TELECOMMUNICATION EXCHANGES

The telecommunications network including both public and private parts is one of the most important systems created by humans in modern civilization. The network enables people to communicate between continents at all times of the year and allows thoughts and ideas to be exchanged between families, companies, and governments. Several phone calls have changed history. The network also saves lives every day. An example is when mobile subscribers in their cars call emergency numbers when witnessing car accidents.

Due to the network's great importance the requirements on the components that constitute the system are several. The more central parts of the network must always be available. The most important type of such a central component in the classic telecommunication network is the telecommunication exchange. The exchange enables many calls to be switched and alternative paths to be taken if a path in the network has failed, and hence should be very reliable. This reliability is probably the most important requirement of the exchange. Other requirements are that the exchange should be able to cope with all signaling standards in a network. It should also be able to coexist with all equipment in the network, irrespective of its age or fabrication. This imposes great requirements on backward compatibility.

Note that a telecommunication exchange is sometimes also called a central office or switch. There are both public exchanges and private branch exchanges. An exchange is a node in a telecommunications network that includes functions for access, control, switching and charging of calls. These parts may be physically separated and distributed, also to places outside the central office site; and act as subnodes connected via signaling protocols, for example between access, control and switch resources. A call in a modern context is not just about phone calls but is a general connectivity service for two or more parties that need access to communication bandwidth over a shorter or longer period of time. Connection services can be established not only between peer users but also from peer users to central servers.

A telecommunications exchange is one of the most complex systems created by humans. State-of-the-art hardware and software technology is used to create very large systems in terms of the number of connected subscribers and number of switched calls. The development cost of an exchange is very large. Several millions of hours of hardware and software development are invested each year. An exchange can connect hundreds of thousands of subscribers and switch a million of calls at the busiest hour during the day. And all of this is in real time, meaning short setup time and low delay of the voice. The quality of service requirements is thus great. Whether or not the call is made between two different continents, the call has to be switched through with practically no delays and provide an acceptable speech quality.

HISTORICAL BACKGROUND

The basic service associated with traditional telephony is the enabling of the bidirectional voice communication dialogue between two persons located at different but fixed places. To enable such voice communication each person uses a telephone set equipped with a transmitting device starting with (1) a microphone that translates acoustic energy to electric energy and (2) a receiving device ending with a loudspeaker that does the reverse transformation.

Each person who has access to telephony also wants to be able to select whom to connect to and have a dialogue with. For the establishment of such a connection service, one must be able to signal from each place to any other selected place that a connection is wanted between one calling and another called person. The traditional call is hence actually directed to a place where a called person is supposed to be located. To enable the sending and reception of alert signals the telephone set is equipped with a device that can generate some form of electrically encoded signal and a device for reception and transformation of such a signal to an acoustic signal. For a basic two party dialogue, two telephone sets are connected, usually electrically via a pair of wires. In the very beginning the second wire was implemented by an earth connection; this, however, resulted in a high level of cross-talk when several such unshielded wires came close to each other.

To start with, telephones were mainly used between two or a few locations such as between a shop owner's office, workshop, and home. A simple n-way switch could be used to select who to talk to within such a small mesh structured private net. Later when several families and organizations in a town had such connections, the communication possibilities were extended. This was done by connecting all wires in a star structure to a central office equipped with an exchange consisting of a manual switchboard and an operator. Hence, each telephone in such a local area was connected via a single wire or a pair of wires to the switchboard. Such a pair of wires is often called a subscriber (or access) line or loop. The switchboards themselves were then interconnected via other pairs of wires, called trunks. To enable voice signals to be carried with less distortion and over longer distances (by use of inductor coils and amplifiers respectively) the trunk lines soon came to use four wires, one twisted pair in each direction.

The rapid growth of the telephony network during the early days may well be explained by the fact that the network technology had large similarities with the technology already used for telegraphy. But perhaps most important, the telephone set replaced the need for a telegraph operator knowing Morse code and thus simplified the human–machine interface. To establish a call, the calling person had first to send a ring signal to the operator and then tell the thus alerted operator whom to be connected to. If the call was local, the operator then could send a ring signal to the person being called and ask this person for permission to set up a connection to the person calling. The operator could at this point make clear how or by whom the call was to be paid. A call could then be established by connecting the two pairs of wires to each other. To simplify the operator's work the manual switchboard was designed to

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 2007 John Wiley & Sons, Inc.

make it simpler for the operator to supervise the lines that were busy and to handle both the request for, establishment and the ending of a call. If the call was non-local, the operator first searched for and selected a free trunk in the right direction and then called one of the operators at the receiving end of the trunk (using it as a signaling trunk) and asked for help to establish the call. Such searching, routing and forwarding tasks presume that that the operator had some knowledge about the networks topology. For a long-distance call a chain of operators had to be involved via a chain of trunk line links before the called person could be reached and asked for permission to set up the call via voice trunks along the same path.

To set up a long-distance call, a number of resources must be free. Those resources mainly were operators and voice trunks. If some resource was lacking, the operator could serve the customer by organizing a waiting list or job queue and then set up calls when resources eventually became free. The operators had to keep track of each job by help of a written job record including time stamps for charging purposes and also provide time supervision of the ongoing calls to guarantee that resources allocated to a call were released even if the calling parties forgot to send an end-of-call alert signal.

When the number of subscribers increased, quite a large number of operators could work in the same exchange office. Different techniques were then used to ensure that they could share the workload and coordinate the setup of calls between subscribers connected to different switchboards. Parallel processing of call attempts could be achieved, for example, by distributing the incoming access attempt from a subscriber to a non-busy switchboard. Another example is that one single operator could handle the setup of a local call within a large exchange consisting of many switchboards by use of multiple point technique — that is, by having switchboards all outgoing lines connected to all other switchboards in the office.

When the demand for telephony increased, the number of operators swelled. Pressure grew to decrease the costs for human switchboard operators but also the time to set up calls. Requirements on increased personal integrity were also a reason to try to automate the call setup procedure. The first automated switches were based on the use of electromagnetic coils, effectuating drive mechanisms and contact points. These electromagnetic devices were, in turn, controlled by signals generated from the telephone set by use of a dial. The dial could generate sequences of current pulses, where the number of pulses corresponded to a dialed decimal digit. A subscriber was given a telephone number related to a corresponding access line. The digits used to represent this number controlled the behavior of the switch. The signals were first decoded and used directly to control the switch movements in a decadic way and somewhat later indirectly via registers. The use of registered signals (1) reduced the requirements on timing of the signals, mechanical precision, and preventive maintenance and (2) increased the flexibility by making number translations possible.

All decisions to be made were not controlled by digits. For example, a number of trunk lines between two exchanges could be treated as a group and the digits used merely to select and seize a trunk group going in the right direction. Then one could do a search (or hunt) for a free trunk line and, when available, select one within this group. When no trunk line was free, another possible route could be tried. However, the number of alternative routes was limited both due to economic reasons and because the networks soon came to be built more or less hierarchically to simplify the coordination work needed to establish non-local calls. If no free route could be found, a blocking situation occurred and the call attempt had to wait for resources to become free. The operator could then use different methods to supervise resource release events and to handle the queue of waiting call attempts.

It is interesting to note that the tasks performed by a human operator, such as searching for, reservation, monitoring and management of resources and charging records, were very similar to what the control system of a modern exchange does. There are also interesting analogies between the call routing, redirection and answering services provided by a human operator and what today can be provided by the control system of an automated exchange.

The network has evolved from very simple bidirectional communication links via small private mesh and star networks that as soon as signal regeneration and amplification technology permitted were interconnected via transit networks to more public, global and hierarchical network structures. The number of individual trunk lines could also later be reduced by multiplexing techniques allowing several calls to share a physical line. This evolution started with frequency division and has evolved via digital time division toward many different combinations of frequency, phase, time and code division. Multiplexing creates a logical network layer on top of the physical transmission media implementing a number of logical lines (or channels) and hence a more efficient use of each physical line.

Logically, signaling has always been separated from the voice connection. A trend has been to clarify this by a separation into a signaling trunk network and a voice trunk network. However, the signaling network may in practice use reserved logical lines or channels multiplexed on top of the same physical lines as the voice network.

Voice encoding, with its influence on transmission and switching, has evolved from analog; via digital to compressed digital representation and the signal encoding has evolved from simple current pulses, via frequency-encoded signals to digital message records.

Manual exchanges handled by human operators were quite flexible and intelligent in many ways since the primitive alert signals simply could be complemented by verbal communication between subscriber and operator — that is, human to human. Automation required a predefined signaling scheme, including not only simple alert signals but also encoding of the phone number of the called line. To make the automated exchanges able to provide large flexibility and more advanced services, the first decadic control of an exchange directly from signals representing digits evolved via register mapped control to stored program control of exchange behavior. The automation possibilities were increased further by introduction of larger signal alphabets and protocols capable of more than just alert signals and digits.

