domains, **3**-10–**3**-11
management paradigms, **3**-7–**3**-11
management platforms, **3**-15–**3**-16

management platforms for enterprisewide management, 3-17–3-19
management requirements, 3-6–3-7
management systems evolution, 3-16–3-19
management systems topological frameworks, 3-15–3-16
management system technical evolution, 3-16–3-17
management views and associated models, 3-8–3-10
manager of managers, 3-15
network management general model, 3-5–3-6
network of managers, 3-15
open management systems, 3-11–3-12
open management systems concept, 3-11–3-12
open systems conceptual model, 3-11
requirements for open management systems, 3-12
single manager, 3-14–3-15
trends, 3-19
management of emerging technologies, 3-19–3-51
acronyms, 3-49–3-51
address and identification schemes, 3-23–3-24
asynchronous transfer mode, 3-25–3-26
cable technology, 3-31–3-35
connection-oriented and connectionless communications, 3-20
control and congestion management, 3-24
digitial subscriber line technologies, 3-29–3-31
discussion of select technologies, 3-25–3-36
foundation concepts for networking technologies, 3-20–3-24
IMS and SIP, 3-42–3-45
IPTV and VoD, 3-45–3-47
management solutions, 3-36
management solutions for emerged technologies, 3-24–3-25
mobile and wireless communication, 3-36–3-38
move to IP, 3-39–3-41
multiplexing technologies, 3-23
multiprotocol label switching, 3-35–3-36
next-generation telecommunications application technologies, 3-39–3-47
next-generation wireless access technologies, 3-36–3-39
physical and virtual circuits, 3-20–3-21
routing technologies, 3-22–3-23
SONET and SDH, 3-26–3-29
switching technologies, 3-21–3-22
telco and Web 2.0 applications, 3-48–3-49
trends, 3-49
triple and quad play, 3-41
Wi-Fi, 3-38
WiMAX, 3-39
management frameworks and applications, 3-169–3-198
acronyms, 3-197
Amdocs, 3-180–3-182
APIs and development toolkits, 3-176
basic infrastructure, 3-170–3-173
consolidation of support systems, 3-193–3-194
evolving management frameworks, 3-170
features and attributes of management frameworks, 3-170–3-177
Hewlett-Packard, 3-182–3-183
IBM, 3-183–3-189
management applications, 3-177–3-178
management operations support services, 3-167–3-177
management services, 3-173–3-176
new vendors on OSS market, 3-192–3-193
Oracle, 3-190–3-192
results of market research, 3-194–3-195
Telcordia, 3-178–3-180
trends, 3-195
vendor profiles, 3-178–3-193
management function, 3-84–3-103
acronyms, 3-103
comparison of management models, 3-90–3-92
event correlation, 3-94–3-97
examples of products, 3-98–3-99
expectations and trends, 3-92–3-93
FCAPS functions, 3-86–3-88
fulfillment, assurance and billing processes (*e*TOM), 3-88–3-90
in-depth considerations of selected management functions, 3-93–3-102
security and risk considerations, 3-99–3-102
telecom expense management, 3-97–3-98
TMN functions, 3-84–3-86
trends, 3-102–3-103
management-related standards, 3-51–3-84
acronyms, 3-83–3-84
architecture, 3-66
business process execution language, 3-79
communication protocol and security, 3-65
description and location services, 3-78–3-79
desktop management interface, 3-63–3-64
functional components, 3-65
fundamental technologies, 3-77–3-78
Lightweight Directory Access Protocol, 3-80–3-82
manager–agent relationship, 3-52
OSS-J, 3-76
procedures, 3-67
protocol encoding and sessions, 3-67–3-69
remote monitoring, 3-59–3-63
Simple Network Management Protocol, 3-52–3-59
summary and trends, 3-82–3-83
Telecommunications Management Network, 3-69–3-73
TOM and *e*TOM, 3-73–3-75
transitioning from IPv4 to IPv6, 3-76–3-77

