Introduction to

Telecommunications Network Engineering

Second Edition

[Tarmo Anttalainen]

Introduction to Telecommunications Network Engineering

Second Edition

Introduction to Telecommunications Network Engineering

Second Edition

Tarmo Anttalainen

Cover design by Gary Ragaglia

Contents

Preface

Telecommunications is one of the fastest growing business sectors of modern information technologies. A couple of decades ago, to have a basic understanding of telecommunications, it was enough to know how the telephone network operated. Today, the field of telecommunications encompasses a vast variety of modern technologies and services. Some services, such as the fixed telephone service in developed countries, have become mature, and some have been exploding (e.g., cellular mobile communications and the Internet). The deregulation of the telecommunications industry has increased business growth, even though, maybe because, tariffs have decreased.

The present telecommunications environment, in which each of us has to make choices, has become complicated. In the past, there was only one local telephone network operator that we chose to use or not use. Currently, many operators offer us ADSL or cable modem for Internet access and we have many options for telephone service as well.

Telecommunications is a strategically important resource for most modern corporations and its importance continues to increase. Special attention has to be paid to the security aspects and costs of services. The ever-changing telecommunications environment provides new options for users, and we should be more aware of telecommunications as a whole to be able to capitalize on the possibilities available today.

The business of telecommunications has been growing rapidly, and many newcomers have found employment in this area. Even if these

newcomers have a technical background, they may feel that they have a very restricted overall view of the telecommunications network as a whole. The first purpose of this book is to provide an overall view of telecommunications networks to newcomers to the telecommunications business. This kind of general knowledge is useful to the users of telecommunications services, the personnel of operators, and the employees of telecommunications system manufacturers.

The professionals working with these complicated technologies very often have extensive knowledge of one very narrow section of telecommunications, but are not familiar with the hundreds of terms and abbreviations that are used in other telecommunication areas by individuals with whom they need to interact. One purpose of this book is to provide content to some of the most common terms and abbreviations used in different areas of telecommunications.

When I was working as a development department manager at Nokia, I noticed that relatively few books are available that provide a good introduction to data, fixed, and mobile networks. This kind of overview is valuable for people entering a technology area in which all of these technologies are emerging. Most of the books on the market explain telecommunications from only one point of view even though there is no longer any distinct separation of the networks that provide data, speech, and mobile services.

Everyone working in the modern business environment, such as the development engineers, testing personnel, and sales managers, must have a common language if they are to work together efficiently, but not many books supply that common language because they do not provide an overview of telecommunications as a whole.

The material included in this book is used in the Telecommunications Networks course for students of information technologies at the Espoo-Vantaa Institute of Technology in Finland. The goal of this course is to give students a basic understanding of the structure and operation of a global telecommunications network. This course provides an overview of telecommunications; the provision of a deeper understanding about each subject, such as the spectral analysis of signals or detailed knowledge of the operation and functions of mobile networks, is left to dedicated courses.

I have tried not to cover too many aspects of modern telecommunications in this book so as to keep its structure clear. The goal is to lay the basis for later studies of telecommunications for which many good sources are available. Some of them are listed at the end of each chapter.

Objectives

Like the first edition of this book, this second edition is designed to provide answers to the fundamental questions concerning telecommunications networks and services, telecommunications as a business area, and the general trends of technical development. These questions include the following:

- What is the structure and what are the main components of a modern telecommunications network?
- What is the importance of standardization and what are the main standardization bodies for telecommunications?
- How are analog signals processed for transmission over digital circuits?
- What are the basic techniques used in a primary pulse code modulation system that transmits analog speech through the digital telecommunication network?
- How does the *Integrated Services Digital Network* (ISDN) differ from the ordinary telephone network?
- What are the fundamental limiting factors of the rate of information transmission through a transmission channel?
- How do cellular mobile networks operate and what are their main components?
- What are the fundamental differences between circuit and packet switching techniques?
- What technical alternatives are available for provision of wideband access to the Internet?
- What are *local area networks* (LANs) and how are connections arranged over LANs?
- How does the Internet carry its traffic? What are its protocols and how do they operate?
- What happens when I click a mouse on a Web page?

Second Edition

In this second edition, the data communication sections, especially those dealing with local-area networks and the Internet, have been greatly expanded. The Internet has become a very important information source for

most of us and we use it daily in the office and at home. Its use for various kinds of commercial service is expanding, and interactive services, including entertainment, are becoming richer. Most new and evolving network technologies for future telecommunications are also based on data communication concepts, especially Internet technology. Examples of these are packet-switched second and third generation cellular systems. Also the core of the fixed telecommunications network will gradually evolve to packet-switched networks carrying both data and speech traffic. Here we try to emphasize this development.

For future development all opinions and comments concerning the book are welcome. You may send them directly to the author at tarmo.anttalainen@evitech.fi. For those readers who will use this book as training material, please contact the author for additional teachers' instructional material.

Acknowledgments

1

Introduction to Telecommunications

1.1   What Is Telecommunications?

Telecommunications has been defined as a technology concerned with communicating from a distance, and we can categorize it in various ways. Figure 1.1 shows one possible view of the different sections of telecommunications. It includes mechanical communication and electrical communication because telecommunications has evolved from a mechanical to an electrical form using increasingly more sophisticated electrical systems. This is why many authorities such as the national *post, telegraph, and telephone* (PTT) companies are involved in telecommunications using both forms.

Our main concern here is electrical and bidirectional communication, as shown in the upper part of Figure 1.1. The share of mechanical telecommunications such as conventional mail and press is expected to decrease, whereas electrical, especially bidirectional, communication will increase and take the major share of telecommunications in the future. Hence, major press corporations are interested in electrical telecommunications as a business opportunity.

1.2   Significance of Telecommunications

Many different telecommunications networks have been interconnected into a continuously changing and extremely complicated global system. We look at telecommunications from different points of view in order to understand
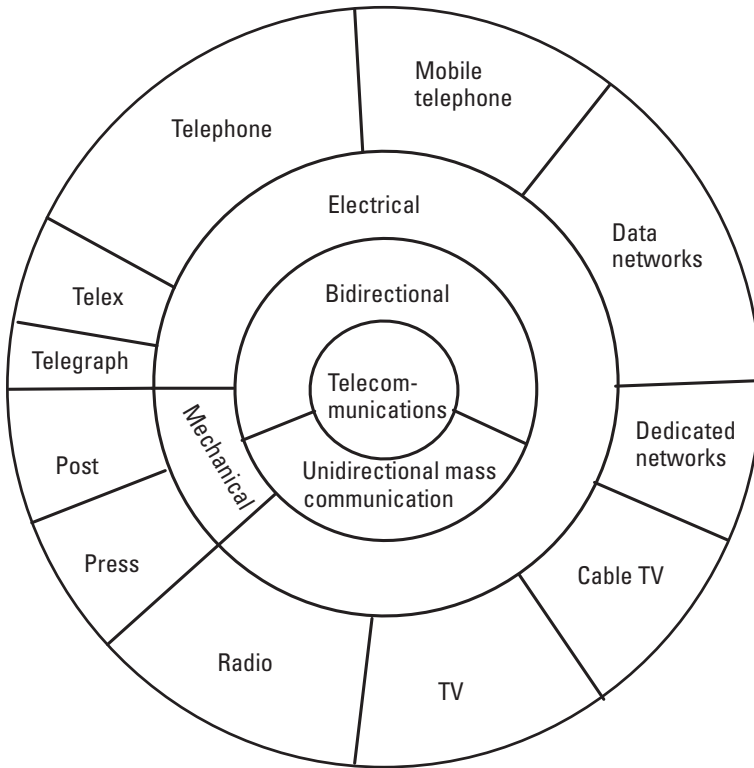
Figure 1.1 Telecommunications.

what a complicated system we are dealing with and how dependent we are on it.

Telecommunications networks make up the most complicated equipment in the world. Let us think only of the telephone network, which includes more than 2 billion fixed and cellular telephones with universal access. When any of these telephones requests a call, the telephone network is able to establish a connection to any other telephone in the world. In addition, many other networks are interconnected with the telephone network. This gives us a view of the complexity of the global telecommunications network—no other system in the world exceeds the complexity of telecommunications networks.

Telecommunications services have an essential impact on the development of a community. If we look at the telephone density of a country, we can estimate its level of technical and economical development. In the developing

countries the fixed telephone density, that is, the *teledensity*, is fewer than 10 telephones per 1,000 inhabitants; in developed countries in, for instance, North America and Europe, there are around 500 to 600 fixed telephones per 1,000 inhabitants. The economic development of developing countries depends on (in addition to many other things) the availability of efficient telecommunications services.

The operations of a modern community are highly dependent on telecommunications. We can hardly imagine our working environment without telecommunications services. The *local area network* (LAN) to which our computer is connected is interconnected with the LANs of other sites throughout our company. This is mandatory so that the various departments can work together efficiently. We communicate daily with people in other organizations with the help of electronic mail, telephones, facsimile, and mobile telephones. Governmental organizations that provide public services are as dependent on telecommunications services as are private organizations.

Telecommunications plays an essential role on many areas of everyday living. Everyday life is dependent on telecommunications. Each of us uses telecommunications services and services that rely on telecommunications daily. Here are some examples of services that depend on telecommunications:

- Banking, automatic teller machines, telebanking;
- Aviation, booking of tickets;
- Sales, wholesale and order handling;
- Credit card payments at gasoline stations;
- Booking of hotel rooms by travel agencies;
- Material purchasing by industry;
- Government operations, such as taxation.

1.3 Historical Perspective

Some of the most important milestones in the development of electrical telecommunications systems according to [1] are discussed in this section. Terms and abbreviations used in the chronology are explained in later chapters of this book. The development and expansion of some telecommunications services is also illustrated in Figure 1.2.
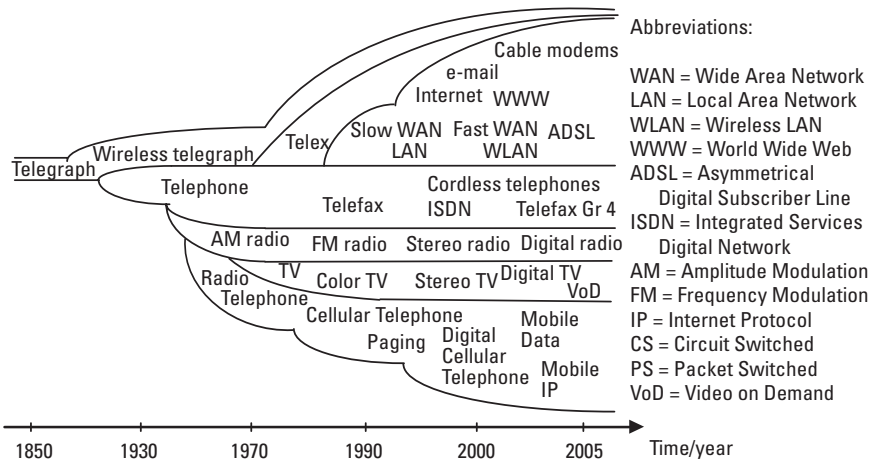
Figure 1.2 Development of telecommunications systems and services.

1800–1837 *Preliminary developments:* Volta discovers the primary battery; Fourier and Laplace present mathematical treatises; Ampere, Faraday, and Henry conduct experiments on electricity and magnetism; Ohm's law (1826); Gauss, Weber, and Wheatstone develop early telegraph systems.

1838–1866 *Telegraphy:* Morse perfects his system; Steinhill finds that the earth can be used for a current path; commercial service is initiated (1844); multiplexing techniques are devised; William Thomson calculates the pulse response of a telegraph line (1855); transatlantic cables are installed.

1845   Kirchoff's circuit laws.

1864   Maxwell's equations predict electromagnetic radiation.

1876–1899 *Telephony:* Alexander Graham Bell perfects acoustic transducer; first telephony exchange with eight lines; Edison's carbon-button transducer; cable circuits are introduced; Strowger devises automatic step-by-step switching (1887); Pupin presents the theory of loading.

1887–1907 *Wireless telegraphy:* Heinrich Hertz verifies Maxwell's theory; demonstrations by Marconi and Popov; Marconi patents complete wireless telegraph system (1897); commercial service begins, including ship-to-shore and transatlantic systems.

1904–1920   *Communication electronics:* Lee De Forest invents the Audion (triode) based on Fleming's diode; basic filter types devised; experiments with AM radio broadcasting; the Bell System completes the transcontinental telephone line with electronic repeaters (1915); multiplexed carrier telephony is introduced: H. C. Armstrong perfects the superheterodyne radio receiver (1918); first commercial broadcasting station.

1920–1928   Carson, Nyquist, Johnson, and Hartley present their transmission theory.

1923–1938   *Television:* Mechanical image-formation system demonstrated; theoretical analysis of bandwidth requirements; DuMont and others perfect vacuum cathode-ray tubes; field tests and experimental broadcasting begin.

1931   Teletypewriter service initiated.

1934   H. S. Black develops the negative feedback amplifier.

1936   Armstrong's paper states the case of *frequency modulation* (FM) radio.

1937   Alec Reeves conceives *pulse code modulation* (PCM).

1938–1945   Radar and microwave systems developed during World War II; FM used extensively for military communications; hardware, electronics, and theory are improved in all areas.

1944–1947   Mathematical representations of noise developed; statistical methods for signal detection developed.

1948–1950   C. E. Shannon publishes the founding papers on information theory.

1948–1951   Transistor devices are invented.

1950   *Time-division multiplexing* (TDM) is applied to telephony. Hamming presents the first error correction codes.

1953   Color TV standards are established in the United States.

1955   J. R. Pierce proposes satellite communication systems.

1958   Long-distance data transmission system is developed for military purposes.

1960   Maiman demonstrates the first laser.

1961   Integrated circuits are applied to commercial production.

1962   Satellite communication begins with Telstar I.

1962–1966	Data transmission service offered commercially; PCM proves feasible for voice and TV transmission; theory for digital transmission is developed; Viterbi presents new error-correcting schemes; adaptive equalization is developed.
1964	Fully electronic telephone switching system is put into service.
1965	Mariner IV transmits pictures from Mars to Earth.
1966–1975	Commercial satellite relay becomes available; optical links using lasers and fiber optics are introduced; ARPANET is created (1969) followed by international computer networks.
1976	Ethernet LAN invented by Metcalfe and Broggs (Xerox) [2].
1968–1969	Digitalization of telephone network begins.
1970–1975	PCM standards developed by CCITT.
1975–1985	High-capacity optical systems developed; the breakthrough of optical technology and fully integrated switching systems; digital signal processing by microprocessors.
1980–1983	Start of global Internet based on TCP/IP protocol [3].
1980–1985	Modern cellular mobile networks put into service, NMT in Northern Europe, AMPS in the United States, OSI reference model is defined by *International Standards Organization* (ISO). Standardization for second generation digital cellular systems is initialized.
1985–1990	LAN breakthrough; *Integrated Services Digital Network* (ISDN) standardization finalized; public data communications services become widely available; optical transmission systems replace copper systems in long-distance wideband transmission; SONET is developed. GSM and SDH standardization finalized.
1989	Initial proposal for a Web-linked document on the *World Wide Web* (WWW) by Tim Berners-Lee (CERN) [2].
1990–1997	The first digital cellular system, *Global System for Mobile Communications* (GSM), is put into commercial use and its breakthrough is felt worldwide; deregulation of telecommunications in Europe proceeds and satellite TV systems become popular; Internet usage and services expand rapidly because of the WWW.
1997–2001	Telecommunications community is deregulated and business grows rapidly; digital cellular networks, especially GSM,

expand worldwide; commercial applications of Internet expand and a share of conventional speech communications is transferred from *public switched telephone network* (PSTN) to Internet; performance of LANs improves with advance of gigabit-per-second Ethernet technologies.

2001–2005   Digital TV starts to replace analog broadcast TV; broadband access systems make Internet multimedia services available to all; telephony service turns to personal communication service as penetration of cellular and PCS systems increases; second generation cellular systems are upgraded to provide higher rate packet-switched data service.

2005–   Digital TV will replace analog service and start to provide interactive services in addition to broadcast service; third generation cellular systems and WLAN technologies will provide enhanced data services for mobile users; location-based mobile services will expand, applications for wireless short-haul technologies in homes and offices will increase; global telecommunications network will evolve toward a common packet-switched network platform for all types of services.

1.4   Standardization

Communication networks are designed to serve a wide variety of users who are using equipment from many different vendors. To design and build networks effectively, standards are necessary to achieve interoperability, compatibility, and required performance in a cost-effective manner.

Open standards are needed to enable the interconnection of systems, equipment, and networks from different manufacturers, vendors, and operators. The most important advantages and some other aspects of open telecommunications standards are explained next.

Standards enable competition.   Open standards are available to any telecommunications system vendor. When a new system is standardized that is attractive from a business point of view, multiple vendors will enter this new market. As long as a system is proprietary, specifications are the property of one manufacturer and it is difficult, if not impossible, for a new manufacturer to start to produce compatible competing systems. Open competition makes products more cost-effective, therefore providing low-cost services to telecommunications users.

Standards lead to economies of scale in manufacturing and engineering.   Standards increase the market for products adhering to the standard, which leads to mass production and economies of scale in manufacturing and engineering, *very large scale integration* (VLSI) implementations, and other benefits that decrease price and further increase acceptance of the new technology. This supports the economic development of the community by improving telecommunications services and decreasing their cost.

Political interests often lead to different standards in Europe, Japan, and the United States.   Standardization is not only a technical matter. Sometimes opposing political interests make the approval of global standards impossible, and different standards are often adapted for Europe, the United States, and Japan. To protect local industry, Europe does not want to accept American technology and America does not want to accept European technology.

One example of a political decision in the 1970s was to define a different PCM coding law for Europe instead of the existing American PCM code. (We will explain this terminology in Chapter 3.) A more recent example is the American decision in the 1990s not to accept European GSM technology as a major digital second generation cellular technology.

International standards are threats to the local industries of large countries but opportunities to the industries of small countries.   Major manufacturers in large countries may not support international standardization because it would open their local markets to international competition. Manufacturers in small countries strongly support global standardization because they are dependent on foreign markets. Their home market is not large enough for expansion and they are looking for new markets for their technology.

Standards make the interconnection of systems from different vendors possible. The main technological aim of standardization is to make systems from different networks "understand" each other. Technical specifications included in open standards make systems compatible and support the provision of wide-area or even global services that are based on standardized technology.

Standards make users and network operators vendor independent and improve availability of the systems.   A standardized interface between a terminal and its network enables subscribers to purchase terminal equipment from multiple vendors. Standardized interfaces among systems in the network enable network operators to use multiple competing suppliers for systems. This improves the availability and quality of systems and reduces their cost.

Standards make international services available. Standardization plays a key role in the provision of international services. Official global standards define, for example, telephone service, ISDN, and facsimile. The standards of some systems may not have official worldwide acceptance, but if the system becomes popular all around the world, a worldwide service may become available. Recent examples of these services are GSM and the Internet with WWW. Internet specifications have no official status, and GSM was originally specified for Europe only. Their specifications have been openly available, which has supported their expansion.

To clarify and understand the influence of standardization on our everyday lives, consider these examples of international standardization:

- *Screw thread pitches (ISO, Technical Committee 1):* This was one of the first activity areas of standardization. In the 1960s, a bolt from one car would not fit another. Currently, bolts are internationally standardized and most often compatible.

- *International telephone numbering and country codes:* Without globally unique identification of subscribers, automatic international telephone calls would not be available.

- *Telephone subscriber interfaces.*

- *PCM coding and primary rate frame structure:* This coding and structure make national and international digital connections between networks possible.

- *Television and radio systems.*

- *Frequencies used for satellite and other radio communications.*

- *Connectors and signals for PC, printer, and modem interfaces.*

- *LANs:* These enable people to use computers from any manufacturer in a company network.

- *Cellular telephone systems:* Enable users to choose a handset from among a large selection with different features from many different vendors.

1.5 Standards Organizations

Many organizations are involved in standardization work. We look at them from two points of view: (1) the players in the telecommunications business

involved in standardization and (2) the authorities that approve official standards.

1.5.1   Interested Parties

Figure 1.3 shows some groups that are interested in standardization and participate in standardization work. Let us look at a list of these parties and their most important interests, that is, why they are involved in standardization work.

Network operators support standardization for these reasons:

- To improve the compatibility of telecommunications systems;
- To be able to provide wide-area or even international services;
- To be able to purchase equipment from multiple vendors.

Equipment manufacturers participate in standardization for these reasons:

- To get information about future standards for their development activities as early as possible;
- To support standards that are based on their own technologies;
- To prevent standardization if it opens their own markets.

Service users participate in standardization for these reasons:

- To support the development of standardized international services;
- To have access to alternative system vendors (multivendor networks);
- To improve the compatibility of their future network systems.
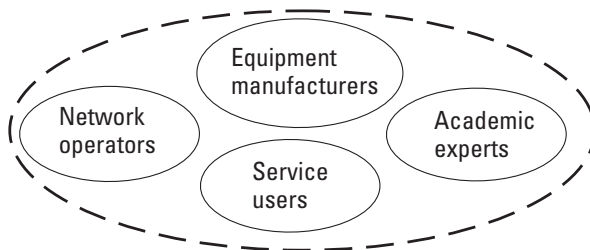
Figure 1.3 Interested parties.

   Other interested parties include governmental officials who are keen on having national approaches adopted as international standards and academic experts who want to become inventors of new technological approaches.

1.5.2   National Standardization Authorities

National standardization authorities approve official national standards. Many international standards include alternatives and options from which a national authority selects those suitable for their own national standards. These options are included in cases for which a common global understanding could not be agreed on. Sometimes some aspects are left open and they require a national standard. For example, national authorities determine the details of their national telephone numbering plan, for which international standards give only guidelines. Another example is frequency allocation. International standards define usage of frequency bands (e.g., which frequency ranges are used for satellite and which for cellular networks), whereas the national authority defines detailed usage of frequencies inside the country; for example, they allocate frequency channels for cellular network operators. Some examples of national authorities are shown in the Figure 1.4. They take care of all areas of standardization, and they set up specialized organizations or working groups to work with the standardization of each specific technical area, such as telecommunications and information technology. These example organizations are shown in Figure 1.4: the British Standards Institute (BSI; United Kingdom), Deutsche Industrie-Normen (DIN; Germany), American National Standards Institute (ANSI; United States), and the Finnish Standards Institute (SFS; Finland).

1.5.3   European Organizations

The most important European standards organizations are shown in the Figure 1.5. They are responsible for developing European-wide standards to
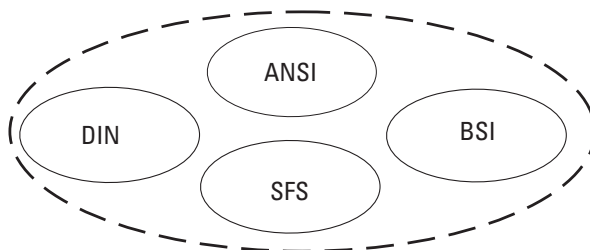
Figure 1.4   Some examples of national standardization authorities.
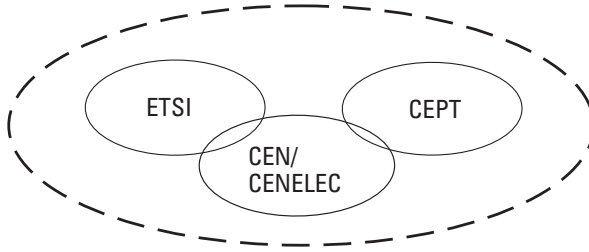
Figure 1.5   European standards organizations.

open national borders in order and improve pan-European telecommunications services.

The *European Telecommunications Standards Institute* (ETSI) is an independent body for making standards for the European Community. Telecommunications network operators and manufacturers participate in standardization work. One example of standards made by ETSI is the digital cellular mobile system GSM, which became a major standard for second generation digital mobile communications all around the world.

The *European Committee for Electrotechnical Standardization/European Committee for Standardization* (CEN/CENELEC) is a joint organization for the standardization of information technology. It corresponds to IEC/ISO on a global level and it handles environmental and electromechanical aspects of telecommunications.

The *Conférence Européenne des Administrations des Postes et des Telecommunications* or European Conference of Posts and Telecommunications Administrations (CEPT) was doing the work of ETSI before the European Commission Green Paper opened competition in Europe within the telecommunications market. The deregulation of telecommunications forced national PTTs to become network operators equal to other new operators and they are not allowed to make standards alone any more.

1.5.4   American Organizations

The U.S. national standards authority American National Standards Institute has accredited several organizations to work for standards for telecommunications. Some of these organizations are shown in Figure 1.6.

The *Institute of Electrical and Electronics Engineers* (IEEE) is one of the largest professional societies in the world and it has produced many important standards for telecommunications. Some of these standards, such as the standards for LANs, have been accepted by the ISO as international
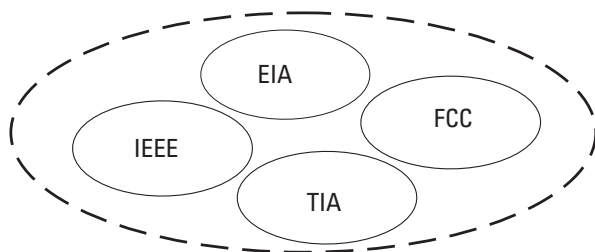
Figure 1.6 American standards organizations.

standards. For example, international standard ISO 8802.*x* for the Ethernet LAN family is currently the same as IEEE 802.*x*.

The *Electronic Industries Association* (EIA) is an American organization of electronic equipment manufacturers. Many of its standards, such as those for connectors for personal computers, have achieved global acceptance. For example, the data interface standard EIA RS-232 is compatible with the V.24/28 recommendations of ITU-T.

The *Federal Communications Commission* (FCC) is not actually a standards body but a regulatory body. It is a government organization that regulates wire and radio communications. It has played an important role, for example, in the development of worldwide specifications for radiation and susceptibility of electromagnetic disturbances of telecommunications equipment.

The *Telecommunications Industry Association* (TIA) has been developing global third generation cellular systems together with ETSI from Europe and the *Association of Radio Industries and Broadcasting* (ARIB) from Japan. Its task is to adapt the global standard to the American environment [4].

1.5.5   Global Organizations

The *International Telecommunication Union* (ITU) is a specialized agency of the United Nations responsible for telecommunications. It has nearly 200 member countries, and standardization work is divided between two major standardization bodies: ITU-T and ITU-R (see Figure 1.7).

The *Comité Consultatif International de Télégraphique et Téléphonique*, or International Telegraph and Telephone Consultative Committee (CCITT/ITU-T) is presently called ITU-T, where the "T" comes from telecommunications. The *Comité Consultatif International des Radiocommunications* or International Radio Consultative Committee (CCIR/ITU-R) is presently known as ITU-R, where the "R" stands for radio.
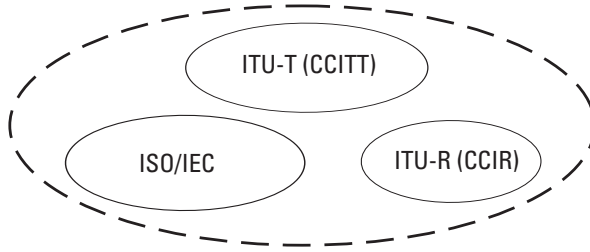
Figure 1.7   Global standards organizations.

ITU-T and ITU-R publish recommendations that are in fact strong standards for telecommunications networks. ITU-T works for the standards of public telecommunications networks (e.g., ISDN), and ITU-R works with radio aspects such as the usage of radio frequencies worldwide and specifications for radio systems. Many parties participate in their work, but only national authorities may vote. ITU-T, formerly CCITT, has created most of the current worldwide standards for public networks.

The International Standards Organization/International Electrotechnical Commission (ISO/IEC) is a joint organization responsible for the standardization of information technology. ISO has done important work in the area of data communications and protocols, and IEC in the area of electromechanical (for example, connectors), environmental, and safety aspects.

The organizations shown in Figure 1.7 work together closely to avoid duplicating effort and to avoid creating multiple standards for the same purpose. As a consequence, some ITU recommendations may contain merely a reference to an ISO standard.

1.5.6   Other Organizations

Many organizations other than those just mentioned are working with standards. Some of these are active in ITU-T and ISO, and many international standards are based on (or may even be copies of) the initial work of these groups. We introduce some of these as examples of standards organizations without official status (see Figure 1.8).

The *Internet Engineering Task Force* (IETF) is responsible for the evolution of the Internet architecture. It takes care of the standardization of the TCP/IP protocol suite used in the Internet.

The *Universal Mobile Telecommunications System* (UMTS) Forum is an open organization of cellular system manufacturers. Its goal is to define a third generation cellular system that will receive worldwide acceptance and
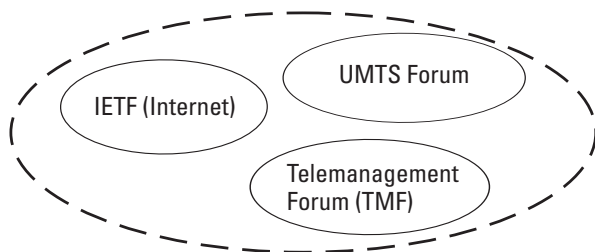
Figure 1.8 Examples of other standards organizations.

ensure compatibility among equipment from different vendors. Unofficial forums are more flexible and can produce necessary standards on a shorter timescale than can official organizations. Their specifications are often used as a basis for later official standards.

The *Telemanagement Forum* (TMF) is an organization of system manufacturers that works to speed the development of network management standards. With the help of these standards, telecommunications network operators will be able to control and supervise their multivendor networks efficiently from the same management center. Proposals are then given to ITU-T and ISO for official international acceptance.

The organizations mentioned here are just examples; many other such organizations and cooperative units exist. New groups appear and some organizations disappear every year.

One important problem in standardization is the question of *intellectual property rights* (IPRs). One company involved in development of a standard may have a patent or copyright for a method or function that is essential for implementation of the standardized system. In such a case, other manufacturers may not be able to implement the standard in an economically feasible manner without interfering with a patent or copyright. There are no fixed rules regarding how to solve this problem, but very often the patent or copyright owner agrees to license the patent or copyright for a standardized system under fair terms [5].

1.6 Development of the Telecommunications Business

In the past, telecommunications has been a protected business area. The national PTTs were once the only national telecommunications operators in most countries. They had control over standardization in international standardization bodies and a monopoly in providing telecommunications

services in their home country. For political reasons domestic manufacturers were preferred as suppliers of the systems needed in the network. Competition was not allowed, and the development of services and networks was slow in many countries.

During the latter part of the 1980s the deregulation of the telecommunications business started in Europe and proceeded rapidly in many other areas of the world. Competitive telecommunications services are important for the development of an economy, and governments supported the development of free markets heavily.

In Europe the European Union has paid much attention to the deregulation of the telecommunications business. New operators have obtained licenses to provide local and long-distance telephone and data services and mobile telecommunications services. Previously many standards, such as analog mobile telephone standards, did not even support a multioperator environment. The initial requirement of the digital mobile telecommunications system (GSM) in Europe was the support of multiple networks in the same geographical area. The deregulation of the telecommunications business has reduced tariffs on long-distance calls and mobile calls to a small fraction of the tariffs paid in the mid-1980s. The reduction of fees has further increased the demand for services, which has prompted reductions in the price of terminal equipment, such as mobile telephones, and the fees for calls.

These developments have demonstrated how dangerous it is for manufacturers to be too dependent on a single domestic customer. Many telecommunications manufacturers that were independent in the past do not exist as independent suppliers anymore. This process still continues. At the same time, new small manufacturers are appearing. Their window of opportunity is to produce special equipment, in which the largest vendors are not interested, or systems for brand new rapidly growing services.

Plain old telephone service (POTS) will still be important in the future, but mobile and data communications grow most rapidly in volume. The two main directions of this development are in the areas of voice communications, which will become mobile, and data communications, which will become wideband, high-data-rate communications. Because of deregulation, subscribers can choose which network operator they want to use to get wideband access to the Internet over ordinary telephone lines. Cable TV operators are also providing similar services in competitive terms.

The provision of developing multimedia services in the future will be especially interesting. The expansion of the Internet, with its improving capability to transmit voice in addition to data, presents a new challenge to the public telecommunications network operators. Wideband access to

homes will be used for telephone calls in addition to Internet surfing. This requires telecommunications network operators, including cellular network operators, to change their strategies from telephone and data transmission to complete service and information content provision. These services will contain Internet portals and location-based services, such as information on the nearest fast-food restaurant, in cellular networks.

For the future development of the telecommunications business, we must pay attention to customer services that technology can provide, not technology itself. Many good technologies, which we explain in later chapters, have not been successful because ordinary subscribers have not viewed them as attractive. Examples of these technologies are ISDN and *wireless application protocol* (WAP) services. On the other hand, some services, such as the WWW, have grown very rapidly. We have to keep in mind that only attractive services make new technologies successful.

1.7   Problems and Review Questions

Problem 1.1

List two or more electrical telecommunications systems that provide (a) bidirectional and (b) unidirectional service.

Problem 1.2

What were the three main developments in communications technologies during the last 20 years? Explain why you think so because this is a matter of opinion.

Problem 1.3

What are the most important advantages of global telecommunications standards?

Problem 1.4

Why is it often difficult to achieve a common understanding of and approve global standards? Explain both political and business interests.

References

[1]   Carlson, A. B., *Communication Systems: An Introduction to Signals and Noise in Electrical Communication,* New York: McGraw-Hill, 1986.

[2]   Tanenbaum, A. S., *Computer Networks,* 3rd ed., Upper Saddle River, NJ: Prentice Hall, 1996.

[3]   Comer, D. E., *Internetworking with TCP/IP: Principles, Protocols, and Architecture,* 4th ed., Upper Saddle River, NJ: Prentice Hall, 2000.

[4]   Steele, R., and L. Hanzo, *Mobile Radio Communications*, 2nd ed., West Sussex, England: John Wiley & Sons Ltd., 1999.

[5]   Egyedi, T. M., "IPR Paralysis in Standardization: Is Regulatory Symmetry Desirable?" *IEEE Communications Magazine,* April 2001, pp. 108–144.

2

The Telecommunications Network: An Overview

This chapter describes the basic operation of a telecommunications network with the help of a conventional telephone. The operation of a conventional telephone, which is easy to understand, is used to clarify how telephone connections are built up in the network. We look at subscriber signaling over the subscriber loop of the telephone network. The same kind of signaling is needed in modern telecommunications networks, such as ISDN and cellular networks. We start with this simple service to lay a foundation for understanding more complicated types of service in later chapters.

In this chapter we divide the network into layers and briefly describe different network technologies that are needed to provide various kinds of service. Some of these, such as mobile and data networks, are discussed in more detail later in this book. The last topic of this chapter is an introduction to the theory of traffic engineering; that is, how much capacity we should build into the network in order to provide a sufficient grade of service for the customers.

2.1   Basic Telecommunications Network

The basic purpose of a telecommunications network is to transmit user information in any form to another user of the network. These users of public networks, for example, a telephone network, are called subscribers. User

information may take many forms, such as voice or data, and subscribers may use different access network technologies to access the network, for example, fixed or cellular telephones. We will see that the telecommunications network consists of many different networks providing different services, such as data, fixed, or cellular telephony service. These different networks are discussed in later chapters. In the following section we introduce the basic functions that are needed in all networks no matter what services they provide.

The three technologies needed for communication through the network are (1) transmission, (2), switching, and (3) signaling. Each of these technologies requires specialists for their engineering, operation, and maintenance.

2.1.1 Transmission

Transmission is the process of transporting information between end points of a system or a network. Transmission systems use four basic media for information transfer from one point to another:

1. Copper cables, such as those used in LANs and telephone subscriber lines;
2. Optical fiber cables, such as high-data-rate transmission in telecommunications networks;
3. Radio waves, such as cellular telephones and satellite transmission;
4. Free-space optics, such as infrared remote controllers.

In a telecommunications network, the transmission systems interconnect exchanges and, taken together, these transmission systems are called the transmission or transport network. Note that the number of speech channels (which is one measure of transmission capacity) needed between exchanges is much smaller than the number of subscribers because only a small fraction of them have calls connected at the same time. We discuss transmission in more detail in Chapter 4.

2.1.2 Switching

In principle, all telephones could still be connected to each other by cables as they were in the very beginning of the history of telephony. However, as the number of telephones grew, operators soon noticed that it was necessary to switch signals from one wire to another. Then only a few cable connections were needed between exchanges because the number of simultaneously ongoing calls is much smaller than the number of telephones (Figure 2.1). The
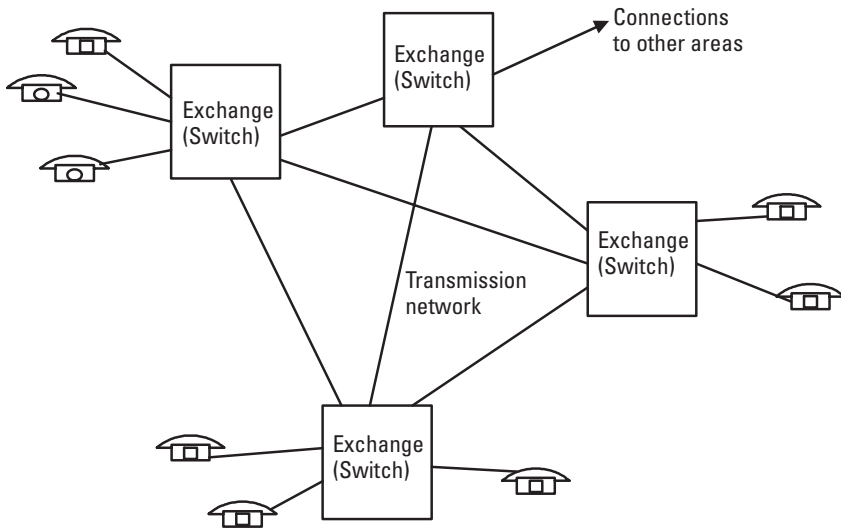
Figure 2.1 A basic telecommunications network.

first switches were not automatic so switching was done manually using a switchboard.