A great step from a functional point of view was the introduction of radio transmission for cellular coverage and wireless access to/from mobile terminals. This allows a call to be directed to a mobile terminal carried by a person rather than just to a fixed place (terminating a wire or fiber). Technically it was not new to reuse the radio spectrum by dividing a geographical area into regions, in this case called cells, but connecting the cells covered by base stations to the switched telecommunication network was new and created many new challenges and opportunities.

Other steps in this direction are: digital subscriber lines; new call services (often for redirection of calls) sometimes substituting what a manual operator previously could give help with; introduction of personal numbers that in conjunction with mobility services can help to make it easier to reach a specific person rather than a phone terminal and also can increase competition among operators if the personal phone number becomes a property of a person rather than an operator; the merger of telecommunication and data communication networks that enables new multimedia communication services.

A recent such network convergence that will simplify the evolution of streamed and real-time multimedia services is the sharing of a more and more common infrastructure for fixed and mobile voice, data and video services. In this new setting some of the functions of a telecommunication exchange become obsolete and other more dedicated nodes become more important. For example, circuit switch functions may be replaced by packet forwarding while new forms of access control, authorization, authentication, address location and charging services will be more important. This development along with international standardization is also believed to increase the competition between different equipment suppliers.

An exchange can include support for almost all functions in a telephone network. However, one can also distinguish specialized network nodes. Early examples of these specialized nodes are local exchanges and transit exchanges. Today one can also distinguish other types of nodes such as network access nodes; switch, call and service control nodes; mobile switching nodes; and network database nodes (for example, databases for: number translations; handling of subscriber service profiles; location information supporting mobility of subscribers, authentication/authorization data, equipment data). Other nodes are different types of information service, media content, e-trade transaction and access right handling servers.

THE EXCHANGE FROM A NETWORK PERSPECTIVE

The main purpose of a telecom network is to interconnect telephone users, devices and services. Since all users does not need to use the network all of the time, many network resources can be shared among several users.

A large exchange is a means to share switching resources by keeping many such resources in one common pool, thus providing statistical gains. A large exchange also will reduce network costs by enabling the sharing of network resources and reducing the number of lines, transmission capacity and external signaling needed if split into smaller exchange nodes or several specialized nodes.

The concept of an exchange can be related to the functions of a network node where many telecommunication lines are connected. However, the traditional functions associated with an exchange are these days sometimes distributed and located also to other node types in the network.

Functional and Structural Perspectives on Networks

A practical approach is to look at networks from the operator's perspective. From this point of view a telecommunications network can be divided into four major networks: an access network, a transport network, a signaling and control network, and a management network. In more recently developed networks one often complements this view with service and media networks.

However, for different purposes and reasons, one may also talk about other types of networks that are related to each other in different ways. A reason to do this is that one often needs to look at networks both from functional and from structural perspectives.

In abstract functional perspectives, one views the network from the external environment as if it were a black box and focuses mainly on the services that the network provides. Functionally, a network can be defined as a set of points at which a set of functions are provided and at which certain properties of each such function can be measured. A network function at one level may be implemented by using a function or a set of functions at the adjacent lower level (see Fig. 1).

In structural perspectives, one takes a closer, more internal, detailed and structural (white box) view and focuses on matters related to structurally and physically measurable properties, including how the network functions can be partitioned, distributed in space, and allocated to network nodes. Structurally, one can hence define a network as a set of network nodes, each with a set of allocated node functions, placed and interconnected via logical interconnection points in this way enabling the node functions to cooperate so as to implement the network functions (or network services) to be provided.

Using the functional perspective of networks, one can distinguish some network concepts related to telecommunication exchanges. For example: an access network provides access related functions at a set of network access points; a service network provides service (e.g., call) related functions at a set of service points; a mobile network provides location and mobility related functions at a set of location points; a signaling network provides signaling functions at a set of signaling points; a connection network provides connection or switch related functions at a set of connection points; a management network provides management functions at a set of management points; a transport network provides transport functions at a set of transport points; and so on a media network provides information functions at a set of media content provisioning points.

The functions at the network points mentioned above are functionally separated but can for economical or other reasons be colocated in the same physical node or site. The

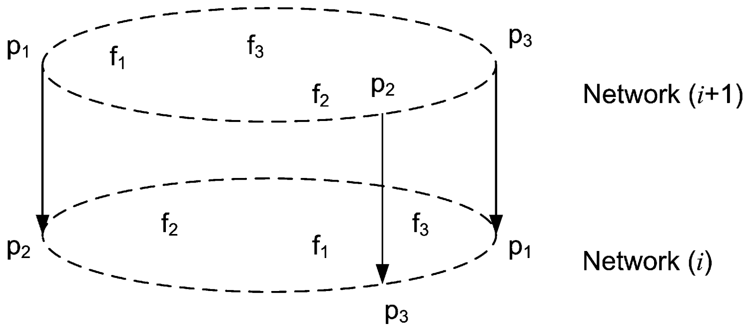

Figure 1. Network using functions from a lower-level network.

networks are related to and use each other in many ways. For example: the signaling network is used to enable functional inter-work between the other networks; a subscriber can, by sending signals via the access network, obtain access to functions provided by the service network that establishes these services using (by sending of signals to) the connection network(s); the transport network is used by all the other networks as a bearer for transport of raw data.

Purpose of Networks and Network Design

As discussed above, there are several types of networks seen from a functional point of view. A purpose of this partitioning or layering into several functional networks is to help us to build network architecture with well-defined functional areas and interfaces between these that will simplify extensions and modifications to be done. Another purpose is to support the building of cost-effective networks, both from a purchase and from a life-cycle point of view, with low operation and maintenance costs.

The end users of a telephone network have a more holistic view: They want to be able to contact each other so that they can talk when they are physically far away. Hence, the main purpose of a telephone network is to establish communication between people. In a mobile multimedia network this interest extends to having mobile access wherever the user is located also to media servers including ordinary databases and file systems as well as audio and video stream content.

End users of telephony not only want to talk, they also want to talk inexpensively and get a good quality of service. This requires efficient use of resources and a good network design. The costs to operate a network depend much on how the subscribers are distributed in the geographical area that the network is aimed to cover. Charging policy, traffic intensity, and traffic patterns such as length and locality of calls are other factors that all will influence the design of the network. Most networks are built in a more or less hierarchical structure to collect, concentrate, transmit, and connect the traffic such that higher-capacity transmission and switching equipment can be utilized. This reduces the costs for transmission and switching since many users can then use the equipment more efficiently and with less risk for blocking due to statistical multiplexing effects. The degree of concentration that can be utilized is influenced by the traffic intensity and the desired service quality level.

Another important cost factor is that the total physical length of the transmission path can be reduced in this way, and this is important since the cost for digging ditches and use of pipes is a large contributor to the overall network costs. Microwave links are used to reduce cost where it is too expensive to dig ditches. For high traffic routes, direct routes are often used, especially if the distances are short, such as in metropolitan areas. Alternative routes are also used to improve reliability, but these increase the cost and for that reason are usually not used in the access network.

Using the structural perspective of networks, one can for a specific area identify a distribution of subscribers that must be interconnected in a reasonably efficient way—for example, in a star- or ring-structured physical access network. One can then allocate a number of network nodes of different basic types that are needed, both to deal with expected maximum traffic loads and to interconnect these in a way that keeps transmission costs close to a minimum. Typical nodes in fixed and mobile telecommunication networks are illustrated in Fig. 2.

We can now go back to the external functional network perspective and hence to the different functions and examine how these are distributed in the physical network. This distribution can be guided by different principles. One principle that has to do with costs is that simple and cheap functions that are used often and for long periods of time should be located close to the subscribers. Complex expensive functions used seldom and for short periods of time should be placed more centrally in the network and thus be shared by many users. However, with today's technology and using mainstream components, complex must not mean expensive and hence a complex function if implemented as an integrated circuit can often be placed close to the subscribers, for example voice coders. Furthermore, the effects of distribution on reliability, signaling, and maintenance must also be considered when deciding on function distribution over network nodes.

From the above discussion one can see that there is no absolute definition of how networks shall be implemented or what a network node shall contain. Rather, there are several possible configurations that can fulfill the requirements, because not only does the distribution of subscribers as well as their quality of service (QoS) requirements and traffic patterns differ significantly but also implementation costs for different types of solutions differ and change over time.