TR-069 CPE WAN management protocol, **3**-64–**3**-69
Web service technologies and SOA, **3**-77–**3**-82
management of sensor networks, **3**-226–**3**-240
acronyms, **3**-241
challenges of monitoring, **3**-227–**3**-229
objectives of sensory monitoring systems, **3**-227
selecting and applying sensory systems, **3**-236–**3**-238
sensory monitoring technologies and alternatives, **3**-229–**3**-235
trends, **3**-238
solution architectures, **3**-241–3-258
acronyms, **3**-258
business drivers require new OSS approach, **3**-242–**3**-243
business strategy drives technology, **3**-241
comprehensive architectural approach, **3**-241–**3**-242
job of OSS, **3**-243–**3**-253
OSS building, **3**-253–**3**-257
trends, **3**-258–**3**-259
support processes for service providers, **3**-132–**3**-169
acronyms, **3**-168–**3**-169
assurance, **3**-136–**3**-137
billing and revenue assurance, **3**-137
customer billing and collections management, **3**-143
customer contact management, retention, and loyalty, **3**-138
customer information management, **3**-138–**3**-139
customer interface management, **3**-139
customer problem handling, **3**-141
customer quality of service and service-level agreement, **3**-141–**3**-142
customer relationship management, **3**-138–**3**-143
customer self-service, customer self-management, **3**-142–**3**-143
eBusiness, **3**-154–**3**-164
end-to-end vertical process groupings, **3**-135
enterprise management, **3**-134–**3**-137
eTOM taxonomy, **3**-154
fulfillment, **3**-136
fulfillment, assurance, billing processes, **3**-135–**3**-136
high-level breakdown of support processes, **3**-134
Information Technology Infrastructure Library, **3**-164–**3**-166
infrastructure lifecycle management processes, **3**-135
inventory management process, **3**-148–**3**-149
IT management frameworks, **3**-164–**3**-166
operations support and readiness, **3**-137
product lifecycle management processes, **3**-135
rating and discounting, **3**-147–**3**-148
resource data collection and processing, **3**-152
Resource Management and Operations, **3**-148–**3**-152
resource, network planning, and development process, **3**-149–**3**-150
resource performance management process, **3**-151–**3**-152
resource provisioning, **3**-150
resource trouble management process, **3**-151
selling/sales process, **3**-139
service configuration and activation process, **3**-144–**3**-145
service design and assign, **3**-145
service-level agreement management, 3-147
Service Management and Operations, **3**-143–**3**-148
service order management, order handling, **3**-140–**3**-141
service planning and development process, **3**-143–**3**-144
service problem management, **3**-145–**3**-146
service quality management process, **3**-146–**3**-147
strategy and commit processes, **3**-135
strategy, infrastructure, and product process, **3**-135
Supplier/Partner Relationship Management, **3**-152–**3**-153
support processes taxonomy, **3**-153–**3**-154
taxonomy by Amdocs, **3**-154
taxonomy by Insight Research, **3**-154
TMN-OSS model, **3**-153
trends, **3**-166–**3**-168
workforce management, **3**-149
support systems for service providers, **3**-104–**3**-131
acronyms, **3**-130–**3**-131
carrier support system issues, **3**-107–**3**-109
convergence and telecom consolidation, **3**-106
convergent billing, **3**-108
developing markets, **3**-106
emergence of complex, multiplatform environments, **3**-106
emphasis on system integration, **3**-106
evaluation of suppliers, **3**-108
future of OSS/BSS, **3**-117–**3**-127
importance of quick ROI, **3**-106
independent software vendors, **3**-107
industry issues of support systems, **3**-105–**3**-107
legacy dilemma, **3**-107
migration dilemma, **3**-107
new service offerings, **3**-108
OSS/BSS growth, **3**-106
OSS/BSS market drivers, **3**-110–**3**-117
outsourcing, **3**-106, **3**-109
product-based, vendor-driven solutions, **3**-105