Strowger developed the first automatic switch (exchange) in 1887. At that time, switching had to be controlled by the telephone user with the help of pulses generated by a dial. For many decades exchanges were a complex series of electromechanical selectors, but during the last few decades they have developed into software-controlled digital exchanges. Modern exchanges usually have quite a large capacity—tens of thousands subscribers—and thousands of them may have calls ongoing at the same time.

2.1.3 Signaling

Signaling is the mechanism that allows network entities (customer premises or network switches) to establish, maintain, and terminate sessions in a network. Signaling is carried out with the help of specific signals or messages that indicate to the other end what is requested of it by this connection. Some examples of signaling examples on subscriber lines are as follows:

- *Off-hook condition:* The exchange notices that the subscriber has raised the telephone hook (dc loop is connected) and gives a dial tone to the subscriber.

- *Dial:* The subscriber dials digits and they are received by the exchange.
- *On-hook condition:* The exchange notices that the subscriber has finished the call (subscriber loop is disconnected), clears the connection, and stops billing.

Signaling is naturally needed between exchanges as well because most calls have to be connected via more than just one exchange. Many different signaling systems are used for the interconnection of different exchanges. Signaling is an extremely complex matter in a telecommunications network. Imagine, for example, a foreign GSM subscriber switching his telephone on in Hong Kong. In approximately 10 seconds he is able to receive calls directed to him. Information transferred for this function is carried in hundreds of signaling messages between exchanges in international and national networks. Signaling in a subscriber loop is discussed in Section 2.3 and signaling between exchanges in Section 2.6.

2.2 Operation of a Conventional Telephone

The ordinary home telephone receives the electrical power that it needs for operation from the local exchange via two copper wires. This subscriber line, which carries speech signals as well, is a twisted pair called a local loop. The principle of the power supply coming from the exchange site makes basic telephone service independent of the local electric power network. Local exchanges have a large-capacity battery that keeps the exchange and subscriber sets operational for a few hours if the supply of electricity is cut off. This is essential because the operation of the telephone network is especially important in emergency situations when the electric power supply may be down.

Figure 2.2 shows a simplified illustration of the telephone connection. Elements of the figure and operation of the subscriber loop are explained later in this chapter. Minor operational differences, particularly in the provision of *private branch exchange/automatic branch exchange* (PBX/PABX) systems, exist around the world, but the principles discussed in this chapter apply to the overwhelming majority of PSTN systems.

2.2.1 Microphone

When we raise the hook of a telephone, the on/off hook switch is closed and current starts flowing on the subscriber loop through the microphone that is connected to the subscriber loop. The microphone converts acoustic energy
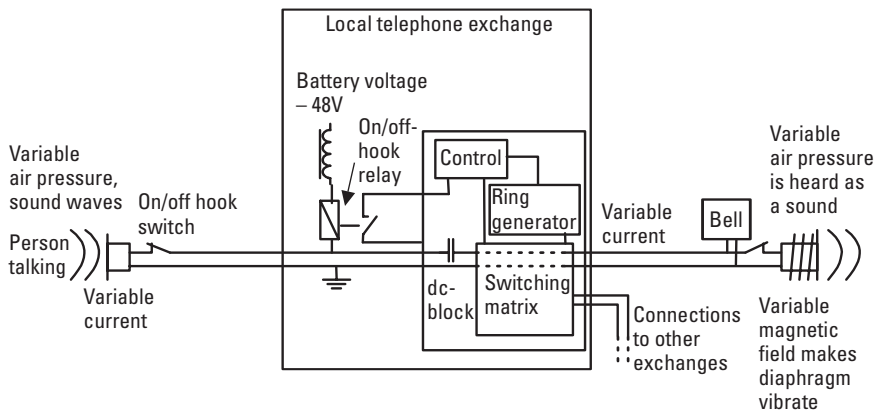
Figure 2.2 Operation principle of a conventional telephone.

to electrical energy. Originally telephone microphones were so-called carbon microphones that had diaphragms with small containers of carbon grains and they operated as variable resistors supplied with battery voltage from an exchange site (see the subscriber loop on the left-hand side of Figure 2.2). When sound waves pressed the carbon grains more tightly, loop resistance decreased and current slightly increased. The variable air pressure generated a variable, alternating current to the subscriber loop. This variable current contained voice information. The basic operating principle of the subscriber loop is still the same today, although modern telephones include more sophisticated and better quality microphones.

2.2.2 Earphone

Alternating current, generated by the microphone, is converted back into voice at the other end of the connection. The earphone has a diaphragm with a piece of magnet inside a coil. The coil is supplied by alternating current produced by the microphone at the remote end of the connection. The current generates a variable magnetic field that moves the diaphragm that produces sound waves close to the original sound at the transmitting end (see the subscriber loop on the right-hand side of Figure 2.2).

2.2.3 Signaling Functions

The microphone generates the electrical current that carries voice information, and the earphone produces the voice at the receiving end of the speech

circuit. The telephone network provides a dialed-up or circuit-switched service that enables the subscriber to initiate and terminate calls. The subscriber dials the number to which she wants to be connected. This requires additional information transfer over the subscriber loop and from the exchange to other exchanges on the connection, and this transfer of additional information is called signaling. The basic subscriber signaling phases are described in the following section.

2.3   Signaling to the Exchange from the Telephone

Telephone exchanges supply dc voltage to subscriber loops, and telephone sets use this supplied voltage for operation. The conventional telephone does not include any electronics, and the supplied voltage and current are directly used for speech transmission in addition to signaling functions that include the detection of on/off-hook condition and dialing. Modern electronic telephones would not necessarily need this if they could take their power from a power socket at home. However, getting the power supply from the exchange is still an important feature because it ensures that the telephone network operates even in emergency situations when the power network may be down.

2.3.1   Setup and Release of a Call

Each telephone has a switch that indicates an on- or off-hook condition. When the hook is raised, the switch is closed and an approximately 50 mA of current starts flowing. This is detected by a relay giving information to the control unit in the exchange (Figure 2.2). The control unit is an efficient and reliable computer (or a set of computers) in the telephone exchange. It activates signaling circuits, which then receive dialed digits from subscriber A. (We call a subscriber who initiates a call subscriber A and a subscriber who receives a call subscriber B.) The control unit in the telephone exchange controls the switching matrix that connects the speech circuit through to the called subscriber B. Connection is made according to the numbers dialed by subscriber A.

When the call is being routed to subscriber B, the telephone exchange supplies to the subscriber loop a ringing voltage and the bell of subscriber B's telephone starts ringing. The ringing voltage is often about 70V ac with a 25-Hz frequency, which is high enough to activate the bell on any telephone. The ringing voltage is switched off immediately when an off-hook condition

is detected on the loop of subscriber B, and then an end-to-end speech circuit is connected and the conversation may start.

Figure 2.3 shows the signaling phases on a subscriber loop. When the exchange detects the off-hook condition of a subscriber loop, it informs us with a dial tone that we hear when we raise the hook that it is ready to receive digits. After dialing it keeps us informed about whether the circuit establishment is successful by sending us a ringing tone when the telephone at the other end rings. When subscriber B answers, the exchange switches off both the ringing signal and the ringing tone and connects the circuit. At the end of the conversation, an on-hook condition is detected by the exchange and the speech circuit is released.

In next sections we explain in more detail one of the subscriber signaling phases, the transmission of dialed digits from a subscriber's telephone to the local exchange.

2.3.2   Rotary Dialing

The telephone set has a switch that is open in the on-hook condition and closed when the hook is off. This indicates to the telephone exchange when a call is to be initiated and when it has to prepare to receive dialed digits. In old telephones, which exchanges still have to support, this method of local-loop connection/disconnection is used to transmit dialed digits as well (Figure 2.4). We call this principle rotary or pulse dialing.
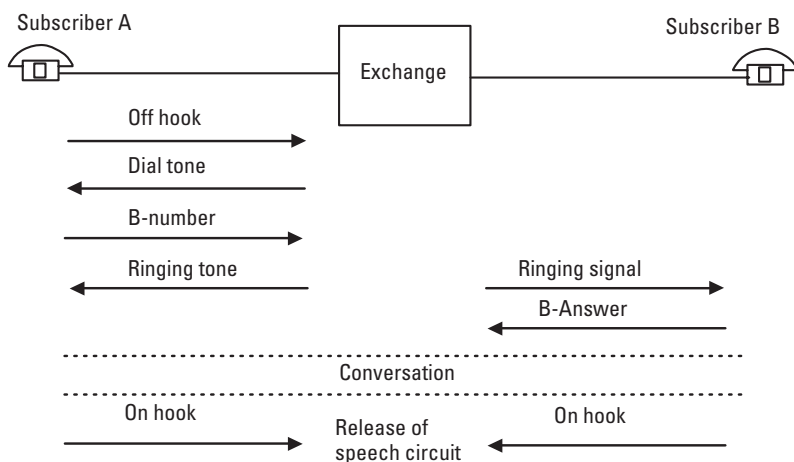
Figure 2.3   Subscriber signaling.
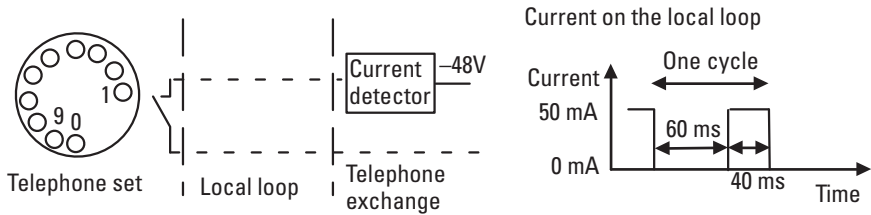
Figure 2.4 Rotary, or pulse, dialing.

In rotary dialing a local loop is closed and opened according to the dialed digits, and the number of current pulses is detected by the exchange. This signaling method is also known as loop disconnect signaling. The main disadvantages of this method are that it is slow and expensive due to high-resolution mechanics and it does not support supplementary services such as call forwarding. The local-loop interfaces in telephone exchanges have to support this old technology though it has been gradually replaced by tone dialing.

When a digit is to be dialed, the dialing plate with finger holes is rotated clockwise to the end and released. While homing, the switch is breaking the line current periodically and the number of these periods indicates the dialed digit. For example, digit 1 has one period, 2 has two periods, and 0 has 10 periods or cycles. Mechanics make the homing speed approximately constant and each period is about 100 ms long with a 60-ms break (Figure 2.4). This method for the transmission of digits has also been used for signaling between exchanges and then it is known as loop disconnect signaling.

The value of the loop current differs slightly from country to country and it is also dependent on line length and supply voltage, for example. Typically it is from 20 to 50 mA, high enough to control old generation electromechanical switches that used pulses to control directly the rotating switches of the switching matrix of an exchange.

2.3.3   Tone Dialing

Currently telephones include electronic circuits that make possible the implementation of better means for signaling. Digital exchanges do not require high-power pulses to drive the selectors as old electromechanical switches did. However, subscriber lines are still, and will be, supplied by a –48- or –60-V battery so that telephones continue to operate independent of the electric power supply. Electronic telephones use 50- to 500-μA current

all the time to supply power to their electronic circuitry, which is needed for number repetition, abbreviated dialing, and other additional features of modern telephone sets.

Modern telephones usually have 12 push buttons (keys A to D of Figure 2.5 are not included in an ordinary subscriber set) for dialing, each generating a tone with two frequencies. One of the frequencies is from the upper frequency band and the other from the lower band. All frequencies are inside the voice frequency band (300–3,400 Hz) and can thus be transmitted through the network from end to end, when the speech connection is established. This signaling principle is known as *dual-tone multifrequency* (DTMF) signaling.

Tones are detected at the subscriber interface of the telephone exchange and, if necessary, signaled further to the other exchanges through which the connection is to be established. All digital local exchanges have a capability to use either pulse or tone dialing on a subscriber loop. The subscriber is able to select with a switch on his telephone which type of dialing is to be used. Tone dialing should always be selected if the local exchange is a modern digital one.

Advantages of tone dialing are as follows:

- It is quicker and dialing of all digits takes the same time.
- Fewer dialing errors result.
- End-to-end signaling is possible.
- Additional push buttons are available (*, #, A, B, C, D) for activation of supplementary services.

Upper frequency band/Hz

1,209 1,336 1,477 1,633

Lower frequency band/Hz				
697	1	2	3	A
770	4	5	6	B
852	7	8	9	C
941	*	0	#	D

Push buttons A, B, C, and D are not available in ordinary telephones

DTMF = Dual tone multifrequency

DTMF telephone keypad
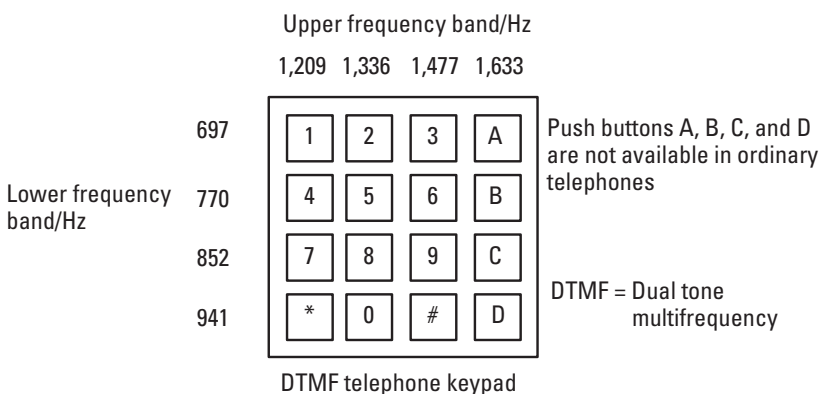
Figure 2.5 Tone dialing.

Supplementary services enable subscribers to influence the routing of their telephone calls. These services, for example, call transfer, are not available with telephones that use pulse dialing. To control these services we need control buttons * and #, which are available only in push-button telephones that use tone dialing.

We use tone dialing also to control *value-added services*. Value-added services are services that we can use via the telephone network but that are usually provided by another service provider, not the telecommunications network operator. One example of value added services is telebanking. Tones are transmitted on the same frequency band as voice, and during a call we are able to dial digits to transmit, for example, our discount number and security codes to the telebanking machine.

The worst disadvantage of a fixed subscriber telephone is still the poor man–machine interface that makes new services difficult to use. Some telephones that have displays are more user friendly, but subscribers still have to memorize command sequences to use the new services offered by a modern telephone network.

2.3.4   Local Loop and 2W/4W Circuits

Any use of telephone channels involves two unidirectional paths, one for transmission and one for reception. The local loop, which connects a telephone to a local exchange is a *two-wire* (2W) circuit that carries the signals in both transmission directions (Figure 2.6). Even ISDN and *asymmetrical digital subscriber lines* (ADSLs) (described in Chapter 6) use this same 2W local
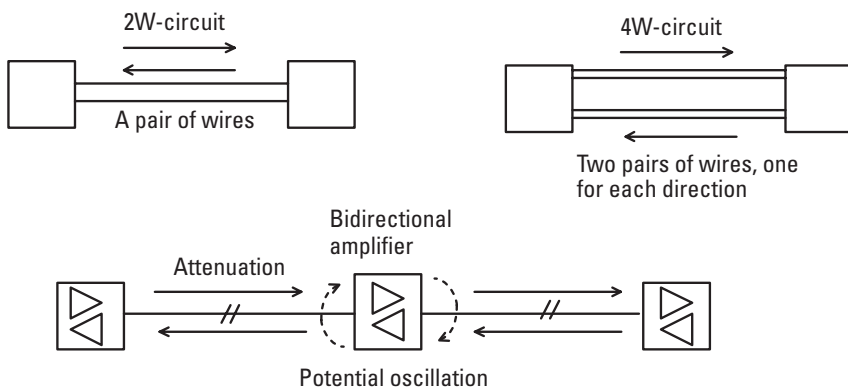
Figure 2.6 2W/4W circuits.

loop. Subscriber loops are and will remain two-wire circuits, because they are one of the biggest investments of the fixed telephone network.

Early telephone connections through the network were two-wire circuits. Longer connections attenuate the speech signal and amplifiers are needed on the line. In two-wire circuits, amplification of a signal may cause oscillation or ringing if the output signal of an amplifier loops back to the input circuit of another transmission direction (Figure 2.6).

The operating principle of electronics in the network is unidirectional and inside the network we use two wires for each direction, or *four-wire* (4W) connections. Four-wire connections are also much easier to maintain than 2W connections because transmission directions are independent from each other and potential oscillation, as shown in Figure 2.6, is avoided. To connect a 2W local loop to a 4W network a circuit called a *2W/4W hybrid* is needed.

We explain the operating principle of the 2W/4W hybrid with the help of transformers. A transformer consists of coils of wires wrapped around an iron object. When an alternating current flows through one coil, it produces a magnetic field in the iron core. This magnetic field generates current to the wires of other coils around the same iron core.

Figure 2.7 shows the 2W/4W hybrid in a subscriber interface of the telephone exchange. Two separate transformers are needed in the hybrid and both of them consist of three similar, tightly coupled windings. In each transformer an alternating current in one coil generates alternating current to all other coils of the same transformer. Spots of coils indicate the direction of the current flow (polarity of the coil). In Figure 2.7 we see that the current of the receive pair generates two currents with opposite polarity through the two coils of transformer T2. These currents have opposite directions in transformer T1; they, or actually their magnetic fields in the iron core, cancel each other, and the signal from the receive pair is not connected to the transmit pair, or at least it is much attenuated. In practice, the balance is not ideal and attenuated signal is connected back, which is heard as an echo from the far end of the telephone circuit if two-way propagation delay of the circuit is long enough. Dashed lines in Figure 2.7 show the main signal paths for received and transmitted speech.

Satellite connections have long propagation delays because of the long propagation distances. Also speech from the digital cellular network to the fixed telephone network suffers long delays because of speech coding (A/D and D/A conversion). The round-trip delays of these connections are longer than 50 to 100 ms, causing a disturbing echo. Hence, in the case of these connections, we have to use special equipment known as *echo cancellers* in the network to eliminate the echo.
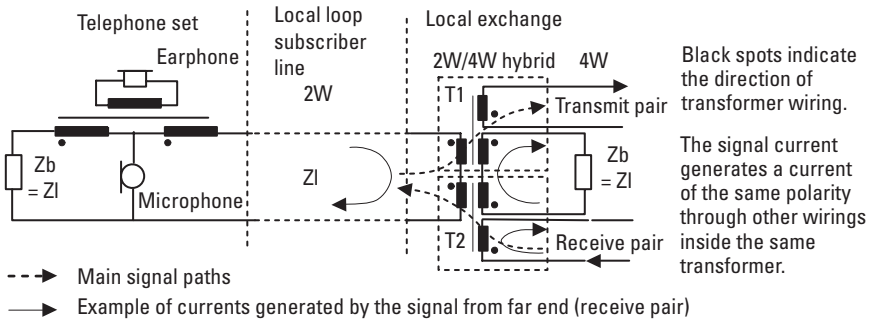
Figure 2.7   Local loop and 2W/4W hybrid.

The 2W/4W hybrid performs the following operations:

- Separates the transmitting and receiving signals.
- Matches the impedance of the 2W local loop to the network circuit.
- Provides a loss to signals arriving on the receiving path, preventing them from entering the transmitting path, which would cause echo.

The ISDN basic rate interface uses bidirectional 160-Kbps data transmission on a 2W circuit (ordinary subscriber loop). There the transmission directions are separated with the help of digital signal processing technology. Many applications use the transformer circuit described earlier together with digital signal processing technology to improve performance.

In every subscriber set quite the same principle as the 2W/4W hybrid is used to attenuate the subscriber's own voice from the microphone to the earphone (Figure 2.7). The reader can imagine what happens when the microphone generates an alternating current in the telephone set of the figure.

2.5   Telephone Numbering

An international telephone connection from any telephone to any other telephone is made possible by unique identification of each subscriber socket in the world. In mobile telephone networks, each telephone set (or subscriber card) has a unique identification number.

The numbering is hierarchical, and it has an internationally standardized country code at the highest level. This makes national numbering

schemes independent from each other. E.164 specifies the structure of international telephone numbers and it is presented in Figure 2.8. In the following sections, we explain the fields of the telephone number shown in Figure 2.8.

2.5.1 International Prefix

An international prefix or international access number is used for international calls. It tells the network that the connection is to be routed via an international telephone exchange to another country. The international prefix may differ from country to country, but it is gradually becoming harmonized. For example, all of Europe uses 00; elsewhere it may be different. If many operators are providing international telephone service, a subscriber may select from among different operators by using an operator prefix instead of 00, for example, in Finland a user would dial 999 for Oy Finnet International.

2.5.2 Country Code

The country code contains one to four numbers that define the country of subscriber B. Country codes are not needed for national calls because their purpose is to make the subscriber identification unique in the world. A telephone number that includes the country code is called an international number and it has a maximum length of 12 digits.

Because there are a few hundred countries in the world, many country codes have been defined by the ITU and the length of them varies from a single digit to four digits (some small areas have an even longer code). Consider these examples of country codes: 1 for the United States and Canada, 49 for
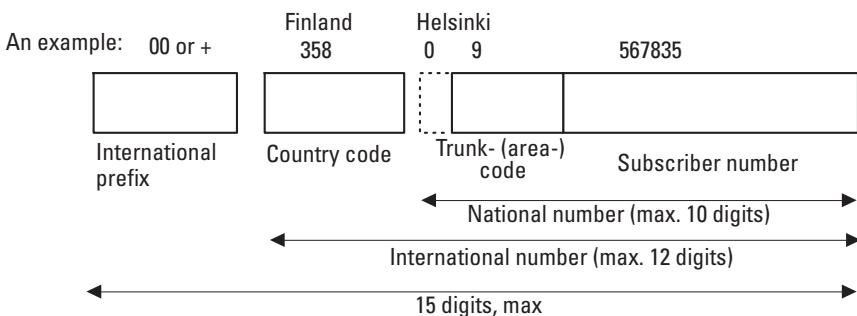
Figure 2.8 The structure of the telephone number hierarchy.

Germany, 44 for the United Kingdom, 52 for Mexico, 358 for Finland, and 1809 for Jamaica.

2.5.3   Trunk Code, Trunk Prefix, or Area Code

The trunk code defines the area inside the country where the call is to be routed. The first digit is a long-distance call identification and other numbers identify the area. The first digit is not needed in the case of an international call because that type of call is always routed via the long-distance level of the destination network.

In the case of cellular service, the trunk code is used to identify the home network of the subscriber instead of the location. With the help of this network code, a call is routed to the home network, which then determines the location of the subscriber and routes the call to the destination.

The trunk code and the subscriber number together create a unique identification for a subscriber at the national level. This is called a national number and its maximum length is 10 digits.

Trunk codes start with a 0 in Europe, but the 0 is not used in calls coming from abroad. In countries where multiple operators provide long-distance telephone service, the subscriber may select an operator by dialing an operator prefix in front of the trunk code. In Finland, two examples of the long-distance operator numbers are 109 for Finnet and 1041 for Song Networks.

2.5.4   Subscriber Number

The subscriber number in a fixed telephone network is a unique identification of the subscriber inside a geographical area. To connect to a certain subscriber, the same number is dialed anywhere in the area. Because of the numbering hierarchy, the subscriber part of the telephone number of one subscriber may be the same as that of another subscriber in another area.

If provision of local telephone service is deregulated (as is the goal in Europe), a subscriber is able to choose a network operator for local calls by dialing a local operator prefix in front of the subscriber number.

2.5.5   Operator Numbers

As the telecommunications business is deregulated, new service providers are beginning to enter on the market. Then in addition to the numbers just described, a subscriber will need to dial additional digits to select a service provider (network operator). As explained earlier, a subscriber may choose a service provider for local calls, long-distance calls, and international calls. The national telecommunications authority defines the operator numbers

used. The national telecommunications authority also defines how calls dialed without an operator number are charged. If the subscriber does not specify the international and long-distance network operators by operator prefix, the network is chosen randomly or according to other rules specified by the national telecommunications authority. The creation of real competition in fixed telecommunications service provision has been successful in many countries. One problem with this situation is that additional dialing of operator prefixes at all levels is required, and another is that the fees for fixed telephone service are too low to make subscribers interested in taking the time to choose a service provider.

For business users, for which monitoring the costs of telecommunications is essential, competition will certainly reduce those costs. To avoid the problem of additional dialing, a business or residential subscriber may make a service agreement with one of the network operators for local, long-distance, and international calls.

2.6 Switching and Signaling

To build the requested connection from one subscriber to another, the network has switching equipment that selects the required connection. These switching systems are called exchanges. The subscriber identifies the required connection with signaling information (dialing) that is transmitted over the subscriber line. In the network, signaling is needed to transmit the control information of a specific call and circuits from one exchange to another.

2.6.1 Telephone Exchange

The main task of the telephone or ISDN exchange is to build up a physical connection between subscriber A, the one who initiates the call, and subscriber B according to signaling information dialed by subscriber A. The speech channel is connected from the time when the circuit was established to the time when the call is cleared. This principle is called the *circuit switching* concept and is different from *packet switching,* which has been used in data networks.

In the past, the switching matrix was electromechanical and controlled directly by pulses from a telephone. Later, the control functions were integrated into a common control unit. Currently, the common control unit is an efficient and reliable computer or a multiprocessor system, including large amounts of real-time software. This kind of exchange is called a *stored program control* (SPC) exchange (Figure 2.9).

Subscriber lines

Exchange

Switching matrix

Common control computer

Lines or channels to the other exchanges (speech channels and signaling channels in the case of channel associated signaling)

Common channel signaling connections to other exchange
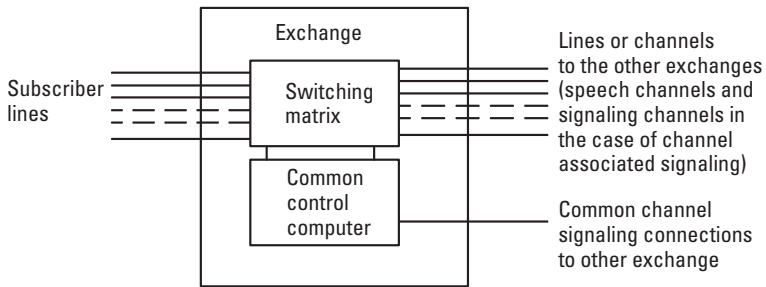
Channel associated signaling: each active telephone call has its own signaling channel between exchanges in addition to the speech channel

Common channel signaling: One data channel between exchanges is used for signaling purposes of all telephone calls. This principle is similar to computer communications where data packets are transmitted between computers.

Figure 2.9 SPC exchange and signaling principles used between exchanges.

Every exchange between subscribers A and B connects a speech circuit according to signaling information that is received from a subscriber or from the previous exchange. If the exchange is not the local exchange of subscriber B, it transmits signaling information to the next exchange that connects the circuit further.

2.6.2   Signaling

The control unit of the local exchange receives the subscriber signaling, such as dialed digits, from the subscriber line and makes consequent actions according to its program. Usually the call is routed via many exchanges and the signaling information needs to be transmitted from one exchange to another. This can be done via *channel associated signaling* (CAS) or *common channel signaling* (CCS) methods (Figure 2.9).

2.6.2.1   CAS

When a call is connected from a local exchange to the next exchange, a speech channel is reserved between exchanges for this call. At the same time another channel is reserved only for signaling purposes and each speech path has its own dedicated signaling channel while the call is connected. This channel can be, for example, a signaling channel in time slot 16 of the primary PCM frame as explained later in Chapter 4. The main phases of signaling between exchanges are shown in Figure 2.10. First the speech channel and the related signaling channel are seized from exchange A to exchange B.
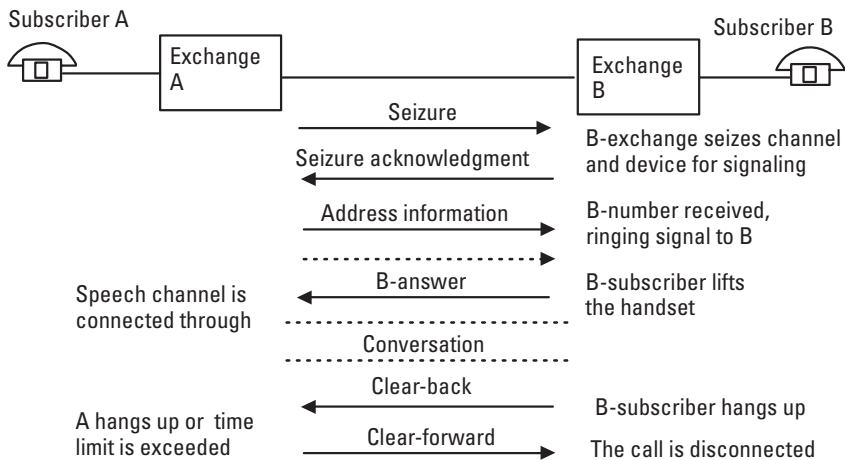
Figure 2.10 CAS between exchanges.

Then the telephone number of subscriber B is transmitted to exchange B, which activates the ringing signal. When subscriber B answers, the speech connection is switched on and the conversation may start.

If subscriber B hangs up first, a *clear-back* (CBK) signal is transmitted from exchange B to A. Exchange A responds with a *clear-forward* (CLF) signal when subscriber A hangs up or when the time constant expires. The call is then disconnected by both exchanges.

Many different signaling systems are used for CAS and some of them include additional signals that are not present in Figure 2.10. Signals that carry signaling information indicated in Figure 2.10 depend on the signaling system in use and they may be, for example, as follows:

- Breaks of the loop between exchanges (loop/disconnect signaling);
- Tones with multiple frequencies, *multifrequency code* (MFC);
- Bit combinations of signaling channel of a PCM frame.

CAS is still used in telephone networks, but it is gradually being replaced with a more efficient standardized method known as CCS.

2.6.2.2 CCS

The modern interexchange signaling system is called CCS. It is based on the principles of computer communications in which data frames containing

information are exchanged between computers only when required. Signaling frames contain, for example, information about the connection to which the message belongs, the address of the destination exchange, dialed digits, and information about whether subscriber B has answered. In most cases only one data channel between two exchanges is required to serve all established calls. This is usually one 64-Kbps time slot of a primary 2- or 1.5-Mbps PCM frame, as explained in Chapter 4, and one channel is usually enough for all call-control communication between exchanges.

A widely used international standard of CCS is called CCS7, also known as *signaling system number* 7 (SS7), CCITT#7, or ITU-T 7, and it is used in all modern telecommunications networks such as ISDN and GSM.

Establishing a call requires the same signaling information as indicated in Figure 2.10, but in the case of CCS the signaling information is carried in data frames that are transferred between exchanges via a common data channel.

In Figure 2.11 we see an example in which an ordinary fixed network subscriber, subscriber A, calls subscriber B when CCS is used between exchanges in the network. The dialed digits are transmitted from subscriber A to the local exchange, as explained in Section 2.3. When a set of digits is received by exchange A, it analyzes the dialed digits to determine to which
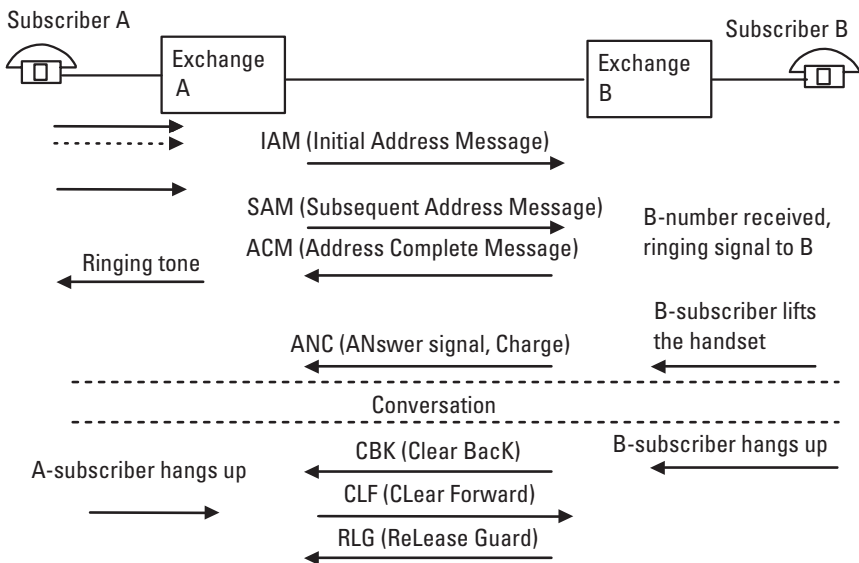
Figure 2.11 CCS between exchanges.

direction it should route the call. From this information it looks up an address of the exchange to which it should send the signaling message for call connection. Then the exchange builds a data packet that contains the address of exchange B. This signaling message, called the *initial address message* (IAM), is then sent to exchange B. The remaining digits that did not fit into the IAM are transmitted in one or more *subsequent address messages* (SAMs).

When all the digits that identify subscriber B are received by exchange B, it acknowledges this with an *address complete message* (ACM), to confirm that all digits have been successfully received. This message also contains information about whether the call is to be charged or not and if the subscriber is free or not. Exchange B transmits the ringing tone to subscriber A and the ringing signal to subscriber B, and telephone B rings.

When subscriber B lifts the handset, an *answer signal charge* (ANC) is sent in order to activate charging. Exchange B switches off the ringing signal and ringing tone. Then both exchanges connect the speech channel through so the conversation can start. When subscriber B hangs up, exchange B detects an on-hook condition and sends a CBK to exchange B. Exchange A responses with CLF signal. All exchanges on the line transmit the CLF message to the next one, and each receiving exchange acknowledges it with a *release guard* (RLG) signal. The RLG message indicates to the receiving exchange that the connection has been cleared and the channel released by the other exchange. It also ensures that both exchanges have cleared the circuit to make it available for a new call.

2.6.3 Switching Hierarchy

During the early years of the telephone, the switching office or exchange was located at a central point in a service area and it provided switched connections for all subscribers in that area. Hence, switching offices are still often referred to as central offices.

As telephone density grew and subscribers desired longer distance connections, it became necessary to interconnect the individual service areas with trunks between the central offices. With further traffic growth, new switches were needed to interconnect central offices and a second level of switching, trunk or transit exchanges, evolved. Currently national networks have several switching levels.

The actual implementation of the hierarchy and the number and names of switching levels differ from country to country. Figure 2.12 shows an example of a possible network hierarchy [1].

The hierarchical structure of the network helps operators manage the network and it makes the basic principle of telephone call routing
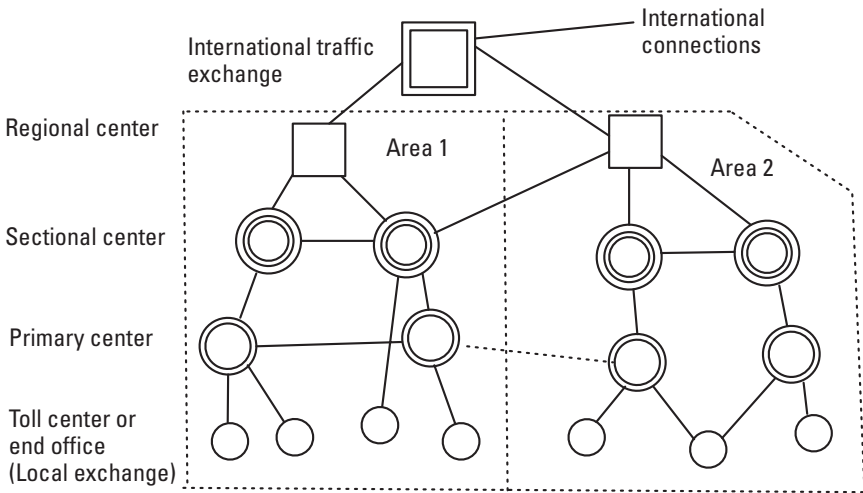
Figure 2.12 An example of switching hierarchy.

straightforward; the call is routed up in the hierarchy by each exchange if the destination subscriber is not located below this exchange. The structure of the telephone number, explained in Section 2.5, supports this simple basic principle of routing up and down in the switching hierarchy.

2.6.4 Telephone Call Routing

Calls that are carried by the network are routed according to a plan, a set of rules. The routing plan includes the numbering plan and network configuration.

2.6.4.1 Numbering Plan

The global rules for the highest-level numbering, country codes, and overall numbering (maximum length and so on) are given by ITU-T. The national telecommunications authority coordinates the national numbering plan. It defines, for example, trunk or area codes and operator prefixes used inside the country. It also defines nationwide service numbers (e.g., emergency numbers). These service numbers are defined to be the same wherever the call is originated and they require additional intelligence from switching systems. Their routing principle is explained later in Section 2.10.4.

At the regional level, the numbering plan includes digits allocated to certain switching offices, exchanges, and the subscriber numbers for subscribers connected to a certain switch.