Going back to and analyzing what an exchange is, one can see that many of the functions carried out in the different logical networks have traditionally been placed in an exchange node and can still be located in such a node. On the other hand, these functions and hence the exchange

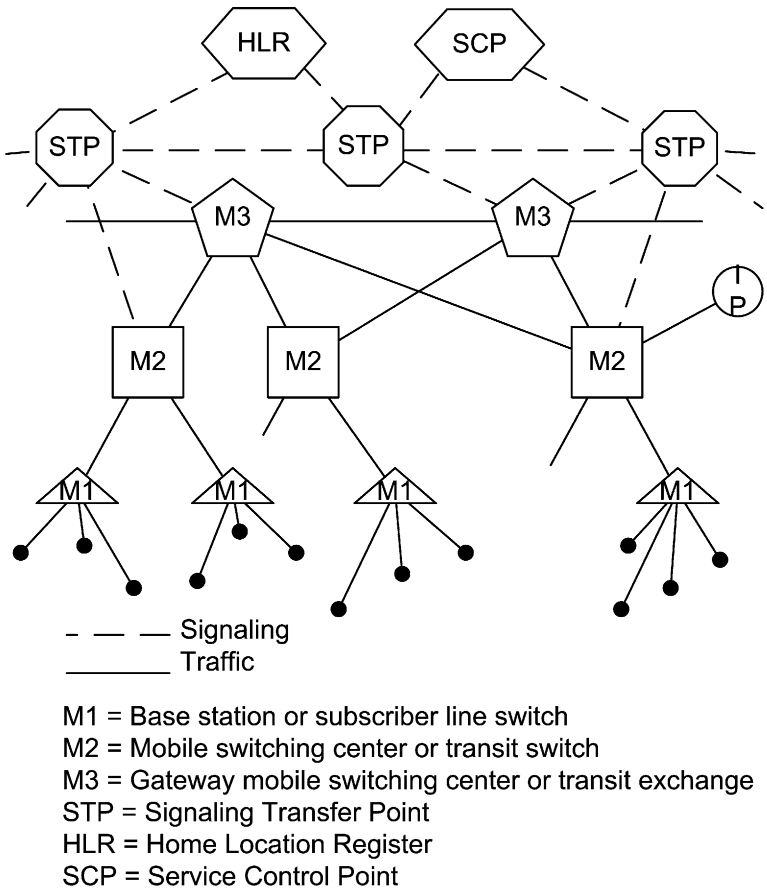

Figure 2. Typical nodes in a telecom network.

can also be more or less distributed and placed in more specialized networks and nodes. With well-defined functional areas and interfaces between these it is possible to configure networks and network nodes in many ways. However, for a node or cluster of nodes to be regarded as an exchange it must have some basic call access, control, switching and connection handling capability.

Routing and Switching Techniques

The task of finding a path from source to destination is called routing. The task to follow such a route from source to destination guided by end-to-end address information is called route selection or forwarding. Each node that participates in this task does not know the whole path but must be able to analyze the destination address and find out in which direction the path should go and also find such a path with a trunk that is free to use.

To forward information (a bit, a voice sample, a message, a cell, a packet or a frame) from a fixed or mobile input terminal (address, channel, line or trunk) to a selected output terminal (address, channel, line or trunk) of an exchange where the route selection or forwarding is controlled by local address information (based on the selected path) is called switching.

The telephone exchange will use network topology or routing information to prepare a connection. Hence, routing precedes the establishment of a full circuit switched connection and is done only once while switching is done for each voice sample (using traditional circuit switching) during a connection.

To guarantee that all resources needed to enable communication in real time are available, they can be reserved by setting up a circuit connection along a route between the circuit end points. Information (voice and data) to be exchanged between such end points in a network can then be sent along the circuit connection that is routed and set up before the actual information exchange starts and then released when the information exchange ends. This guarantees that all resources needed to enable voice and data communication in real time are reserved beforehand.

For historical reasons, circuit connection techniques are often associated with the synchronous transfer mode (STM) and pulse code modulation (PCM) using a 125 □s frame rate for transmission of byte (8-bit) encoded voice samples.

Another technique, called asynchronous transfer mode (ATM), uses small packets called cells with a given fixed size (53 bytes) divided into a small header (5 bytes) and payload (48 bytes). The header does in this case not contain the address but rather path and channel identifiers, where a path is a bundle of channels. Routing is done in a setup mode where the destination address is sent as payload in a cell (hence used to carry control information). The traffic can then be switched at the path or channel level cell by cell in each node along the paths and channels routed and set up in advance from sources to destinations.

A third technique usually used for data communication is to send information in packets of a reasonable size, e.g. as short Internet Protocol (IP) packets, add addressing information and other control information to the packet, and send it via a route toward its destination address. Each node on the way to the destination participates in the forwarding of the packet. For information that needs to be divided into several packets, forwarding is done for each packet in each node. To get deterministic packet transport delays circuit like reserved routes can be established using multi protocol label switching (MPLS) techniques.

Network Standards

The classic telecommunications network available all over the world is the public switched telephone network (PSTN). It is basically designed to allow the transmission of speech between two or more users and services related to that. However, this network is also used for facsimile traffic and data traffic via modems. Examples of services are alarm calls and abbreviated dialing; call forwarding and three-party calls.

Integrated services digital network (ISDN) is an evolution of PSTN that gives the subscribers access to integrated or combined services. ISDN integrates different telecommunication services into the same network that transports voice and data in digital form between network access points. The main advantage of the evolution from analog to digital end-to-end communication is safer and more flexible transfer of information. ISDN provides a wide range of services divided into bearer services and teleservices. ISDN is based on the digital telephony network using ordinary two-wire subscriber lines, 24-or 32-channel PCM link structures, and Signaling System No. 7. Integrated access implies that the user has access to both voice and non-voice services through a single subscriber line, whereas combined access implies the use of several subscriber lines. Services include voice, facsimile, and computer connections.

There are two types of user-network accesses defined by ITU-T:

- Basic Rate Access. A basic rate access is used for low traffic load. It normally includes one 16 kbps signaling channel (D) and two 64kbps communication channels (B).
- Primary Rate Access (T1/E1). The primary rate access handles higher traffic loads. It normally includes 23 or 30 communication channels (B) and one signaling channel (D).

The public land mobile network (PLMN) is used here as an acronym for a set of networks based on standards such as advanced mobile phone system (AMPS) and its digital version (DAMPS), Nordic mobile telephony (NMT), global system for mobile communication (GSM), and personal digital cellular (PDC) with the primary objective to provide communication to and from mobile subscribers connected to the fixed network via radio. The radio interface is implemented by the mobile terminals and the base stations. The base stations are end points in the fixed (wired) network.

Analog systems such as AMPS and NMT are sometimes called first generation mobile networks while DAMPS, GSM and PDC are referred to as second generation mobile networks, and are all based on cellular digital technology.

The third generation systems based on code division multiple access (CDMA) and wideband code division multiple access (WCDMA) IMT-2000 and UMTS are designed to support variable speed multimedia communication.

Typical network nodes in the switching part of the PLMN are:

- The mobile services switching center (MSC) controls the radio base stations and the calls within the PLMN and calls to and from other telephony and data communication networks such as PSTN and ISDN.
- The home location register (HLR) contains subscriber information such as which supplementary services are activated and information regarding in which MSC-area the subscriber is currently located.
- The visitor location register (VLR) is a database with information of the locations of the mobile stations in the area controlled by the MSC. The VLR also fetches information from the HLR so that the call setup can be performed without using the HLR each time.
- The media gateway (MGW) is a switching/routing/transferring node in the UMTS transport network, to facilitate communication between RNCs, between RNCs and the core network nodes, and between RNCs and O&M nodes.
- The base station controller (BSC), called radio network controller (RNC) in UMTS, coordinates and controls a number of radio resources usually located in base stations and some interwork functions such as handover between the cells, covered by the these base stations.
- The Radio Base Station (RBS) is responsible for radio transmission/reception in one or more cells to/from the User Equipment (UE).

The intelligent network (IN) is an architecture aimed at making a telecom network work as one uniform system where new network services can be easily developed, introduced, and made available in the network from a central service control point (SCP). IN aims at a logical separation of signaling, call, connection, and transport. The idea is that local and transit exchanges use number analysis and ask the SCP to handle the call, unless it is a simple IN service that can be handled locally. The SCP then executes a corresponding service script that results in orders to the exchange regarding how to proceed with the call. For this communication the intelligent network application protocol (INAP) is used. New services are described as service scripts or building blocks consisting of functional components and developed by the operator in a service creation environment.

The telecommunication management network (TMN) is an architecture and standard portfolio in the area of operation and maintenance. It defines functional areas and protocols for management in general terms and also more specific information models for how to monitor and manage these areas.

The open systems interconnection (OSI) reference model is a standardized layered model of how computer systems can be interconnected and interoperate that has had an influence on the way signaling networks and protocols are viewed and built. The model defines seven layers: application, presentation, session, transport, network, link, and physical.

The telecommunication information networking architecture (TINA) is an international collaboration for defining an open architecture for telecommunication systems. It focuses on the software architecture. To some extent this effort can be seen as an attempt to put together some other standardization efforts such as OSI, IN, and TMN from a software architecture point of view.

THE FUNCTIONS OF AN EXCHANGE

The telecommunications exchange is a multi application digital switching product that offers its main services to its subscribers but also services to the operator of the exchange. One example of a service offered to the subscribers is telephony calls, and a service offered to the operator is the ability to charge for such services by registering of charging data in the exchange.

A telecommunications network offers various services to the users and the operator. ITU-T has divided these services into two main categories:

- A bearer service for transport of speech or data in the network between the user interfaces. The transport of speech should be done in real time and without distortion or alteration. The function of the bearer service corresponds to the OSI levels 1–3 for transport, routing, and safeguarding of the information through the network.
- A teleservice is a complete communication service that combines the information transfer of the bearer service with terminal services, such as information processing functions. A teleservice corresponds to the OSI levels 1–7. Some teleservices are tied to a special bearer service, whereas others can utilize different bearer services. Examples of teleservices are telephony, facsimile, and computer connection.