professional services, **3**-107, **3**-109
rapid fulfillment of new services, **3**-108
service differentiation, **3**-108
service providers' investment strategies, **3**-106
status, definitions, and markets for support systems, **3**-105–**3**-110
trends, **3**-129
upgrading of cycles in support systems, **3**-105
Network neutrality, **5**-21
Network organization and governance, **4**-1–**4**-147
best practices benchmarks for service providers, **4**-123–**4**-147
acronyms, **4**-146
benchmarking, **4**-123–**4**-134
Control Objectives for Information and Related Technology, **4**-140–**4**-142
dashboards and balanced scorecards, **4**-142–**4**-145
Information Technology Infrastructure Library, **4**-134–**4**-137
ISO/IEC 17799, **4**-137–**4**-139
trends, **4**-145–**4**-146
business intelligence and analytics, **4**-27–**4**-51
acronyms, **4**-51
customer intelligence, **4**-42–**4**-46
customer is king, **4**-29–**4**-31
disappointments from past and barriers to business intelligence, **4**-39–**4**-41
four Cs of telecommunications industry, **4**-28–**4**-29
impact of BI and BPM on telecommunications industry, **4**-46–**4**-50
overcoming barriers, **4**-41–**4**-42
strategy at work, **4**-37–**4**-39
telecommunications industry, **4**-32–**4**-37
trends, **4**-50–**4**-51
information life cycle management, **4**-3–**4**-16
acronyms, **4**-16
critical success factors of document management, **4**-14–**4**-15
document life cycle, **4**-4–**4**-7
hot topics, **4**-7–**4**-14
terms, standards, and statistics, **4**-3–**4**-4
trends, **4**-15–**4**-16
information technology alignment with businesses, **4**-17–**4**-26
acronyms, **4**-26
baseline IT plan, **4**-18–**4**-19
directions for service orientation, **4**-20–**4**-22
importance of IT, **4**-17–**4**-18
real-time enterprise, **4**-19–**4**-20
role of IT infrastructure, **4**-22–**4**-25
support of business by IT, **4**-18
trends, **4**-25–**4**-26
management services and outsourcing, **4**-77–**4**-107, **4**-102–**4**-103
acronyms, **4**-106–**4**-107
contract management, **4**-103–**4**-106
multiprovider collaboration and peering, **4**-87–**4**-102
policies and tasks of governance, **4**-77–**4**-86
trends, **4**-106
network management organization, **4**-107–**4**-123
assigning subject matter experts to processes and support systems, **4**-109
building management teams, **4**-109–**4**-111
job profiles for human resources of telecommunications service providers, **4**-114–**4**-123
keeping management teams, **4**-111–**4**-113
organization structure of average provider, **4**-108–**4**-109
trends, **4**-123
service-level management, **4**-51–**4**-76
acronyms, **4**-76
certification of SLAs, **4**-68–**4**-72
principal terms and metrics, **4**-51–**4**-58
process, **4**-58–**4**-61
role of SLAs in settlements between service providers, **4**-72–**4**-75
sample SLA, **4**-61–**4**-67
trends, **4**-75
New product and service creation, **5**-33–**5**-38
drivers and constraints, **5**-33–**5**-35
increasing bandwidth, **5**-37–**5**-38
new service creation, **5**-36–**5**-37

O

Open management systems, **3**-11–**3**-12
concept, **3**-11–**3**-12
conceptual model, **3**-11
requirements, **3**-12
Open source software (OSS), **2**-115–**2**-125
application servers, **2**-122–**2**-123
browsers, **2**-121–**2**-122
building, **3**-251–**3**-255
applications, **3**-254–**3**-255
SOA approach, **3**-252–**3**-254
business process management, **2**-124
commercial and noncommercial applications and tools, **2**-116–**2**-117
content management systems, **2**-123
cost considerations, **2**-119–**2**-121
CRM and ERP, **2**-124
database management, **2**-122
e-mail servers, **2**-123
examples, **2**-121–**2**-124
job of, **3**-243–**3**-253
assurance functions, **3**-250–**3**-252
fulfillment, assurance, and usage, **3**-246
fulfillment functions, **3**-248–**3**-250
OSS/BSS integration functions, **3**-247–**3**-248