2.6.4.2   Switching Functionality for Routing

From the received signaling information (dialed digits), a switching system must be able to interpret the address information, determine the route to or toward the destination, and manipulate the codes in order to advance the call properly. This includes the deletion of certain digits and automatic alternate routing. Number conversion may also be needed when, for example, the emergency call dialed with a nationwide short emergency number has to be routed to a regional center that has a different physical telephone number. Some of this intelligence for routing may be stored in a centralized control system from which the exchanges request routing information. This modern network structure is called an *intelligent network* (IN) and is described in Section 2.10.

2.6.4.3   Route Selection Guidelines

The basic routing principle is hierarchical: If the destination does not belong to the subscribers of the switch or of the switches under it, the call is routed upward; otherwise, it is routed to the port toward that destination (Figure 2.13).

In the example of Figure 2.13, a Finnish subscriber makes a call to Stockholm, Sweden, and dials the international prefix "00," country code "46" for Sweden, area code "(0)8" (leaves out zero) for Stockholm, and subscriber number "*xxxxx.*" The international prefix is actually all that the lower-level exchanges in Finland need to know. When exchanges in the
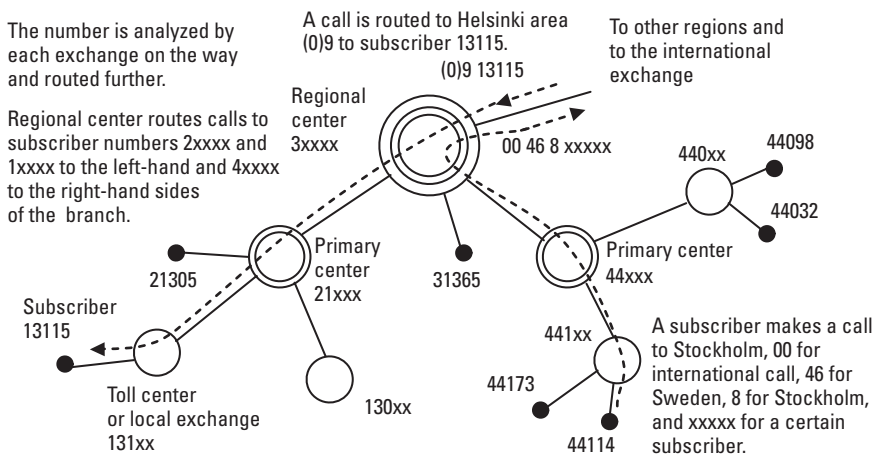
Figure 2.13   Telephone call routing.

switching hierarchy detect it, they route this call up toward the international exchange. The international exchange then analyzes the country code and selects an outgoing route to Sweden.

Another example in Figure 2.13 illustrates routing of a long-distance call from a subscriber in another region. A subscriber in another region dialed "09 13115" for a long-distance call to Helsinki. The first digit "0" tells the exchanges that this is a long-distance call and is to be routed to the regional exchange. The regional center is connected to other regional centers and then routes this call, with the help of other regional centers, to Helsinki according to the next digit, "9." The regional center of Helsinki analyzes the next two numbers, "13," and selects the route down to the primary center where these subscribers are located. (Operator has defined in his numbering plan that the subscriber numbers 2*xxxx* and 1*xxxx* are placed on the left-hand branch from the regional center.) The primary center then checks the following numbers, "131," and notices that this is not "my subscriber" but the destination subscriber is located "below me" and routes the call to the corresponding lower-level exchange, in this example, the local exchange. The local exchange selects the subscriber loop of the telephone number 13115 and connects a ringing signal to the subscriber.

However, modern exchanges can do more than the simple strictly hierarchical routing just introduced. If there is a sufficient volume of traffic, calls may pass by a hierarchy level or may be connected directly to another low-level switch, as illustrated in Figure 2.12. This may be reasonable, for example, if the local exchanges of subscribers A and B are on the opposite sides of the regional border. The telecommunications operator is free to define the detailed actual routing to optimize the usage of the network.

In this section we have described the switching hierarchy of the telephone network and the telephone call routing principle through the exchanges in this hierarchy. In modern networks the actual implementation may be different from this strictly hierarchical routing principle we described. Local telephone exchanges may analyze the whole telephone number, bypass the switching hierarchy, and route the call directly if the destination is a subscriber of a neighbor local exchange. Also, some sets of the telephone numbers have no fixed connection to the physical location of a subscriber loop. The IN technology, which we discuss later in this chapter, connects a dialed logical number and a certain physical telephone number (i.e., subscriber loop).

Deregulation of the fixed telephone business has created another need for increased intelligence in the network. Local network operators have to be able to connect calls to other parallel networks belonging to competing

network operators when subscriber A requires that. The need for this is indicated by the operator prefix dialed by the subscriber as discussed in Section 2.5.

In the next section, we divide the global telecommunications network into three simplified layers in order to clarify their structure and the technologies that are used to implement their required functions.

2.7 Local-Access Network

The local-access network provides the connection between the customer's telephone and the local exchange. Ordinary telephone and ISDN subscribers use two wires, a pair, as a subscriber loop, but for business customers a higher capacity optical fiber or microwave radio link may be required. Many different technologies are used in a local-access network to connect subscribers to the public telecommunications network. Figure 2.14 illustrates the structure of the local-access network and shows the most important technologies in use.

Most subscriber connections use twisted pairs of copper wires. Subscriber cables contain many pairs that are shielded with common aluminum foil and plastic shield. In urban areas cables are dug into the ground and may be very large, having hundreds of pairs. Distribution points that are installed in outdoor or indoor cabinets are needed to divide large cables into smaller
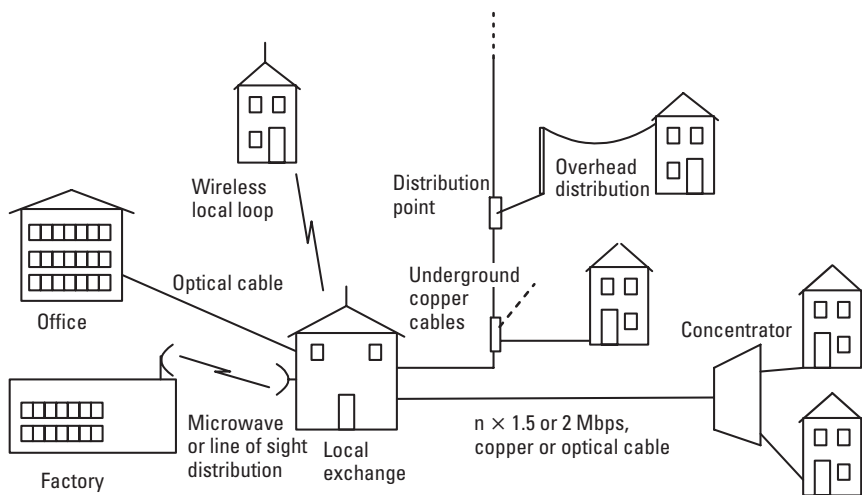
Figure 2.14 An example of a local-access network.

ones and distribute subscriber pairs to houses as shown in Figure 2.14. In suburban or country areas, overhead cables are often a more economical solution than underground cables.

An optical connection is used when a high transmission capacity (more than 2 Mbps) or very good transmission quality is required. A microwave radio relay is often a more economical solution than optical fiber when there is a need to increase data capacity beyond the capacity of an existing cable network. Installation of optical or copper cables takes more time because permissions from landowners and city authorities are required. Installation of cables is also very expensive when they must be sunk into the ground.

One technology for implementation of ordinary subscriber loops for fixed telephone service is known as *wireless local loop* (WLL). WLL uses radio waves and does not require installation of subscriber cables; it is a quick and low-cost way to connect a new subscriber to the public network. With the help of this technology, new operators can provide services in an area where another old operator owns the cables. WLL is also used for replacement of old fixed overhead subscriber telephone lines in rural areas.

When cable network capacity for subscriber connections needs to be increased, it may be more economical to install concentrators, remote subscriber units, or subscriber multiplexers so as to utilize existing cables more efficiently. We use one of these terms to describe the switching capability of the remote unit. Concentrators may be capable to independently switch local calls among the subscribers connected to them. A remote subscriber unit is basically the subscriber interface part of the exchange that is moved away close to the subscribers. Subscriber multiplexers may only connect each subscriber to a time slot (channel) in the PCM frame. The detailed functionality of these systems depends on the manufacturer, but we can say that only those subscribers who have picked up their handsets reserve a channel to the local exchange. Digital transmission between an exchange and a concentrator further improves cable utilization so that two cable pairs serve tens of subscribers.

We have explained the access alternatives shown in Figure 2.14 mainly from a fixed telephone service point of view, but they can also be used to provide access to the Internet. Technologies used for Internet access are explained in Chapter 6.

2.7.1   Local Exchange

Local or subscriber loops connect subscribers to local exchanges, which are the lowest-level exchanges in the switching hierarchy. These are the main tasks of the digital local exchange:

- Detect off-hook condition, analyze the dialed number, and determine if a route is available.

- Connect the subscriber to a trunk exchange for longer distance calls.

- Connect the subscriber to another in the same local area.

- Determine if the called subscriber is free and connect ringing signal to her.

- Provide metering and collect charging data for its own subscribers.

- Convert 2W local access to 4W circuit of the network.

- Convert analog speech into a digital signal (PCM).

The size of local exchanges varies from hundreds of subscribers up to tens of thousand subscribers or even more. A small local exchange is sometimes known as a *remote switching unit* (RSU) and it performs the switching and concentration functions just as all local exchanges do. A local exchange reduces the required transmission capacity (number of speech channels) typically by a factor of 10 or more; that is, the number of subscribers of the local exchange is 10 times higher than the number of trunk channels from the exchange for external calls. The number of required trunk circuits is analyzed in Section 2.12. Figure 2.15 shows some different subscriber connections to a local exchange and the way they are physically installed.

2.7.2   Distribution Frames

All subscriber lines are wired to the *main distribution frame* (MDF), as shown in Figure 2.15, which is located close to the local exchange. It is a large construction with huge number of connectors. Subscriber pairs are connected to one side and pairs from the local exchange to the other. Between these connector fields there is enough space for free cross-connections. Cables and connectors are usually arranged in a logical way considering the subscriber cable network structure and switching arrangements. This fixed cabling stays the same over long periods of time, but connections between sides change daily, for example, because a subscriber moves to another house in the same switching area.

A cross-connection in the MDF is usually done with twisted open pairs that are able to carry data rates up to 2 Mbps. Ordinary subscriber pairs are used for analog telephone subscribers, analog and digital PBX/PABX connections, ISDN basic rate connections and ADSL. ADSL and ordinary analog telephone circuits use the same 2W subscriber loop. Data and speech connections may be used simultaneously and they are separated in the
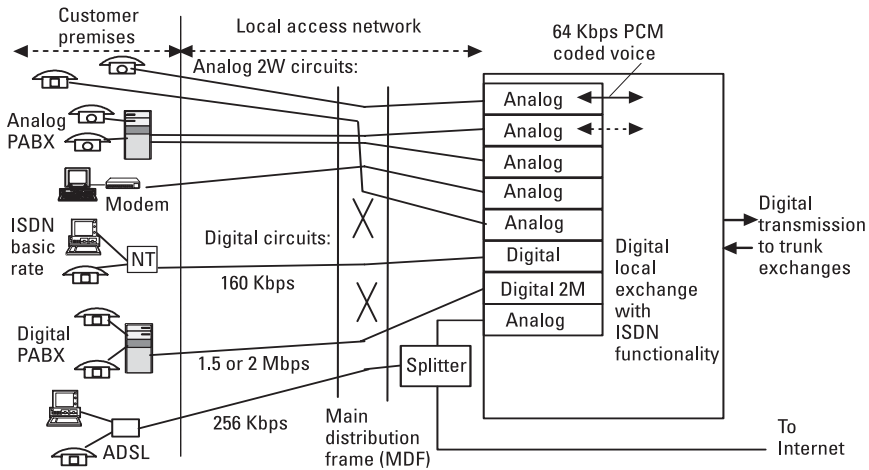
Figure 2.15 Local-access network and digital local exchange site.

exchange where speech is connected to an ordinary analog exchange interface and data are routed to the Internet, as shown in Figure 2.15.

A digital exchange may include both analog and digital subscriber interfaces. For digital private (automatic) branch exchange (PBX/PABX) applications, 1.5- or 2-Mbps digital interfaces are available. If the local switch has ISDN capability, basic rate and primary rate interfaces are available. Ordinary subscriber pairs are used for ISDN basic rate connections (160-Kbps bidirectional) and a *network terminal* (NT) is required on customer premises. The primary rate interface of ISDN (1.5 or 2 Mbps) is used for PABX connections. It requires two pairs, one for each transmission direction, and supports many simultaneous external calls.

In addition to MDF, network operators may use other distribution frames for transmission network management and maintenance. An *optical distribution frame* (ODF) contains two fields of optical fiber connectors. The optical cables of the network are connected to one connector field and the other one is connected to optical line terminal equipment. Cross-connections between two connector fields are created with optical fibers. This allows maintenance personnel, for example, to replace a faulty optical cable connection with a spare one. A *digital distribution frame* (DDF) is a cross-connection system to which digital interfaces from line systems and the exchange (or other network equipment) are connected. With the help of a 1.5- or 2-Mbps DDF, an operator may easily change transmission connections between equipment sites.

DDF may be implemented by *digital cross-connect equipment* (DXC) to which many high-data-rate systems are connected. DXC is managed via its network management interface and an operator may change its cross-connection configuration from a *network management system* (NMS) site. From a remote NMS he may, for example, define to which of the 1.5- or 2-Mbps interfaces a certain 64-Kbps channel from one 1.5- or 2-Mbps interface is connected. Operation of DXC is discussed in Chapter 4.

2.8   Trunk Network

As we saw in Section 2.6, the national switching hierarchy includes multiple levels of switches above local exchanges. Figure 2.16 shows a simplified structure for a network where higher levels than local exchanges are shown as a single level of trunk exchanges. The local exchanges are connected to these trunk exchanges, which are linked to provide a network of connections from any customer to any other subscriber in the country.

High-capacity transmission paths, usually optical line systems, with capacities up to 10 Gbps, interconnect trunk exchanges. Note that a transport network has alternative routes. If one of these transmission systems fails, switches are able to route new calls via other transmission systems and trunk exchanges to bypass the failed system (Figure 2.16). Connections between local and trunk exchanges are usually not fault protected because their faults affect on a smaller number of subscribers.
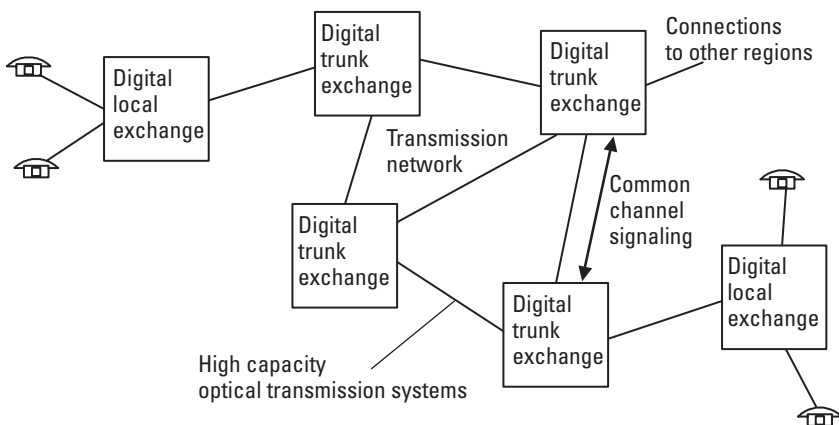
Figure 2.16   Two-layer network and links between trunk and local exchanges.

The transmission systems that interconnect trunk exchanges make up a transmission or transport network. Its basic purpose is simply to provide a required number of channels (or data transmission capacity) from one exchange site to another. Exchanges use these channels of the transport network for calls that they route from one exchange to another on subscriber demand.

The trunk exchanges are usually located in major cities. They are digital and use the international common channel signaling standard SS7 to exchange routing and other signaling information between exchanges. The transmission lines between exchanges have conventionally carried TDM telephone channels, as explained in Chapter 4. Currently the use of IP networks for connections among exchanges is increasing and it requires *media gateways* (MGWs) between the exchange and IP network to take care of signaling and real-time transmission through the IP network.

2.9   International Network

Each country has at least one international switching center to which trunk exchanges are connected, as shown in Figure 2.17. Via this highest switching hierarchy level, international calls are connected from one country to another and any subscriber is able to access any of the other more than 2 billion subscribers around the world.

High-capacity optical systems interconnect international exchanges or switching centers of national networks. Submarine cables (coaxial cable or optical cable systems), microwave radio systems, and satellites connect continental networks to make up the worldwide telecommunications network.

The first submarine cable telephone system across the north Atlantic Ocean was installed in 1956, and it had the capacity of 36 speech channels. Modern optical submarine systems have a capacity of several hundred thousand speech channels and new high capacity submarine systems are put into use every year. In addition to speech, submarine systems carry intercontinental Internet traffic, which is estimated to take most of the capacity of the new systems under installation. Submarine systems are the main paths for intercontinental telephone calls and Internet communication. Satellite systems are sometimes used as backup systems in the case of congestion.

We described the common structure of the global telecommunications network without separating the different network technologies. We need different network technologies to provide different types of services, and the telecommunications network is actually a set of networks, each of them having characteristics suitable for the service it provides. In the next section we
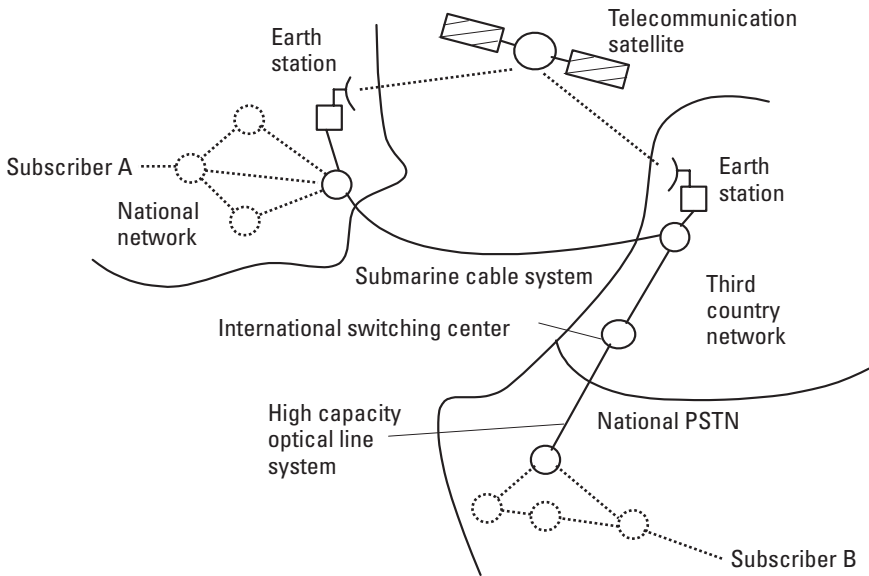
Figure 2.17 The international network.

describe briefly the most important network technologies, some of which are discussed in more detail in later chapters.

2.10 Telecommunications Networks

Up to this point, we have explained the operation of the public switched telecommunications network and used the conventional telephone networks as an example. However, the public network contains many other networks that are optimized to provide services with different characteristics. We review these different network technologies in this section.

We can divide telecommunications networks into categories in any of many different ways. If we consider the customers of networks and the availability of services, there are two broad categories: public networks and private or dedicated networks.

2.10.1 Public Networks

Public networks are owned and managed by telecommunications network operators. These network operators have a license to provide telecommunications services and that is usually their core business. Any customer can be

connected to the public telecommunications network if he has the correct equipment and an agreement with the network operator.

2.10.1.1 Telephone Network

The PSTN is the main public network in use. Sometimes we refer its service to as POTS if we want to distinguish ordinary fixed telephone service from other services provided by telecommunications networks today. In addition to voice communications between fixed telephones, data can be substituted for speech with the help of a voice-band modem. ISDN, introduced later, is considered the next evolutionary step after PSTN.

2.10.1.2 Mobile Telephone Networks

Mobile or cellular telephone systems provide radio communications over the local access part of the network. They are regional or national access networks and connected to the PSTN for long-distance and international connections. We introduce mobile networks in Chapter 5.

2.10.1.3 Telex Network

This is a telegraph network that allows teleprinters to be connected by means of special dedicated switches. The bit rate of telex is very slow, 50 or 75 bps, which makes it robust. It was once widely used but its importance has been reduced as other messaging systems such as electronic mail and facsimile have reduced its market share.

2.10.1.4 Paging Networks

Paging networks are unidirectional only. Pagers are low-cost, lightweight wireless communication systems for contacting customers without the use of voice. Simple pagers just say "beep," but sophisticated pagers can receive large amounts of text and display the e-mail message on a screen. The importance of paging systems has been reduced in countries where penetration on cellular systems, providing text-messaging service, is high.

2.10.1.5 Public Data Networks

These networks provide leased point-to-point connections or circuit-switched or packet-switched connections. Leased point-to-point lines are often an economical solution for connections between the LANs of corporate offices in a region. Circuit-switched networks dedicated to data transmission are not widely used today. Packet-switched data service is provided by the X.25 network worldwide. It operates according to the X-series recommendation of ITU-T, but the marketing names of X.25 networks differ from

country to country, for example, Auspak is used in Australia and Finpak in Finland. These networks were developed to provide commercial data communication service and they provide charging functionality so that the customer bill may be based on the amount of transferred data. The importance of these networks has been reduced because of expansion of the Internet. Internet e-mail has replaced X.25 e-mail. Public wireless data networks, such as *general packet radio service* (GPRS), have been implemented to provide data services for mobile users. *Wireless LAN* (WLAN) is another technology that is used to provide data service in hot spots, such as airports.

2.10.1.6 Internet

The Internet is a worldwide packet-switched network developed from the ARPANET, which in turn was developed in the late 1960s by the U.S. Department of Defense. The ARPANET grew until it became a wide-area computer network called the Internet, which was used in the 1970s and 1980s mainly by academic institutes such as universities. Because of its history the Internet does not provide charging functions, and customer billing is usually based on the access data rate and fixed monthly fee. In the first half of the 1990s the user-friendly graphical user interface WWW was introduced; since then the use of the Internet has expanded very rapidly. Currently, the Internet is the major information network in the world, and many *Internet service providers* (ISPs) have sprung up to provide Internet services for both businesses and residential customers. The expansion of the Internet continues, and the evolving commercial services (e.g., electronic shopping), the new access technologies (such as *x*DSL, discussed in Chapter 6), and integrated speech and video services will further increase its importance in the future.

2.10.1.7 ISDN

The current telephone network is gradually developing into ISDN, in which all information is transmitted in digital form from end to end. With the help of some hardware and software updating, modern digital telephone exchanges are able to provide ISDN service. The main hardware modification required is the replacement of analog subscriber interface units with digital ones, as shown in Figure 2.18.

The ordinary two-wire subscriber loop of the telephone network is upgraded to the basic rate access of ISDN by an NT on the subscriber premises and by a basic rate interface unit and ISDN software in the local exchange. The bidirectional data rate in the subscriber loop is 160 Kbps, which carries 144 Kbps of user data and additional framing information. With the help of

Ordinary 2W subscriber line
with bidirectional 160 Kbps
digital transmission

NT

ISDN
exchange

Network
terminal

ISDN-
basic
rate
interface
unit

Max. 8 subscriber
equipment, two of
which can communicate
at the same time

Routing of B-channels, 64 Kbps, is independent.
D-channel, 16 Kbps, is used for signaling.
Total information rate is 144 Kbps, which makes
160 Kbps when framing information is added.

Figure 2.18 ISDN basic rate interface.

framing information, the receiving end is able to distinguish different channels from the data stream. User data contain two independent 64-Kbps circuit-switched user channels, B channels, and a 16-Kbps signaling channel, the D channel. Subscribers may use user channels, B channels at 64 Kbps, for ordinary speech transmission, data, facsimile, or videoconferencing connections.

Subscribers may use both B channels independently at the same time and dial them up independently, for example, using one of these channels for a telephone call and another for an Internet connection. For Internet surfing B channels can be combined to provide a single 128-Kbps data rate connection. ISDN provides a reliable 64/128-Kbps connection end to end, which is much more than that available for subscribers using a voice-band modem over an ordinary analog telephone circuit.

Users may connect up to eight terminals to a network terminal and two of them may be in use at the same time. The advantages of ISDN over the analog telephone service are a higher data rate and the availability of two connections at the same time. ISDN technology has been available for some time but its usage has been low because of high tariffs in the past. Today operators offer attractive tariffs and the increased demand of better Internet connections in particular has increased ISDN's popularity to some extent. On the other hand, higher rate access technologies, such as *x*DSL and cable modems, provide better performance and they have cut the growth of ISDN.

However, the existing low-cost ISDN technology makes it feasible for network operators to provide ISDN connections sometimes at a lower cost than two conventional analog telephone connections.

2.10.1.8 Radio and Television Networks

Radio and television networks are usually unidirectional radio distribution networks for mass communications. Traditionally, the operators of these networks have not provided dial-up bidirectional telecommunications services. Access to these networks is currently available in urban areas via cable TV networks built by cable TV operators. These operators have not been allowed to provide other telecommunications services and their wideband cable network to homes has not supported bidirectional communication. As the deregulation of the telecommunications business has proceeded, these operators have become active in providing other telecommunications services as well, especially fixed telephone service and high-data-rate Internet access.

To provide interactive services, the cable TV networks need to be upgraded with the technologies that allow subscribers not only to receive TV and radio signals, but to transmit data to the network. Most of the investment was already made when wideband cables were installed. This existing medium is especially attractive for providing Internet service to every home connected to a cable TV network. Typically, a data connection made via a cable TV network is shared between many home users; that is, there is no physically separate connection to every home as we have in the case of ISDN or *x*DSL. This service is has often attractive tariffs because of shared investments, but it may suffer from temporary congestion when many users happen to be active at the same time.

2.10.2 Private or Dedicated Networks

Private networks are built and designed to serve the needs of particular organizations. They usually own and maintain the networks themselves. Services provided are a tailored mix of voice, data, and, for example, special control information.

2.10.2.1 Voice Communication Networks

Examples of private dedicated voice networks are those used by the police and other emergency services and taxi organizations. They are called *private* or *professional mobile radio* (PMR). Railway companies also have private telephone networks that use cables that run alongside the tracks.

2.10.2.2 Data Communication Networks

Data communication networks are dedicated networks especially designed for the transmission of data between the offices of an organization. They can incorporate LANs with mainframe computers feeding information to the branch offices. Banks, hotel chains, and travel agencies, for example, have their own separate data networks to update and distribute credit and reservation information.

2.10.3 Virtual Private Networks

It is very expensive for an organization to set up and maintain its own private network. Another choice is to lease resources, which are also shared with other users, from a public network operator. This *virtual private network* (VPN) provides a service similar to an ordinary private network, but the systems in the network are the property of the network operator.

In effect, a VPN provides a dedicated network for the customer with the help of public network equipment. As companies concentrate more and more on their core businesses, they are willing to outsource the provision, management, and maintenance of their telecommunications services to a public network operator that has skilled professionals dedicated to telecommunications.

The principle of VPN is used for voice services such as corporate PBX/PABX networks. In this case the network that interconnects the offices of a company uses (voice or 56/64 Kbps) channels from the public network that are leased from a public network operator.

An important application of VPN is intranet use. An *intranet* is a private data network that uses open Internet technology. Physically, an intranet may be made up of many LANs at different sites. To interconnect these LANs, a VPN is established to provide data transmission between sites through the public Internet network. Note that the Internet uses the packet-switching principle and there are no physically separate channels for each VPN as in the previously explained voice VPN. Because the packets are not separated into dedicated point-to-point channels, security risks arise when the public Internet is used for interconnections instead of leased lines or a circuit-switched network such as ISDN. To overcome this problem, *firewalls* are used in an intranet at the interface between each LAN and the public Internet. The firewalls perform the authentication duties for the communicating parties and they encrypt and encapsulate data for transmission through the public Internet from one office to another. A dedicated secure data pipe through the Internet is established with the help of encapsulation

and ciphering and then the Internet can be used instead of a more expensive leased or circuit-switched data connection.

Another network related to an intranet is an extranet. An *extranet* is connected between selected users of the Internet and an intranet. These external users of a private intranet may be, for example, customers or material suppliers. Like an intranet, an extranet uses Internet technology, and for security reasons firewalls or other security gateway arrangements are used for user authentication purposes and data encryption.

2.10.4 INs

A conventional telephone network is able to establish a connection only to a socket that is identified by the number of a B subscriber. There is no "intelligence" in this kind of operation; dialing a certain number makes every time a connection to a certain socket. Connection setup is always done in the same way, whether the intended B subscriber is available or not.

In the old days, a human operator performed the switching process manually on a switchboard. If an operator knew that the called party was presently visiting his neighbor, she might connect the call directly to the neighbor's phone. There was some "intelligence" in the network that improved accessibility. In a modern telecommunications network this intelligence is implemented with help of IN technology.

The IN is an ordinary digital telephone network with some additional capabilities like flexible routing of calls and voice notifications. Traditionally, a telephone number has been the identifier of a certain physical subscriber line and a socket. In an IN the physical number and service number have no fixed relation and may change with time. For example, emergency service may be available at daytime in multiple locations but at nighttime only in one location of the area.

2.10.4.1 Distributed Intelligence

Network operators implement supplementary services, such as call forwarding, to assist subscribers in making successful calls. This increases the number of successful calls, the utilization of the network, and, as a consequence, the network operator's revenue from call fees. We can implement these services by updating corresponding functions to each local exchange. Examples of supplementary services include the following:

- *Call forwarding* permits you to direct incoming calls to another telephone. Forwarded calls are regarded as being made from your home telephone and will therefore be charged to the telephone bill of the subscriber who has forwarded the call.

- *Call waiting* means that, during a call in progress, a subscriber is notified of an incoming call. You hear the message as a faint tone in the receiver, while the caller simultaneously hears a normal ringing tone. You can alternate between these two calls.

- *Automatic callback* can be used when the number you are trying to call is busy. A subscriber notifies the system that you want to have a call established when the called party becomes free and she will be informed when this happens. When the subscriber then lifts the receiver, the number will be automatically dialed again.

- *Abbreviated dialing* permits a subscriber to specify short numbers that correspond to complete telephone numbers you use most frequently. These short numbers can be used by all home telephones that are connected to the same subscriber loop.

- *Screening of incoming and outgoing calls* allows a subscriber to specify which telephone numbers he does not want to receive calls from or make calls to. This service is implemented by the telephone service provider according to a customer request. A subscriber may, with the help of this service, avoid charges that may be very high when expensive service numbers are called from his telephone.

Implementation of supplementary services in local exchanges is reasonable because these services are related to only one subscriber connected to one exchange. A subscriber is also able to modify the service and there is no need to transfer service information to other exchanges. However, some services should be available in all exchanges. Examples of this include use of the same emergency number all over the country and establishment of nationwide service numbers. Calls to these numbers are to be routed to one physical telephone number depending on where the call is initiated or time of day. As more and more of these kinds of services have been introduced, the updating of new services to many exchanges has become a great burden to the network operator. The IN structure was developed to help network operators and service providers introduce, update, and develop new services in a more efficient way.

2.10.4.2 Centralized Intelligence

The basic structure of an IN, illustrated in Figure 2.19, is based on centralized intelligence. With central intelligence, control information is stored in a central place and the same information is available for all exchanges in the network. Exchanges request information when they need it for call handling. The great advantage of the IN concept is that when a new service is

The structure of intelligent network

SMS: service management center, for the updating of services or the introduction of new ones

SCP: service control point, which gives routing and charging information to switches

DB: database, stores the service information, for example, number conversion for call transfer

SSP: service switching point, telephone exhange which requests routing information from SCP if IN-number is detected

IP: intelligent peripheral, gives the voice notifications if required

Figure 2.19 The structure of the IN.

introduced or a service is updated, all exchanges in the network are able to provide the modified service immediately.

2.10.4.3 Structure of the IN

IN technology makes provision of new services efficient with the help of control data that are centralized and available to all switches. Otherwise, service information would need to be updated to all exchanges when a change is made. Figure 2.19 shows the main network elements of an IN.

The *service management system* (SMS) provides tools for introduction of new services and service updates. The *database* (DB) contains control information, such as emergency numbers and corresponding physical numbers, for the *service control point* (SCP), which controls *service switching point* (SSP) exchanges. The *intelligent peripheral* (IP) is a system that provides voice notifications when required, and the *service transfer point* (STP) is an intermediate exchange, which routes signaling messages between the SSP and STP.

A certain range of telephone numbers is reserved for IN services only. When a SSP, which performs the functions of an exchange, detects an IN service number, it requests routing information from the SCP. The SCP then provides information about how that call should be handled.

In principle, we could implement all intelligence in the SCP and its database could store all the routing information. This would require heavy signaling between the switching points and the SCP. In practice, the services that do not require a centralized database are implemented in switching points to reduce the load on the SCP and the signaling connections between SCP and SSPs.

Some examples of IN services follow:

- *Universal access number:* A company with several offices in different parts of a country may have the same number throughout the country. Each call is automatically connected to the office closest to the calling subscriber (SSP transfers caller's number to SCP). The cost of the call is the same no matter to which office the call is connected.

- *Premium rate services:* Information provision over the phone, for instance, doctor and layer services. The service provider charges subscribers via the telephone bill. The charge is dependent on the called service number.

- *Freephone:* Companies that want to provide free customer service use this service in which the receiver pays for the call.

- *Credit card call:* A service user can pay with his or her credit card by dialing his or her account number and identity code.

The modern telecommunications networks using IN technology provide many other services and a few new ones appear annually. An example of these is inexpensive home-to-mobile and mobile-to-home calls for which you dial a specific number given by an operator. Another example is a card service for which a serviceperson dials a specific service number and security code and the network operator charges his or her employer instead of the telephone from which he or she is calling.

One category of services implemented with the help of IN technology is value-added services. This term refers to the services that give additional value, not just point-to-point telephone conversation. Separate service providers, not the telecommunications service provider, often provide these services. Examples of value-added services are telebanking, telephone doctor or lawyer services, and participation to TV games. IN technology provides flexible routing and service-specific charging for these services.

In previous sections we described the structure and operation of the telephone network and we have also looked at different network technologies that we need to provide different services. In the following section we look at how all of this fits together.

2.10.5 Public Switched Telecommunications Network Today

The overview of the modern public switched telecommunications network is presented in Figure 2.20. The structure and functionality of the network are only reviewed here because most of the elements in the figure are discussed in
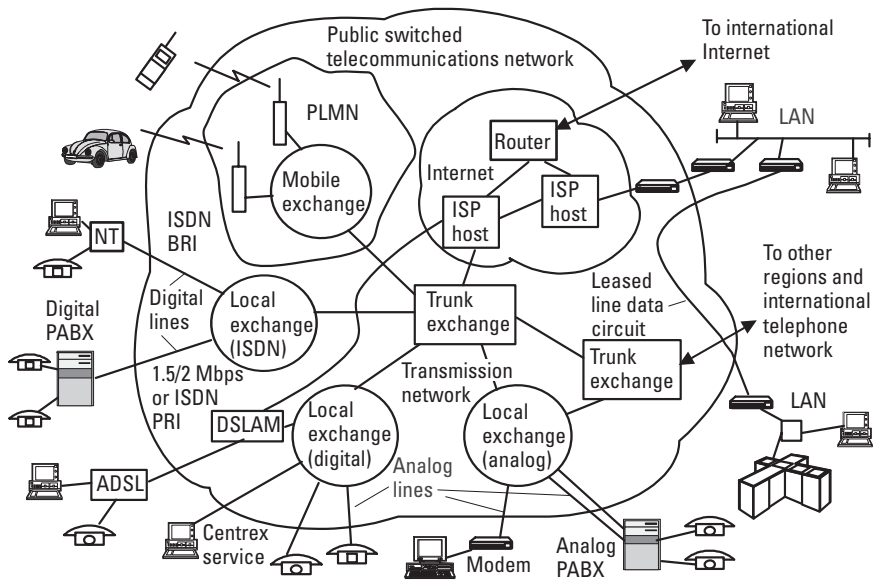
Figure 2.20 Overview of the public switched telecommunications network.

other sections of this book. Figure 2.20 presents a simplified diagram of a regional or national PSTN that has connections to the global Internet and PSTN. The network contains the *public land mobile network* (PLMN), which provides wireless access for cellular subscribers and is connected to the PSTN/ISDN network at the trunk exchange level.

Internet users are connected to the global Internet via the hosts of their ISPs. Networks of national ISPs are connected and this interconnection is extended to the networks of ISPs of neighboring countries, and these networks together make up the global Internet. Figure 2.20 shows two main methods for accessing the Internet. A telephone or ISDN network is used for dial-up connections and ADSL provides permanent higher rate Internet service.

Some different means of accessing telecommunications networks are also shown in Figure 2.20. Digital PBX/PABX is connected to a local exchange with a 1,544/2,048-Kbps digital line that has the capacity of 23/30 simultaneous calls. This connection is called the primary rate interface in the case of ISDN. PBX/PABX is a dedicated small exchange that provides telephone service to the personnel of a company. Analog PBX/PABX uses analog telephone lines, one for each simultaneous external call. Each analog line (twisted pair) carries one telephone call with signaling. This analog signaling is close to the ordinary analog subscriber loop signaling that we described previously.