The bearer services and the teleservices are divided into basic and supplementary services. Telephony is an example of a basic teleservice, and call waiting is an example of a supplementary service that gives users additional functionality. In general, the supplementary services provide additional capabilities that rely on basic services to be used.

Examples of teleservices are:

- Telephony. The normal two-way voice communication between two users is the most fundamental service.
- Facsimile. This teleservice allows the connection of facsimile machines.
- Voice Mail. This service offers the subscribers the possibility to forward calls to a central location in the network. The subscriber can later check the system for unanswered calls and listen to recorded voice messages.

Basic Telephony

Below is a brief description of the main functions required by the exchange in order to set up, maintain, and disconnect a basic telephone call between two mobile or fixed subscribers.

Subscriber Signaling. In order to set up a call, the calling subscriber alerts the exchange that there is a new call attempt and then sends the dialed number. For a digital (mobile or ISDN) access, the alert and the digits are all sent in one message, in order to save bandwidth resources and decrease the delay for call setup (see Fig. 3). For an analog access, the alert is made by lifting the handset, and the exchange replies by a dial tone. The dialed numbers are then sent one by one. In both cases, the exchange replies with a tone to the calling subscriber when the status of the called subscriber has been checked.

Number Analysis. The A-number (the calling subscriber) and the B-number (the called subscriber) are analyzed. The result is then used as input to charging and routing analysis. For analog subscribers, each digit in the B-number arrives as the subscriber dials, whereas all digits can be sent at once for digital mobile or fixed subscribers.

Subscriber Category and Service Analysis. The exchange must check whether the calling subscriber has any particular service invoked. Some of the services must be analyzed early in the call setup, such as blocking of outgoing calls.

The services implemented in software (programs and data) are executed either in the local exchanges used by the calling and called subscribers, in other cases in a transit exchange or in a separate exchange that handles intelligent network (IN) services, a service control point (SCP). In the latter case, the local exchange will need to check whether to request the service from the SCP or not.

Charging. There are two classic charging methods, pulse metering and detailed billing. Detailed billing, also called toll ticketing, enables an operator to specify the characteristics of each call very extensively.

Charging can be divided into two steps: analysis and output. The result of the analysis is the charging method (toll ticketing, pulse metering or flat rate) and the charging rate, depending on a number of call data set by the operator. The output includes formatting of the charging data along with output on a reliable storage medium, locally or in a charging and maintenance center.

Routing Analysis. Finding a path from source to destination is called routing and is made mainly on the B-number. Usually there are two or three (and sometimes up to ten) alternative routes to select among. The selection of a route (also called trunk group) is guided by priority and load status information. If the first route fails or is overloaded, the next alternative is selected. There are more sophisticated

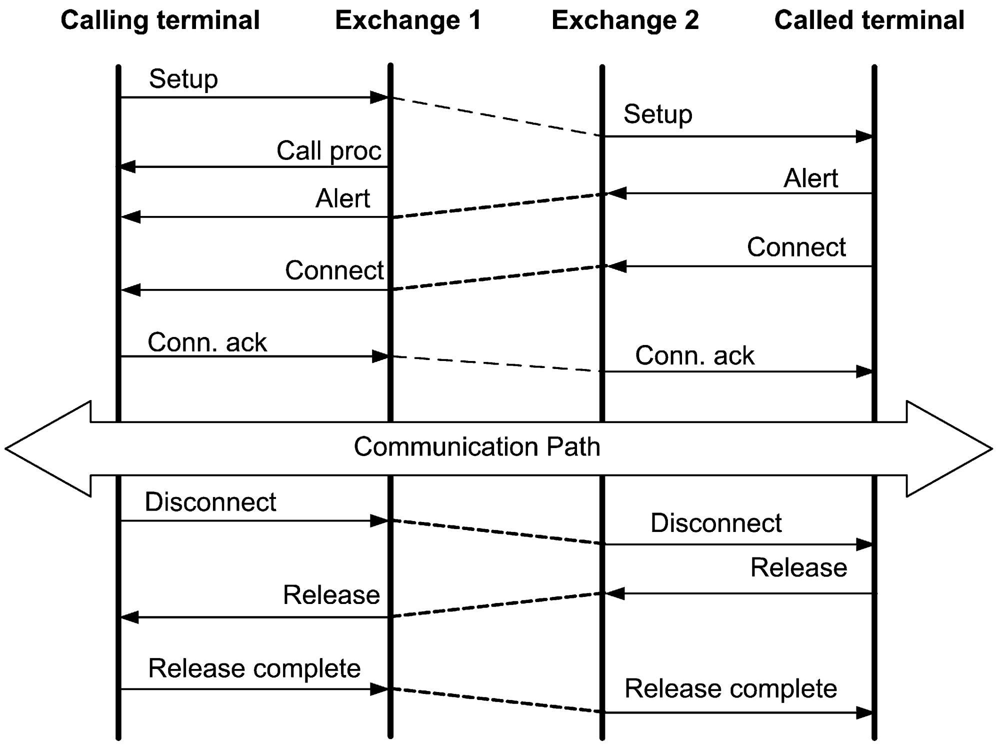

Figure 3. Subscriber signaling, ISDN.

routing algorithms that dynamically choose a link in order to minimize the congestion in the network. These dynamic routing algorithms can be either local or central; the local provides results by using data available in its own exchange such as previous success rates on different link choices, while the central algorithms collect input data from other exchanges in the network.

Connection. The connection is the through-connection of two normally 64 kb/s circuits, one in each direction, in the hardware devices, and particularly in the switch fabric. The connection is required to be with limited probability of blocking, from end-to-end. This means that the switch fabric must add very low blocking probabilities, in order to fulfill the end-to-end requirements for calls that pass several transit exchanges. The connection also must be well synchronized with the rest of the exchange and with the rest of the network, in order to handle the digital speech connections properly.

Trunk Signaling. Trunk signaling enables a call to be connected between subscribers in separate exchanges. The basic data in all trunk signaling systems are alert messages that a call is to be connected or disconnected, along with routing information, mainly the relevant parts of the dialed digits. Modern signaling systems can transmit all types of data—for instance, in order to support detailed billing, advanced network services, and transparent user data.

Early signaling was made on the same line where the speech was transmitted, first by decadic pulses and later by tones of different frequencies. A still common such in-band signaling system is multi-frequency signaling, where a combination of two tones is sent to a tone-receiver in the exchange. Modern signaling is based on digital message passing. The globally dominant signaling system is the Signaling System No. 7.

Subscriber Services

When an exchange receives digital data and is computer-controlled, almost any communication service can be performed, and a large number have also evolved. The service software is located in the terminals, in the ordinary exchanges, service control points and network databases. The most cost-efficient location depends on the type of service. For some services where the logic is local (such as abbreviated number, also called speed dialing), it is most efficient to store the translation between abbreviated number and real number in the calling terminal. For more intelligent network services such as virtual private networks, freephones, or universal personal numbers, it is preferable to locate the service logic in a network node, in either (1) the ordinary exchanges such as the mobile switching center or local exchange or (2) a network database such as the SCP or HLR. The advantage of central service control is that the introduction of new services and features is simplified. In addition, some features require consistent data for the entire network, such as the information in the HLR regarding where a called mobile subscriber is located.

More powerful protocols enable more advanced services to be implemented in the network. At the same time, there are an increased number of services that are implemented in the terminals, and related data is sent transparently through the network between the end users. As an example, ISDN has spread slowly while data traffic over the telecom network has increased much more rapidly, where the services are executed in the end-users' computers.

A few common telephony subscriber services, implemented in an exchange, are as follows:

- Freephone. The call is free of charge and paid by the called party. Often the freephone service can be directed to various physical subscribers depending on time, date, and traffic.
- Conference Call. More than two parties take part in a call.
- Transfer Services. The call is transferred to another telephone immediately or when the called number is busy or not replying.
- Universal Personal Number. One phone number is used regardless of which mobile or fixed physical connection a person is using.
- Call Completion Services. When the called party is no longer busy or nonreplying, the call is reinitiated.
- Virtual Private Network. A group of subscribers, for instance a corporation, form a private network with their own charging and telephone numbers.

These services are also effective over a network, not only when both parties reside in one exchange.

Cellular Mobile Telephony

Cellular mobile telephony differs from basic telephony since the subscribers can move freely within areas covered by the radio access network. As a result, the exchanges in a cellular system must keep track of where the subscribers are located and find free radio channels to use for new calls and during calls since these are shared among all subscribers in an area.

The radio frequency spectrum made available for mobile telephony is a scarce resource that is reused by dividing the space in small areas called cells. Usually the frequency spectrum is also divided into frequency bands, and these bands can in turn be time or code divided into channels. Cells that are not too close to each other can, due to the use of limited power levels, share frequency bands and channels without disturbing each other.