service lifecycles, **3**-244–**3**-246
usage, **3**-252–**3**-253
languages and development environments, **2**-122
licensing, **2**-119
market, new vendors on, **3**-192–**3**-193
nature of open source, **2**-115–**2**-116
office suites, **2**-123
opportunities and vulnerabilities, **2**-117–**2**-119
public domain, **2**-119
security software, **2**-122
trends, **2**-124–**2**-125
virtualization, **2**-123–**2**-124
Web servers, **2**-121
wikis, **2**-123
OSS, *see* Open source software
OSS/BSS, future of, **3**-117–**3**-127
billing convergence, **3**-120–**3**-121
customer experience management, **3**-120
customer network management, **3**-119
customer orientation, **3**-119
fulfillment, **3**-121–**3**-123
inventory management, **3**-124–**3**-125
management convergence, **3**-121
OSS transformation, **3**-117–**3**-118
promising market, **3**-125–**3**-127
service assurance, **3**-123–**3**-124
OSS/BSS market drivers, **3**-110–**3**-117
convergence, **3**-112–**3**-116
device, **3**-115–**3**-116
digital **3**-116
fixed-mobile, **3**-114
industry and communications, **3**-113
network, **3**-114–**3**-115
service, **3**-113–**3**-114
telecom-media, **3**-116
deregulation and privatization, **3**-110–**3**-111
growth of global telecommunications market, **3**-111
increasing network complexity, **3**-111–**3**-112
Outsourcing, management services and, **4**-77–**4**-107
contract management, **4**-103–**4**-106
benefits of contract management tools, **4**-105–**4**-106
selection and setup issues and concerns, **4**-106
management services and outsourcing, **4**-102–**4**-103
multiprovider collaboration and peering, **4**-87–**4**-102
detailed process steps of multiprovider collaboration, **4**-89–**4**-102
peering point and administration boundaries, **4**-87–**4**-88
process of multiprovider collaboration, **4**-88–**4**-89
policies and tasks of governance, **4**-77–**4**-86
governance models, **4**-77–**4**-85
security frameworks of governance, **4**-86
tasks of governance, **4**-85–**4**-86

P

Packet switching, **2**-8
Paper-less office, **2**-3
Passwords, **2**-28
PGP, *see* Pretty good privacy
Point-to-Point Tunneling Protocol, **2**-31–**2**-32
POP, **2**-110
Pretty good privacy (PGP), **2**-26
Product lifecycle management processes, **3**-135
Programming language
Java, **2**-12
NesC, **1**-57
Public-key encryption, **2**-25

Q

QoS, *see* Quality of service
Quality of service (QoS), **1**-65

R

Radio frequency identification (RFID), **1**-39
Real-time enterprise (RTE), **4**-19–**4**-20
Resource Management and Operations (RM&O), **3**-148–**3**-152
inventory management process, **3**-148–**3**-149
resource data collection and processing, **3**-152
resource, network planning, and development process, **3**-149–**3**-150
resource performance management process, **3**-151–**3**-152
resource provisioning, **3**-150
resource trouble management process, **3**-151
workforce management, **3**-149
RFID, *see* Radio frequency identification
RFID architecture and protocols, **1**-39–**1**-50
architecture, **1**-39–**1**-43
back end software subsystem, **1**-42–**1**-43
front end communication subsystem, **1**-40–**1**-42
Gen-2 performance improvement, **1**-49–**1**-50
Gen-2 RFID protocol, **1**-43–**1**-49
adaptive Q algorithm, **1**-48–**1**-49
basic operations of identifying tags, **1**-43–**1**-47
RM&O, *see* Resource Management and Operations
RTE, *see* Real-time enterprise