The corporate-wide PBX/PABX service can also be implemented without any equipment investments in the company, that is, without physical PABX equipment. Network operators provide a service called Centrex and for that the public exchange is programmed to behave as a PBX/PABX. One of the subscriber lines is set to operate as a switchboard line and the others make up a user group with abbreviated dialing and other PBX/PABX services.

For data communication via an analog network or digital network with analog subscriber interfaces, a modem is required. The term *modem* comes from modulator/demodulator and it transmits data through a speech channel in voice frequency tones. If a subscriber has ISDN service, which is fully digital, no modem is needed and an end-to-end bidirectional 64- or 128-Kbps digital circuit is available with the help of a network terminal that takes care of the digital bidirectional transmission over the subscriber loop. For active Internet users who require continuous connection or higher data rates, circuit-switched services are expensive because the cost is based on the duration of the call and they do not provide high enough performance. An attractive access method for these types of users is ADSL, which provides data rates up to a few megabits per second with a fixed monthly fee.

In Figure 2.20 one office site of a company has high-data-rate access to its ISP. All employees have access to the Internet via the company's private LAN. Leased lines, which interconnect two offices in Figure 2.20, are often the most economical solution for high-data-rate circuits that are needed, for example, for LAN interconnections. Different options for data connections are discussed in Chapter 6.

As we have seen, telecommunications networks contain a huge number of different complex systems that are located in multiple sites. In the old days, when the structure of the network was simple, most of the equipment sites had personnel to keep systems operational and they carried out fault location and performed needed maintenance operations. Nowadays systems are so numerous and so complicated that this way of network operations and maintenance is not possible anymore and implementation of automated network management tools is mandatory for all network operators. The following section gives an overview of the importance of network management and of the standardized structure of network management.

2.11 Network Management

The importance of network management has grown together with the size and complexity of the telecommunications network. The standardization of

this area is not as advanced as the standardization of telecommunications systems that carry the actual traffic and provide the services. Efficient network management is a key tool in helping a network operator improve services and make them more competitive.

2.11.1 Introduction

Traditionally, systems that take care of control and supervisory functions in a telecommunications network have been known as *operation and maintenance* (O&M) systems. Nowadays we prefer to use the term network management system because the functions performed by network management systems include much more than those supported by the conventional O&M systems.

Operation functions cover subscriber management functions and enable the network operator, for example, to collect charging data and move and terminate subscriptions. Operation also includes traffic monitoring and controlling the network in such a way that the risk of overload is minimized, for example, by switching traffic from overloaded connections to other systems.

Maintenance includes monitoring of the network and, when a fault occurs, corrective actions are performed. Bit error rates and other parameters are continuously measured for the early detection of faults. When a fault is detected, the operator's staff starts troubleshooting in order to localize the fault. This used to be quite a difficult task because it was done manually and many systems may detect a fault even when the actual fault may be in only one of them or even somewhere else. Maintenance, like other network management functions, has become more and more computerized, making fault location easier and quicker with the help of centralized management systems that provide graphical information about the network's condition.

2.11.2 Who Manages Networks?

Corporate networks are private networks containing LANs interconnected by circuits provided by a public telecommunications network operator. We can divide corporate networks into two main areas of network management responsibility: local networks in corporate sites and interconnections between sites implemented in a public network that provides interconnections as shown in Figure 2.21. The corporate networks, LAN1 and LAN2 in Figure 2.21, are managed by people responsible for network operation inside a company.
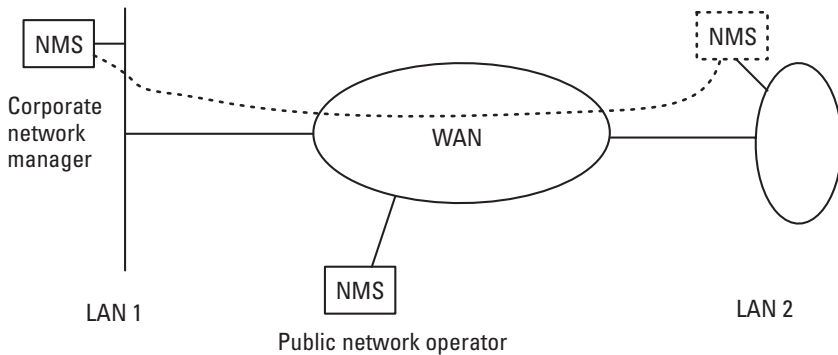
Figure 2.21 Management responsibility of a corporate network.

Network management responsibility is often divided hierarchically. Local or site managers only take care of LAN networks at each office. A centralized organization of the company manages the usage and availability of *wide-area network* (WAN) connections between sites. A centralized organization offers service to business units at various sites and optimizes the utilization of expensive long-distance or even international WAN connections.

The main concerns of network managers of a company include these:

- Network change management (hardware updates);
- The location and repair of malfunctions;
- Software updates and version control;
- Network security.

Most network elements of LANs provide network management functions via a standardized management interface. This open standard is known as the *Simple Network Management Protocol* (SNMP). Software packages for centralized management workstations for LANs are commercially available.

The public network operator manages the public network in order to be able to provide reliable service to customers. Network optimization to avoid unnecessary investments as well as quick repairs in the case of faults is important. Short delivery times of leased-line circuits are an important competitive advantage today, and a network operator can make delivery time shorter with the help of sophisticated network management tools.

In addition to private network management needs, accounting functions are needed in a public network for switched circuits. For example, in

the case of packet-switched service, the amount of transferred data is recorded to generate bills to customers. Accounting functions of the Internet are very limited but in packet-switched cellular networks, such as in the *general packet data service* (GPRS) of the GSM, accounting based on the amount of transferred data is implemented.

Public networks contain many different technologies and the operator's organization is usually divided into different responsibility areas, such as transmission, telephone exchanges, leased-line data networks, and packet-switched data services. Today these organizations usually have their own dedicated and incompatible network management systems, probably with some kind of geographical hierarchy, and the integration of these is an important issue for the future. At least some level of integration is needed because, for example, all services usually use the same transmission network. To solve this problem, ITU-T has defined a common management concept that is known as the *telecommunications management network* (TMN). In the following section we describe the *data communications network* (DCN), which belongs to the TMN concept and is responsible for the transmission of management data.

2.11.3 DCN

Not only different networks, but even network elements (equipment), may have their own O&M systems that may be incompatible today. As a consequence, if a fault occurs in the network, the network operator's personnel may have to use several different O&M systems for fault localization. ITU-T has worked a long time to define a vendor-independent network management concept. It is called TMN.

In ITU-T's TMN concept, the transmission of management data between management workstations and network elements is separated from the transmission of user data as shown in Figure 2.22. The transportation network of management data is called the DCN.

Even though DCN is supposed to be a logically separate network from the actual telecommunications network, the management messages often use the same network as the actual telecommunications services. Most transmission systems, for example, *synchronous digital hierarchy* (SDH) as described in Chapter 4, provide data channels for network management purposes. This requires careful planning of the DCN because a fault on a transmission link may disturb management messages that are necessary for fault localization. Therefore, the DCN should be designed to be as independent as possible from the network that transmits user data.
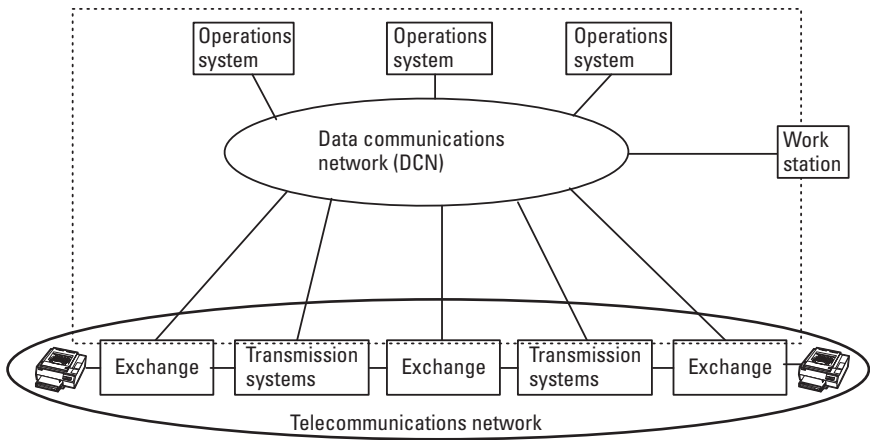
Figure 2.22 DCN.

Sometimes a network operator can physically separate management data from user data by using another independent network for management links. For example, the packet-switched X.25 network may be used for telephone network management. The use of another network may also be feasible to implement redundant routes to DCN, that is, the management data are sent via another connection when the one in use fails.

2.11.4 TMN

The overall management concept that ITU-T has defined is known as TMN. The standardization of TMN is aimed at covering all aspects to make the centralized O&M of telecommunications networks possible in a multivendor environment.

The complete standardization of TMN is designed to cover the following specifications:

- *Physical architecture of TMN:* what systems are needed in TMN and how they are interconnected;
- *Interface protocols:* how network elements and management systems exchange information (the structure and types of messages);
- *Management functions:* what functions in the network elements the network management system should be able to access;
- *Information model:* for each different system in the network, how each manageable function (in detail) is described in management messages.

The recommendations for the TMN concept, approved by ITU-T, define the physical architecture of TMN as shown in Figure 2.23. TMN is understood to be separate from the actual telecommunications network, though network systems have to provide the management interfaces and management functions that they are able to perform. The physical architecture of TMN (Figure 2.23) contains these elements:

- *Operations system* (OS) for centralized network management;
- *Data communications network* for management data transfer;
- *Mediation devices* (MD) to adapt proprietary management interfaces to Q3 interfaces under standardization;
- Management functions integrated in the *network elements* (NEs) of the telecommunications network.

Management areas that TMN takes care of are called FCAPS functions, as listed next. The management system performs or is used to perform following actions:

- *Fault management:* collects alarm information and takes corrective action; detects a system malfunction and carries out measurements to locate the fault.
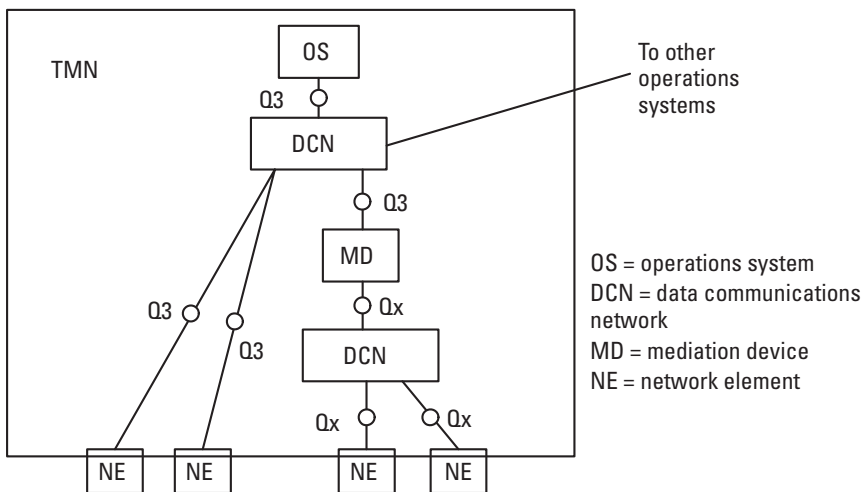
Figure 2.23 Physical architecture of the TMN.

- *Configuration management:* changes the configuration of network elements, for example, disconnects a subscriber who has not paid a bill.

- *Accounting:* sets accounting functions in network elements.

- *Performance:* measures performance of the network to detect faults and bottlenecks in advance.

- *Security:* detects security threats, for example, collects data about users of a corporate network that frequently provide wrong security codes to detect hackers.

The most important and most difficult standardization issue has been the specification of the highest layer of the management interface, Q3. Lower-level protocols, like the physical network that carries actual data and formats messages, are already standardized, but detailed information models are not. The specification work of information models is an endless task, because new systems require their own models and an update to a system often requires a revision of the information model. The information model defines the managed objects (manageable resources) of a system and their relationships. The specification of an information model is mandatory before we can talk about vendor-independent network management.

The information model is specified by the *management information tree* (MIT) or *management information base* (MIB), which defines all managed objects in a system. The managed objects contain all resources that the management system can access. Each managed object has a unique identification that consists of a sequence of names (or numbers) starting from the root and having multiple options at each level. For example, at the second level after the root we have one branch for the ISO (1) and another for the ITU-T/CCITT (0) and an object identifier contains ISO if this is specified to be the right path to our system and its managed objects. The highest levels of the MIT are standardized, but the compatibility of the systems from different vendors requires detailed standardization down to the managed object and its behavior.

For example, if we want to get information about whether subscriber 1 of an exchange is busy, we must have a complete specification of what kind of message, transmitted to the exchange, will produce the wanted response regardless of the manufacturer of that exchange. The lower-level protocols define the structure of the messages, and the information model must specify in detail the information content of the management message with which the

network element responds. For example, all exchanges should respond with exactly the same message if subscriber 1 is busy.

Much work remains to standardize the network management functions of the present systems in the public telecommunications network and new systems require their own standards for network management. However, the Internet community has achieved detailed MIB specifications for many LAN and Internet systems. This has made many NMS software tools, which can manage multivendor local networks, available for LAN environments.

In this chapter we have looked at telecommunications networks, their structure, and functionality; we also introduced network management, which network operators use to improve the performance of their networks and to maintain their network in an effective way. Telecommunications network operators who build up and maintain their network have to provide good performance service at as low an investment level as possible if they are to be competitive. Their problem is how to minimize investment but still keep customers happy. To find out where they should invest and what the bottlenecks of the network are, they continuously perform traffic engineering, which is introduced in the next section.

2.12 Traffic Engineering

Traffic engineering is a key issue for telecommunications network operators trying to keep customers (subscribers) happy while minimizing network investments. Nowadays, network operators have to pay more and more attention to these aspects because of increasing competition in the telecommunications services market. The capacity of the network (e.g., number of channels between exchanges, exchange sizes, number of radio channels in a cellular network) should be increased where the bottlenecks of the network are found. Therefore, the utilization of the network is continuously measured and traffic demand in the future is estimated. Then, based on these estimates, the capacity of the network can be increased before severe problems occur.

An important capacity planning method is based on theoretical analyses of capacity demand and introduction to these calculations is given next.

2.12.1 Grade of Service

How happy subscribers are depends on the *grade of service* (GoS, availability or quality of the service) they receive. The GoS depends on the network capacity that should meet the service demand of the customers. Here we

analyze only GoS for circuit-switched service and the most important factor in our study is whether the call is successful or blocked. System faults, error rates, and other quality measures are not considered here. We instead concentrate only on the evaluation of the blocking probability. In Figure 2.24 blocking occurs if more than n subscribers make external calls at a time. For the probability of unsuccessful calls, operators define the target value, the highest probability of an unsuccessful call that they assume to be acceptable for their customers. The smaller this probability is, the more capacity they have to build into the network.

Another factor we could use to define GoS is how long the subscriber has to wait until the service becomes available. We could design the network to keep customers in a queue until, for example, a transmission channel becomes free. This factor is also essential to those who plan the telephone service where a person answers incoming calls (e.g., switchboard service of an enterprise, customer service telephone).

2.12.2 Busy Hour

Network capacity planning is based on the so-called busy hour traffic intensity, and at other times the GoS is typically much better. *Busy hour* is an hour in the year when the average traffic intensity gets the highest value. To be accurate, the busy hour is determined by first selecting the 10 working days in a year with the highest traffic intensity; four consecutive 15-minute periods (of those 10 days) with the highest traffic intensity make up the busy hour.

The basic goal is to find a minimum capacity that gives the defined grade of service. Figure 2.24 shows a local exchange with a number of subscribers and a much smaller number n of trunk lines to the next exchange. If more than n subscribers make an external call at a time, some of them are
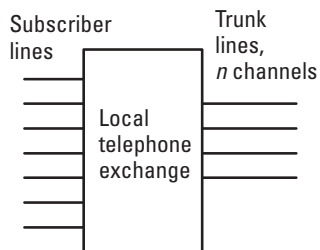
Figure 2.24 Local exchange and blocking.

blocked and they have to try again. The number of external calls varies in a random manner and to be sure that blocking never occurs *n* should be equal to the number of subscribers. This is a far too expensive solution because the number of subscribers connected to a local exchange is usually very large and on average only a small portion of them place external calls at the same time. The principle of how to find the capacity, that is, the number of lines *n* in our example, that is economically feasible but acceptable from subscribers' points of view is explained next.

2.12.3 Traffic Intensity and the Erlang

The measure of traffic intensity for circuit-switched connections is called the *erlang* in honor of the Danish mathematician A. K. Erlang, the founder of traffic theory. The erlang unit is defined as (1) a unit of telephone traffic specifying the percentage of average use of a line or circuit (one channel) or (2) the ratio of time during which a circuit is occupied and the time for which the circuit is available to be occupied. Traffic that occupies a circuit for 1 hour during a busy hour is equal to 1 erlang. Consider these examples:

- If the traffic intensity of a subscriber line is 1 erlang, the line is occupied for 60 minutes in an hour.
- If a subscriber line is in use 6 minutes out of an hour (on average), the traffic intensity is 6 minutes/60 minutes or 100 mErl.
- The maximum traffic intensity of a 2-Mbps (30 PCM channels) line system is 30 erlangs, that is, all channels are in use 60 minutes during the busy hour.

The typical average busy-hour traffic volume generated by one subscriber is in the range of 10 to 200 mErl. Low values are typical for residential use and high values for business subscribers.

2.12.4 Probability of Blocking

The problem in traffic engineering is determining the capacity if the average offered traffic intensity is known (or estimated). The term *offered traffic* refers to the average generated total traffic including the traffic that is blocked in the system. Clearly the capacity should (at least usually) be higher than offered traffic; otherwise, many users would not be able to get service because all lines would be occupied all the time (on average). If all trunk lines are occupied, new users are blocked, they receive a busy tone, and they have to

try again. The essential question is this: How much higher should the capacity be for the subscribers to feel that the grade of service is acceptable?

The starting point is how often subscribers are allowed to be blocked and receive a busy tone. This probability of blockage for an acceptable GoS is usually set to be in the range of 0.2% to 5%, which means that every 500th to 20th call is blocked during a busy hour. When the average traffic load is estimated to increase to a certain volume, the network operator should increase the network capacity to keep the blocking probability below the defined GoS level.

The Poisson distribution is used as a probability model for these calculations and it gives a probability for occurrence of x events when the average number of events is A according to this formula:

$$P(x) = \frac{A^x e^{-A}}{x!} \qquad (2.1)$$

where $e = 2.71828$ and $x!$ is the factorial of x, $1 \cdot 2 \cdot 3 \ldots \cdot x$. Now the average number of occupied channels is A in erlangs and (2.1) gives the probability that x number of channels is occupied at a time when a subscriber makes a call. Blocking occurs if all n channels are occupied or there may even be a need for a larger number of channels. This probability is given by:

$$P(x \geq n) = P(n) + P(n+1) + P(n+2) + \ldots \qquad (2.2)$$

On the other hand, one number of channels is always in use, giving this probability for

$$P(0) + P(1) + P(2) + \ldots P(n) + P(n+1) \ldots = P(x < n) + P(x \geq n) = 1 \qquad (2.3)$$

and we change (2.2) into this form:

$$P(x \geq n) = 1 - P(x < n) \qquad (2.4)$$

Substituting (2.1) gives

$$P(x \geq n) = 1 - \left[P(0) + P(1) + \ldots + P(n-1) \right]$$
$$= 1 - \left[\frac{A^0 e^{-A}}{0!} + \frac{A^1 e^{-A}}{1!} + \cdots + \frac{A^{n-1} e^{-A}}{(n-1)!} \right] \qquad (2.5)$$

Now we have the Poisson formula, which is also known as the *Molina lost calls held trunking formula,* for blocking probability and it is as follows:

$$P(x \geq n) = 1 - \sum_{x=0}^{n-1} \frac{A^x e^{-A}}{x!} \qquad (2.6)$$

To determine the grade of service, blocking probability, we compute (using the Poisson distribution) the probability that not one channel is free when a subscriber makes a call. For this we take a value of the average (total) offered traffic as A and calculate the probability that traffic occupies all n channels or is even higher at that point in time. (The offered traffic load may be higher than n even though actual traffic can never exceed n.) We get this by subtracting from 1 the probability that traffic is smaller than n according to (2.6).

Figure 2.25 illustrates the procedure we just carried out and shows an example where average offered traffic intensity $A = 1$ erlangs and the number of available channels $n = 3$. The probability density function $P(x)$ in the figure tells the probability for each value of x, that is, the number of occupied channels. The probability that all channels are free ($x = 0$) is $P(0) = 0.37$, one channel is occupied is $P(1) = 0.37$, and two channels are occupied (and one is free) is $P(2) = 0.18$. We subtract the sum of these probabilities from 1 and get the blocking probability, that is, the probability that the number of occupied channels $x \geq 3$. We get the result, when there are three channels available and average offered traffic is 1 Erl (i.e., on average, one call on all the time), that the blocking probability is 8%. This means that every twelfth call the user makes is blocked and a busy signal is received. Note that this blocking rate allows that only one channel is in use, two channels are free, and only one-third of the capacity can be utilized on average.

Equation (2.6) is based on following assumptions [2]:

- Poisson arrival rate; Poisson-distributed call attempts;
- Equal traffic volume per source;
- Lost calls held; calls that are blocked stay in the system and wait for the free channel (subscriber dials and dials again) [3];
- Infinite number of sources; if some sources are blocked or making a call, this does not affect the total offered traffic.

Let us consider another example: Total offered average traffic during the busy hour is 2 Erl ($A = 2$); and the number of servers, for example,

Blocking does not occur;
0, 1, or 2 channels are
occupied and at least one
of 3 channels is free

Blocking occurs; there are
3 or more active users and
all 3 channels are occupied.
Blocking probability:

$P(x < n) = P(0) + P(1) + P(2)$

$P(x > = n) = P(3) + P(4) + ... = .$

$1 - P(x < n) = 1 - [P(0) + P(1) + P(2)]$
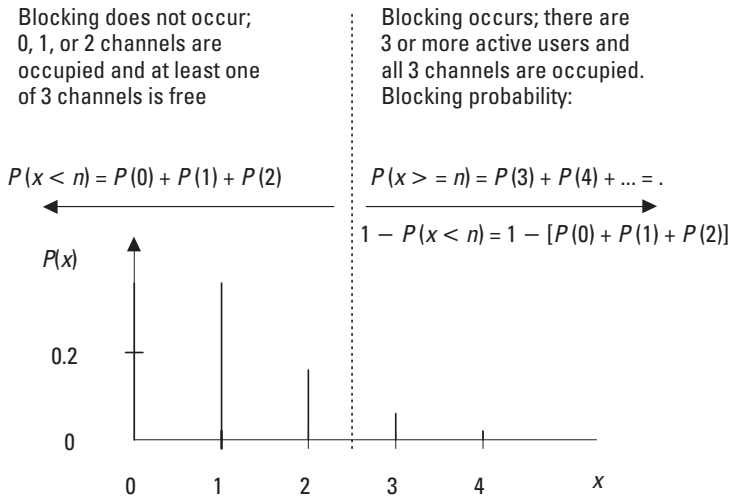
Figure 2.25 Probability of blocking.

transmission lines, is 5 ($n = 5$). Then the probability of blockage is 5.3 % [$P(x \geq 5) = 0.053$]. This means that, on average, during the busy hour every ninteenth call is blocked, a busy tone is heard, and the subscriber has to redial.

When the number of channels or servers n is high, precalculated tables like Table 2.1 are used for network planning. Such a table gives the required number of servers n when the GoS (= blocking probability in our study) and estimated offered traffic intensity A are given. For example, if the GoS is set to be 2% and offered traffic is 5 Erl (e.g., 100 subscribers with average offered traffic intensity of 50 mErl per subscriber), the network capacity should support at least 10 simultaneous calls ($n = 10$). If the capacity is smaller, for example, $n = 9$, we get 2% with offered traffic 4.34 Erl; an offered traffic 5 Erl would give a higher blocking rate.

When we look at Table 2.1, we note that when the number of servers is small, offered traffic intensity is of the order of one-tenth of the maximum traffic intensity. For example, with two channels, offered traffic intensity at blocking probability 1% is only 150 mErl. Only one 9-minute call during an hour is allowed and both lines may be occupied on average only 4.5 minutes in an hour. The utilization of channels is less than 10%.

When the number of servers is high, allowed average traffic intensity is close to the maximum or even higher. Even when most of the channels are occupied, some channels are still free for a new call because the number of channels is large.

Table 2.1
Network Capacity Planning Blocking Probability, GoS

n:	0.5% *A*	1.0% *A*	2.0% *A*	3.0% *A*	5.0% *A*	10% *A*	20% *A*	50% *A*
1	0.01	0.01	0.02	0.03	0.05	0.11	0.25	1.00
2	0.11	0.15	0.22	0.28	0.38	0.60	1.00	2.73
3	0.35	0.46	0.60	0.72	0.90	1.27	1.93	4.59
4	0.70	0.87	1.09	1.26	1.52	2.05	2.95	6.50
5	1.13	1.36	1.66	1.88	2.22	2.88	4.01	8.44
6	1.62	1.91	2.28	2.54	2.96	3.76	5.11	10.4
7	2.16	2.50	2.94	3.25	3.74	4.67	6.23	12.4
8	2.73	3.13	3.63	3.99	4.54	5.60	7.37	14.3
9	3.33	3.78	4.34	4.75	5.37	6.55	8.53	16.3
10	3.96	4.46	5.08	5.53	6.22	7.51	9.69	18.3
12	5.28	5.88	6.61	7.14	7.95	9.47	12.0	22.2
15	7.38	8.11	9.01	9.65	10.6	12.5	15.6	28.2
20	11.1	12.0	13.2	14.0	15.3	17.6	21.6	38.2
25	15.0	16.1	17.5	18.5	20.0	22.8	27.7	48.1
30	19.0	20.3	21.9	23.1	24.8	28.1	33.8	58.1
35	23.2	24.6	26.4	27.7	29.7	33.4	40.0	68.1
40	27.4	29.0	31.0	32.4	34.6	38.8	46.2	78.1
45	31.7	33.4	35.6	37.2	39.6	44.2	52.3	88.1
50	36.0	37.9	40.3	41.9	44.5	49.6	58.5	98.1
55	40.4	42.4	44.9	46.7	49.5	55.0	64.7	108.1
60	44.8	46.9	49.6	51.6	54.6	60.4	70.9	118.1
65	49.2	51.5	54.4	56.4	59.6	65.8	77.1	128.1
70	53.7	56.1	59.1	61.3	64.7	71.3	83.3	138.1
75	58.2	60.7	63.9	66.2	69.7	76.7	89.5	148.1
80	62.7	65.4	68.7	71.1	74.8	82.2	95.8	158.1
85	67.2	70.0	73.5	76.0	79.9	87.7	102.0	168.0
90	71.8	74.7	78.3	80.9	85.0	93.2	108.2	178.0
95	76.3	79.4	83.1	85.9	90.1	98.6	114.4	188.0
100	80.9	84.1	88.0	90.8	95.2	104.1	120.6	198.0
110	90.1	93.5	97.7	100.7	105.5	115.1	133.1	218.0
140	118.0	122.0	127.0	130.6	136.4	148.1	170.5	278.0

Table 2.1 (continued)
Network Capacity Planning Blocking Probability, GoS

n:	0.5% A	1.0% A	2.0% A	3.0% A	5.0% A	10% A	20% A	50% A
200	174.6	179.7	186.2					
300	270.4	277.1	285.7					
400	367.2	375.2	385.9					
500	464.5	474.0	486.4					

Note in Table 2.1 that when high blocking probability is allowed, offered traffic may be higher than the number of available channels. A part of offered traffic is blocked and actual traffic, that part which is not blocked, naturally never gets a higher value than the number of channels in erlangs.

Blockage probability can be calculated in many different ways. Table 2.1 is calculated according to erlang B formula, which assumes that blocked calls are immediately cleared and a subscriber waits and makes a new call later [3]. It gives slightly more optimistic results than the Poisson formula in (2.6). Erlang B formula is used in Europe and the Poisson formula in used in the United States for network planning. To compare the results of these two slightly different approaches, we consider an example where average offered traffic $A = 2$ Erl. If the number of circuits $n = 5$, we get blocking probability $P = 0.0367$ according to erlang B formula instead of the 0.053 we got previously with (2.6). The erlang B formula is [2] as follows:

$$P = \frac{\dfrac{A^n}{n!}}{\displaystyle\sum_{x=0}^{x=n} \dfrac{A^x}{x!}} \tag{2.7}$$

2.13 Problems and Review Questions

Problem 2.1

Describe how dialed digits are transferred from a subscriber's telephone to the local exchange.

Problem 2.2

Explain how the telephone attenuates the speaker's voice from the microphone to the earphone. (*Hint:* Draw the current coming from the microphone in Figure 2.7 and imagine what happens to the magnetic field in the iron core of the transformer.)

Problem 2.3

What is a 2W/4W hybrid and why is it needed at the end of the subscriber line?

Problem 2.4

Explain how a 2W/4W hybrid prevents the signal from the network (receiving pair) from looping back to the transmitting pair.

Problem 2.5

Explain the basic principle of telephone call routing through the switching hierarchy to another region of the country.

Problem 2.6

A network has N subscribers. Each subscriber is connected directly to all other subscribers.

    (a) What is the total number of lines L in the network?
    (b) What is the value of L for $N = 2$, 10, 100, and 1,000?
    (c) How many lines must be built to each subscriber?
    (d) Is this kind of network structure suitable for a public telecommunications network? Explain.

Problem 2.7

What are the basic differences between the public and private telecommunications networks? List a few examples of both public and private networks.

Problem 2.8

What is ISDN? How does the service and structure of the subscriber interface differ from the conventional analog telephone service?

Problem 2.9

How does an IN differ from conventional fixed telephone network? List some examples of IN services.

Problem 2.10

A PBX/PABX has seven telephone channels to a public exchange. During the busy hour, on average, 3.4 lines are occupied. (a) What is the traffic intensity during the busy hour? (b) Estimate, with the help of the Table 2.1, the GoS (blocking probability).

Problem 2.11

What is the total offered traffic intensity from a PBX/PABX to PSTN if 10 calls are made, each with a duration of 6 minutes during 1 hour?

Problem 2.12

A subscriber makes one 6-minute call in one day between 10:00 and 10:06. What is the average traffic intensity of her subscriber line during (a) 10:00–10:06, (b) 10:00–10:15, (c) 10:00–11:00, and (d) 00:00–24:00 of that day?

Problem 2.13

Use the Poisson (or "Molina lost calls held") trunking formula to calculate the blocking probability (GoS) when the total offered traffic is 2 Erl and the number of available transmission channels in the network is 5.

Problem 2.14

Draw two curves for GoS levels of 1% and 10%. Use the vertical axis as a ratio A/n from 1% to 100% and the horizontal axis as a number of circuits n from 1 to 20. Use traffic engineering Table 2.1. What can you say about network utilization when the number of circuits n is small? How does the utilization of the circuits depend on the allowed probability of blocking?

Problem 2.15

What will the approximate capacity of a network be (i.e., how many channels should be available) if there are 100 subscribers and each of them generates offered traffic of 40 mErl? The probability of blocking is (a) 20% and (b) 1%. Use traffic engineering Table 2.1.

Problem 2.16

There are 20 users of a keyphone system that has two lines to a public network. What is the blocking probability when each user generates a 100-mErl offered traffic?

Problem 2.17

A keyphone system with three lines to the local exchange is used in an office of 10 persons. Each of them uses the phone for an external call of 15 minutes in a busy hour. How many lines are reserved on average during an hour? What is the blocking probability? What do you think about the capacity of this system?

Problem 2.18

Subscribers of a local exchange generate 100 mErl of traffic through the exchange to the network. What should the number of trunk channels be if the number of subscribers in the area is (a) 10, (b) 100, (c) 1,000, and (d) 4,000? The allowed blocking level is 1%. Use Table 2.1 to estimate the required number of circuits.

References

[1]   *Telecommunications Transmission Engineering,* Bellcore Technical Publications, 1990.

[2]   Freeman, R. L., *Telecommunication System Engineering,* 3rd ed., New York: John Wiley & Sons, 1996.

[3]   Freeman, R. L., *Fundamentals of Telecommunications,* New York: John Wiley & Sons, 1999.

3

Signals Carried over the Network

Services that the telecommunications networks provide have different characteristics. Required characteristics depend on the applications we use. To meet these different requirements, many different network technologies that are optimized for each type of service are in use. To understand the present structure of the telecommunications network, we have to understand what types of signals are transmitted through the telecommunications network and their requirements. In this chapter we look at the requirements of various applications, characteristics of analog voice channels, fundamental differences between analog and digital signals, analog-to-digital conversion, and a logarithmic measure of signal level, the decibel.

3.1   Types of Information and Their Requirements

Modern digital networks transmit digital information transparently; that is, the network does not necessarily need to know what kind of information the data contain. This information that is transmitted through the network may be any one of the following:

- Speech (telephony, fixed, or cellular);
- Moving images (television or video);
- Printed pages or still picture (facsimile or multimedia messaging);
- Text (electronic mail or short text messaging);

- Music;
- All types of computer information such as program files.

For digital transmission, analog signals such as speech are encoded into digital form and transmitted through the network as a sequence of bits in the same way computer files are transmitted. However, although all information is coded into digital form, the transmission requirements are highly dependent on the application; because of these different requirements, different networks and technologies are in use. Video and e-mail applications, for example, require different architectures. Network technologies have taken two main development paths: one for speech services and another for data services. The telephone network and ISDN have been developed for constant-bit-rate voice communication that is well suited to speech transmission. Data networks such as LANs and the Internet have been developed for bursty data transmission.

The constant-bit-rate requirement for speech follows from the principle that digitized voice signals have traditionally been transmitted in digital form as samples at regular intervals, as we will see in Section 3.6. Data transmission is bursty by nature. Sometimes we may copy a file across the network, whereas at other times we may work locally with our workstation.

When many different applications are integrated into multimedia communications, both basic types of service requirements of constant-bit-rate voice and bursty data have to be fulfilled and we need a concept that is able to meet both types of requirements.

In Table 3.1 different applications are compared from the communication requirements point of view. The applications are ordinary speech, *computer-aided design* (CAD) (a service in which high-resolution graphical information is transmitted), moving images (video), file transfer, and multimedia with integrated video, voice, and data. The importance of the transmission requirements for each application is explained next.

Data Rate or Bandwidth Requirement

Voice communication usually requires a constant data rate of 64 Kbps or less and high-resolution video a constant data rate of 2 Mbps or higher over the network. Characteristics of data communication are very different, for example, file transfer requires high-bit-rate transmission only during download, and high-resolution graphics on a Web page require high-data-rate transmission only when we download a new page. When we are reading a Web page we do not need transmission capacity at all. To define data transmission

Table 3.1
Communication Requirements of Different Applications

Transmission Characteristics	Voice	Video	File Transfer	Interactive Media
Bandwith requirement	Low, fixed	Very high, fixed	High, variable	High, variable
Data loss tolerance	Tolerant	Tolerant	Nontolerant	Tolerant or nontolerant
Fixed delay tolerance	Low delay	Tolerant	Tolerant	Low delay
Variable delay tolerance	No	No	Tolerant	No
Peak information rate	Fixed	Fixed	High	Very high

capacity, we sometimes use the term *bandwidth* instead of *data rate* because these terms are closely related to each other, as we will see in Chapter 4.

Data Loss Tolerance

Noise and other disturbances in the network may cause errors in the transmitted data. If errors occur, some amount of data may be lost. Voice and video transmission services are used by human beings, and they can tolerate accidental short disturbances. In computer communications a single erroneous bit usually destroys a whole data frame, which may contain a large amount of data. The loss of one frame destroys the transmission of a large file that is transferred in multiple frames. Most of the data communication systems are able to retransmit data frames in error. Systems designed for voice or video transmission do not use retransmission schemes because temporary retransmission delay is even more disturbing for human users than the loss of some data.

Fixed Delay Tolerance

When communication is interactive, as voice communication usually is, the two-way transmission delay should be very short for good quality. In the case of voice it should be of the order of some tens of milliseconds. Otherwise, we feel that quality is degraded because the response from the other party is

delayed. We tolerate much longer delays in the case of ordinary data applications when we are waiting for a response to our "click" command.

Variable Delay Tolerance

Voice and video information is traditionally transmitted as samples at regular periods of time. The reconstruction of images and voice requires that all sample values be received sequentially and suffer the same delay. Conventional data networks recover from errors with the help of retransmission of the frames in error. This is a very efficient error recovery scheme, but it introduces some additional and variable delay. For voice applications this variable delay is often a worse solution than that of losing some data.
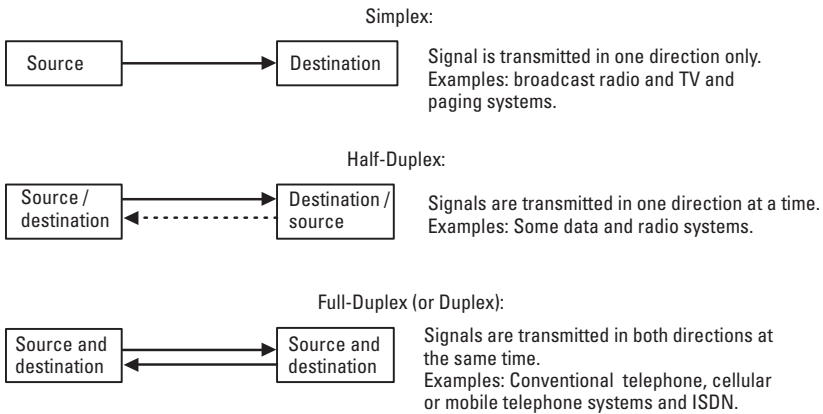
Peak Information Rate

Encoding of analog voice and video often produces a constant information data rate. Values of the samples with constant length contain voice or video information and they are transmitted at a constant rate. In data communication applications we usually work locally and every now and then a high data rate is needed to load graphical information or files. A peak load is typically of the order of 1,000 times higher than the average transmission capacity we use.