Handover. Handover means to change or switch connection from one cell to another with better radio transmission quality during an ongoing call. The handover decision is based on measurements of received signal quality in up and down links. Handover can be made several times during a call. This and the fact that handover decisions require the collection and analysis of measurement data contribute to making cellular mobile telephony quite processing resource consuming compared to fixed telephony.

Channel Allocation. Channel allocation, aims at finding free frequency, time and/or code divided channels within a cell and then allocating such free channels to calls. The allocation logic gives an ongoing call higher priority than a new call. Channel allocation is closely related to the handover function and especially intercell handover, which does handover between channels in the same cell.

Location Update. When the mobile phone is turned on and running in idle mode, it listens to control messages indicating which area the closest base station sending control messages belongs to. When a border is passed it sends a message to the mobile switching center (MSC), indicating that the subscriber has moved to another location. The information is stored in the MSC, in a visitor location register (VLR), and it is also stored in the home location register (HLR) if the area is handled by a new MSC.

Paging. Locating a mobile subscriber within the network is called paging. This is done by requesting the HLR in which VLR/MSC and location area where the subscriber is located, and then sending a paging message on the page channel to the cells in that area.

Roaming. When subscribers move to another operator's network than their own, the network can page the subscriber and then set up a call to their new location.

Operation and Maintenance

A modern telecommunications exchange should offer operation and maintenance functions that guarantee a high quality of service to the operators and the subscribers. Operation is the normal everyday running of the exchange. This includes activities to adapt the exchange to continuously changing demands. Examples of operational activities are:

- Connection and disconnection of subscribers
- Change of subscriber data
- Collection of charging data
- Collection of statistics

Maintenance is the prevention, detection, localization, and correction of faults. The faults can be detected automatically by the exchange or reported to the operator by the subscribers or other exchanges in the network. Examples of maintenance activities are:

- Fault detection, testing, and repair of exchange hardware, for example, trunk lines or subscriber lines
- Fault detection, auditing, recovery and correction of exchange software and data
- Checking of disturbance indicators in various parts of the exchange such as the power system or the control system

In a modern telecommunications exchange there are several approaches to maintenance. One is preventive maintenance that involves a set of routine tasks to check for faults before they occur and requires a high level of effort to achieve a certain grade of service. Another is corrective maintenance where faults are dealt with as they occur. This requires a more selective and limited effort but may result in a less consistent quality of service. The best is to have a balance between the preventive and corrective maintenance.

Statistics are used to supervise traffic and performance in the network and to reconfigure an exchange to handle more or less traffic. Particularly for the configuration of location areas and cells within a mobile network, large amounts of traffic data are used as facts to support corrective actions.

All operation and maintenance activities should have minimal impact on the traffic handling of the exchange. The ideal situation is an exchange where every subscriber can make a call at any time no matter what happens to the system. In order to achieve this, the system should be robust to operator errors and allow the performance of maintenance and software and hardware upgrades without affecting the execution of traffic events.

An exchange should be able to be operated and maintained remotely. The ability to access exchanges remotely using an operation and support system (OSS) ensures a high level of service to subscribers and low costs for the operator. A centralized operation also contributes to a more reliable operation of the exchange and a reduction in personnel.

ARCHITECTURE OF COMPUTER CONTROLLED EXCHANGES

The architecture of computer controlled exchanges is influenced to a large extent by the architecture and technology of both the switching system and the (central) control system and how they are related via decentralized (regional) control systems (see Fig. 4).

A modular architecture can lower the costs of system handling and make it easier to adapt the system to the changing world of telecommunications. In a truly modular system each module is fully decoupled and independent of the internal structure of other modules. There are different forms of modularity, for example:

- **Application modularity**. Make it easier to combine several larger applications in one node.
- **Functional modularity**. The system defined in terms of functions rather than implementation units. Functions should be possible to add, delete, and change without disturbing the operation of the system.
- **Software modularity**. The software modules should be programmed independently of each other, and they should interact only through defined interfaces and protocols. In this way, new or changed modules can be added without changing existing software.
- **Hardware modularity**. Supports that new hardware can be added or changed without affecting other parts of the exchange.

On the highest level the system architecture of the exchange can be divided into various application modules in analogy to how telecommunications nodes interact and communicate, using protocols enabling modules to be added or changed without affecting the other modules. Typically the implementation of a telecommunications exchange can be divided into:

- **Application modules**. These implement various telecommunication applications much like virtual nodes using standardized interfaces to other application modules. Application modules act as clients to resource modules.
- **Resource modules**. These modules coordinate the use of common resources available to applications by means of well-defined interfaces to the users. Resource modules act as servers to application modules.
- **Control modules**. These modules are responsible for the operating system functions, input-out-put functions, basic call service functions, and so on.

Application Modules

The application modules implement various telecommunication applications and have standardized interfaces to resource modules. In general an application consists of access and services. Examples of application modules are:

- Analog access module
- Digital access module
- Mobile access module
- PSTN user services module
- ISDN user services module
- MSC user service module
- Home location register (HLR) module

Resource Modules

The resource modules typically handle and coordinate the use of common resources and may contain both software and hardware. The most important part is the group switch. Trunks and remote and central subscriber switches (RSS and CSS respectively) are connected to the group switch. The trunks are used to connect the switch to other switches, to data networks, mobile base stations, etc. The subscriber switch handles the subscriber calls and concentrates the traffic (see Fig. 5).

Group Switch. The main function of the group switch is selection, connection, and disconnection of concentrated speech or signal paths. The group switch often has a general structure.

The overall control of the group switch is performed by the central processor system. The regional processors take care of simpler and more routine tasks, such as periodic scanning of the hardware, whereas the central control system handles the more complex functions. Associated functions included in the group switching resource module are network synchronization devices and devices to create multiparty calls.

Subscriber Switch. The subscriber switch handles selection and concentration of the subscriber lines, its main functions are as follows:

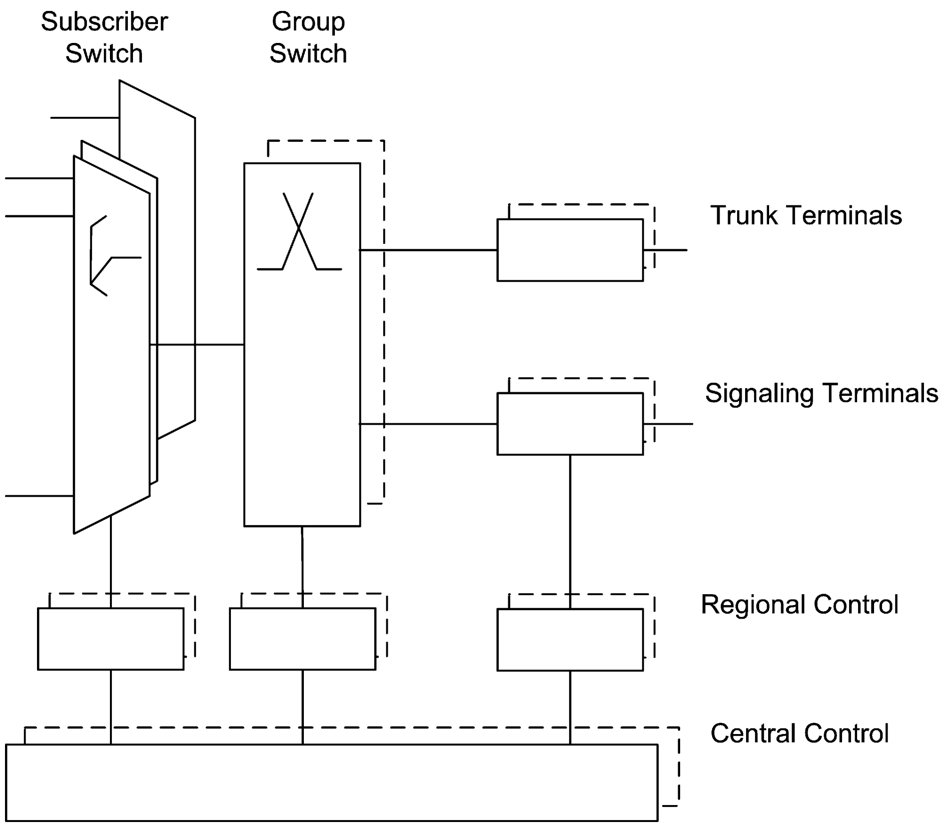

Figure 4. Typical architecture of stored program controlled exchange.