S

SDPs, *see* Service Delivery Platforms
Secure HTTP (S-HTTP), **2**-26
Secure Socket Layer (SSL), **2**-26
Sensor networks, management of, **3**-226–**3**-240
challenges of monitoring, **3**-227–**3**-229
IP Multimedia Subsystem, **3**-229
IP services, **3**-228

merging of traditional signaling onto IP networks, 3-228–3-229
objectives of sensory monitoring systems, 3-227
selecting and applying sensory systems, 3-236–3-238
signaling vs. packet flow, 3-238
visibility, 3-236–3-238
sensory monitoring technologies and alternatives, 3-229–3-235
agent-based sensor networks, 3-231–3-232
CDR-based sensor networks, 3-231
device-based sensor networks, 3-229–3-230
packet-flow-based sensor networks, 3-234–3-235
signaling-based sensor networks, 3-232–3-233
trends, 3-238
Service Delivery Platforms (SDPs), 2-93–2-96
challenges with traditional network management tools, 2-94–2-96
performance management for next generation, 2-93–2-96
SDP overview, 2-93–2-94
strategic investment, 2-94
Service level agreement
application performance management and, 2-85
certification, 4-68–4-72
attributes of professional solution, 4-68
difficulties with different techniques and tools, 4-68
measurement procedures, 4-68–4-69
QoS procedures, 4-69–4-70
SLA procedures, 4-70–4-72
customer, 3-141–3-142
management, 3-147
role of in settlements between service providers, 4-72––475
assistance of SLAs in settlements, 4-73–4-74
carriers fighting for stronger SLAs, 4-72
present difficulties, 4-72
tasks of overseer, 4-74–4-75
sample, 4-61–4-67
closing comments, 4-65
horizontal and vertical SLAs, 4-65–4-66
QoS techniques, 4-67
template, 4-62–4-65
traffic classes, 4-66–4-67
Service-Level Management (SLM), 4-51–4-76
certification of SLAs, 4-68–4-72
attributes of professional solution, 4-68
difficulties with different techniques and tools, 4-68
measurement procedures, 4-68–4-69
QoS procedures, 4-69–4-70
SLA procedures, 4-70–4-72
principal terms and metrics, 4-51–4-58
classification of metrics, 4-55
general considerations, 4-55–4-58
service-dependent metrics, 4-54–4-55
service-independent metrics, 4-52–4-54
process, 4-58–4-61
role of SLAs in settlements between service providers, 4-72–4-75
assistance of SLAs in settlements, 4-73–4-74
carriers fighting for stronger SLAs, 4-72
present difficulties, 4-72
tasks of overseer, 4-74–4-75
sample SLA, 4-61–4-67
closing comments, 4-65
horizontal and vertical SLAs, 4-65–4-66
QoS techniques, 4-67
template, 4-62–4-65
traffic classes, 4-66–4-67
Service Management and Operations (SM&O), 3-143–3-148
rating and discounting, 3-147–3-148
service configuration and activation process, 3-144–3-145
service design and assign, 3-145
service-level agreement management, 3-147
service planning and development process, 3-143–3-144
service problem management, 3-145–3-146
service quality management process, 3-146–3-147
Service-oriented architecture (SOA), 3-77, 4-21
S-HTTP, *see* Secure HTTP
SLM, *see* Service-Level Management
SM&O, *see* Service Management and Operations
SOA, *see* Service-oriented architecture
Social networking, 5-21–5-22
Spoofing
description of, 2-27
IP, 2-28
S/PRM, *see* Supplier/Partner Relationship Management
SSL, *see* Secure Socket Layer
Supplier/Partner Relationship Management (S/PRM), 3-152–3-153

T

Tariffing, 5-38–5-42
impact of new technologies, 5-41–5-42
regulatory trends, 5-38–5-39
service pricing trends, 5-39–5-41
TCP/IP, 2-8
desktop clients running, 2-15
encryption, 2-26
purpose of, 2-29
security, 2-28–2-29
denial of service, 2-29
IP spoofing, 2-28
Telecommunications Management Network (TMN), 3-69–3-73
mediation function, 3-71