The different requirements just explained have supported development of the circuit-switched networks, such as PSTN and ISDN, for voice communications and packet-switched networks, such as LANs and the Internet, for data communications. *Asynchronous transfer mode* (ATM) technology was developed by ITU-T to be suitable and efficient for transferring all types of information. However, the expansion of the Internet has reduced its importance and Internet technology will be developed further to provide a platform for all kinds of communications.

3.2   Simplex, Half-Duplex, and Full-Duplex Communication

In telecommunications systems the transmission of information may be unidirectional or bidirectional. The unidirectional systems that transmit in one direction only are called *simplex,* and the bidirectional systems that are able to transmit in both directions are called *duplex* systems. We can implement bidirectional information transfer with *half-* or *full-duplex* transmission as shown in Figure 3.1.

In simplex operation the signal is transmitted in one direction only. An example of this principle is broadcast television, where TV signals are sent

Simplex:

Source	Destination

Signal is transmitted in one direction only.
Examples: broadcast radio and TV and
paging systems.

Half-Duplex:

Source /
destination	Destination /
source

Signals are transmitted in one direction at a time.
Examples: Some data and radio systems.

Full-Duplex (or Duplex):

Source and
destination	Source and
destination

Signals are transmitted in both directions at
the same time.
Examples: Conventional telephone, cellular
or mobile telephone systems and ISDN.

Figure 3.1 Simplex, half-duplex, and full-duplex transmission.

from a transmitter to TV sets only and not in the other direction. Another example is a paging system that allows a user to receive only alphanumerical messages.

In half-duplex operation the signal is transmitted in both directions but only in one direction at a time. An example of this is a mobile radio system where the person speaking must indicate by saying the word *over* that she is done transmitting and the other person is allowed to transmit. LANs use a high-speed, half-duplex transmission over the cable even though users may feel that the communication is continuously bidirectional, that is, full duplex.

In full-duplex operation signals are transmitted in both directions at the same time. An example of this is an ordinary telephone conversation where it is possible for both people to speak simultaneously. Most modern telecommunications systems use the full-duplex principle, which we call *duplex operation* for short.

3.3 Frequency and Bandwidth

To understand the requirements of different applications for a telecommunications network, we must understand the fundamental concepts of frequency and bandwidth. The information that we transmit through a telecommunications network, whether it is analog or digital, is in the form of electrical voltage or current. The value of this voltage or current changes through time, and this alteration contains information.

The transmitted signal (the alteration of voltage or current) consists of multiple frequencies. The range of frequencies is called the *bandwidth* of the

signal. The bandwidth is one of the most important characteristics of analog information and it is also the most important limiting factor for the data rate of digital information transfer.

3.3.1 Frequency

We can see the telecommunications signal as a combination of many cosine or sine waves with different strengths and frequencies. The frequency refers to the number of cycles through with the wave oscillates in a second. As an example of the concept of frequency, we hear the oscillation of air pressure as sound. We are able to hear frequencies in the range of approximately 20 Hz to 15 kHz, where Hz (hertz) represents the number of cycles in a second. An example of the different frequencies is heard in the keys of a piano. The right-hand keys generate basic frequencies of the order of 1,000 Hz and the left-hand keys of the order of 100 Hz.

In electrical terms, an *alternating current* (ac) changes its direction of flow several times per second. This variation in direction is known as a *cycle,* and the term *frequency* refers to the number of cycles in a second that is measured in hertz. If a signal has 1,000 complete cycles in a second, then its frequency is 1,000 Hz or 1 kHz. A pure sine wave, like that shown in Figure 3.2, is generated with a loop of wire rotated in a magnetic field at a constant rate. This fundamental waveform can be seen as a cosine of the
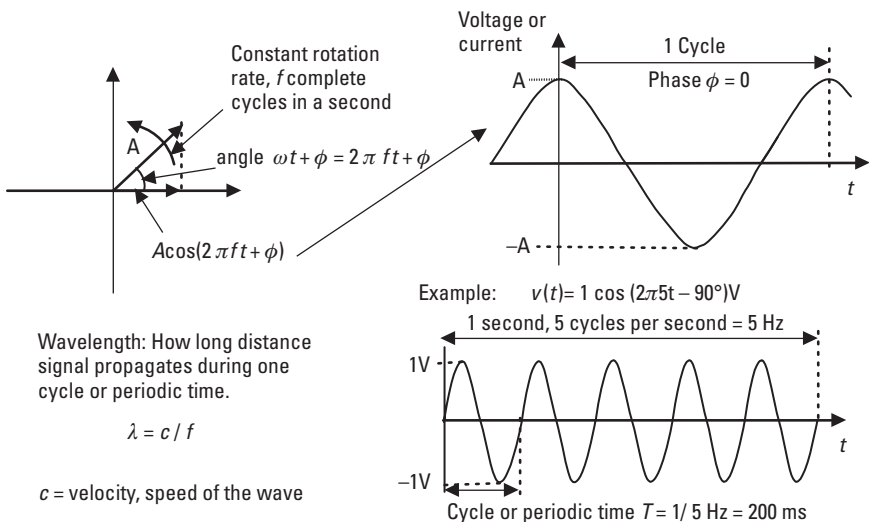
Figure 3.2 Cosine wave and frequency.

angle of the phasor rotating at a constant rate. The strength of the voltage or current alters according to the cosine curve when time increases. The length of the phasor corresponds to the maximum value of the signal and it is called *amplitude,* shown as *A* in Figure 3.2.

We can see any telecommunications signal as a sum of these fundamental waveform cosine waves that are expressed as

$$v(t) = A\cos(\omega t + \phi) = A\cos(2\pi ft + \phi) \tag{3.1}$$

where f is frequency, the number of complete cycles in a second expressed in hertz, 1 Hz = 1/sec; t is time in seconds, and ϕ is the phase shift (phase of the cosine wave at time instant t = 0). The angular frequency ω in radians per second is $\omega = 2\pi f$, which comes from the fact that one complete cycle of a phasor makes up an angle of 2π radians.

The *periodic time* or *period T* in seconds represents the time of one complete cycle:

$$T = 1/f \text{ and } f = 1/T \tag{3.2}$$

Wavelength λ represents the propagation distance in one cycle time, thus,

$$\lambda = c/f = cT \tag{3.3}$$

where c is the velocity of the signal. For a sound wave, the velocity in the air is approximately 346 m/s; for light or radio waves, approximately, c = 300,000 km/sec.

The example in Figure 3.2 shows a waveform with a frequency of 5 Hz and amplitude of 1V. It corresponds to a phasor with length of A = 1V making five complete cycles in a second. At time instant t = 0, the waveform has a value of 0 and the phase or angle of the phasor is –90°. As the time increases and the phasor rotates, its projection at the horizontal axis of the phasor diagram increases, corresponding to an increase in the value of the wave with time. The equation for this example waveform is then $v(t)$ = A cos $(\omega t + \phi)$= 1 cos$(2\pi 5t - 90°)V$.
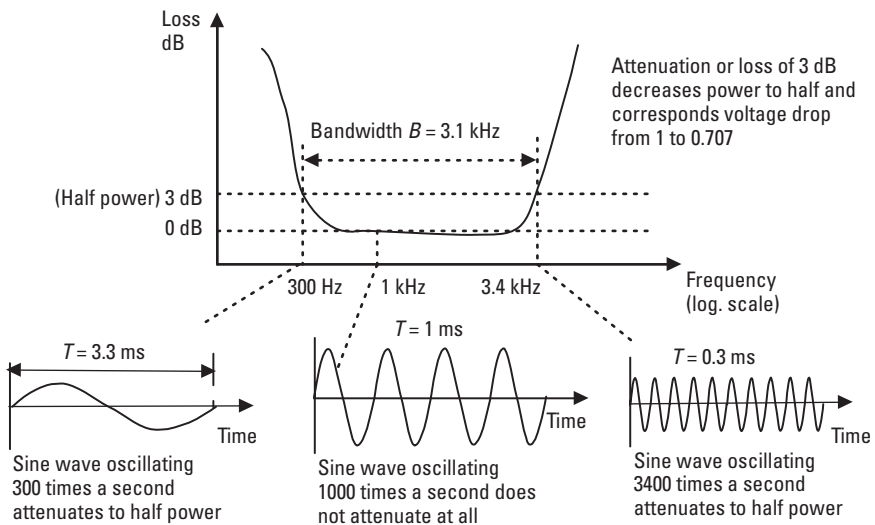
3.3.2 Bandwidth

The voice signal, which is the most common message in telecommunications network, does not look similar to a pure cosine wave in Figure 3.2. It

contains many cosine waves with different frequencies, amplitudes, and phases combined together. The range of frequencies that is needed for a good enough quality of voice, so that the speaker can be recognized, was defined to be the range from 300 to 3,400 Hz. This means that the bandwidth of the telephone channel through the network is 3,400 – 300 Hz = 3.1 kHz, as shown in Figure 3.3. A human voice contains much higher frequencies, but this bandwidth was defined as a compromise between quality and cost. It is wide enough to recognize the speaker, which was one requirement for telephone channel.

Bandwidth is not strictly limited in practice, but signal attenuation increases heavily at the lower and upper cutoff frequencies. For speech, channel cutoff frequencies are 300 and 3.4 Hz, as shown in Figure 3.3. The bandwidth is normally measured from the points where the signal power drops to half from its maximum power. Attenuation or loss of channel is given as a logarithmic measure called a decibel (dB), and half power points correspond to a 3-dB loss. Decibels are discussed later in this chapter.

Bandwidth, together with noise, is the major factor that determines the information-carrying capacity of a telecommunications channel. The term *bandwidth* is often used instead of *data rate* because they are closely related, as we will see in Chapter 4.

Figure 3.3 Bandwidth of the telephone speech channel.

3.4 Analog and Digital Signals and Systems

Most of the systems in the modern telecommunications network are digital instead of analog. In this section we look at the fundamental characteristics of analog and digital signals and how they influence the performance and operation of telecommunications systems.

3.4.1 Analog and Digital Signals

The difference between analog and digital form is easily understood by looking at the two watches in Figure 3.4. A true analog watch has hands that are constantly moving and always show the exact time. A digital watch displays "digits" and the display jumps from second to second and shows only discrete values of time.

Another example could be the slope of analog voltage where all values of voltage can be measured as shown in Figure 3.4. In "digital slope," only discrete values may be measured. In the example of the figure, we have eight discrete values, 0 to 7, in the digital slope. This does not mean that the digital systems perform worse than analog systems. If we want to improve the accuracy of the digital system, we just increase the number of steps and, in principle, any voltage level can be represented with the digital system as well.
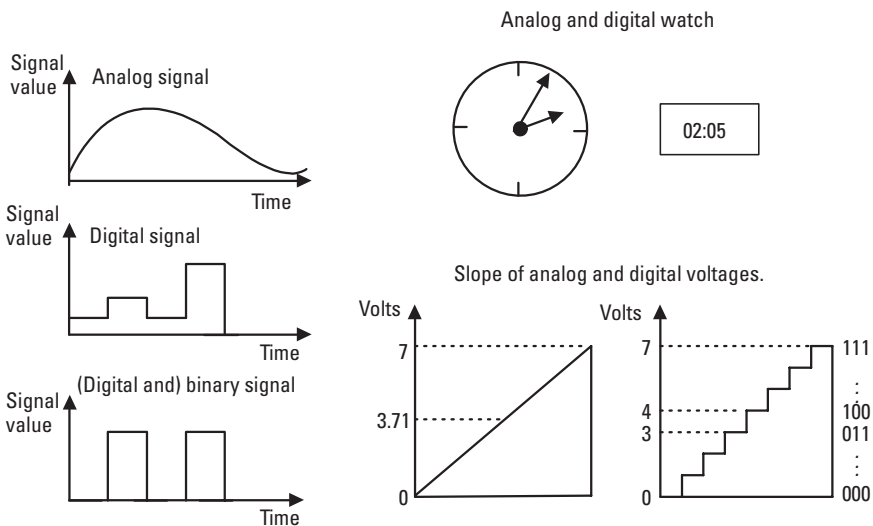
Figure 3.4 Analog and digital signals.

A special and very important case of digital signals is a binary signal where only two values, binary digits 0 and 1, are present as illustrated in Figure 3.4. Examples of binary signals are light on and off, voltage versus no voltage, and low current versus high current.

Binary signals are used internally in computers and other digital systems to represent any digital signal. For example, we can encode eight voltage levels of the slope in Figure 3.4 into three binary bits and each of these three bit words then represents one of the $2^3 = 8$ (0 (000) to 7 (111)) different values. As another example, a digital signal with eight-bit words or bytes (often called *octets* in digital telecommunications systems) can represent $2^8 = 256$ discrete values of a signal. These kinds of digital numbers are used to represent analog voice, in which each sample of a voice signal is encoded into eight-bit words, as we will explain in Section 3.6.

3.4.2   Advantages of Digital Technology

Analog systems in a telecommunications network have gradually been replaced with digital systems. Development of digital circuits and software technologies has made digital systems more and more attractive. The most important advantages of digital technology over analog technology are as follows:

- Digital functions make a high scale of integration possible.
- Digital technology results in lower cost, better reliability, less floor space, and lower power consumption.
- Digital technology makes communication quality independent of distance.
- Digital technology provides better noise tolerance.
- Digital networks are ideal for growing data communication
- Digital technology makes new services available.
- Digital system provides high transmission capacity.
- Digital networks offer flexibility.

An analog system requires the accurate detection of signal values inside its dynamic range, that is, between the maximum and minimum values of the signal. Digital systems use binary signals internally. A binary signal has only two values, and the only problem is to distinguish these two values from each other. The dynamic range is well defined and linearity is not required.

This makes the elements of digital circuits simple, and the utilization of compact technology for very complicated functions, such as integrated circuits, is feasible.

As a consequence, circuit integration leads to a smaller number of electronic components, smaller equipment, lower manufacturing costs, lower maintenance costs because of better reliability, and less power consumption. More and more complex integrated circuits are replacing many lower scale integrated circuits. This decreases system costs, because the increased complexity of components does not cost much in volume. When integrated circuits are manufactured in volume, complex ones do not cost much more than less complex circuits. In addition, the smaller number of separate components gives better reliability.

In long-distance connections, we have to amplify or regenerate the signal on the line many times. When we amplify an analog signal on the line, we amplify noise at the same time. This added noise decreases the quality of an analog signal, that is, decreases the *signal-to-noise* (S/N) ratio.

In the case of a digital system we use regenerators or repeaters instead of amplifiers. Repeaters regenerate the signal symbol by symbol, that is, transmit further the value that is closest to the received value. The regenerated signal is a sequence of digital symbols with nominal values and thus it contains no noise. If the noise is low in the input of each regenerator, symbols of the digital signal are regenerated without errors and we receive exactly the same digital message on the other side of the world as it was at the transmitting end. The operation of a digital repeater or regenerator is described in Chapter 4.

Modern switches digitize speech in the subscriber interface. If the path through the network is fully digital, conversion back to analog is done only at the far end. There is only one analog-to-digital and one digital-to-analog conversion regardless of the communication distance, that is, whether we make a call to our neighbor or to other side of the world.

The digital systems have to identify only signals from a set of discrete values. If symbols are not mixed because of too high a noise level, noise does not have any impact on the operation. Analog communication usually requires a much better S/N than low error rate digital communication. As a consequence, digital systems can utilize channels with much higher noise levels and they can tolerate higher interference than analog systems.

If the network is analog, a digital message has to be modulated into the frequency band of the analog telecommunications channel. This reduces the capacity available for the user. For example, a voice channel in the digital telephone network has a data capacity of 64 Kbps. If we use it via an analog

subscriber loop with a voice-band modem, the data rate is restricted in practice to approximately 30 Kbps. With a *digital subscriber line* (DSL) (e.g., ISDN), the user data are exactly the same 64 Kbps used inside the network.

Digital systems are ideal for control via software because digital circuits operate in a numerical way. Integrated software makes systems flexible and new functions needed for new services are easier to implement. Intelligent network services, reviewed in Section 2.10, are good examples of these new services. As another example, we would not have cellular telephone service if we did not have digital software-controlled systems in the network.

The digital processing of information makes better utilization of channels possible; for example, several digital broadcast television channels fit into the band of one analog broadcast channel. In Chapter 4 we will see that digital signals tolerate higher disturbances than analog signals and this is one reason behind the better frequency efficiency. Low-cost multiplexing (no analog filtering and modulation circuitry required) and efficient use of optical transmission media make high-capacity digital systems feasible. Optical systems transmit digital signals as a series of short light pulses. The distortion of these digital pulses does not influence the quality of the message because distorted pulses are regenerated, which eliminates distortion.

All types of analog signals can be converted into digital signals. When this is done, the digital network is able to carry any information. Bits are handled in the same way whether they represent voice, video, or data.

Analog systems are different for each application because of different performance requirements. For example, a telephone connection requires channels with approximately 4-kHz bandwidth, but television signals require 5-MHz bandwidth with a much better S/N. In digital systems the corresponding characteristic is the data rate. For example, an analog telephone signal requires 64 Kbps and video with a much wider bandwidth requires 2 to 140 Mbps depending on the coding scheme in use. We can use one high-data-rate system for a single video channel or a large number of speech channels.

Digital technology provides efficient multiplexing for sharing capacity in high-data-rate connections. This makes high-capacity digital networks and systems flexible. The same system, if it provides a high enough data rate, can be used for any application.

3.4.3   Examples of Messages

In the previous sections, we described the characteristics of the digital and analog signals and systems. Now we look at some simple examples of

information sources that produce messages that are transmitted through the network. There are many different information sources, including machines as well as people, and messages or signals appear in various forms. As for signals we can identify the two main distinct message categories: analog and digital.

3.4.3.1   Information, Messages, and Signals

The concept of information is central to communication. However, *information* is a loaded word, implying schematic and philosophical notions and, therefore, we prefer to use the word *message* instead. *Message* means the physical manifestation of information produced by a source. Systems handling messages convert them into electrical signals suitable, for example, for a certain transmission media.

3.4.3.2   Analog Message

An analog message is a physical quantity that varies through time, usually in a smooth and continuous fashion. Examples of analog messages are acoustic pressure produced when you speak or light intensity at one point in an analog television image. One example of an analog message is the voice current on a conventional subscriber telephone line as illustrated in Figure 3.5. In Section 2.2 we explained how the current is produced.

Because the information resides in a time-varying waveform, an analog communication system should deliver this waveform with a specific degree of fidelity. Because the strength of signals may vary in a range from 30 to 100 dB, depending on the application, the analog systems should have good linearity from the weakest signal to 1,000 to 10,000 million times stronger signal values.

3.4.3.3   Digital Message

A digital message is an ordered sequence of symbols selected from a finite set of discrete elements. Examples of digital messages are the letters printed on this page or the keys you press at a computer keyboard. When you press a key at your computer keyboard, each key stroke represents a digital message that is then encoded into a set of bits for binary transmission.

Because the information resides in discrete symbols, a digital communication system should deliver these symbols with a specified degree of accuracy in a specified amount of time. The main concern in the system design is that symbols remain unchanged, which is the final requirement for transmission accuracy.
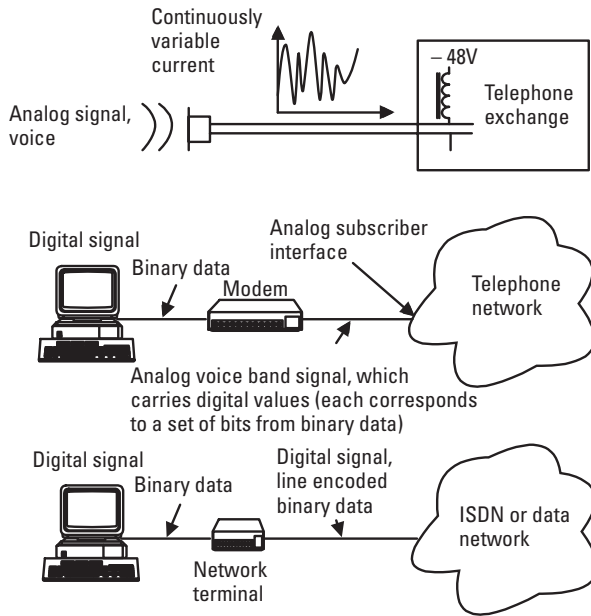
Figure 3.5 Examples of messages.

We need modems for the transmission of digital messages over analog channels. The modems receive a message from the terminal in the form of binary data and send it as an analog waveform to the speech channel as shown in Figure 3.5. Current modems do not modulate or change the analog waveform at the rate of the binary data they receive from the terminal. Instead they encode a set of bits into a digital symbol that may get many more values than just two. Each multilevel symbol corresponds to a set of bits and it is sent as one analog waveform to the line. When receiving a certain analog signal on the other end, the receiver detects a set of bits defined to correspond to that signal. Use of more than two signals increases the data rate through the speech channel compared with the binary principle, in which only two different signals are used. Speech channels have quite a narrow bandwidth, but a good S/N, which allows use of many different signals, as we will explain in Chapter 4.

When a digital network is used to transmit digital messages, signals are in digital form from end to end. Instead of a modem, a network terminal is needed at the subscriber's premises to encode binary signals into digital pulses suitable for cable transmission to an exchange site; see the ISDN example in Figure 3.5.

3.5 Analog Signals over Digital Networks

In this section we look at how analog signals are handled before transmission through a digital network. In the next section we concentrate on the pulse code modulation, which is performed in the network on our voice during a telephone call, and in Section 3.7 we present a brief review of other voice-coding schemes.

If a digital signal is to be transmitted through an analog network, it has to be converted into an analog signal suitable for the frequency band of the channel, as we saw in Figure 3.5. Digital networks provide communication only with a set of discrete symbols (in the binary case these symbols are called bits) at a certain data rate and the analog signal has to be converted into a series of these symbols for digital communication. The data rate of a digital network corresponds to the channel bandwidth of an analog network. The higher the data rate, the wider the required bandwidth and vice versa.

If the network is fully digital, analog voice is encoded into digital form at the transmitting end and decoded into analog form at the receiving end, as shown in Figure 3.6. This coding is performed in the subscriber interface of a digital telephone exchange and, in the case of ISDN service, in the subscriber's ISDN telephone or network terminal.

This process has two main phases, as shown in Figure 3.6:

1. *Analog-to-digital conversion (A/D):* An analog signal is sampled at the sampling frequency and the sample values are then represented as numerical values by the encoder. These values, presented as
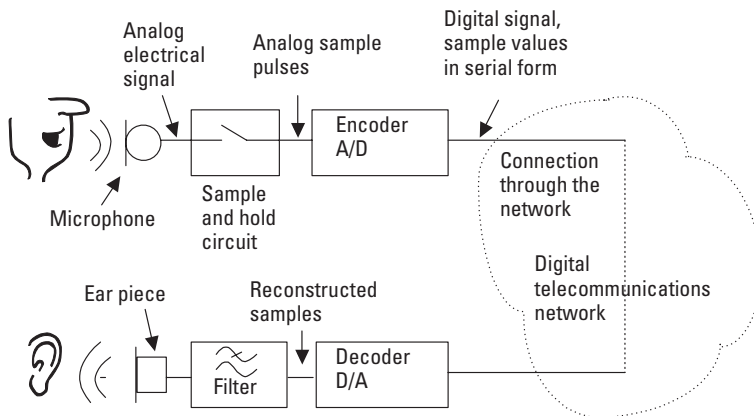
Figure 3.6 Analog voice signal through a digital network.

binary words, are then transmitted within regular time periods through the digital channel.

2. *Digital-to-analog conversion (D/A):* At the other end of the channel, the decoder receives numerical values of the samples that indicate the values of the analog signal at sampling instants. The sample pulses that have amplitudes corresponding to the values of the original signal at sampling instants are reconstructed and the series they form is filtered to produce an analog signal close to the original one.

The methods for these A/D and D/A conversions have to be specified in detail so that the reproduction of the analog signal is compatible with the production of the digital signal that may have occurred on the other side of the world. In the next section we describe the method that is used in the telecommunications network and internationally standardized by the ITU.

3.6 PCM

PCM is a standardized method that is used in the telephone network to change an analog signal to a digital one for transmission through the digital telecommunications network. The analog signal is first sampled at a 8-kHz sampling rate; then each sample is quantized into 1 of 256 levels and then encoded into digital eight-bit words. This encoding process is illustrated in Figure 3.7. The overall data rate of one speech signal becomes $8,000 \times 8 = 64$ Kbps. This same data rate is available for data transmission through each speech channel in the network. In the United States one bit of eight in every sixth frame is "robbed" for in-band signaling and the available transparent data capacity of a single speech channel in the network is reduced to $8,000 \times 7 = 56$ Kbps.

Now we take a more detailed look at the three main processing phases of the PCM in the telecommunications network. Note that this principle is employed by all systems when there is a need to process analog signals with a digital system. Sampling rates and the number of quantizing levels vary from application to application, but the basic principle and phases of the process remain the same.

3.6.1 Sampling

The amplitude of an analog signal is sampled first. The more samples per second there are, the more representative of the analog signal the set of samples
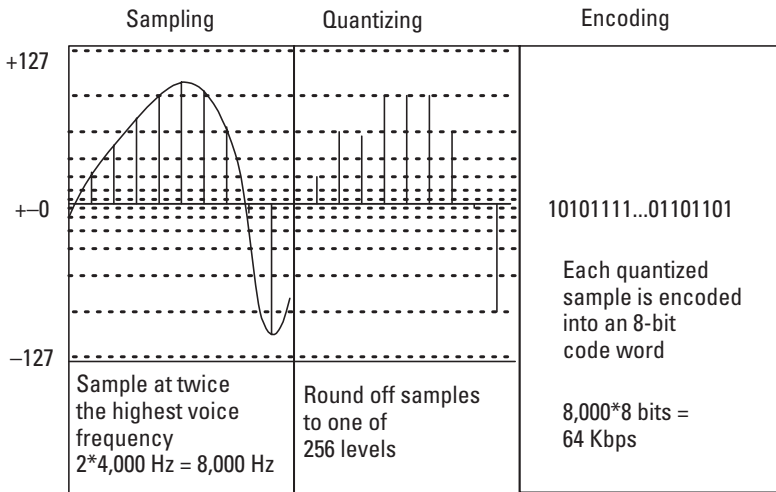
Figure 3.7 PCM.

will be. After sampling, the signal value is known only at discrete points in time, called sampling instants. If these points have a sufficiently close spacing, a smooth curve drawn through them allows us to interpolate intermediate values to any degree of accuracy. We can therefore say that a continuous curve can be adequately described by the sample values alone.

In a similar fashion, an electrical signal can be reproduced from an appropriate set of instantaneous samples. The number of samples per second is called the sampling *frequency* or sampling rate, and it depends on the highest frequency component present in the analog signal. The relation of sampling frequency and the highest frequency of the signal to be sampled is stated as follows:

> If the sampling frequency, f_s, is higher than two times the highest frequency component of the analog signal, W, the original analog signal is completely described by these instantaneous samples alone; that is, $f_s > 2W$.

This minimum sampling frequency is sometimes called the *Nyquist rate*. We can describe it in other words as an analog signal with the highest frequency component as W Hz. It is completely described by instantaneous sample values uniformly spaced in time within a period:

$$T_S = 1/f_S < 1/(2W) \tag{3.4}$$

Figure 3.8 represents the operating principle of a sampling circuit and an analog signal before and after sampling in both the time and frequency domains. The sampling circuit contains a generator, *G,* that produces short sampling pulses at the sampling frequency f_s. These sampling pulses close the switch of a relay at each sampling instant for a short period of time. The original analog signal $x(t)$ is sampled each time the switch is closed and a sampled signal $y(t)$ is produced. The sampled analog signal $y(t)$ contains short pulses that represent signal $x(t)$ values at discrete points in time. This sampling process that produces $y(t)$ is known as *pulse amplitude modulation* (PAM) because the amplitudes of the pulses contain the values of $x(t)$.

The time-domain curves in Figure 3.8 show the original continuous analog signal $x(t)$ and the sampled signal $y(t)$. The sampled signal $y(t)$ contains values of an analog signal at sampling instants. We can imagine that if the sampling frequency f_s is high, that is, the distance between sampling instants T_s is short, the sample pulses describe the original signal quite well. We could draw a line that connects the peak values of the pulses and the shape of this curve would be close to the original signal shape of $x(t)$.

The changes in $x(t)$ are related to the frequency content of $x(t)$. The more rapidly $x(t)$ changes, the higher frequency the components it contains. This explains why the sampling frequency is related to the highest frequency of the analog signal to be sampled. From the time-domain figure we understand that the sampling frequency must be much higher than the highest
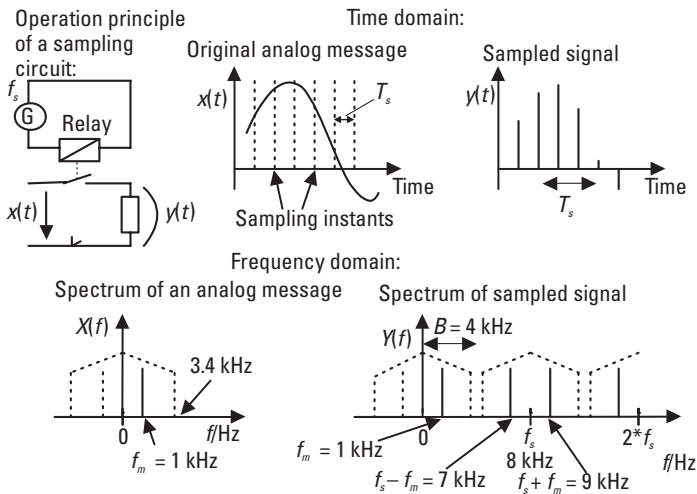
Figure 3.8 Sampling.

frequency of the analog message. Otherwise, rapid changes of signal $x(t)$ between sampling instants could not be described by sample values. The accurate answer to how much higher it should be can be understood more easily via the frequency domain.

The frequency-domain descriptions in Figure 3.8 show the spectrum of $x(t)$ and the sampled signal $y(t)$. Before sampling, the spectrum $X(f)$ of $x(t)$ contains speech frequencies up to 3.4 kHz, shown as a dashed line in the figure. As an example of the frequency components of speech we drew the spectrum of a 1-kHz cosine wave as a solid spectral line at the 1-kHz point on the frequency axis.

After sampling, the spectrum of the message also appears around the sampling frequency. If the message contains a single 1-kHz frequency component, after sampling we will have components at 1 kHz, 8 kHz – 1 kHz = 7 kHz, and at 8 kHz + 1 kHz = 9 kHz, as seen in the figure. In addition to these components, sampling also generates components around double sampling frequency, three times sampling frequency, and so forth.

The reproduction of an original signal from a sampled signal is performed by a lowpass filter and in the case of voice the bandwidth $B = 4$ kHz, that is, half the sampling frequency. We see from Figure 3.8 that this filter would let through only a 1-kHz component of the spectrum, that is, the actual original analog signal. With the help of the lowpass filter we have successfully reproduced the original analog message from the samples alone.

If we increase the frequency of an analog message $x(t)$ from 1 to 2 kHz we will have the lowest component of the sampled signal at 2 kHz, the solid spectral line at 1 kHz is moved to the right, the next spectral component at 8 kHz – 2 kHz = 6 kHz, and the solid line at 7 kHz is moved to the left. Lowpass filtering will still give the original 2-kHz message. Now if we increase the frequency beyond 4 kHz to, say, 5 kHz, we will get components at 5 kHz and 8 kHz – 5 kHz = 3 kHz, and lowpass filtering will give a 3-kHz signal instead of the original 5-kHz signal. Reproduction will not work anymore because the frequency of the analog signal has exceeded half of the sampling frequency.

We have seen that the sampling frequency must be more than twice the highest frequency component of the original signal to be encoded; otherwise, the message spectra around zero frequency and sampling frequency will overlap. This can be seen from the spectrum $Y(f)$ in Figure 3.8 if we imagine what happens if $W > f_s/2$. From the spectrum of the sampled signal $Y(f)$ in Figure 3.8, we also see that the message can be completely reconstructed from a PAM signal with a 4-kHz lowpass filter if $W < f_s/2$. This requirement is fundamental for all digital signal processing.

The highest frequency of voice that will be transmitted is chosen to be 3,400 Hz and the sampling frequency is standardized at 8,000 Hz, leaving enough guard band for filtering. Samples are then taken at intervals of $T_s = 125\,\mu$s.

In the sampling process a PAM signal $y(t)$ is created. The amplitudes of PAM pulses follow the original analog signal. Note that the samples are still analog, having any analog value between the minimum and maximum values of the original signal.

3.6.2   Quantizing

In the previous section we utilized sampling that produces a PAM signal that represents discrete but still analog values of the original analog message at the sampling instants. To transmit the sample values via a digital system, we have to represent each sample value in numerical form. This requires quantizing where each accurate sample value is rounded off to the closest numerical value in a set of digital words in use. Figure 3.9 represents the original and the quantized signal. The latter stays at the sample value until the next sampling instant.
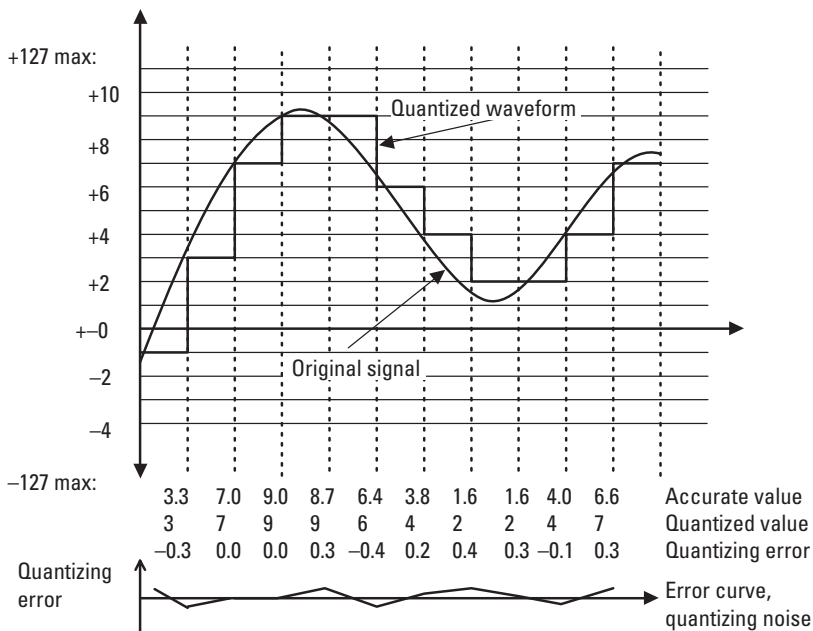
Figure 3.9 Quantizing and noise.

In this quantizing process the information in accurate signal values is lost because of rounding off and the original signal cannot be reproduced exactly any more. The quality of the coding depends on the number of quantum levels that is defined to provide the required performance. The more quantum levels we use, the better performance we get. For example, for a voice signal 256 levels (8-bit binary words) are adequate, but for music encoding (CD recording) 65,536 levels (16-bit binary word) are needed to give sufficient performance.

In the case of binary coding, the number of quantum levels is $q = 2^n$, where q denotes the number of quantum levels and n is the length in bits of the binary code words that describe the sample values.

The better quality we require, the more quantum levels we need and the longer sample words we have to use. This leads to the requirement of a higher bit rate for transmission of the data representing the original message. The data rate must be so high that the digital word of the previous sample will be transmitted before the next one is available for transmission. In each system, a certain compromise has to be made between quality and the data rate.

In uniform quantizing, the quantum levels are uniformly spaced between certain minimum and maximum values of the analog signal. In the next section we consider quantizing noise that the rounding off produces in the case of uniform quantizing.

3.6.3 Quantizing Noise

Quantizing causes signal distortion because the sample values no longer represent accurate values of the analog signal. Usually this distortion caused by rounding off in quantizing is small compared to the signal value. The maximum distortion, that is, maximum quantizing error, is half of the distance between quantum levels. This distortion is heard and theoretically modeled as noise; see the quantizing error curve in Figure 3.9. We can imagine that the decoder first receives accurate sample values and produces a perfect original signal. Then quantizing error is added on top of the perfect signal just as we hear, for example, background noise on top of an ideal voice or music signal.