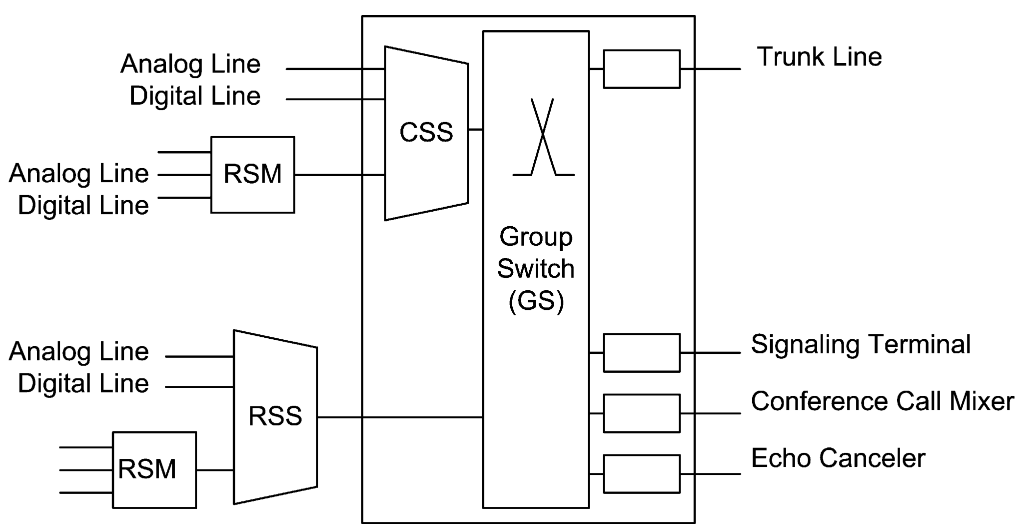

Figure 5. Switch architecture.

- Transmission and reception of speech and signaling data to and from the subscriber equipment (for example, on-and off-hook detection).
- Multiplexing and concentration of the subscriber lines, to save hardware and make more efficient use of the communication links between the subscriber stage and the group switch.

The architecture should be modular and enable to combine PSTN and ISDN access in the subscriber stage. The subscriber switch can be colocated with the group switch in the exchange (central subscriber switch, CSS) or located at a distance from the exchange (remote subscriber switch, RSS).

Remote Subscriber Multiplexer. The remote subscriber multiplexer (RSM) is an add-on subscriber access node, used in the access network, which can cater small groups of subscribers. It provides both mobile and standard telephony connections. The RSM multiplexes and concentrates the traffic to the central or remote subscriber switch but does not carry out traffic switching functions.

Trunk and Signaling. This resource module includes the circuits for connecting trunks and signaling devices to the group switch. The module should handle the adaptation to different signaling systems, namely common channel signaling as well as various register and line signaling systems.

Traffic Control. This resource module contains the traffic handling and the traffic control functions of the exchange. This module is responsible for finding the most suitable route between calling and called subscribers and of verifying that call establishment is allowed.

Operation and Maintenance. This resource module enables tasks such as supervision of traffic, testing of the transmission accessibility and quality, and diagnostics and fault localization of devices or trunks.

Common Channel Signaling. This resource module includes the signaling terminals and the message transfer part (MTP) functions for common channel signaling systems such as SS7.

Charging. This resource module is used in exchanges that act as charging points. Both pulse metering and specified billing (toll ticketing) can be offered. It should be possible to charge both calls and services, and the charging should be based on:

- Usage
- Provision/withdrawal of subscriber services and supplementary services
- Activation/deactivation of subscriber services and supplementary services

Control Modules

The primary function of the control module is to provide the real-time processing and execution environment required to execute software in application modules and resource modules used to perform traffic-handling functions and call services. The processing can be centralized where one processor takes care of all tasks, or distributed where the processing of information is distributed over several processors.

Execution of telecom software imposes stringent real-time requirements on the control system. Calls appear stochastically, short response times are needed, and overload situations must be handled. The main control modules are: the central processor(s); the data store to store call data; and the program store to store the actual programs.

In order to achieve an efficient overall control system, it can be divided into:

- **Central control.** One or several processors that perform the non-routine, complex program control and data-handling tasks such as execution of subscriber services, collection of statistics and charging data, and updating exchange data and exchange configuration.
- **Regional control.** A set of distributed processors that perform routine, simple and repetitive tasks but also some protocol handling. They are of different types optimized for their main tasks, for example, input/output processing. They often have strict real-time and throughput requirements, for instance for protocol handling, and may have customized hardware support in the form of ASIC, FPGA and DSP circuits for this purpose.

Switching Techniques

Until about 1970 most switches were analog and based on electromechanical rotor switches and crossbar switches. Since then the utilization of digital techniques has become dominant. Digital switches have been based on synchronous time and space multiplexed circuit switch technology. For pure data communication, packet switching technology is often used. In order to be able to support voice and data sharing a common infrastructure, an asynchronous transfer mode (ATM) has been developed. Synchronous time and space multiplexed circuit switch technology is based on synchronous transfer mode (STM) carrier and synchronization technology (STM transport technique is also often used as a carrier for ATM).

A typical STM-based switch architecture is made up of a combination of time and space switches (T and S, respectively). A space switch connects physical lines by changing of positions in the space domain, see Fig 6a and a time switch changes the ordering sequence of data (voice) samples by changing of positions in the time domain, illustrated in Fig. 6b.

The elements T and S can be combined in several ways to realize a switch and make it configurable in many ways. Usually time switching is used in input and output stages and space switching is often used in central parts of a switch. This basic time-space-time (TST) switch structure can be used both in subscriber and in group switching stages. The first part of a TST switch is a time switch, which interchanges time slots between the external incoming digital paths and the space switch. The space switch connects the time switches at the input and the output. The last part of the TST switch is a time switch, which connects the time slots between the external outgoing digital paths and the space switch (see Fig. 7).

The time switch moves data contained in each time slot from an incoming bit stream to an outgoing bit stream but with a different time slot sequence. To accomplish this, the time slot needs to be stored in memory (write) and read out of the data store (DS) memory and be placed in a new position (read). The reads and the writes need to be controlled, and the control information needs to be stored in a control store (CS) memory as well. The timing of DS and CS is controlled by a time controller (TC). Examples of control actions are time slot busy or time slot idle.

A typical space switch consists of a cross-point matrix made up of logic gates that realize the switching of time slots in space. The matrix can be divided into a number of inputs and a number of outputs and is synchronized with the time switching stages via a common clock and a control store.

Switch architectures based on asynchronous transfer mode (ATM) handle small packets called cells with a given

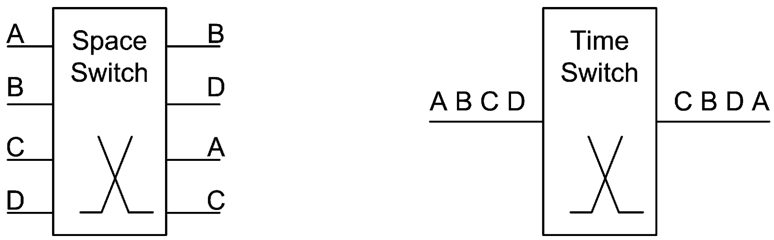

Figure 6. Examples of switching in space (a) and time (b).

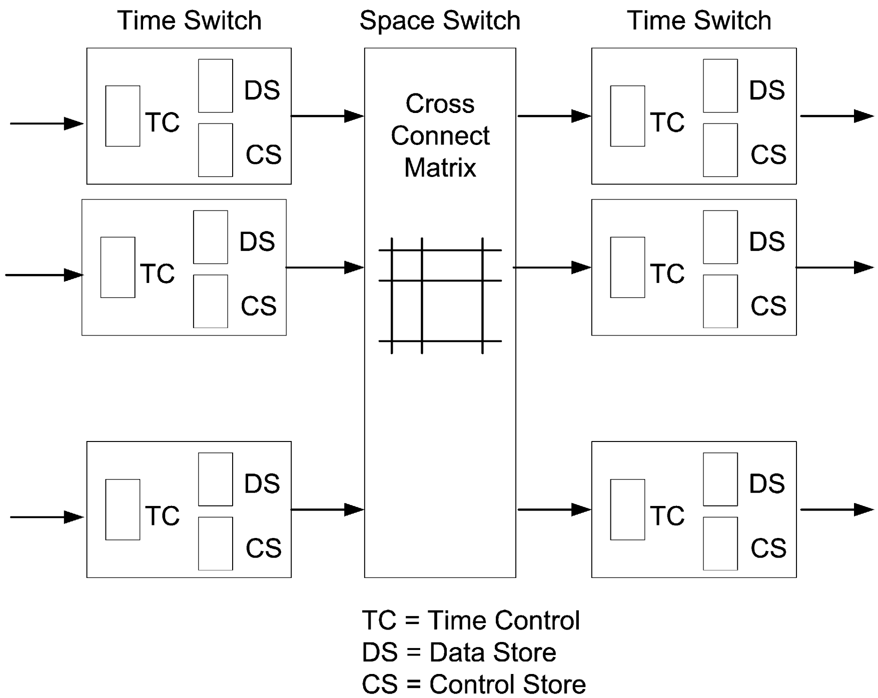

Figure 7. Typical STM switch with TST structure.