network element function, **3**-71–**3**-73
Q adapter function, **3**-71
workstation function, **3**-71
Telecommunications services, future, **5**-1–**5**-48
application trends, **5**-13–**5**-22
application functionality, **5**-14–**5**-18
e-mail applications, **5**-16
enhancements to traditional services, **5**-18
functionality implementation, **5**-18–**5**-20
Internet applications, **5**-14–**5**-16
Internet functionality implementation, **5**-20
network neutrality, **5**-21
semantic Web, **5**-21
social networking, **5**-21–**5**-22
video conferencing applications, **5**-16
wireless applications, **5**-17–**5**-18
wireless functionality implementation, **5**-20
new product and service creation, **5**-33–**5**-38
drivers and constraints, **5**-33–**5**-35
increasing bandwidth, **5**-37–**5**-38
new service creation, **5**-36–**5**-37
systems and service integration for management, **5**-22–**5**-30
acronyms, **5**-30–**5**-31, **5**-32
drivers for integration, **5**-22–**5**-24
integration for business users, **5**-25–**5**-27
integration for mobile professionals, **5**-27
integration for residential users, **5**-28–**5**-29
integration for service providers, **5**-24–**5**-25
integration for SOHO users, **5**-27–**5**-28
mobile device management, **5**-32–**5**-33
network behavior analysis, **5**-31–**5**-32
unified threat management, **5**-29–**5**-30
telecommunications strategies, **5**-42–**5**-46
goals, **5**-46
green computing, **5**-46–**5**-48
service providers, **5**-42–**5**-46
software as service, **5**-48
telecommunications tariffing, **5**-38–**5**-42
impact of new technologies, **5**-41–**5**-42
regulatory trends, **5**-38–**5**-39
service pricing trends, **5**-39–**5**-41
user needs, **5**-2–**5**-13
business end-user needs, **5**-4–**5**-6
different users have different needs, **5**-4–**5**-11
end-user requirements summary, **5**-12–**5**-13
instant messaging, **5**-11–**5**-12
location services, **5**-12
mobile professional end-user needs, **5**-6–**5**-7
residential end-user needs, **5**-9–**5**-11
SOHO end user needs, **5**-8–**5**-9
types of users, **5**-3–**5**-4
unified messaging, **5**-11
TMN, *see* Telecommunications Management Network
Total Quality Management (TQM), **2**-6–**2**-7
TQM, *see* Total Quality Management

U

Uniform Resource Locator (URL), **2**-10
UNIX, **2**-13–**2**-14
UNIX Web servers, **2**-15
URL, *see* Uniform Resource Locator

V

Virtual private network (VPN), **2**-30–**2**-38
frame relay, **2**-37
Layer 2 or Layer 3 comparison, **2**-38
Layer 2 protocols, **2**-30–**2**-35
Layer 2 Forwarding, **2**-32–**2**-34
Layer 2 Tunneling Protocol, **2**-34–**2**-35
Point-to-Point Tunneling Protocol, **2**-31–**2**-32
Layer 3 tunneling protocols, **2**-35–**2**-37
IPSec, **2**-36
Mobile IP, **2**-36–**2**-37
Voice and data communications, **1**-1–**1**-70
acronyms, **1**-21
cognitive radio networks, multimedia applications for, **1**-58–**1**-67
cognitive radio devices, **1**-62–**1**-63
cognitive radios and cognitive radio networks, **1**-59–**1**-60
dynamic spectrum access, **1**-60–**1**-62
policies for cognitive radio operation, **1**-63–**1**-64
pricing schemes for multimedia applications, **1**-65–**1**-67
quality of service, **1**-65
spectrum management, **1**-61
spectrum mobility, **1**-62
spectrum sensing, **1**-60–**1**-61
spectrum sharing, **1**-61–**1**-62
Computer Telephone Integrated, **1**-2–**1**-12
applications and trends, **1**-11–**1**-12
basic definitions, **1**-2–**1**-3
brief history, **1**-3–**1**-4
call control, **1**-7–**1**-8
components and models, **1**-5–**1**-11
first-party and third-party CTI, **1**-8–**1**-11
media processing, **1**-5–**1**-7
IEEE 802.3 (CSMA/CD) specifics, **1**-23–**1**-36
adding collision detection, **1**-26–**1**-29
building cabling specifications, **1**-36–**1**-38
carrier sense multiple access, **1**-25–**1**-26
collision backoff scheme, **1**-27–**1**-29
example implementations, **1**-30–**1**-32
Fast Ethernet, **1**-33–**1**-34
fiber optic media, **1**-35
frame structure, **1**-23–**1**-24
full-duplex operation, **1**-36
Gigabit Ethernet, **1**-34, **1**-35
higher speed extensions, **1**-33–**1**-35
Logical Link Control layer, **1**-36
multisegment guidelines, **1**-33