The rounding off causes an error that is independent of the message because quantizing levels are close to each other and we can assume that the signal has the same probability to be anywhere between two levels at a certain sampling instant as shown in Figure 3.10. This error can be assumed to have a uniform probability density function and a zero mean. When we define the
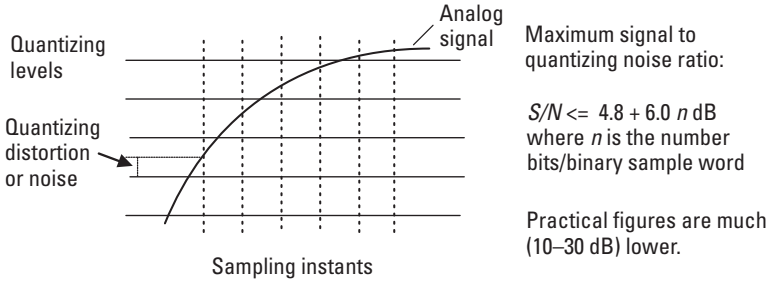
Figure 3.10 Quantizing noise and SQR.

signal to have values between –1 ... +1, it can be shown that the quantum noise power is equal to the variance of quantizing error and is given by

$$N = \sigma_q^{\,2} = \frac{1}{3q^2} \qquad (3.5)$$

where $N = \sigma_q^2$ = quantization noise power and q = the number of quantum levels. [Equation (3.5) gives the variance σ_q^2 of uniform distribution with a value of $q/2$ from $–1/q$ to $1/q$. The variance corresponds to the noise power N when the mean is zero.]

We see that if the number of quantum levels is increased, quantizing noise power decreases rapidly. We get the maximum *signal-to-quantizing noise ratio* (SQR) of linear quantizing when the maximum signal power is equal to one (power is a square of the signal value that was defined to be between –1 and +1):

$$SQR = S/N \leq 3q^2 \qquad (3.6)$$

where S = signal power, $N = \sigma_q^2$ = power of quantization noise, and q = number of quantum levels. The only noise we consider here is generated by quantizing and then SQR = S/N.

We can easily show further that in the case of linear quantizing and binary words, the absolute maximum S/N in decibels in the case of linear quantizing is

$$S / N \leq 10 \log_{10}\left(3q^2\right) = 10 \log_{10}\left(3 \cdot 2^{2n}\right) = 4.8 + 6.0n \text{ dB} \qquad (3.7)$$

where n = the number of bits/word. The maximum S/N is achieved with the maximum signal power that is 1. The logarithmic measure decibel is described at the end of this chapter. The preceding formula gives the absolute maximum S/N of a system that uses uniform quantizing and codes sample values into n-bit binary words.

If we add one bit to the data word representing a linear sample value, we double the number of quantizing levels, which cuts the maximum quantizing error in half. On the other hand, from (3.7) we see that each bit increases the S/N by 6 dB. This means that the quantizing noise power is reduced by a factor of 4 corresponding to error voltage reduction by a factor of 2.

However, we assumed that the average power of the analog signal equals the maximum power, that is, all sample words have the maximum value. In practice, this cannot be the case and the average S/N is some tens of decibels lower than the maximum value given by (3.7). How much lower an average S/N we have in a practical system depends on the dynamic range that we reserve for the highest signal levels (the distance between the average signal power and the maximum signal power) to avoid the clipping of the signal and consequent severe distortion. As an example, if average signal power is 20 dB below maximum, the average S/N (or SQR) is 20 dB below its maximum value given by (3.7).

We have seen that in the quantizing process accurate information about analog signals is lost and we cannot reproduce a perfect original signal anymore. Quantizing errors are heard as noise and to maintain the quality (S/N) of the signal adequately we need to use a large enough number of quantizing levels. The more levels we use, the better the S/N, the longer the binary words used to describe samples, and the higher the data rate required for information transmission.

3.6.4 Nonuniform Quantizing

The goal in the coder design is to get as good an average S/N as possible when the sampling rate and the number of bits for each sample are given. Linear quantizing is not the optimum solution because at low signal levels the quantizing noise is high and the S/N is very low. At high signal levels the quantizing noise is the same even though we would tolerate a high noise level. We should define quantizing levels in such a way that performance is acceptable over a wide dynamic range of the voice. This requires that quantum levels are not uniformly spaced and we call this nonuniform quantizing.

In nonuniform quantizing we use more code words and we have a shorter distance between quantum levels for low-level samples and allow higher quantizing distortion at high-level samples. This is reasonable,

because higher noise is not so disturbing when the signal level is higher as well. To do this, we may compress the voice signal in an encoder and expand it in a decoder. This expanding/compressing process is known as *companding* and is shown in Figure 3.11.

One way to understand the companding process is to think of compressing the dynamic range of the analog signal first by compressor circuitry, which amplifies low levels more than higher levels (Figure 3.11). After this we may use linear quantization, and the signal values after compression and linear quantizing will actually be nonuniformly quantized. In the decoder of the receiver, we use linear quantizing to reproduce the compressed sample values. Then we lowpass filter the sample sequence to reproduce the compressed analog signal. We then expand this analog signal by amplifying low levels less than high levels to cancel the distortion that was produced by the compressor in the encoder. After linear decoding in the receiver, the noise level is the same at any sample level. In expansion a low-level signal is reduced to its original value and quantizing noise is attenuated. This makes the noise level lower at low signal levels than at high signal levels and improves the S/N at low signal levels.

The integrated codec (encoder/decoder) chips that are available for PCM coding include both encoder and decoder circuits. They use signal processing technology to perform companding and we cannot find separate analog nonlinear amplifiers in real-life chips.

An example of a PCM compressor curve for positive analog signal values is presented in Figure 3.12. The horizontal axis represents the original value of an analog voice signal, and the vertical axis gives the output value of the compressor. Uniformly spaced levels of the linear quantizer are shown on the left-hand side. At a certain sampling instant, an analog signal value x is quantized according to the curve into one of the quantum levels of $Z(x)$ and this level is then transmitted as a digital word unique to that level.

When a high signal value changes (see change "b" in Figure 3.12), only a couple of quantizing levels are involved. This is adequate because the quantizing noise does not disturb the listener very much if the signal level is high
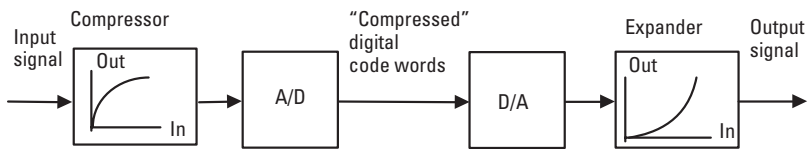
Figure 3.11 Nonuniform quantizing.
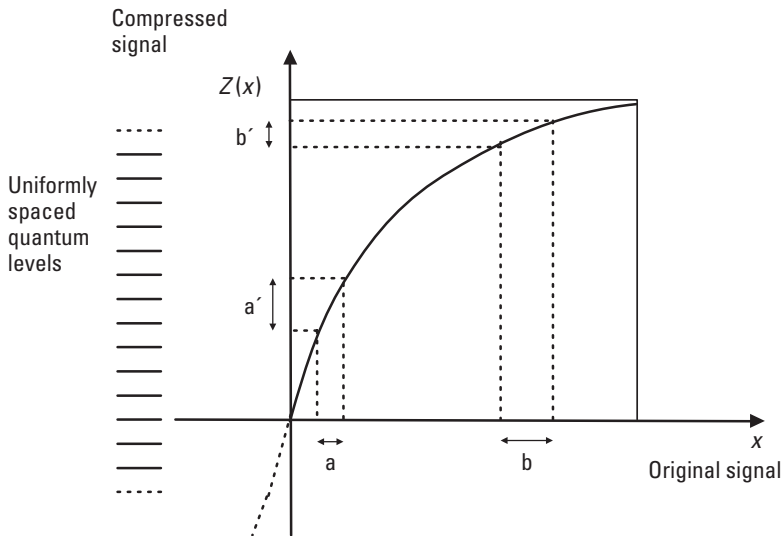
Figure 3.12 Compressor characteristics.

as well. At low levels (see change "a" in Figure 3.12), a small change of signal level uses many quantizing levels; this results in a smaller quantizing error or noise. This improvement of the average S/N at low analog signal levels is essential because noise is most disturbing at low signal levels.

In the decoder, the inverse process is carried out. We can imagine the same curve as in Figure 3.12 but the input values are samples at quantum levels of the vertical axis and the output signal of the expander of the decoder is given as "*x*" on the horizontal axis. Alternatively, we can see an expansion curve as presented in Figure 3.11, where reproduced samples are on the horizontal axis and the output analog signal is given by the values of the vertical axis according to the response curve of the expander.

3.6.5 Companding Algorithms and Performance

As we saw before, we can improve coding performance if the quantization intervals are not uniform but are allowed to increase with respect to the sample value. If we let quantization intervals be directly proportional to the sample value, the SQR will be constant for all signal levels. When the quantization intervals are not uniform (nonlinear quantizing), a nonlinear relationship exists between code words and the sample values that they represent. Two main different nonlinear coding schemes have been standardized

for speech by ITU; they are known as A-law and μ-law coding. Here are some key points about these coding schemes:

- Companding curves are based on the statistics of human voice and many good solutions can be found.
- The two approaches that are standardized internationally are the A-law, which is used in European standard countries (Recommendation G.732 of ITU-T), and the μ-law, which is used in North America and Japan (Recommendation G.733 of ITU-T).
- These schemes provide quite the same quality, but they are not compatible. A conversion device, a transcoder, is needed between countries using different standards.
- Nowadays conversion is a straightforward digital mapping process, in which one digital sample value corresponds to another digital value of another coding scheme.

Various compression–expansion characteristics can be chosen to implement the compander. By increasing the amount of compression, we increase the dynamic range at the expense of the S/N for high signal amplitudes. One family of compression characteristics (Recommendation G.733) used in North America and Japan is the μ-law companding, which is defined as follows:

$$Z(x) = \text{sgn}(x) \cdot \frac{\ln(1 + \mu|x|)}{\ln(1 + \mu)} \qquad (3.8)$$

where x is the signal value, $Z(x)$ represents the compressed signal, $\text{sgn}(x)$ is the polarity (+ or −) of x and μ is the constant with a standard value of 255.

Another approach is A-law companding (Recommendation G.732) used as a European standard, where the curve is divided into linear and logarithmic sections:

$$Z(x) = \begin{cases} \text{sgn}(x) \cdot \dfrac{1 + \ln A|x|}{1 + \ln A} & \text{for } \dfrac{1}{A} < |x| < 1 \\[3mm] \dfrac{Ax}{1 + \ln A} & \text{for } \dfrac{-1}{A} < x < \dfrac{1}{A} \end{cases} \qquad (3.9)$$

where x is the signal value, $Z(x)$ represents the compressed signal, sgn(x) is the polarity (sign) of x, and A is a constant with a standard value of 87.6.

In the case of A-law companding, the SQR is constant in the logarithmic section and directly proportional to the signal value in the linear section; see the dashed line in Figure 3.13. ITU-T/CCITT recommendations define the continuous curves given by the preceding formulas but approximate them with a curve with linear segments for easier implementation.

As an example of the performance of a nonlinear coding scheme, Figure 3.13 represents the SQR dependence on the signal level for A-law companding. The signal level is measured in dBm0, which we explain at the end of this chapter and may vary within the range of 40 dB while SQR remains nearly unchanged. However, when signal level is high, linear quantizing would give better performance, as the "without companding" dashed line shows.

We see from Figure 3.13 that at low levels the SQR of A-law companding is more than 20 dB better than linear coding. The curve gives this performance when the signal is a sine wave and the ripple of the curve is a consequence of the approximation of the compression curve with linear segments.

3.6.6 Binary Coding

Finally, in the PCM encoding process each sample is represented as one in the set of eight-bit binary words. As an example of binary coding, the
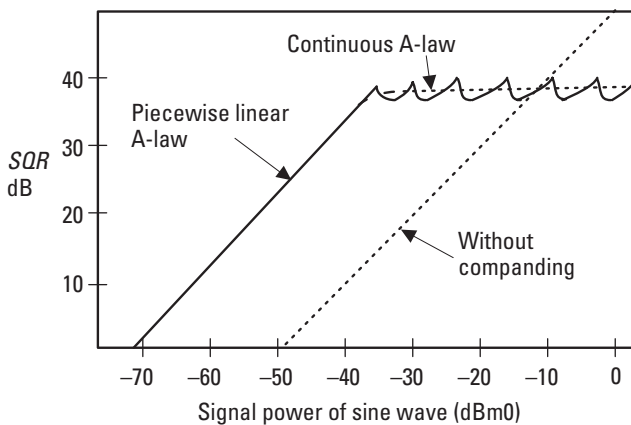
Figure 3.13 Companding performance.

structure of the eight-bit binary word in the case of European PCM coding, the A-law, is defined in the following way:

- *Bit 1, the most significant bit (MSB):* The MSB is the first bit and it reveals the polarity of the sample. Value 1 represents positive polarity and 0 represents negative polarity. The sample value zero may create two different code words depending on whether it has a positive or negative polarity.

- *Bits 2, 3, and 4:* These bits define the segment where the sample value is located. Segments 000 and 001 together form a linear curve for low-level positive or negative samples. Thus an A-law curve has 13 linear sections as shown in Figure 3.14.

- *Bits 5, 6, 7, and 8:* These are the *least significant bits* (LSBs) and they reveal the quantized value of the sample inside one of the segments. Thus each segment is divided in a linear fashion into 16 values (quantum levels).

The structure of the encoded binary word together with the nonlinear relationship between signal values and binary words is shown in Figure 3.14 [1]. Note that both the previously described nonlinear compression and linear coding are combined in the same figure. The vertical axis is linear and each binary word corresponds to one of the quantum levels at equal distance from each other, and linear quantizing is performed for a compressed signal. For compression, the vertical and horizontal axes have a logarithmic relationship according to (3.9). We see, for example, that half of the quantum level is used for signal levels smaller than 6.25% (1/16) of its maximum value to reduce quantizing noise at low signal levels.

Finally, after this encoding process, every other bit of the code words is inverted before multiplexing. This inversion was specified to "mix" the digital signal for easier timing of line systems and equipment interfaces. We can, for example, imagine that the signal value stays at a small negative value that produces the encoded word 00000000, and inversion of every other bit produces the word 01010101. Without the inversion of every other bit, we would transmit continuous zero and might have difficulties synchronizing the receiver with the received data stream. Other coding schemes used to ensure proper synchronization are discussed in Chapter 4.
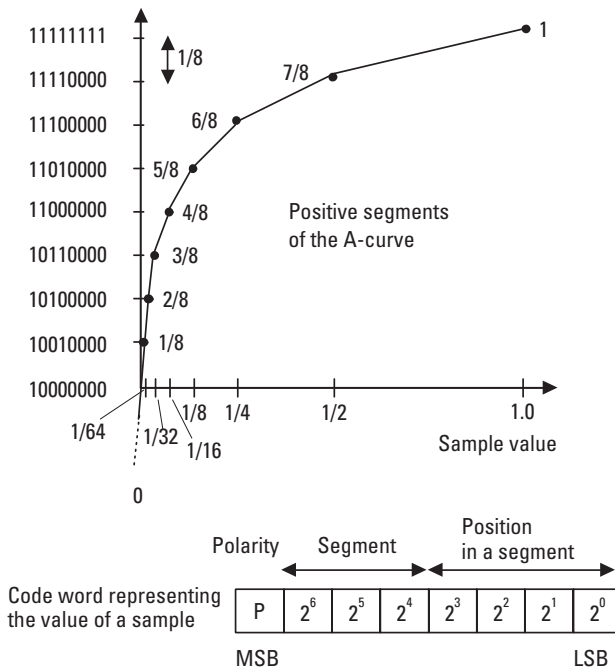
Figure 3.14 Binary coding.

3.6.7 PCM Encoder and Decoder

The PCM coding schemes for digital voice communications were standardized by CCITT (now ITU-T) in the early 1970s. The standards were based on the technology of those days. The European standard was defined to be slightly different from the American standard, which is why conversion equipment is needed when communicating over the Atlantic or from Europe to Japan. Most countries in the world use the European A-law standard. As a conclusion to our discussion about PCM coding, we now look at the block diagrams of the PCM encoder and decoder that contain the processes that we have discussed in previous sections.

3.6.7.1 PCM Encoder

Figure 3.15 presents a block diagram of a PCM encoder based on the European standard. Before actual encoding, the analog signal is filtered into the frequency band from 300 to 3,400 Hz. This bandwidth was defined to be acceptable for sufficient quality human voice so that the speaker can be recognized at the other end. This filtering is mandatory to ensure that the
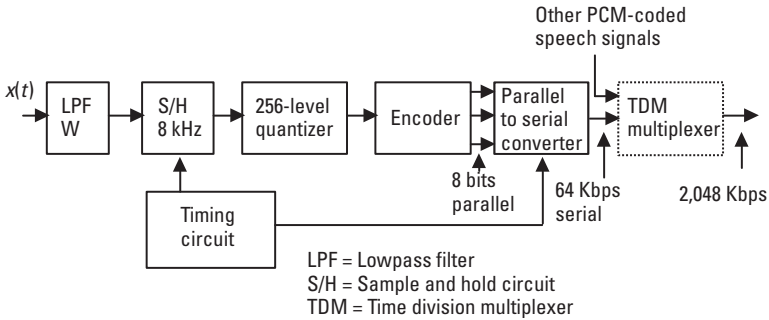
Figure 3.15 PCM encoder.

sampling theorem is satisfied, that is, that the analog signal does not contain frequencies higher than half of the sampling frequency. Then the analog signal is sampled at an 8-kHz sampling frequency and the samples are nonlinearly coded into 8-bit words by a quantizer and an encoder.

Words are then converted into serial form and multiplexed with other PCM-coded voice signals into a 2,048-Kbps primary rate signal that contains 30 voice channels according to the European standard. This 2-Mbps rate is a very common data rate in telecommunications networks. For example, digital exchanges build up 2-Mbps streams with 30 PCM-coded subscriber interfaces for internal transmission inside the equipment. The multiplexing process is described in Chapter 4.

In the United States the corresponding data rate is 1.544 Mbps instead of 2.048 Mbps. In this DS1 system, each frame contains 24 speech channels and a framing bit. The sampling rate is the same 8 kHz and we get

$$8,000 \cdot \left\{ (8 \cdot 24) + 1 \right\} = 1.544 \text{ Mbps} \qquad (3.10)$$

3.6.7.2   PCM Decoder

At the receiver the demultiplexer separates 64-Kbps individual channels that are then converted into 8-bit parallel sample values, as shown in Figure 3.16. Sample pulses are reconstructed and the resulting series is filtered to create a voice signal that closely resembles the original.

3.7   Other Speech-Coding Methods

PCM was standardized during the 1970s and implementation of many more efficient coding methods has become feasible. By "more efficient" we mean
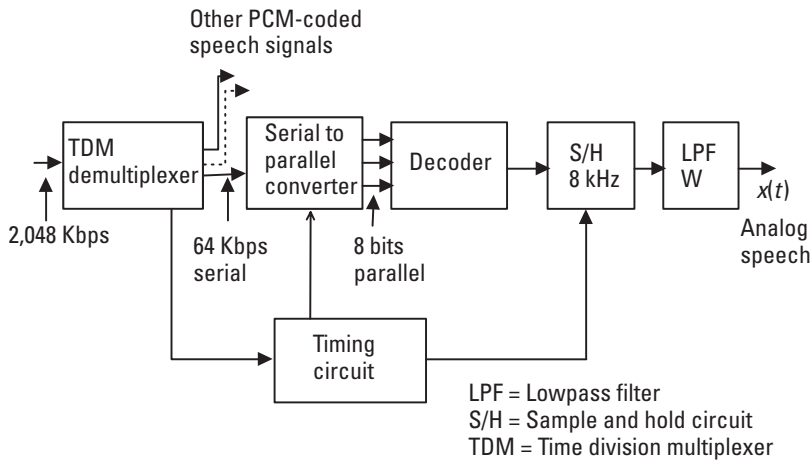
Figure 3.16 PCM decoder.

that we may get better quality at the same data rate or equal quality at a lower data rate. More sophisticated coding schemes are used, for example, in ISDN, where an ISDN telephone may transmit a better quality 7-kHz speech band at 64 Kbps than before. Another example that we will briefly review is GSM, where speech requires only 13 or 7 Kbps.

In the following sections, we review some methods that are used in telecommunications networks in addition to the PCM discussed in the previous section. We can divide voice coding methods into two categories: waveform and voice coding (vocoders) [1]. In waveform coding, such as PCM, we transmit information that describes a signal waveform in time domain.

In vocoders we use characteristics of human voice. To understand the basic principle of vocoders, imagine that we have a set of signal models each identified by a code. We divide speech into, for example, 50-ms segments and choose one of the models that is closest to the signal to be encoded and send its identification code to the other end. The decoder reproduces the signal corresponding to the received code. Vocoders may also split voice signals into several "components" in the frequency domain, each of them modeled separately for better quality. Vocoders introduce additional delay because each speech segment has to be analyzed before encoding. Waveform coding does not add delay and it usually give better quality but requires a higher data rate than vocoders. To achieve a suitable compromise between the quality and the data rate, the two basic principles are sometimes combined into hybrid coders.

In conventional PCM we encode all samples independently. We can improve encoding performance by assuming that the next sample value is not independent from the previous one, which is the case in practice.

3.7.1   Adaptive PCM (APCM)

APCM is a variation of conventional PCM in which signal strength information is transmitted periodically in addition to sample values. Now a smaller number of bits is needed for samples and they define the quantum level inside a given scale. If the signal level is high, the quantizing error is high because the same number of levels is used for all samples. On the other hand, for low signal levels the quantizing error is small and SQR can be kept high enough over a wide range of signal levels. This principle is used, for example, in original GSM as part of the voice coding process.

3.7.2   Differential PCM (DPCM)

In DPCM only the difference between a sample and the previous value is encoded as shown in Figure 3.17. Because the difference is typically much smaller than the overall value of the sample, we need fewer bits for the same accuracy as in ordinary PCM and the required bit rate is reduced [1]. In the example shown in Figure 3.17, PCM requires 5 bits (polarity and 4 bits for 16 quantum levels). DPCM, in which the only difference from the previous sample is encoded, 4 bits is clearly enough to describe the difference between subsequent samples.

For better quality or to further reduce the data rate, DPCM may use several of the preceding samples to predict the next sample. The example in Figure 3.17 shows that if two previous samples are used for prediction the encoder and decoder assume that the next sample follows the same slope. Now, 3 bits would be enough for the encoder to describe the difference between the prediction and the actual sample value. The decoder performs the same prediction and only the difference between predictions shown in Figure 3.17 need to be transmitted. Further improvement can be achieved if three previous sample values are used for prediction, but more than three samples do not add much advantage [2]. Actually, the first simple form of DPCM in Figure 3.17, which encodes the difference between preceding sample values, uses prediction as well but that prediction is based on only one sample and it equals the previous sample value.
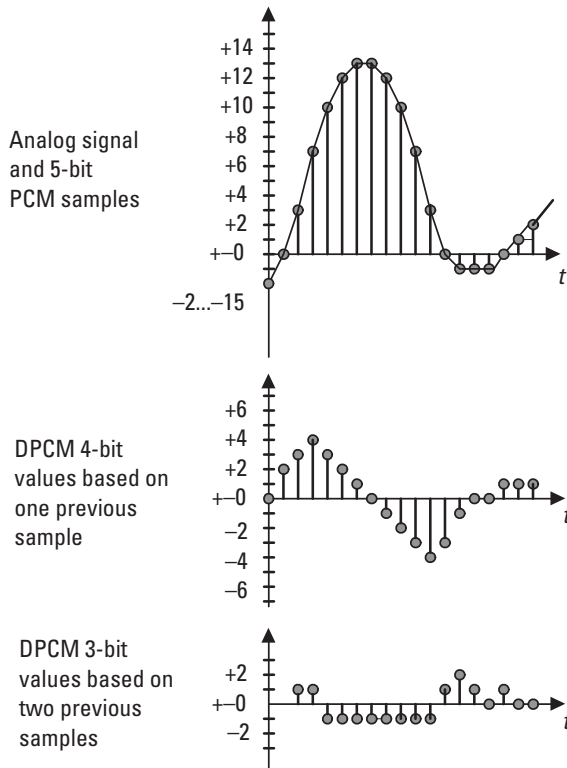
Figure 3.17 DPCM.

These waveform coding methods do not introduce much delay because prediction is based on previous sample values. DPCM methods require that absolute sample values be transmitted periodically to prevent propagation of errors. DPCM is sometimes used for digitized video transmission.

3.7.3 DM

DM is a very simple type of DPCM that transmits the binary value 1 if the sample is higher than the previous one. Binary value 0 is transmitted if the signal value has decreased. A variation of DM uses large quantizing steps when the signal contains steep slopes and small steps when the signal does not change much. This method is called *continuous variable slope delta* (CVSD) modulation and it is an alternative to ordinary PCM in Bluetooth speech transmission. CVSDM is also used in military voice applications [3].

3.7.4   Adaptive DPCM (ADPCM)

ADPCM combines two previously described methods, APCM and DPCM. Further compression is achieved by adapting the predictor and the quantizer to the characteristics of the signal. Both the encoder and the decoder use the same algorithm to estimate the values of the following samples with help of the preceding samples, and only the error to this estimate is transmitted as in DPCM in Figure 3.17. To further reduce the number of bits per sample, ADPCM adapts quantizing levels to the characteristics of the analog signal. Figure 3.18 shows a simplified example in which the prediction error is initially small and all bits can be used for half of the full quantizing error scale. Then the prediction error increases and the quantizing step size is doubled to describe higher values of prediction error. When the prediction error decreases, the quantizing step size is reduced again to describe properly small errors. Adaptation information is transmitted from encoder to decoder in addition to the prediction error.

In the original 32-Kbps ADPCM method, the difference between the predicted and actual sample value is coded with four bits, that is, into 15 quantum levels, and the data rate is half that of conventional PCM. If several subsequent samples vary widely, the quantizing steps are adapted to that change so that four bits are enough for prediction error. If prediction errors tend to increase, quantizing steps are increased and vice versa.
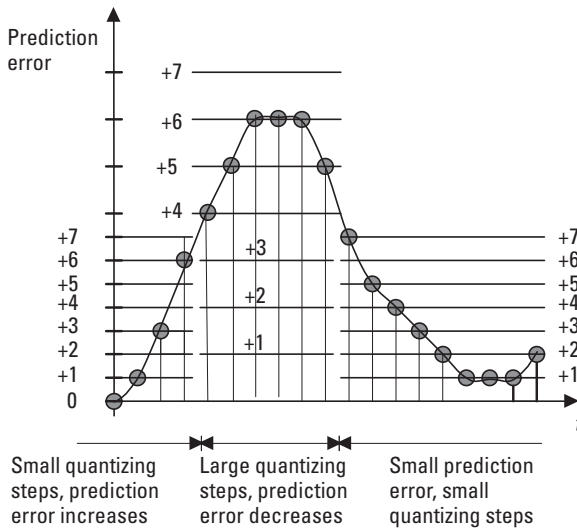
Figure 3.18 ADPCM principle.

According to the ADPCM standard, commercial voice quality is coded into 32 Kbps or even a lower (24 or 16 Kbps) bit rate. Samples are still taken at 8 kHz but transmitted with four bits (in the case of 32-Kbps ADPCM) and the quality is equal, or at least close, to the quality of ordinary PCM.

Recommendation G.728 for 16/24/32/46-Kbps ADPCM was approved by the ITU-T in 1990 and has been adopted worldwide for digital voice transmission between countries or within a country. It can partly resolve the current compatibility problem between North America's and Europe's PCM formats, due to their different companding schemes, by acting as a common language between the two PCM schemes.

There is also a recommendation for an ADPCM algorithm (G.722) that will code 7.1-kHz bandwidth audio signals into 64 Kbps. This coding scheme improves the quality of speech and it can be used for good quality voice over ISDN networks.

ADPCM systems are available on the market that convert two primary rate PCM streams into one data stream at the same rate by using ADPCM. Two 32-Kbps ADPCM channels occupy one ordinary PCM channel. Network operators use ADPCM to utilize long-distance transmission systems, for example, submarine systems, more efficiently. Another application example is in PABX networks where the offices of private enterprises are interconnected by leased-line 64-Kbps channels. ADPCM doubles the capacity of these expensive leased lines between PBX/PABXs. One application for ADPCM is also in cordless telephones such as *digital enhanced cordless telecommunications* (DECT).

The ADPCM coding scheme is based on the statistics of speech and it does not support modem or facsimile signals at higher data rates than 4,800 bps. Because of this, telecommunications network operators cannot use ADPCM instead of PCM coding for all calls. This is a problem if ADPCM systems are used inside a telecommunications network. One way to overcome this problem is to have the ADPCM encoder detect whether a data or facsimile connection is to be established and in that case disable the PCM/ADPCM transcoder for that channel.

Up to this point we have discussed primarily waveform coding methods. The phrase *waveform coding* refers to attempts to describe the shape or the waveform of the original analog signal, just as PCM, DPCM, and ADPCM do. In a more efficient coding scheme in terms of data rate, such as *voice coding,* which is implemented by vocoders, we divide speech into segments with lengths of some tens of milliseconds. Then we analyze each segment to find a model that describes it best and send parameters of the model instead of trying to imitate the shape of the signal. An example of the so-

called hybrid methods that use both of the two main principles discussed earlier is the voice coding of a cellular network, as briefly reviewed next.

3.7.5   Speech Coding of GSM

In cellular networks an efficient coding scheme is needed in order to make maximum use of radio frequencies. The lower our data rate, the narrower the frequency band we need for each call and the more simultaneous calls a given frequency band supports, as we will see in Chapter 4. As an example of these efficient coding schemes we now briefly review the principle that is used in the GSM.

During efforts to standardize the speech-coding algorithm for GSM, the goal was to achieve a 16-Kbps data stream with the same speech quality as ordinary PCM. Waveform coding, such as PCM or ADPCM, did not give sufficient quality at this low data rate. Voice coding methods did give a low enough data rate but not good enough quality. In a voice coder or vocoder, the signal is modeled and the codes of the sound elements are sent. In the decoder the speech is reproduced.

A combination of these two basic principles was selected. The maximum processing delay was restricted to less than or equal to 65 ms, which requires the use of echo cancellers in the network. The original data rate became 13 Kbps, which was further reduced to 7 Kbps in 1995 with a more efficient coding algorithm.

The selected efficient speech coding is always used at the radio path where efficient utilization of transmission channels is more important than in the wireline network. We will see in Chapter 5 that to increase the radio interface capacity we need to make cells smaller and build more base stations, which is very expensive. For switching and interconnection to a fixed telecommunications network, GSM coding is changed into ordinary PCM.

GSM's operating principle is as follows. The voice signal is first divided into 20-ms slices. Each slice of the signal is analyzed and the periodicity is noticed. The periodical component is subtracted by an analysis filter from the original signal and the amplitude of the voice signal level is considerably reduced (Figure 3.19).

The periodical high-power component is transmitted as a set of parameters, and the low-level error or difference signal at the output of the analysis filter is waveform coded. This waveform coding does not require a high bit rate because the amplitude of the error signal is low.

At the receiving end, a synthesis filter is used and, with the help of the transmitted coefficients, it adds the periodical component to the error signal, which is reproduced from waveform-coded samples.
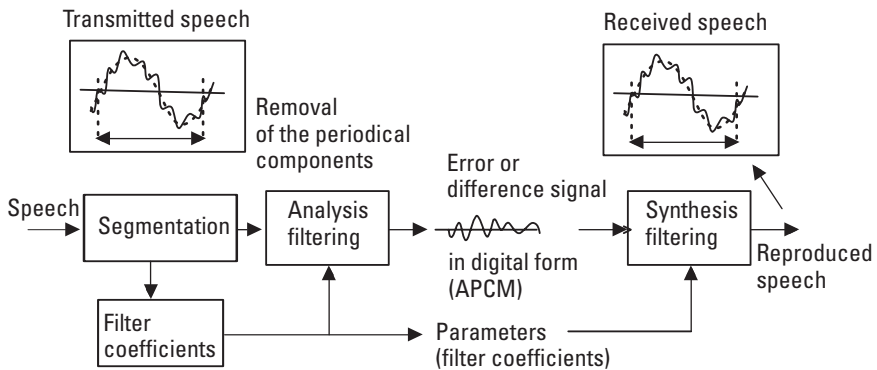
Figure 3.19 The principle of GSM speech coding.

3.7.6   Summary of Speech-Coding Methods

We have introduced some important standardized coding methods such as PCM, ADPCM, and GSM radio channel voice coding schemes that are widely used in public telecommunications networks. However, in private PABX networks more efficient coding schemes are sometimes attractive, because the charge for leased lines between office sites is based on the chosen bit rate capacity and we can accept worse quality than in public networks.

One way to radically reduce the bit rate required is to use voice coding implemented by vocoders. The speech is first synthesized (or modeled) and the resulting parameters are then encoded for transmission instead of the actual signal. This method is also used for speech synthesis (speech generation). These types of algorithms actually try to imitate the human vocal tract, utilizing codebooks of common phonetic sounds and transferring these codes between the encoder and decoder.

The quality of vocoder is worse than the quality of waveform coders. Vocoders sound synthetic. They do not meet the quality requirements of the telephone network, one of which is speaker recognition, but they can be used in private networks. This principle is also used as a part of the GSM speech-coding scheme together with waveform coding as we saw in the previous section.

The service quality of a telephone channel is governed by many factors, including volume, distortion, background noise, round-trip delay, and echo loss. We have many ways in which to measure quality. The results in Figure 3.20 are based on a so-called *mean opinion score* (MOS) measurement, in which many people have given their opinions about the quality.

The interactive nature of human conversation places a demand on the coder in terms of an acceptable path delay. Subjectively, noticeable deterioration is perceived in channel quality once the round-trip delay exceeds 180 ms. Note that one-way delay via a geosynchronous satellite is approximately 250 ms. For high-quality voice, the round-trip delay should be less than 150 ms.

Another problem is an echo, which is noticeable if the delay is more than 30 to 50 ms. Delays of 10 to 20 ms are generally undetectable. Some echo is always produced at the far end by a 4W/2W hybrid of the far-end subscriber loop described in Chapter 2 because of the nonideal return loss of the hybrid. This is why, in the case of a long delay (e.g., on satellite channels), echo cancellers are needed. The long coding delay (e.g., GSM) also requires echo cancellers.

Figure 3.20 provides a comparison of the coding schemes discussed in this chapter. The measure of quality is the MOS, which indicates the average opinion expressed by a number of people about the quality of each coding scheme.

We have introduced just some of the available speech-coding methods; many other standardized speech-coding schemes are in use at different data rates. ITU standards cover constant-bit-rate coders at data rates down to 5 Kbps. For cellular networks many different lower rate coders are defined and
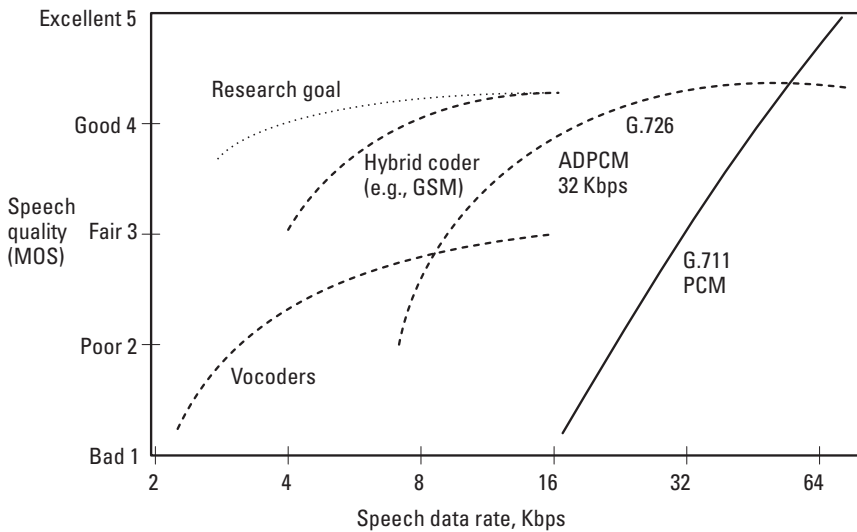
Figure 3.20 Comparison of speech-coding techniques.

they operate at a fixed data rate from 3 to13 Kbps. In the United States a variable-bit-rate coder is used in the *code division multiple access* (CDMA) cellular network. It varies the data rate between 1 and 9 Kbps depending on the speech characteristics.

3.8   Power Levels of Signals and Decibels

In this final section on signals we explain the *decibel*, a measure of signal level and its change. We use this logarithmic measure or its variants in the telecommunications network for many purposes, for example, to express the voice level or the transmission and reception power of radio systems, such as mobile telephones, or an optical line system.

3.8.1   Decibel, Gain, and Loss

Along the long-distance communication connection or channel, the power of the signal is reduced and amplified over and over again. The signal power needs to be rigidly controlled to keep it high enough in relation to background noise and low enough to avoid system overload and resulting distortion.

The reduction of signal strength, loss or attenuation, is expressed in terms of *power loss*. When the signal is regained, this is expressed in terms of *power gain*. Thus the absolute gain of ten corresponds to the loss of 1/10.

Alexander Graham Bell was the first to use logarithmic power measures. This was found to be handy and the unit for power gain was named in Bell's honor as decibel (dB). The gain in decibels is defined as follows:

$$g_{dB} + 10 \log_{10} g = 10 \log_{10} \left(\frac{P_{out}}{P_{in}} \right) \qquad (3.11)$$

If the output and input powers are the same, the absolute gain and loss both have values of 1 and the corresponding gain and loss in decibels are each 0 dB. If the gain is 10, the corresponding decibel value of gain is 10 dB. The loss is correspondingly 1/10, that is, equal to –10 dB. Thus if the power is reduced, the gain in decibels results in a negative value. Figure 3.21 presents an element in a telecommunications network with a certain input power and an output power. The formulas of loss (attenuation) and gain are given in the figure as well.