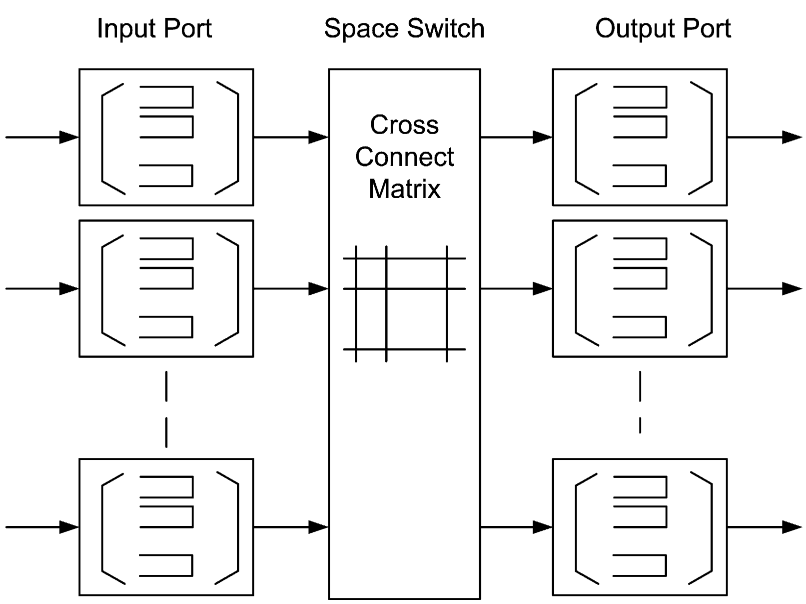

Figure 8. Typical ATM switch structure with input and output ports and control space switch.

fixed size (53 bytes) divided into header (5 bytes) and payload (48 bytes). The header contains virtual path identifiers (VPIs) and virtual channel identifiers (VCIs), where a virtual path (VP) is a bundle of virtual channels (VC). Traffic can be switched at the VP or VC level cell by cell. Associated with a VP or a VC is a quality of service (QoS) contract. In order to be able to guarantee the switching of cells according to the contract without unacceptable cell loss a number of queues are used at the input ports and output ports of the switch. Between input and output ports (including buffer queues and cell multiplexers and demultiplexers) a space matrix often is used, as in Fig. 8. However, other types of central switching structures sometimes are used, such as fast fiber buses and Banyan networks.

Other switching techniques are used in a telecommunication exchange in addition to STM and ATM. These include Ethernet and Rapid IO switches as well as IP routers, each with their own characteristics. RapidIO is highly efficient internally in nodes with hard latency and real-time requirements and for small packets, while Ethernet switches show good enough performance for high throughput for large packets. Embedded IP routers, for IP forwarding, are useful for nodes that border IP networks or are part of these. Sometimes, several switches can be needed inside a node.

CHARACTERISTICS REQUIREMENTS ON THE EXCHANGE

Availability

High availability of the telecom network and associated services is the single most important operator and subscriber requirement. Normal requirements on maximum unavailability are in the order of one or a few minutes of subscriber unavailability per year. This includes downtime due to faults in the exchange and in the transmission equipment and software, but also unavailability due to planned software upgrades and often also accidents outside the control of the vendor, such as fires, damaged transmission cables, and incorrect operation of the exchange. Several methods are used to increase the availability of the exchange to the subscriber:

- Redundancy, including fault-tolerant processors
- Segmentation
- Diagnostics of hardware and software faults
- Recovery after failure
- Handling of overload
- Disturbance-free upgrades and corrections
- Robustness to operator errors

Each of these is treated briefly below.

Redundancy. In order to cope with hardware faults, redundant hardware is used for those parts of the switch that are critical for traffic execution. Requirements are in the order of 1000 years for the mean time between system failures (MTBSF).

Specifically, current technology requires that a redundant processor is available in synchronized execution (hot standby), ready for a transparent takeover if a single hardware fault occurs in one processor. An intelligent fault analysis algorithm is used to decide which processor is faulty. In a multiprocessor system, n + 1 redundancy is normally used, where each processor can be made of single, double, or triple hardware. When one processor fails, its tasks are moved to the idle (cool standby) processor. A similar redundancy method is based on load sharing, where the tasks of the failed processor are taken over by several of the other processors that are not overloaded themselves.

The group switch hardware is also normally duplicated or triplicated, because it is so vital to the exchange functions. The less central hardware devices, such as trunk devices, voice machines, transceivers, signal terminals, and code receivers, are normally pooled, so that a faulty device is blocked from use and all users can instead access the remaining devices until the faulty device is repaired.

Segmentation. To avoid system failure a fault must be kept isolated within a small area of the exchange. This is done by segmentation of hardware, with error supervision at the interfaces. In software, the segmentation is made by partitioning of and restricted access to data structures; only owners of data can change the data, where the owner can be a call process or a function.

Diagnostics of Faults. After the occurrence of a fault in hardware or software, the fault must be identified and localized, its effect restricted, and the exchange moved back to its normal state of execution. For this to work, the diagnostics must be extensive and automatic. The exchange must be able to identify the faulty software and hardware and must be able to issue an alarm, usually to a remotely located operator.

Recovery After Failure. After a fault has been detected, the effect should be restricted to only the individual call or process (for instance, an operator procedure or a location update by a mobile subscriber) or an individual hardware device. This call or process is aborted, while the rest of the exchange is not affected. The recovery must be automatic and secure. In a small fraction of events, the fault remains after the low-level recovery, or the initial fault is considered too severe by the fault handling software, so that a more powerful recovery procedure must be used. The process abort can be escalated to temporary blocking of hardware devices or software applications and, if required, result in the restart of an entire processor or a number of processors. If the restart fails in recovering the exchange into normal traffic handling, new data and software are loaded from internal or external memory.

Handling of Overload. The exchange is required to execute traffic literally nonstop and when offered more traffic than it can handle, the rejection of overflow traffic should be made gracefully. ITU requires, that an exchange that is offered 150% of what it was designed for should still have 90% of its maximum traffic handling capacity. The exchange must also be able to function without failure during extreme traffic loads. Such extreme loads can be both short peaks lasting a few milliseconds or sustained overload due to failures in other parts of the network. Overload handling is accomplished by rejecting excess traffic very early in the call setup, before it has used too much processor time or any of the scarce resources in the switching path. Figure 9 shows the overload performance with and without an overload control function.

Disturbance-Free Upgrades. Both upgrading of software packages and the replacement or extension of hardware to

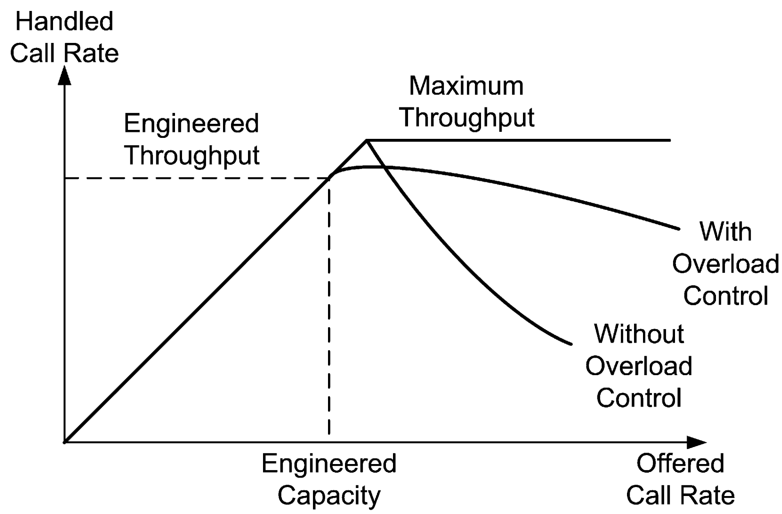

Figure 9. Overload control.

the exchange must not disturb ongoing traffic execution. This should be true both for fault corrections and when new functions are introduced. The architecture must thus allow traffic to be executed in redundant parts, while some parts of the exchange are upgraded.

Robustness to Operator Errors. Security against unauthorized access is accomplished by use of passwords and physically locked exchange premises. The user interface part of the exchange can supervise that valid operation instructions are followed for operation and maintenance, and it can issue an alert or prohibit other procedures. Logging of operator procedures and functions for undoing a procedure can be used. If a board is incorrectly removed from the exchange, the exchange should restrict the fault to that particular board, use redundant hardware to minimize the effect of the fault, and then indicate that this board is unavailable. A simple user interface with on-line support makes operator errors less probable.

Grade of Service. The real-time delays in the exchange must be restricted to transmit speech correctly. Packet-switched connections have problems achieving good real-time speech quality for this reason, especially during heavy usage. Circuit-switched networks have so far given the best real-time performance regarding delays and grade of service for voice and video, compared with packet data networks. ATM switching (due to its short and fixed size cell/packets) also fulfills the grade of service requirements and also defines service classes for different requirements.

Scalability

There is a need for scalable exchanges regarding capacity, from the very small (such as base stations and local exchanges in desolate areas) to the very large, mainly the hubs (transit exchanges) and MSC:s of the networks and exchanges in metropolitan areas. Furthermore, there is sometimes a requirement for downward scalability regarding physical size and power consumption, particularly for indoor or inner city mobile telephony.

The following are the common system limits for downward scalability of an exchange:

- The cost to manufacture and to operate in service will be too high per subscriber or line for small configurations.
- The physical size is limited by the hardware technology and by the requirements for robustness to the environment, and what is cost efficient to handle.
- The power consumption is limited by the hardware technology chosen.