repeaters, **1**-32
sample frame transmission, **1**-24–**1**-25
shielded copper cable, **1**-35
switched hubs, **1**-32–**1**-33
system components, **1**-29–**1**-30
topology extensions, **1**-32–**1**-33
local area networks, **1**-21–**1**-38
IEEE 802.3 (CSMA/CD) specifics, **1**-23–**1**-36
major MAC standards **1**-22–**1**-23
overview, **1**-21–**1**-23
reference model, **1**-21–**1**-22
standards, **1**-21
RFID architecture and protocols, **1**-39–**1**-50
adaptive Q algorithm, **1**-48–**1**-49
architecture, **1**-39–**1**-43
back end software subsystem, **1**-42–**1**-43
basic operations of identifying tags, **1**-43–**1**-47
front end communication subsystem, **1**-40–**1**-42
Gen-2 performance improvement, **1**-49–**1**-50
Gen-2 RFID protocol, **1**-43–**1**-49
summary, **1**-70
voice over IP, **1**-13–**1**-20
applications, **1**-13–**1**-15
component-based overview, **1**-15–**1**-20
integration of voice and IP data, **1**-13
keys to successful deployment, **1**-20
wireless sensor network applications, hardware and software (design), **1**-51–**1**-57
BTnode, **1**-54
design of WSNs, **1**-52–**1**-53
event-based radio model, **1**-56–**1**-57
execution model, **1**-56
hardware components, **1**-54–**1**-55
Imote2, **1**-54
Mica2, **1**-53–**1**-54
Mica2Dot, **1**-53
Micaz, **1**-54
motes, **1**-53–**1**-54
NesC programming language, **1**-57
processors, **1**-54
radio transceivers, **1**-54–**1**-55
TinyOS, **1**-55–**1**-57
Tmote Sky, **1**-53
WSN research, **1**-53
WSN versus conventional networking, **1**-51–**1**-52
Voice over IP (VoIP), **1**-13–**1**-20
applications, **1**-13–**1**-15
component-based overview, **1**-15–**1**-20
integration of voice and IP data, **1**-13
keys to successful deployment, **1**-20
VoIP, *see* Voice over IP
VPN, *see* Virtual private network

W

Web browsers, **2**-9–**2**-10, **2**-15
Web-enabled data warehousing, **2**-38–**2**-46
advantages, **2**-39
benefits, **2**-41
concepts, **2**-39
data marts, **2**-40
future growth, **2**-40
future trends, **2**-45–**2**-46
making it happen, **2**-41–**2**-43
obstacles and limitations, **2**-43
overview, **2**-39–**2**-40
vendors, **2**-43–**2**-45
established DSS vendors, **2**-44–**2**-45
major database vendors, **2**-44
start-up DSS vendors, **2**-44
Web farming, **2**-45
Web performance management, **2**-46–**2**-81
application performance management, **2**-75–**2**-79
architectures, **2**-77–**2**-78
products and product suites, **2**-78–**2**-79
response time measurements, **2**-76
trends, **2**-79
generic intranet management challenges, **2**-49–**2**-55
accounting management, **2**-53–**2**-54
configuration management, **2**-54
fault management, **2**-54–**2**-55
performance management, **2**-49–**2**-51
security management, **2**-51–**2**-53
Internet, intranets, and extranets, **2**-47–**2**-48
intranet performance management, specific challenges, **2**-55–**2**-62
content-smart flow admission control, **2**-58–**2**59
content-smart link management, **2**-59–**2**-60
content-smart load balancing, **2**-60–**2**-61
content-smart quality of service and resource management, **2**-58
load-balancing firewall, **2**-61
load-balancing switches, **2**-61
load-balancing traffic shapers, **2**-61–**2**-62
load distribution and balancing, **2**-59
managing content, **2**-56–**2**-57
technologies of access networks, **2**-62
Web server management, **2**-57–**2**-59
load balancing, **2**-72–**2**-75
issues in deploying load-balancing products, **2**-74–**2**-75
load-balancing tools, **2**-75
move to Internet/intranet-based business, **2**-73
need for bandwidth, service quality, and granularity, **2**-73–**2**-74
need for granularity, **2**-74
need for guaranteed bandwidth, **2**-73
need for service-level agreements, **2**-73–**2**-74
log file analysis, **2**-62–**2**-68
drawbacks of pure log file analyzers, **2**-67–**2**-68
issues, **2**-64–**2**-67
tools, **2**-68
usage analysis, **2**-63–**2**-64