In telecommunications systems we usually have many elements in a chain. If the overall gain or loss needs to be calculated, all gain figures (which

Gain

$g = \dfrac{P_{out}}{P_{in}}$

Gain in dB

$g_{dB} = 10 \log 10 \dfrac{P_{out}}{P_{in}}$

Loss

$L = \dfrac{P_{in}}{P_{out}}$

Loss in dB

$L_{dB} = 10 \log 10 \dfrac{P_{in}}{P_{out}}$

Overall gain: $g = g_1 {}^* g_2$

Overall gain in dB: $g_{dB} = g_{1,dB} + g_{2,dB}$
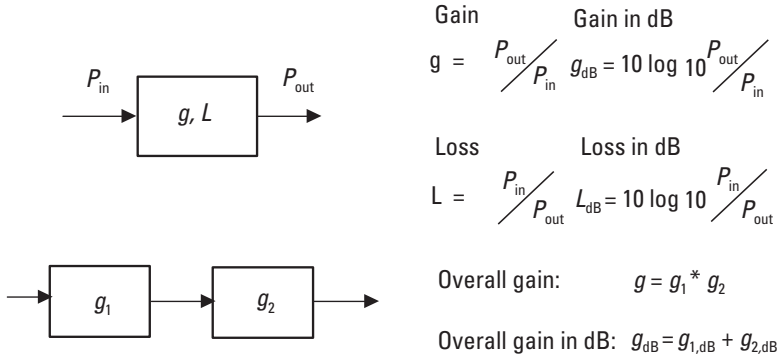
For example, a gain of 100000000 corresponds to the gain of 80 dB

Figure 3.21 Gain, loss, and decibels.

often are very large or small numbers) must be multiplied. If the gain of each element is presented in decibels, the figures (which usually have values of less than 100) are added along the chain to determine the overall gain in decibels as shown in Figure 3.21.

Decibels allow us to add small positive or negative numbers instead of multiplying with very large or very small numbers. For example, a gain of 100,000,000 corresponds to a gain of 80 dB.

Note that the decibel is the measure of power gain and, if we are interested in how voltage level changes, the impedances must be considered. The voltage and power gains are the same only if the impedances at the points where the power and voltage are measured are the same. The following formula gives the power gain if input and output voltages and impedance are known:

$$g_{dB} = 10 \log_{10} \left(\frac{P_{out}}{P_{in}} \right) = 20 \log_{10} \left(\frac{V_{out}}{V_{in}} \right) + 10 \log_{10} \left(\frac{Z_{out}}{Z_{in}} \right) \quad (3.12)$$

The impedances in the preceding equation are assumed to be real numbers.

3.8.2   Power Levels

In the previous section we expressed power ratios in decibels. That does not tell us anything about the actual power in watts, only the ratio. Instead of expressing the actual power in watts, we can use the decibel-based figures for

this measurement also. Power levels in practical systems may vary from picowatts to tens of watts, corresponding to variations from 1 to 1,000,000,000,000. Power measures based on decibels can be used to express this wide power range in an easy way.

The level of absolute power is often expressed in dBm, where the actual power is compared to 1-mW power. The power level in dBm is given by the expression $10 \log_{10}(P/1 \text{ mW})$ dBm. If we need to know absolute power in watts, we can easily calculate it from the given dBm value. Absolute power level dBm is commonly used instead of the absolute power in watts to express, for example, the optical output and received power of optical line systems or the received radio signal strength of a mobile telephone.

It is very handy to use power levels in dBm together with gain or attenuation in decibels. Assume that the input power level in Figure 3.21 is given in dBms and we know gain in decibels. Then we obtain the output power level in dBm simply by adding input level and gain. This comes from

$$
\begin{aligned}
P_{out} = g \cdot P_{in} &\rightarrow \quad P_{out}/1 \text{ mW} = g \cdot P_{in}/1 \text{ mW} \\
&\rightarrow \quad 10 \log_{10}\left(P_{out}/1 \text{ mW}\right) = 10 \log_{10} g + 10 \log_{10}\left(P_{in}/1 \text{ mW}\right) \\
&\rightarrow \quad P_{out,dBm} = g_{dB} + P_{in,dBm}
\end{aligned}
$$

$$(3.13)$$

To illustrate the use of decibels, we will look at some examples. Let us consider the radio relay system shown in Figure 3.22. Antenna gains and radio link loss are usually measured or given in decibels and receiver sensitivity in dBm. To determine the received power level, we first change
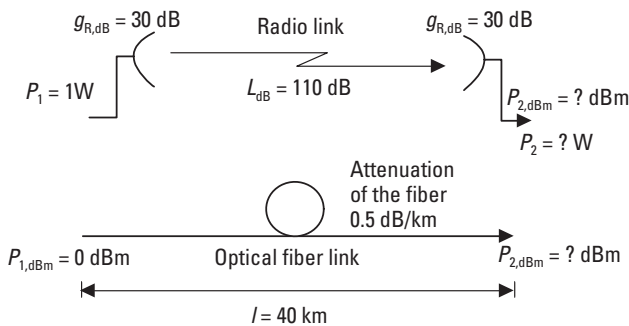
Figure 3.22 Example radio relay and optical fiber systems.

transmission power P_1 = 1W into dBm power level according to 10 $\log_{10}(P_1/1 \text{ mW})$ dBm = +30 dBm. Then we simply derive the received power level as $P_{2,\text{dBm}}$ = +30 dBm + 30 dB –110 dB +30 dB = –20 dBm. If we need the received power expressed in watts, we solve the equation –20 dBm = 10 $\log_{10}(P_2/1 \text{ mW})$ dBm to get P_2 = 10 µW.

Another example in Figure 3.22 shows an optical line system where transmission power level and length and attenuation of the fiber are given. Total fiber attenuation becomes L_{dB} = 40 km · 0.5 dB/km = 20 dB and then received power level will become $P_{2,\text{dBm}}$ = $P_{1,\text{dBm}}$ – 20 dB = –20 dBm.

We have reviewed the main two types of decibel measures used in the engineering of a telecommunications network. Many others are in use, but they will be easy to understand if the need arises if the reader is familiar with the most important ones just discussed: decibel and dBm. Our goal was merely to introduce the reader to the decibel measure.

3.8.3   Digital Milliwatt

As we have seen, PCM systems have a strictly limited operational range. The upper limit is defined by the code word representing the maximum sample value. If an analog signal has a higher amplitude, it is severely distorted because of clipping. The other limiting factor is quantizing noise, which reduces performance as the signal level decreases.

In a digital international connection, the PCM equipment at both ends has to be compatible and it has to convert digital information into the same analog signal level and vice versa. Therefore, control of the power level at the PCM encoder input is extremely important. For this purpose, the ITU-T has defined a digital sequence of code words. By decoding this sequence, a 1-kHz sine wave is produced at the 0-dBm power level (Figure 3.23).

A digital milliwatt can be understood to be the reference signal for the analog signal levels in the network. The actual measured signal power level is written as a dBm0 value when it is compared with the reference level generated by the digital milliwatt, which then generates a 0-dBm0 level to all analog signal points. If the nominal signal level at a measurement point is designed to be lower, the measured dBm value is lower as seen in Figure 3.23. There a PCM decoder contains a 4-dB attenuator and the actual signal level is –4 dBm when the digital milliwatt is applied. However, the measured decoder output level is the same as the level generated by the digital milliwatt and thus it can also be written as 0 dBm0.

Figure 3.13 showed the SQR of a PCM codec, and the measurement signal level was compared to the reference level generated by a digital milliwatt and thus written as a dBm0 value. The overload threshold of the PCM coder
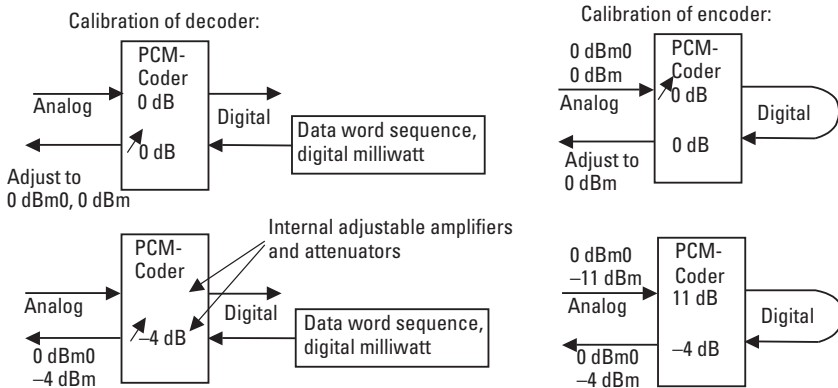
Figure 3.23 Digital milliwatt.

is +3.14 dBm0 (1-kHz sine wave) and the higher level signal is distorted. Note that 0 dBm0 is only a reference level for testing and measurement purposes and the actual average level of speech in an analog speech channel is of the order of –15 dBm.

Measuring systems that generate a bit sequence of the digital milliwatt are used for decoder calibration as shown in Figure 3.23. When this is done, we can loop a digital signal back from the encoder to the decoder and adjust the encoder so that 0 dBm at the input of the encoder produces 0 dBm at the output of the decoder.

The digital milliwatt for European PCM is defined by the 8-word data sequence shown in Table 3.2. Note that before decoding we have to invert every other bit, namely, bits 2, 4, 6, and 8. The PCM decoder produces a 1-kHz sine wave at the 0-dBm power level when this sequence is inserted into the digital input of the decoder.

We often find it necessary to adjust the power level at the interface of PCM equipment so that the following system will not be overloaded. For this purpose PCM equipment contains adjustable amplifiers and attenuators. Figure 3.23 shows an example in which an analog input signal is amplified by 11 dB before encoding and attenuated by 4 dB after decoding.

3.9   Problems and Review Questions

Problem 3.1

Explain how the characteristics of digital data and voice communications differ.

Table 3.2
Data Sequence for Digital Milliwatt

Word	Bit Number							
	1	2	3	4	5	6	7	8
1	0	0	1	1	0	1	0	0
2	0	0	1	0	0	0	0	1
3	0	0	1	0	0	0	0	1
4	0	0	1	1	0	1	0	0
5	1	0	1	1	0	1	0	0
6	1	0	1	0	0	0	0	1
7	1	0	1	0	0	0	0	1
8	1	0	1	1	0	1	0	0
1	0	0	1	1	0	.	.	.
.	0	0	.	.	.	.	.	.

Problem 3.2

What is the wavelength λ of the radio signal for (a) a 100-MHz FM radio and (b) a 10-GHz microwave radio relay system?

Problem 3.3

A voltage waveform of a signal follows the equation $x(t) = 5\cos(1 \cdot 10^3\ t)$ V, where t = time. What are the frequency, amplitude, radian frequency, and periodic time (period) of this signal?

Problem 3.4

Draw the signal $v(t) = 5\cos(1 \cdot 10^3 t + \pi/2)$ V. The vertical scale should be in volts and the horizontal scale in milliseconds.

Problem 3.5

Compare digital telecommunications technology with analog technology and list the most important advantages of digital technology.

Problem 3.6

What are the main three phases of PCM encoding (A/D conversion)? Explain how they are performed.

Problem 3.7

What is nonuniform quantizing and why is it used?

Problem 3.8

What is the minimum sampling rate of speech when the frequency band is 300 to 3,400 Hz and what is the minimum sampling frequency for high-fidelity music of 20 Hz to 20 kHz?

Problem 3.9

Draw the spectrum of an analog signal after sampling when the sampling frequency is 8 kHz and the signal that is sampled is a sine wave with a frequency of 1 kHz. Draw the spectrum for each case when the analog signal frequency is 2, 5, and 6 kHz. What happens in each case when we reconstruct the original signal from the sampled signal with a lowpass filter that has a bandwidth of 4 kHz?

Problem 3.10

The digital *compact disc* (CD) player is designed for a sound bandwidth of 20 kHz. Linear encoding with 16 bits per sample is used. Define (a) the minimum sampling rate, (b) the minimum binary data rate per channel (left or right), (c) the maximum SQR, and (d) the average SQR if the average signal level is 30 dB below the maximum value.

Problem 3.11

Estimate what bit rate would be needed for each voice channel in the digital telephone network if linear PCM coding is used. The same performance, with SQR at least 40 dB at signal levels higher than –40 dBm0 (sine wave), is required. (*Hint:* Estimate with the help of Figure 3.13 how much quantizing noise should be reduced at signal level –40 dBm0 and how much longer sample words would be required for this.)

Problem 3.12

How much PCM voice or stereo music (assume that CD-quality music requires 700 Kbps for both channels) can be stored in (a) a 1.44-MB (B = byte = 8 bits) disc and (b) a 20-GB memory space of a hard disk?

Problem 3.13

Explain how (a) DPCM and (b) ADPCM reduce data rates compared to ordinary PCM.

Problem 3.14

Explain the basic principle of GSM speech coding.

Problem 3.15

The input power of an amplifier is 2 mW and output power is 1W. What are the power levels (dBm) at the input and output and what is the gain of the amplifier in decibels?

Problem 3.16

The input and output powers of a circuit are in listed in Table 3.3. What is the absolute attenuation *L*, absolute gain *g*, attenuation in decibels, gain in decibels, and output power level for each case (a–e)?

Problem 3.17

What is the power in watts that corresponds to power levels of (a) 0 dBm, (b) 3 dBm, (c) −3 dBm, (d) 10 dBm, (e) 20 dBm, (f) 100 dBm, and (g) −100 dBm?

Problem 3.18

Figure 3.24 illustrates a telecommunications connection using a geostationary satellite. Calculate the input and output powers of the satellite amplifier and output power of the antenna at the receiving Earth station. Define both power levels in dBm and absolute power in watts. Use decibels and derive power levels, in dBm values, first.

Problem 3.19

The input power of a 40-km cable system is 2W (power at the beginning of the cable). An amplifier with a 64-dB gain is installed 24 km from the

Table 3.3
Input and Output Powers of a Circuit

	P_{in}	P_{out}
(a)	1 mW	1 mW
(b)	1 mW	0.5 mW
(c)	1 mW	4 mW
(d)	10 mW	10 W
(e)	10 W	10 mW
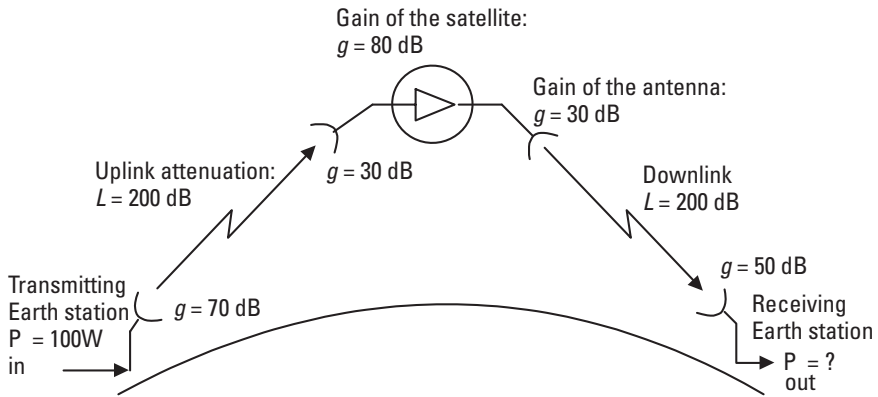
Figure 3.24 Satellite transmission link.

input. Define the signal power level, dBm, and absolute power at (a) the input of the amplifier and (b) the output of the system. The attenuation of the cable is 2.5 dB/km.

Problem 3.20

Explain the meaning and purpose of the decibel units dB, dBm, and dBm0.

Problem 3.21

Sound pressure level is defined as decibels, $L_p = 20\log \dfrac{p}{p_0}$ dB, where $p =$ sound pressure (in Pascals) and $p_0 = 20\ \mu Pa$ (20 micropascals, reference level). The threshold for human hearing is about 0 dB, and threshold for pain about 140 dB. How many times stronger is the sound pressure of the strongest sound that we can hear without pain compared to the weakest one?

Problem 3.22

Explain the term *digital milliwatt.*

Problem 3.23

Draw the analog waveform generated by a PCM decoder with a digital milliwatt as the input signal of the decoder. Use Figure 3.14 to determine the approximate analog signal value. What is the periodic time and what is the frequency of the analog signal produced?

References

[1]   Ericsson Telecom, *Understanding Telecommunications,* Vol. 1, Lund, Sweden: Ericsson Telecom, Telia, and Studentlitteratur, 1997.

[2]   Sklar, B., *Digital Communications: Fundamentals and Applications,* Upper Saddle River, NJ: Prentice Hall, 1988.

[3]   Bray, J., and C. F. Sturman, *Bluetooth Connect Without Cables*, Upper Saddle River, NJ: Prentice Hall, 2001.

4

Transmission

Transmission is the process of transporting information between end points of a system or network. As we have seen in previous chapters, the end-to-end communication distance is often very long and there are many electrical systems on the line. These systems, network elements such as exchanges, are connected to the other elements with connections provided by the transmission systems. In this chapter we discuss the basic restrictions and requirements for transmission and the characteristics of various transmission media and equipment used in the telecommunications core network. The transmission systems for access networks for high-data-rate customer access to the Internet are discussed in Chapter 6.

4.1   Basic Concept of a Transmission System

In this first section we look at the basic elements present in all transmission systems. We introduce the basic functions of these elements and discuss their roles for the successful transmission of information.

4.1.1   Elements of a Transmission System

The main elements of a communication system are shown in Figure 4.1. The transducers, such as a microphone or a TV camera, that we need to convert an original signal to an electrical form are omitted; unwanted disturbances such as electromagnetic interference and noise are included. Note that

Input   Transmitted   Received   Output
signal   signal   signal   signal

| Transmitter | | Transmission Channel | | Receiver |

Source   Destination

Noise, distortion,
and interference
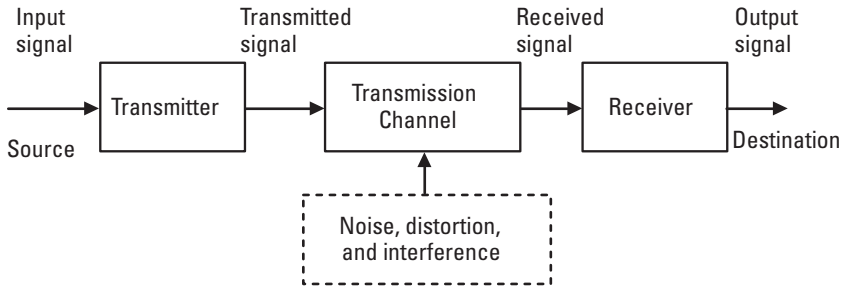
Figure 4.1   Basic concept of transmission system.

bidirectional communication requires another system for simultaneous transmission in the opposite direction.

4.1.1.1   Transmitter

The transmitter processes the input signal and produces a transmitted signal suitable to the characteristics of a transmission channel. The signal processing for transmission often involves encoding and modulation. In the case of optical transmission, the conversion from an electrical signal format to an optical one is carried out in the transmitter.

4.1.1.2   Transmission Channel

The transmission channel is an electrical medium that bridges the distance from the source to the destination. It may be a pair of wires, a coaxial cable, a radio path, or an optical fiber. Every channel introduces some amount of transmission loss or attenuation and, therefore, the transmitted power progressively decreases with increasing distance. The signal is also distorted in the transmission channel because of different attenuation at different frequencies. Signals usually contain components at many frequencies and if some are attenuated and some are not, the shape of the signal changes. This change is known as *distortion.* Note that a transmission channel often includes many speech or data channels that are multiplexed into the same cable pair or fiber. The principle of multiplexing is explained later in this chapter.

4.1.1.3   Receiver

The receiver operates on an output signal from the channel in preparation for delivery to the transducer at the destination. Receiver operations include filtering to take away out-of-band noise, amplification to compensate for

transmission loss, equalizing to compensate for distortion (different attenuation of frequency components), and demodulation and decoding to reverse the signal processing performed at the transmitter.

4.1.1.4 Noise, Distortion, and Interference

Various unwanted factors impact the transmission of a signal. Attenuation is undesirable because it reduces signal strength at the receiver. Even more serious problems are distortion, interference, and noise, the last of which appears as alterations of the signal shape. To decrease the influence of noise, the receiver always includes a filter that passes through only the frequency band of message frequencies and disables the propagation of out-of-band noise.

4.1.2 Signals and Spectra

Electrical communication signals are time-varying quantities such as voltage or current. Although a signal physically exists in the *time domain*, we can also represent it in the *frequency domain* where we view the signal as consisting of sinusoidal components at various frequencies. This frequency-domain description is called the *spectrum.*

Any physical signal can be expressed in both domains. In the time domain we draw the amplitude along the time axis and in the frequency domain we draw the amplitude (and phase) along the frequency axis. Although both of them give a perfect description of the signal, both presentations are needed for easier understanding of the different phenomena. The time-domain signal is the sum of the spectral sinusoidal components. Fourier analysis gives the mathematical connection between the time- and frequency-domain descriptions. Here we merely introduce the connection between the time-and frequency-domain descriptions with a couple of examples. The reader is referred to [1] for mathematical treatment of the transformation between the time and frequency domains.

In Figure 4.2, two examples of time-domain signals and corresponding spectrums are presented. In the first example we see an ordinary rectangular digital pulse with duration of T seconds and its corresponding spectrum. If, for example, the pulse a duration $T = 1$ ms, the strongest spectral content lies below 1 kHz ($1/T = 1/1$ ms $= 1,000$ 1/s $= 1$ kHz), as shown in Figure 4.2. From this we get a thumb rule that we can send 1,000 pulses of this kind in a second through a channel with a bandwidth of 1 kHz, which corresponds to a 1-Kbps binary data rate.

To increase the data rate, we should decrease T and the spectral width and the required bandwidth is increased correspondingly. For example, for a

A pulse and its spectrum

Radio frequency burst and its spectrum
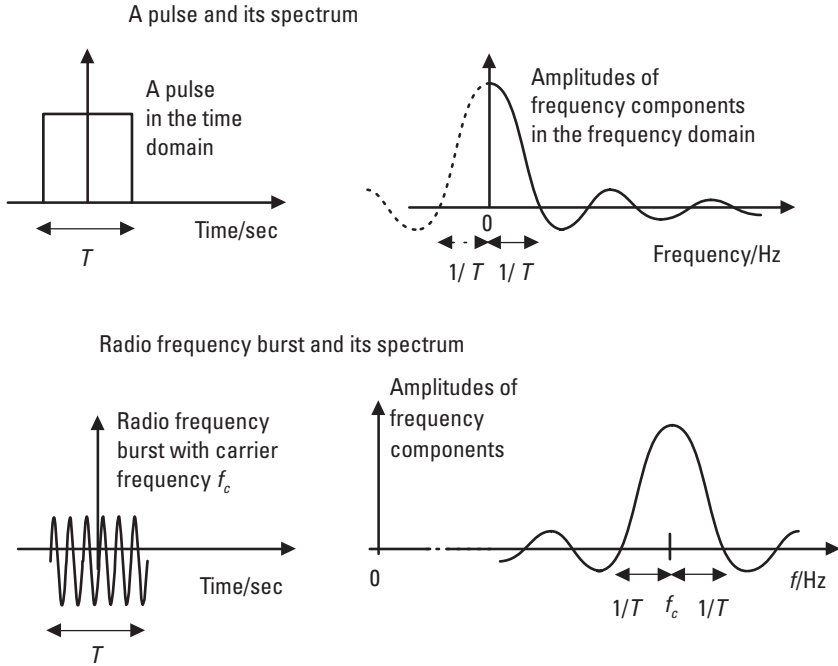
Figure 4.2 Signals in the time domain and the spectrum.

10 times higher data rate, we must use a 10 times shorter pulse, which would require a 10 times wider bandwidth.

In the other example in Figure 4.2, a digital pulse is sent as a radio-frequency burst. This is one example of digital *amplitude modulation* (AM) known as *amplitude shift keying* (ASK). Now the spectrum is concentrated around the carrier frequency, f_c, instead of zero frequency. Note that the spectral width around carrier frequency depends only on the pulse duration T, as in the previous example. If we now increase the data rate (decrease pulse duration), we make the spectrum wider, and a wider radio-frequency band is required. Note that if we let T increase without limit, the spectral width decreases and we finally have only one component in the spectrum, the carrier frequency.

Bandwidth is one of the main restricting factors for transmission. The goal of the two preceding examples was to help us understand the connection between the data rate and the required bandwidth. By understanding this we can understand, for example, why efficient speech-coding schemes are required in cellular systems. By reducing the data rate we increase the

network capacity in terms of maximum number of simultaneous calls via a radio-frequency band available for the system.

4.2   Radio Transmission

In radio transmission we have to transfer the spectrum of the message into the radio-frequency band for transmission. For this we use *continuous* or *carrier wave* (CW) modulation.

4.2.1   CW Modulation Methods

The primary purpose of CW modulation in a communication system is to generate a modulated signal suited to the characteristics of a transmission channel. Modulation is needed in the transmission systems to transfer the message spectrum into high radio frequencies that propagate over radio channels. CW modulation is also used in voice-band modems where digital data modulate the carrier frequencies inside the voice frequency band.

In CW modulation the message alters the amplitude, frequency, or phase of the high-frequency carrier (Figure 4.3). This alteration is detected in the demodulator of the receiver and the original message is reproduced.

We saw in Section 3.3 that a cosine wave such as a carrier is defined by three characteristics: amplitude, frequency, and phase. In the CW modulation that we use in radio systems, we insert the message into the carrier wave by altering these three factors of the carrier wave according to the message to be transmitted.

4.2.2   AM

The original carrier wave has a constant peak value (amplitude) and it has a much higher frequency than the modulating signal, the message. In AM the
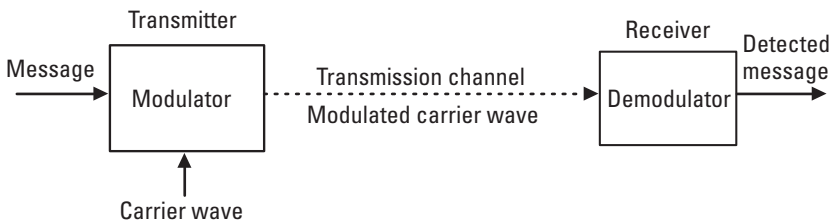
Figure 4.3 CW modulation.

peak value of the carrier varies in accordance with the instantaneous value of the modulating signal and the outline wave shape, or envelope, of the modulated wave follows the shape of the original modulating signal as shown in Figure 4.4. Thus, the unique property of AM is that the envelope of the modulated carrier has the same shape as the message.

We can show with the help of a simple mathematical analysis that when a sinusoidal wave at carrier frequency f_c Hz is amplitude modulated by a sinusoidal modulating signal at message frequency f_m Hz, the modulated wave contains the following three frequencies, as shown in Figure 4.4:

- The original *carrier frequency*, f_c Hz;
- The *sum* of the carrier and modulating signal frequencies, $(f_c + f_m)$ Hz;
- The *difference* between the carrier and modulating signal frequencies, $(f_c - f_m)$ Hz.

These sum and difference frequencies are new, produced by the AM process and they are called *sideband* frequencies. In this case the bandwidth of the modulated signal is
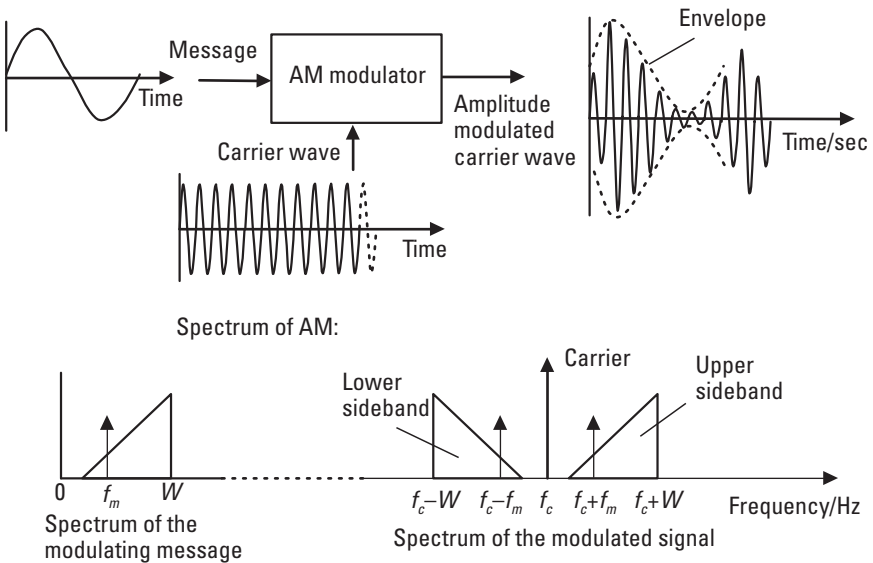
Figure 4.4 Amplitude modulation and its spectrum.

$$\left(f_c + f_m\right) - \left(f_c - f_m\right) = 2f_m \qquad (4.1)$$

If the modulating signal contains multiple frequency components, a band of frequencies such as those in speech or music, the AM process transfers the message spectrum with the carrier. The message spectrum appears after the modulation on both sides of the carrier and the required bandwidth is doubled. Figure 4.4 shows an example in which the original message with baseband bandwidth W modulates a carrier at the frequency f_c. Each individual frequency component that the message contains produces upper and lower sideband frequencies around the carrier frequency, and complete upper and lower sidebands that contain all frequencies of the message are obtained.

If the message is in digital format, the amplitude of the carrier is changed rapidly from one value to another. This is called "keying" because in early wireless telegraph systems the carrier was switched on and off with each keystroke by an operator. This type of digital AM is called *amplitude shift keying* and its spectrum was presented previously in Figure 4.2.

AM is the oldest modulation method but it is still used in radio broadcasting. The original AM has further developed into the *suppressed carrier double-sideband* (SCDSB), *single-sideband* (SSB), and *vestigal-sideband* (VSB) versions, which are briefly introduced next. These principles are explained in the frequency domain, because they are more difficult to understand in the time domain (Figure 4.5).

Suppressed carrier double-sideband (SCDSB) modulation
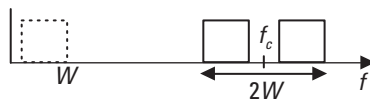
Single-sideband (SSB) modulation
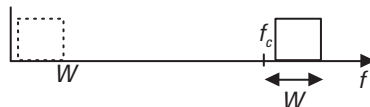
Vestigial-sideband (VSB) modulation
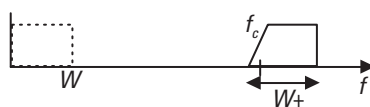
Figure 4.5 Modulation methods SCDSB, SSB, and VSB.

4.2.2.1   SCDSB

In the case of AM, the carrier is in the air even when there is no information to be transmitted. It can be shown that even with the maximum information amplitude, at least 50% of the total transmission power is spent on the carrier wave in AM. Constant amplitude, frequency, and phase carrier wave do not carry any information and transmission of the carrier wave is a waste of power from a performance point of view. In the SCDSB, or DSB for short, modulation scheme, the carrier wave is suppressed and all the power is used for sidebands that carry the information as shown in Figure 4.5.

The cost incurred to save power with the help of SCDSB is that more complex transmitters and receivers are required, but this is no longer important with current technology. The detector in the receiver cannot find the message by following the envelope only. The received carrier wave reverses phase every time the message crosses zero and, in addition to the amplitude, the phase also has to be detected. SCDSB is used, for example, for stereo information processing in analog FM radio broadcasting systems, and together with phase modulation it is used in many modern systems, such as digital radio and TV broadcast systems.

4.2.2.2   SSB Modulation

Conventional AM doubles the bandwidth of the message wasting bandwidth in addition to power. Suppressing one of the sidebands reduces the transmission bandwidth and leads to SSB modulation, as shown in Figure 4.5.

The bandwidth of a transmission channel is an especially important restriction of the carrier systems in the telecommunications networks. SSB modulation is used in the analog carrier systems that are designed to transmit as many telephone channels as possible through a bandwidth-limited channel such as a cable. SSB modulation doubles the capacity (the number of speech channels) compared with AM and SCDSB.

4.2.2.3   VSB Modulation

Consider a modulating signal, for example, the video portion of a television signal, that has a very wide bandwidth and significant low-frequency content. The bandwidth conservation principle argues in favor of SSB modulation, but practical SSB systems have a poor low-frequency response because of the filtering of the other sideband. The SCDSB would be better for this kind of application but it requires a double bandwidth. Clearly, a modulation scheme that negotiates a compromise between SSB and SCDSB is required and this is called VSB modulation.

VSB modulation is derived by filtering SCDSB (or AM; VSB is often used with a carrier) in such a fashion that one sideband is passed on almost completely while just a trace, or vestige, of the other sideband is included. In the receiver detection circuitry the vestige of the lower sideband is added to the upper sideband. This improves the quality, making the frequency response flat to very low frequencies of the message. This method is used in analog TV video transmission.

All of the modulation methods described in this section belong to the class of linear CW modulation method. Consider their common properties:

- The modulated bandwidth never exceeds twice the message spectrum.

- The transmission spectrum is basically the transferred message spectrum.

- The destination S/N is never better than if the baseband transmission was used (no modulation at all). This means that the noise power added to the transmitted signal on the line is detected in the receiver together with the desired modulating signal and the S/N is not improved in detection.

The exponential modulation methods of *frequency modulation* (FM) and *phase modulation* (PM) that we will discuss next differ on all of these counts.

4.2.3 FM

In contrast to linear modulation, exponential modulation is a nonlinear process and therefore the modulated spectrum is not related to the message spectrum in a simple fashion.

The modulated waveform after exponential modulation can be expressed by the following equation:

$$x_c(t) = A_c \cos[\omega_c t + \phi(t)] = A_c \operatorname{Re}\left\{ e^{j[\omega_c t + \phi(t)]} \right\} \qquad (4.2)$$

where $\phi(t)$ represents the varying phase or the frequency containing the message, A_c is the constant amplitude, $\omega_c = 2\pi f_c$ is the angular frequency of the carrier wave, and Re means that we take the real part of the exponential function in brackets. As we can see, the message is inserted into the angle of the carrier wave or in the exponent of the function describing a cosine wave. This

is why these modulation methods are called either *angle* or *exponential modulations.*

In FM the instantaneous frequency of the carrier is varied according to the message and its amplitude is kept constant. Figure 4.6 shows an example in which the frequency of the carrier is increased when the value of the modulating message is increased and vice versa. We can assume that FM has good noise performance, because if the amplitude is distorted we can cut it back to the constant value in the receiver, thus eliminating most of the external disturbances. In the detector of the receiver only the instants when the signal curve crosses zero voltage need to be detected. The disturbances are highly attenuated because a large amplitude change has only a slight impact on the position of the crossing points. This helps us understand that the noise added to the transmitted signal on the line does not reduce the postdetection S/N as much as in the case of linear modulation. Actually the S/N can be improved in detection. This advantage is paid for by a wider transmission bandwidth. For example, commercial FM broadcasting uses more than 200 kHz of bandwidth for the transmission of a 15-kHz message band.

The characteristics of the spectrum of an FM signal are not as simple as those for linear modulation methods. However, in digital FM we use one carrier frequency for each digital symbol value. In the binary case we may transmit 0 for a lower frequency and 1 for the higher frequency, and each
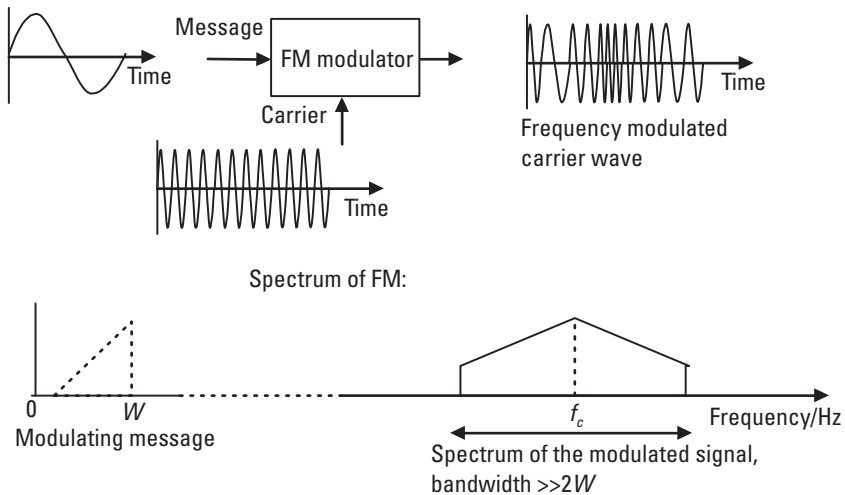
Figure 4.6 FM.

transmitted bit generates spectrum similar to the radio-frequency burst shown in Figure 4.2 around its center frequency. Now we see that width of the spectrum also depends here on the data rate, which defines length of the burst in Figure 4.2, and the distance between higher and lower frequencies used.

As an example of digital FM, some older generation voice-band modems use the digital form of FM called *frequency shift keying* (FSK). For example, a 1,200-bps V.23 modem uses two frequencies, 1,300 Hz for binary 0 and the 2,100 Hz for binary 1. Another example is digital frequency modulation of GSM in which two frequencies, 67.7 kHz above and below the nominal carrier frequency, are used for binary transmission [2].