The following are the common system limits for upward scalability of an exchange, each treated briefly below:

- (Dynamic) real-time capacity
- (Static) traffic handling capacity
- Grade of service (delays)
- Memory limits
- Data transfer capacity
- Dependability risks

Processing Capacity. New more advanced services and techniques requires more processing capacity, this trend has been valid for (a) the replacement of analog technology with digital processor controlled functions and (b) the development of signaling systems from decadic and multifrequency to packet mode digital signaling, including trunk signaling protocols such as ISDN user part (ISUP) together with mobile application part (MAP) and transaction capability application part (TCAP) and the development of mobile telephony. In the charging area, the trend from pulse metering to detailed billing affects the call capacity.

The number of calls per subscriber has also increased due to lower call costs from deregulation and due to the use of subscriber redirection and answering services.

Traffic Capacity. It is very important to design and configure hardware and software correctly, in order to minimize hardware costs, and at the same time ensure sufficiently low congestion in the exchange and in the network. Normally, the switch fabric is virtually non-blocking, and the congestion occurs in other resources, such as the access lines, trunk lines, and other equipment. The relation between congestion probability and the amount of traffic is well known, if all devices are accessible and the traffic follows a Poisson process; that is, the times between offered calls are independent and exponentially distributed. In some cases, the probability can be calculated explicitly. In more complex device configurations with non-Poisson traffic, the congestion probabilities are most easily calculated by simulation techniques.

Memory. The amount of memory per subscriber line or trunk line is another way to measure the complexity of a telecom application. The trend in this area is similar to that of processing capacity, and the same factors are responsible for the large increase in memory needs. Due to the real-time requirements, fast memory is used extensively, and secondary memory is only used for storage of backups, log files and other data where time is not critical.

Transfer Capacity. A third part of the switching system capacity is the data transfer from the exchange to other nodes, for example, to other exchanges and network databases, billing centers, statistical post processing, and nodes for centralized operation. There has been a growing demand for signaling link capacity due to large STPs, for transfer capacity from the exchange to a billing center due to detailed billing and large amounts of real-time statistics, and to transfer capacity into the exchange due to the increased amount of memory to reload at exchange failure.

Dependability Risks. Although dependability has increased for digital exchanges, there is a limit as to how large the nodes in the network can be built. First, the more hardware and the more software functions assembled in one exchange, the more errors there are. The vast majority of these faults will be handled by low-level recovery, transparent to the telecom function, or only affect one process. However, a small fraction of the faults can result in a major outage that affects the entire exchange during some time.

As an example, assume that the risk of a one-hour complete exchange failure during a year for one exchange is 1%. If we add the functionality of an SCP to an HLR node, then we more than double the amount of software, and presumably the number of faults, in the node. The risk of a major outage should be larger in a new exchange introducing new software with new faults. Only if unavailability due to completely stopped traffic execution is much less than the total effect of process abortions and blocked hardware devices can we build exchanges of unlimited software complexity.

The second reason for a limited complexity from a dependability point of view is that network redundancy is required and only can be used if there are several transit exchanges in the network.

Life-Cycle Cost

Since the 1980s, the operating cost has become larger than the investment cost of an exchange. Thus, the emphasis on efficient operation and maintenance has increased, regarding both ease of use and utilization of centers that remotely operate a number of exchanges that are not staffed. For ease of use, the telecommunication management network (TMN) was an attempt by ITU to standardize the operator interface. After several years this standard is still not used much. Instead, the operator interface is to a large extent dependent on the exchange manufacturer as well as the requirements from the telecom operator company. Several open and proprietary user interfaces are common.

For central operation, more robust methods of remote activities have evolved. Software upgrades and corrections, alarm supervision and handling, collection of statistics and charging data, and handling and definition of subscriber data are all made remotely. Transmission uses a multitude of techniques and protocols. Open standard protocols have taken over from proprietary protocols.

In addition, important parts of the life-cycle cost are (a) product handling for ordering and installation and (b) spare part supply.

EVOLUTION TRENDS

Technology Trends

Due to the effects of semiconductor process scaling, improved chip fabrication yield, and increasing numbers of connectivity layers – the storage capability of memory and the execution speed of processors has doubled every 18 months during the last 40 years, following the so called Moore's law. This exponential growth of transistors-per-chip will continue, but will force new hardware architectures such as chip multiprocessors and systems on chip in order to keep energy use within reasonable limits. The development of optical fibers, including the introduction of wavelength multiplexing, is perhaps even faster. Thus there are many factors that lead to cheaper nodes and higher bit rates in the network. At the same time, digital coding and compression techniques have been improved that makes it possible to transmit voice with traditional telecom quality using only a fraction of the bandwidth that is used today. These developments are changing the design of both nodes and networks.

It is also important to provide increased interoperability between network standards since the end users do not want to be concerned about where a person is physically located or to which network the person is connected. The introduction of universal personal numbers can solve this and lead to a convergence of fixed and mobile telephony. The possibility of accessing short message services (SMS), fax, and e-mail via fixed and mobile devices and so-called Internet telephony are examples of services that illustrate the needs for interoperability and convergence of telecommunication and datacommunication.

High bit rates at a low price, combined with the demand for real-time multimedia services, indicates that the network must either become more flexible or consist of sev-

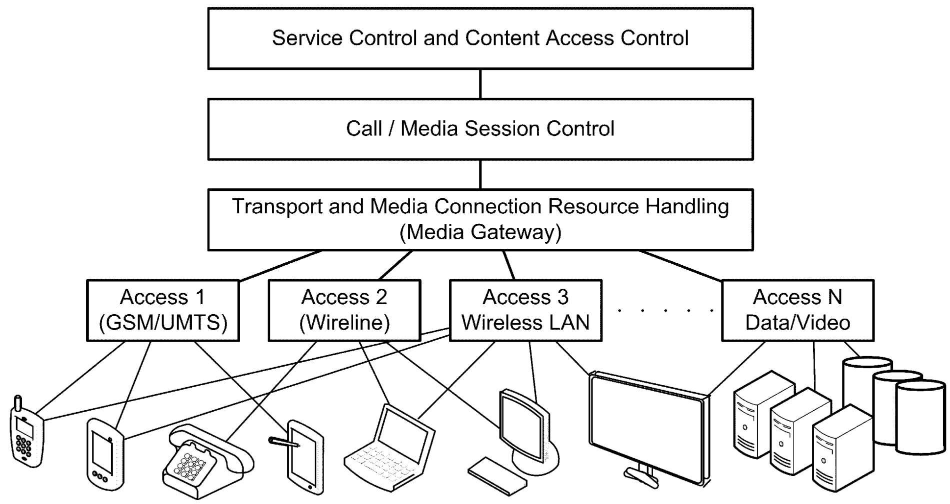

Figure 10. Converged network architecture with decoupling of access, transport, control and service functions.

eral different but interoperating networks. Packet, cell, and new, more flexible circuit switching techniques supported by new signaling protocols (such as MPLS) can solve these needs. In order to integrate service execution, control and connectivity horizontally across multiple access networks, a layered architecture approach can be used using a common transport layer based on IP and Ethernet technology over fiber rather than delivering single services such as voice telephony or data access (vertically integrated networks). This architecture supports efficient IP packet based transport of both signaling and payload, which is not possible with classic switches. By doing this one introduces a single IP infrastructure that can handle all network services, such as fixed and mobile communications.

Network Convergence

The evolution and convergence towards a common core infrastructure is sometimes called the New Generation Network (NGN) architecture, see figure 10.

This network evolution is supported by techniques for separation of control and switching such as the media gateway control protocol H.248. Call session control functions and protocol collections such as H.323 enable call setup including coded and compressed voice calls and choice of coding standard to be used. The Session Initiation Protocol (SIP) has a similar but more limited scope for the setup of communication sessions between two parties and selection of coding standard using the Session Description Protocol (SDP).

The IP Multimedia Subsystems (IMS) defined by the European Telecommunications Standards Institute (ETSI) and the 3rd Generation Partnership Project (3GPP) allow video and other media forms to be exchanged and charged on a session by session basis from peer to peer in a way similar to classic phone calls.

So-called "softswitches" or media gateway controllers include call session control functions for handling of voice calls and other session oriented services. They are also responsible for that sessions can be connected via a physical switch or media gateway (MGW) and handle the signaling between network nodes and other networks. The call session control function establishes a call or session, and further manages its reserved connection path resources end to end, for example through an ATM or IP backbone network and for media stream processing. The MGM provide physical switching and interfaces to access nodes and other networks.

BIBLIOGRAPHY

1. R. L. Freeman, *Telecommunication System Engineering*, 3rd ed., New York: Wiley, 1996.
2. D. Minoli, *Telecommunications Technology Handbook*, Boston: Ar-tech House, 1991.
3. M. Schwartz, *Telecommunication Networks: Protocols, Modelling and Analysis*, Reading, MA: Addison-Wesley, 1987.

TONY LARSSON
ALEXANDER KOTSINAS
BJÖRN KIHLBLOM
Halmstad University, Halmstad, Sweden
3 Scandinavia, Stockholm, Sweden
Ericsson Networks, Stockholm, Sweden