web performance management trends, **2**-79–**2**-80
wire monitors and network analyzers, **2**-68–**2**-72
changes in networking infrastructures, **2**-69–**2**-70
issues in data collection, **2**-70–**2**-72
traffic monitoring tools, **2**-72
Web server software, **2**-14–**2**-15
Web service delivery challenges, **2**-103–**2**-105
building relationships, **2**-104
high cost of small errors, **2**-103–**2**-104
security, **2**-104
speed, **2**-104
technology evolution, **2**-105
Web site activity reporting, **2**-62
Windows NT, **2**-14
Wireless sensor network (WSN), **1**-51–**1**-57
design of WSNs, **1**-52–**1**-53
ease of deployment, **1**-52
latency, **1**-52
quality, **1**-52–**1**-53
system lifetime, **1**-52
hardware components, **1**-54–**1**-55
processors, **1**-54
radio transceivers, **1**-54–**1**-55
motes, **1**-53–**1**-54
BTnode, **1**-54
Imote2, **1**-54
Mica2, **1**-53–**1**-54
Mica2Dot, **1**-53
Micaz, **1**-54
Tmote Sky, **1**-53
TinyOS, **1**-55–**1**-57
event-based radio model, **1**-56–**1**-57
execution model, **1**-56
NesC programming language, **1**-57
other operating systems, **1**-55–**1**-56
requirements, **1**-55
WSN research, **1**-53
WSN versus conventional networking, **1**-51–**1**-52
World Wide Web, **2**-9
WSN, *see* Wireless sensor network

Information Technology / Telecommunications

CRC Handbook of MODERN TELECOMMUNICATIONS

Second Edition

Edited by PATRICIA MORREALE · KORNEL TERPLAN

Addressing the most dynamic areas of the ever-changing telecommunications landscape, this new edition of the bestselling **CRC Handbook of Modern Telecommunications** once again brings together the top minds and industry pioneers in wireless communication networks, protocols, and devices.

In addition to new discussions of radio frequency identification (RFID) and wireless sensor networks, including cognitive radio networks, this important reference systematically addresses network management and administration, as well as network organization and governance, topics that have evolved since the development of the first edition.

Extensively updated and expanded, this second edition provides new information on

- Wireless sensor networks
- RFID architectures
- Intelligent support systems
- Service delivery integration with the Internet
- Information life cycle and service level management
- Management of emerging technologies
- Web performance management
- Business intelligence and analytics

The text details the latest in voice communication techniques, advanced communication concepts, network organization, governance, traffic management, and emerging trends. This comprehensive handbook will provide telecommunications professionals across all fields with ready access to the knowledge they require and arm them with a clear understanding of the role that evolving technologies will play in the development of the telecommunications systems of tomorrow.

CRC Press
Taylor & Francis Group
an informa business

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

78003
ISBN: 978-1-4200-7800-8
90000
9 781420 078008
www.crcpress.com