4.2.4 PM

PM is another method in the class of exponential modulations. In PM the instantaneous phase, instead of frequency, is varied linearly according to the message. Therefore, if the message has discontinuities, there will be discontinuities in the modulated carrier wave as well (Figure 4.7). The spectral characteristics are nearly the same as in the case of FM. Figure 4.7 shows an example where the phase of the carrier is increased with the strength of the message. When message returns to zero there is a sudden phase change when the carrier returns to its nominal phase.

In digital binary PM, which is called *binary phase shift keying* (BPSK), the phase of the carrier is varied according to whether the digital signal is 1 or 0. Figure 4.8 shows an example of BPSK where the digital sequence of 0011011... is transmitted. In binary phase modulation we need only two carrier phases, which are chosen to be 0° for binary 0 and 180° for binary 1 in Figure 4.8.
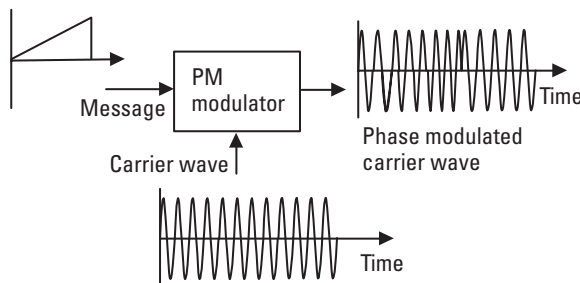
Figure 4.7 Principle of PM.
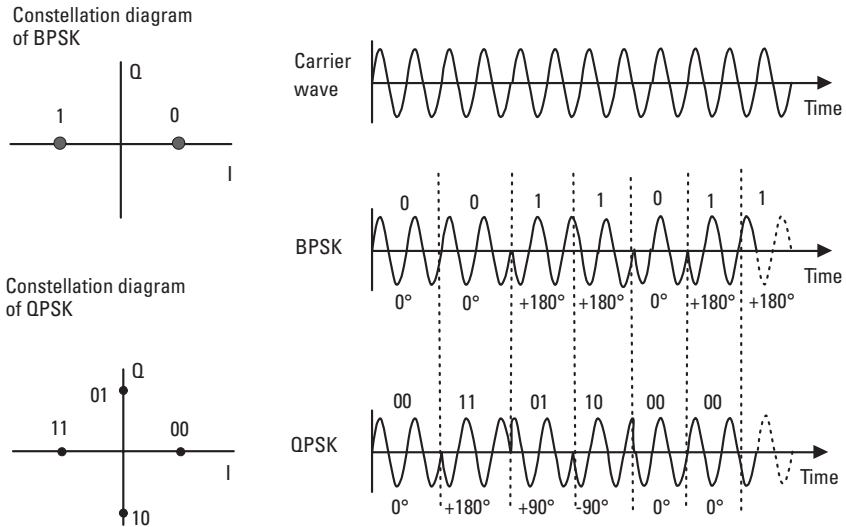
Figure 4.8 Digital PM.

Often we use more than these two phases of the carrier in digital modulation. When four carrier phases are used, each phase transmits the value of two binary bits and we talk about *quadrature phase shift keying* (QPSK). Figure 4.8 illustrates an example of QPSK. An original carrier wave and the modulated one are drawn in the figure. At a point in time a pair of bits is taken from the incoming bit stream (110001101111...) of the modulator and the carrier phase is shifted according to the value of these two bits until the next two bits are received.

One easily understandable way to describe digital phase modulation is by means of a constellation diagram, as shown in Figure 4.8. In the constellation diagram, the *I* axis refers to the in-phase carrier wave and *Q* stands for the quadrature carrier, that is, the carrier in 90° phase shift. Each signal point in the diagram represents one possible transmitted "symbol" or waveform that represents binary values of one or two bits in the examples of Figure 4.8. We can see easily from the constellation diagram for QPSK that, for example, the bit combination 01 is sent as a carrier with a +90° phase shift. The distance of the signal point from the origin tells the carrier amplitude that is the same for all symbols in our examples in Figure 4.8.

To get an idea about the spectral requirements of digital phase modulation, we can consider a single BPSK carrier burst representing for example a single 0-bit. Its spectral width depends on the duration of the symbol, which

equals T in Figure 4.2. Symbol rate or modulation rate $1/T$ is expressed in bauds. Then most of the spectrum resides in the frequency range from $f_c - 1/T$ to $f_c + 1/T$ as shown in Figure 4.2. Binary 1 differs only by the carrier phase and the amplitude spectrum is the same. If we double the data rate of BPSK we have to cut symbol duration T to half, which doubles the required bandwidth. On the other hand, we can double the bit rate without increasing the bandwidth by using QPSK, in which each symbol carries two bits instead of one as shown in Figure 4.8. If symbol duration remains the same, the spectral width remains the same as well.

We could increase the data rate further by using eight different carrier phases as in 8-PSK in Figure 4.9. If the modulation rate is the same for BPSK in Figure 4.8 and for 8-PSK in Figure 4.9, both methods occupy the same frequency band but the bit rate of 8-PSK is three times that of BPSK. The cost we pay for this increased data rate is lower noise tolerance. If the transmission power of both systems is the same, the distance of signal points from origin is the same in Figures 4.8 and 4.9. Then the 8-PSK signals are much closer to other signals than are those in BPSK, and much lower noise or interference can cause errors in the receiver. The 8-PSK is used in cellular networks to increase the data rate in low-interference environments. If interference increases, modulation is changed to binary modulation, which tolerates higher interference.

Use of more phases than in 8-PSK is usually not feasible because of reduced noise tolerance. Instead we can combine AM and PM as shown in Figure 4.9 to become 16-QAM. This combination of phase and amplitude
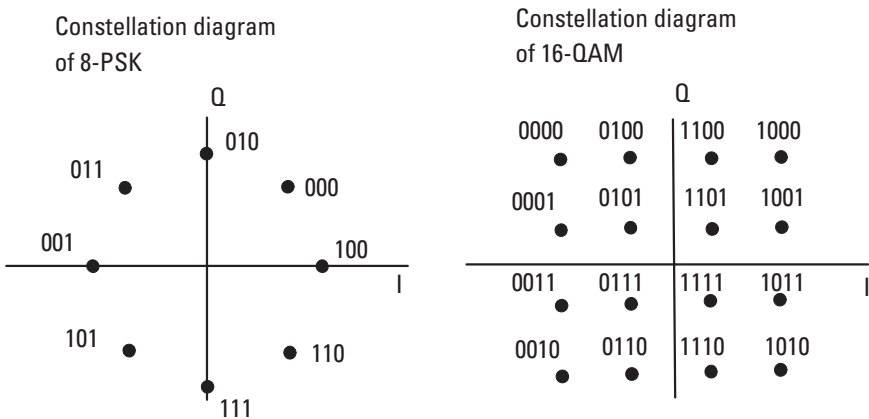
Figure 4.9 8-PSK and 16-QAM.

modulations is called *quadrature amplitude modulation* (QAM). In Figure 4.9, 16-QAM uses three amplitudes and 12 different phases to create 16 different carrier waveforms, each representing one combination of four bits. If the symbol or modulation rate is the same in 16-QAM as in BPSK, the spectral width of the radio signal remains the same but the bit rate of 16-QAM is four times that of BPSK. If we prefer to save spectrum instead of increasing the data rate, 16-QAM could use four times longer radio bursts than BPSK for the same bit rate. This would reduce the radio-frequency band to one-fourth of BPSK as we can see from Figure 4.2.

The optimum modulation method for a particular system depends on the quality of the transmission channel. In voice-band modems, which use low-noise speech channels, very large constellations with hundreds of different combinations of phases and amplitudes are feasible. In bad quality channels, such as in cellular networks, binary modulation may be the best choice.

Phase modulation together with amplitude modulation is used in many modern digital transmission systems, such as in digital radio relay systems, voice-band modems, and *digital video broadcasting* (DVB) systems, which use 64-QAM.

4.2.5   Allocation of the Electromagnetic Spectrum

Signal transmission over an appreciable distance always involves the traveling of an electromagnetic wave, with or without a guiding medium. The efficiency of any particular transmission method depends on the frequency of the signal being transmitted. With the help of CW modulation, the spectrum of the message is transferred to the suitable frequency band of the medium.

The use of frequency bands is controlled internationally by the ITU-R and nationally by national telecommunications authorities. Radio systems are often the most economical solution when new connections are required and there are no free cables or fibers between the end points of the connection. Figure 4.10 illustrates the frequency range that is used in telecommunications and it also shows some examples of the usage of different frequencies.

In Figure 4.10 the electromagnetic spectrum used in telecommunications is shown together with typical transmission media, the propagation modes, and some application examples.

However, radio systemss have one important problem that restricts the use of radio communication, namely, lack of frequency bands. The most suitable bands are overcrowded, and new technical inventions are needed in order to overcome this problem. Among these are, for example, cellular
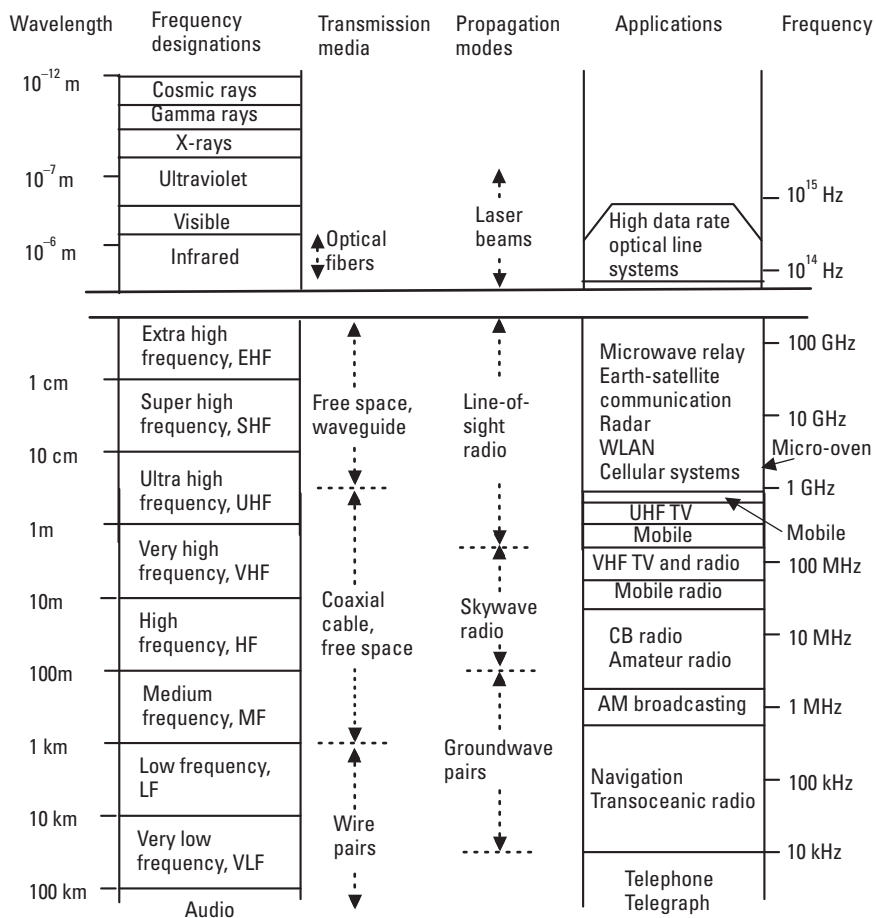
Figure 4.10 Allocation and applications of electromagnetic spectrum.

mobile systems and WLANs with small cell areas that enable them to use frequencies again in other cells of the same network, narrow beam radio relay systems, sophisticated modulation schemes for radio relays, and digital broadcasting systems. We saw in Section 4.2.4 that we can decrease the modulation rate and, correspondingly, the required bandwidth with the help of more complicated modulation schemes.

4.2.5.1 Wavelength and Frequency

The wavelength shown on the left-hand side of Figure 4.10 indicates the propagation distance during one cycle of the radio wave. It is related to the

frequency and speed of light and electromagnetic wave according to $\lambda = c/f$, where λ is the wavelength in meters; c is the propagation speed of light or radio wave in meters per second, approximately 300,000 km/sec; and f is the frequency in Hz = 1/sec.

4.2.5.2   Propagation Modes

Radio waves at different frequency bands propagate in different propagation modes. They are very briefly explained as follows:

- *Ground wave:* The radio wave follows the surface of the Earth, and thus communication over the horizon is possible.

- *Skywave:* The radio wave is reflected from the ionosphere back to Earth. The wave is reflected back from the Earth's surface and back to the Earth again making long-distance communication possible. The communication quality is not stable because the characteristics of the ionosphere vary with time.

- *Line of sight:* The radio wave propagates along the straight line from the transmitter to the receiver. A general requirement for good performance is that the receiving antenna be visible from the transmitter. The radio frequencies above 100 MHz that propagate in line-of-sight mode are used in most modern communication systems.

As the demand for radio communications has increased, higher and higher frequencies have been put into use. However, as we will see in the next section, the attenuation of the radio wave increases with frequency and at extremely high frequencies, beyond 10–100 GHz, even weather conditions affect attenuation. This is why there are no applications at frequencies higher than the *extra high frequency* (EHF) band (Figure 4.10).

4.2.5.3   Optical Communications

At the infrared light frequencies just below visible light (wavelength 400–700 nm) a controlled transmission medium, optical fiber, provides very low attenuation. Optical fiber is the most important media for high-capacity long-distance transmission. It is used in national long-distance networks as well as in international and intercontinental submarine systems.

The commercial optical communication systems of today use binary light pulses for transmission. The transmitted information is usually in binary form, which means that the receiver either detects light or does not. The present optical systems are not able to use transmitted light as a carrier

wave in the same way that radio systems do. Radio systems are able to change phase and frequency of the carrier wave, not just intensity. Traditionally, one optical signal occupies the whole fiber although a small portion of its very wide frequency would be feasible. Characteristics of optical fibers are introduced in Section 4.7.

However, development of narrowband optical transmitters and optical filters has made it possible to increase the data transmission capacity by inserting multiple optical channels into the same fiber with the help of the *dense wavelength division multiplexing* (DWDM) system, which is introduced later in this chapter.

As technology is developed, we will be able use light at a certain frequency as a carrier wave. Then we can increase fiber capacity further by utilizing the CW modulation methods discussed previously. The utilization of this so-called coherent optical technology will increase the transmission capacities of optical fibers dramatically in the future.

4.2.6   Free-Space Loss of Radio Waves

Most radio systems of today operate well above 100 MHz where the radio wave travels a direct path from the transmitting antenna to the receiving antenna. This propagation mode is called *line-of-sight propagation.*

The power of the radio wave is reduced with distance just as a cable attenuates propagating electrical signals. The attenuation of a radio wave, free-space loss, on the line-of-sight path is due to the spherical dispersion of the radio wave. Here we assume that both antennas are isotropic antennas, which radiate to and receive from all directions equally. The transmitted power from isotropic antennas is distributed over a spherical surface and the radiated power per unit area decreases in proportion to the square of the radius because the area of the spherical surface increases in proportion to the square of the radius. The area of the spherical surface follows the equation $A = 4\pi l^2$, where l is the radius. The power density flow F through the surface of a sphere at distance l from isotropic antenna becomes $F = P_T/(4\pi l^2)$ [W/m²] as shown in Figure 4.11 [3].

The receiving antenna is able to receive the power that passes through its *effective aperture area* or *capture area* [4]. The effective aperture area of the receiving isotropic antenna is proportional to the square of the wavelength according to $A_{ei} = \lambda^2/(4\pi)$, and received power becomes $P_R = A_{ei} F$ [W]. From these two facts we can easily derive that the free-space loss, that is, the ratio of transmitted power and the received power in the case of isotropic antennas, where antenna gains $g_T = g_R = 1$, is
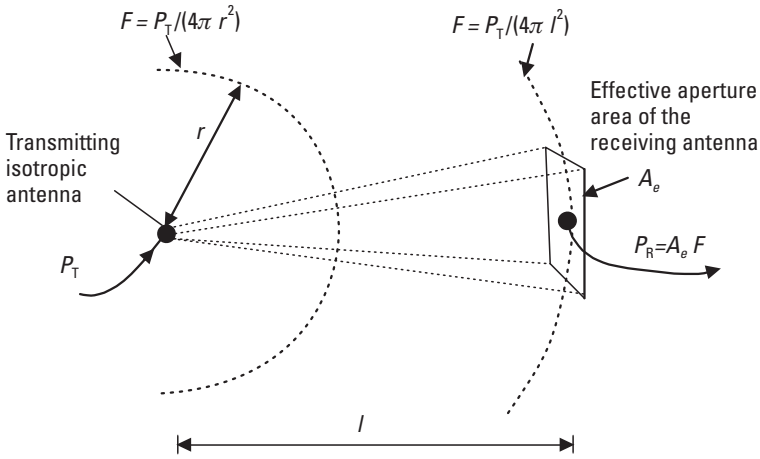
Figure 4.11   Radio wave loss with isotropic transmitting antenna.

$$L = \frac{P_T}{P_R} = \left(\frac{4\pi l}{\lambda}\right)^2 = \left(\frac{4\pi f\, l}{c}\right)^2 \qquad (4.3)$$

where λ is the wavelength, f is the frequency of the signal, c is the speed of light, and l is the transmission distance (Figure 4.11).

We usually prefer to describe attenuation or loss in decibels instead of by the absolute value as given in the previous equations. We obtain the formula that gives decibel values by taking $L_{dB} = 10 \log_{10} L$. Now if we express the frequency f in gigahertz ($f = f_{GHz} \cdot 10^9$) and l in kilometers ($l = l_{km} \cdot 10^3$), we get the free-space loss of a radio wave in decibels as follows:

$$L_{dB} = 92.4 + 20 \log_{10} f_{GHz} + 20 \log_{10} l_{km} \text{ dB} \qquad (4.4)$$

We see that the loss or attenuation is proportional to 20 times the logarithm of frequency and distance. So if the distance or frequency is doubled, the attenuation increases by 6 dB. If we want to maintain the received power, we have to increase the transmitted power by 6 dB, which requires a four times higher transmission power. This comes from the fact that the power ratio in decibels is $10 \log_{10}(P_n/P_o)$ dB, as we saw in Chapter 3.

The free-space loss shown in (4.4) may give results that are too optimistic by as much as 30 dB in actual conditions. Additional attenuation is introduced if there is a hill, a building, or a wall on or close to the straight line between the transmitting and receiving antennas. This is most often the case

in mobile radio communication where actual attenuation may be of the order of 30 dB higher than free-space loss. To estimate the impact of the environment, several propagation models have been developed for cellular network planning. However, free-space loss in (4.4) clearly explains the impact of frequency and distance on radio wave attenuation.

4.2.7 Antennas

Link loss was calculated assuming that antennas are isotropic, which means that they transmit and receive equally to and from all directions. This assumption keeps the attenuation independent of the antennas in use. However, practical antennas have a focusing effect that acts like amplification, compensating for some of the propagation loss. This focusing effect can be expressed as a gain of an antenna, although a passive antenna cannot actually amplify the signal. The maximum transmitting and receiving gain (to direction of maximum radiation or sensitivity) of an antenna with effective aperture area A_e is [1]

$$g = \frac{A_e}{A_{ei}} = \frac{4\pi A_e}{\lambda^2} = \frac{4\pi A_e f^2}{c^2} \tag{4.5}$$

The value of A_e for a dish or horn antenna approximately equals its physical area and large parabolic dishes may provide gains in excess of 60 dBi, where dBi stands for gain in decibels compared with an isotropic antenna. The received power and overall radio link loss when antenna gains are considered becomes (Figure 4.12)

$$P_R = \frac{g_T g_R}{L} P_T; \quad L_{Tot} = \frac{P_T}{P_R} = \frac{L}{g_T g_R} \tag{4.6}$$

In decibel format received power levels and link loss become

$$P_{R,dBm} = P_{T,dBm} + g_{T,dBi} + g_{R,dBi} - L_{dB}$$
$$L_{Tot,dB} = L_{dB} - g_{T,dBi} - g_{R,dBi} \tag{4.7}$$

Note that both antennas have equal impact on the received power level and use of a directional receiving antenna at, for example, the base station site of a cellular network reduces link loss and required transmission power of the mobile station.
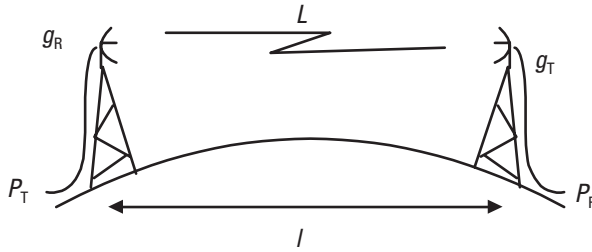
Figure 4.12 Attenuation of the radio wave.

In this section we have reviewed radio transmission at different frequencies and modulation methods that are used to transfer a message to the radio-frequency band for transmission. We have also examined the propagation loss of radio waves. Many other things must considered in radio system engineering but they are beyond the scope of our brief introduction to radio transmission.

In the following section we look at the general characteristics of transmission channels and how the maximum transmission data rate depends on the bandwidth and noise of the channel.

4.3   Maximum Data Rate of a Transmission Channel

A fundamental limit exists for the data rate through any transmission channels, as we will see later in this section. The main restricting factors are the bandwidth and the noise of the channel.

4.3.1   Symbol Rate (Baud Rate) and Bandwidth

Communication requires a sufficient transmission bandwidth to accommodate the signal spectrum; otherwise, severe distortion will result. For example, a bandwidth of several megahertz is needed for an analog television video signal, whereas the much slower variations of a telephone speech signal fit into a 4-kHz frequency band.

Every communication channel has a finite bandwidth. The higher the data rate to be transmitted, the shorter the digital pulses that can be used, as we saw in Section 4.1. The shorter the pulses used for transmission, the wider the bandwidth required, as we saw in Figure 4.2. When a signal changes rapidly in time, its frequency content or spectrum extends over a wide frequency range and we say that the signal has a wide bandwidth.

Figure 4.13 shows the shape of a rectangular pulse with duration T before and after it passed through an ideal lowpass channel of bandwidth B. For example, if the duration of the pulse $T = 1$ ms, distorted pulses are shown in the figure for the channel with bandwidths $B = 2 \cdot 1/T = 2$ kHz, $B = 1/T = 1$ kHz, $B = 1/2 \cdot 1/T = 500$ Hz, and $B = 250$ Hz. If the next pulse is sent immediately after the one in the figure, the detection of the pulse value will be impossible if the bandwidth is too narrow. The spread of pulses over other pulses, which disturbs detection of other pulses in the sequence, is called *intersymbol interference.*

In baseband transmission, a digital signal with r symbols per second, bauds, requires the transmission bandwidth B to be in hertz:

$$B \geq r/2 \tag{4.8}$$

Thus the available bandwidth in hertz determines the maximum symbol rate in bauds. Note that the symbol is not necessarily the same as the bit, but it can carry a set of bits if it is allowed to get many different values.

We can find the theoretical maximum of the symbol or baud rate with the help of a special pulse called the *sinc pulse.* The shape of the sinc pulse is drawn in Figure 4.13 and it has zero crossings at regular intervals $1/(2W)$. With the help of Fourier analysis, we can show that this kind of pulse has no spectral components at frequencies higher than W. If the channel is an ideal

Figure 4.13 Symbol rate (baud rate) and bandwidth.

lowpass channel with a bandwidth higher than W, it is suitable for transmitting sinc pulses that have their first zero crossing at $t = 1/2W$ without distortion. The shape of the pulse remains the same because all frequency components are the same at the output as at the input of the channel.

The sinc pulses have zero crossings at regular periods in time. These periods are $1/(2W)$ seconds for a sinc pulse with a spectrum up to frequency W as shown in Figure 4.13. We can transmit the next pulse at the time instant $1/(2W)$ so that the previous pulse has no influence on the reception because it crosses zero at that time instant. The decision for the value of the pulse is made in the receiver exactly at time instants $n \cdot 1/(2W)$, where $n = \pm 1, \pm 2, \pm 3, \ldots$. The time between pulses $T = 1/(2W)$, which makes data rate $r = 1/T = 2W$. If we now increase the data rate so that $W \to B$, the time between pulses becomes $T \to 1/(2B)$; $r \to 1/T = 2B$, which gives the theoretical maximum rate for transmission of symbols and we can say that the symbol rate and bandwidth are related according to $r \leq 2B$ or $B \geq r/2$.

This kind of pulse does not exist in reality, but the result gives the theoretical maximum symbol rate, which we can never exceed, through a lowpass channel. In real-life systems quite similar pulse shapes are in use and typically a 1.5 to 2 times wider bandwidth is needed.

4.3.2 Symbol Rate and Bit Rate

In digital communications a set of discrete symbols is employed. Binary systems have only two values represented by binary digits 1 and 0. In the previous section we found that the fundamental limit of the symbol rate is twice the bandwidth of the channel. With the help of the symbols with multiple values the data rate, in bits per second, can be increased. As an example, with four pulse values we could transmit the equivalent of 2-bit binary words 00, 01, 10, and 11. Thus each pulse would carry the information of 2 bits and one symbol per second (1 baud) would correspond to 2 bps.

If we use a sinc pulse as in Figure 4.13, the preceding and following pulses do not influence the detection of a transmitted pulse, because each received pulse is measured at a zero crossing point $n \cdot 1/(2W)$ of the other pulses. We may increase the number of peak values of sinc pulses from two to four, from four to eight, for example, in order to increase the bit rate while keeping the symbol rate unchanged. Figure 4.14 shows a simple example where symbols are rectangular pulses with four symbol values and each symbol carries two bits ($k = 2$) of information. Generally, the bit rate depends on modulation rate according to

$$r_b = k \cdot r \text{ bps} \tag{4.9}$$

where k represents the number of bits encoded into each symbol. Then the number of symbol values is $M = 2^k$ and the bit rate is given as $r_b = r \log_2 M$ bps. In the example of Figure 4.14, the number of symbol values is $M = 2^k = 2^2 = 4$, and the bit rate $r_b = k \cdot r$ bps $= 2r$ bps. Then the symbol rate of 1 kbaud makes the bit rate 2 Kbps.

The unit of symbol rate, sometimes called the modulation rate, is bauds (symbols per second). Note that the transmission rate in bauds may represent a much higher transmission rate in bits per second. Table 4.1 shows how the bit rate of a system depends on the number of symbol values.

Figure 4.14 also shows a data sequence of sinc pulses with four amplitude values. The required bandwidth for pulses in this sequence is the minimum bandwidth $B = r/2 = 1/(2T)$ according to (4.8) and Figure 4.13. When pulses are detected by sampling as shown in Figure 4.14 each pulse can be detected independently because values of all other pulses are equal to zero.

In the preceding examples, the amplitudes of the pulses contain the information. This is the principle of PAM, as discussed earlier. This is not the only alternative. We can use other characteristics of the signal as well to create multiple symbol values, for example, the phase of a carrier, as we did in the case of QPSK and 8-PSK in Figures 4.8 and 4.9. There we used a certain modulation rate r in bauds (how many times the phase can change in a second), which defines a required bandwidth. For QPSK 2 bits ($k = 2$) are
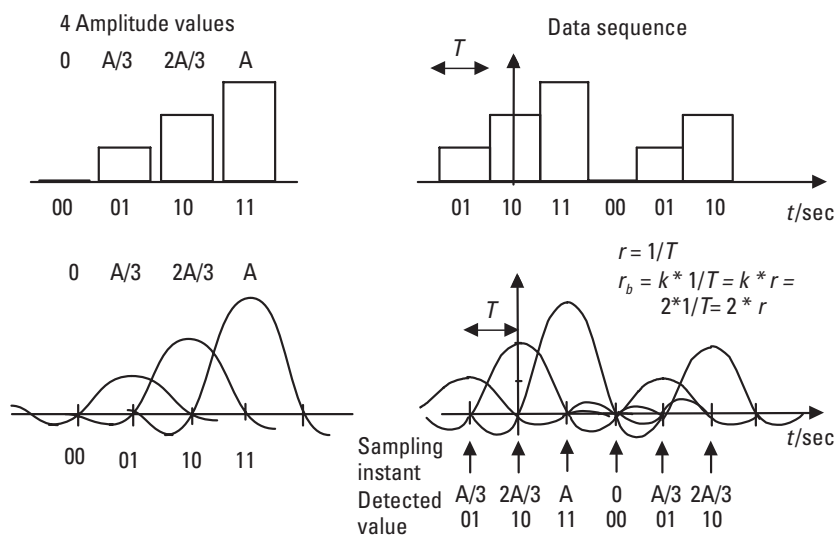
Figure 4.14 Symbol rate and bit rate.

Table 4.1
Bit Rate of a System Using Multiple Symbol Values

Number of Bits, k, Encoded into Each Symbol	Number of Symbol Values, M	Bit Rate Compared with Symbol Rate
1	2	Same as symbol rate
2	4	2 × symbol rate
3	8	3 × symbol rate
4	16	4 × symbol rate
5	32	5 × symbol rate
…		
8	256	8 × symbol rate
…		

encoded into each symbol and the bit rate is two times the modulation rate. For 8-PSK, k =3 and r_b = $3r$. The 16-QAM example in Figure 4.9 used 16 combinations of carrier amplitude and phase amplitude values and the bit rate is four times the modulation rate.

As we can see from Table 4.1, by increasing the number of different symbols used in the system the data rate could be increased without a limit if there were no other limitations than bandwidth. This is not possible in practice because of the noise. The influence of noise is discussed next.

4.3.3   Maximum Capacity of a Transmission Channel

We saw previously that the bandwidth of a channel sets the limit to the symbol rate in bauds but not to the information data rate. In 1948, Claude Shannon published a study of the theoretical maximum data rate in the case of a channel subject to random (thermal) noise.

We measure a noise relative to a signal in terms of the S/N. Noise degrades fidelity in analog communication and produces errors in digital communication. The S/N is usually expressed in decibels as

$$S/N_{dB} = 10 \log_{10} \left(S/N \right) \text{ dB} \tag{4.10}$$

Taking both bandwidth and noise into account, Shannon stated that the error-free bit rate through any transmission channel cannot exceed the maximum capacity C of the channel given by:

$$C = B \log_2 \left(1 + S/N\right) \qquad (4.11)$$

where C is the maximum information data rate in bits per second; B, the bandwidth in hertz; S, the signal power; N, the noise power, and S/N, the S/N power ratio (absolute power ratio, not in decibels).

Equation (4.11) gives a theoretical limit for the data rate with an arbitrarily low error rate when an arbitrarily long error correction code is used. It also assumes that the signal has a Gaussian distribution as does the noise, which is not the case in practice. The influence of bandwidth and noise in the case of binary and multiple value signaling is summarized in Figure 4.15.

The signal power and, thus, the highest value of the signal are always restricted to a certain maximum value. Then, the more symbol values we use, the closer they are to each other, and the lower noise level can cause errors. Thus, a higher bit rate requires a wider bandwidth that allows a higher symbol rate. Alternatively, a better S/N is required to allow for more symbol values.

The example in Figure 4.15 shows what happens to the distance between symbol values when the maximum amplitude is A and four symbol values are used instead of binary symbols that have only two values. In our examples we have used symbols with different amplitudes. This transmission scheme is called PAM, as discussed earlier. Transmission of this type of pulses without CW modulation is called baseband transmission.

In radio systems or modems that use CW modulation, different phases of a carrier wave represent different symbol values. In the Figure 4.8 and 4.9 digital phase modulation methods, BPSK, QPSK, and 8-PSK all require the same bandwidth if symbol rate is the same, but the information data rate for QPSK is double and for 8-PSK triple compared with BPSK. The cost we
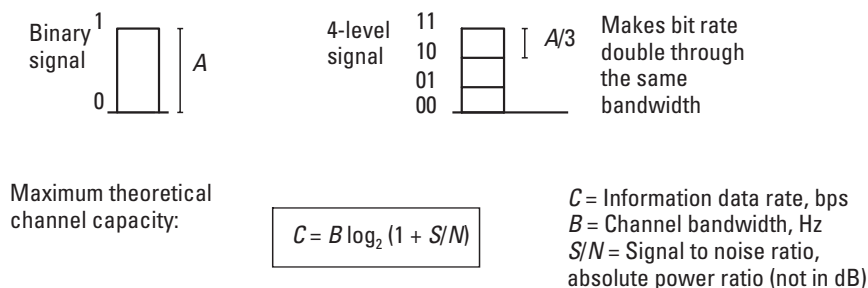
Figure 4.15 The maximum capacity of a transmission channel.

have to pay is reduced noise tolerance because signals become closer to each other as more symbol values or different signals are used. It is not usually reasonable to use more than eight phases; instead, we use different amplitudes as in 16-QAM in Figure 4.9. The 16-QAM tolerates more noise than 16-PSK because with the same average signal power distances between signals can be increased. However, if we would analyze noise tolerance in more detail, we could form a general rule stating that the increase in the number of signals in use reduces noise tolerance. In low-noise channels, such as telephone voice channels, many different signals can be used but in high-interference channels, such as radio channels for cellular systems, binary symbols are often the best choice.

However, modulation moves the spectrum of the pulse from low frequencies to carrier frequencies, and the bandwidth is typically doubled when compared with baseband systems as was shown in Figure 4.2. This is why the symbol rate in radio systems is less than or equal to the transmission bandwidth, that is:

$$r \leq B_T \qquad (4.12)$$

where r is the symbol rate in bauds and B_T is the transmission bandwidth in hertz.

The accurate requirement of bandwidth depends on the modulation scheme in use, the study of which is beyond the scope of this book.

Example 4.1

Assume that the transmission channel is an ideal lowpass channel with a bandwidth of 4 kHz. The maximum symbol rate via this channel is $r \leq 2 \cdot B$ = 8 kbauds; that is, we can transmit up to 8,000 independent signals, symbols, in a second. [To transmit the same symbol rate through a bandpass channel, we would need a bandwidth of 8 kHz according to (4.12); see also Figure 4.2.]

Example 4.2

Assume that the S/N of a lowpass channel is 28 dB and its bandwidth is 4 kHz. Then $S/N_{dB} = 10 \log_{10} S/N$, $S/N = 10^{2.8} \approx 631$. The maximum bit rate according to (4.11) is $C = B \log_2(1 + S/N) = 4,000 \log_2(432) = 4,000 (\log_{10} 632)/\log_{10} 2 = 37.2$ Kbps. In Example 4.1 we learned that the maximum symbol rate is 4 kbauds, which depends only on the bandwidth. To achieve the maximum bit rate, we transmit 4,000 symbols in a second and each of them carries 3 bits (with 4 bits, the maximum bit rate would be

exceeded). The number of different symbols that can be used is $2^3 = 8$ and this depends only on the S/N maximum, not on bandwidth.

4.4   Coding

We have described modulation as the processing of a signal for efficient transmission in a different frequency band than where the information originally exists. Coding is a digital symbol processing operation in which the digital form of the information is changed for improved communication. In general, coding contains many different processes, such as ciphering, compression, and error control coding. For ciphering, the transmitter and the receiver may simply perform an exclusive-or operation with data and a ciphering sequence known only by the transmitter and receiver. An eavesdropper is not able to detect information content without knowing the ciphering sequence.

Most modern systems use error control codes for handling of transmission errors. By appending extra check digits to the transmitted data, we can detect or even correct errors that occur on the line. Error control coding increases both the required bandwidth (data rate increases) and the hardware and software complexity, but it pays off in terms of nearly error-free digital communication even when the S/N is low.

Still another purpose for coding is for compressing information. By using data compression we can reduce the disk space needed to store data in a computer. In the same way we can decrease the required data rate on the line to a small fraction of the original information data rate. We could, for example, use very short codes for the most common characters instead of the full seven-bit ASCII code. Rarely needed characters would use long codes and the total data rate would be reduced. Some compression schemes for voice and video information were introduced in Chapter 3. The study of compression methods is a complex matter and will not be covered in any detail here.

From now on we concentrate only on line coding, which changes source symbols into another form for transmission. The operation of line *encoding* transforms a digital message into a new sequence of symbols. *Decoding* is the opposite process that converts the encoded sequence back into the original message (Figure 4.16).

Consider a computer terminal with a keypad. Each key represents a discrete digital symbol. Uncoded transmission would require as many different waveforms as there are keys, one for each key (or more, one for shift, one for Alt, and one for Alt Gr). Alternatively, each symbol can be an encoder into a binary code word consisting of a number of binary digits for binary transmission.

Binary message → Encoder (+ modulator) → Encoded line signal → Decoder (+ demodulator) → Binary message

Line coding:
  Extracts the dc-content from the message;
  Adds synchronization information into the line signal;
  Increases information data rate through the channel;
  Changes the spectral shape of the message so that
  it suits the channel better
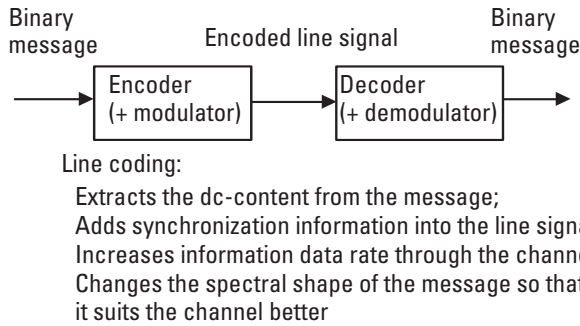
Figure 4.16 Line coding.

4.4.1   Purpose of Line Coding

One purpose of line coding is to make the form of the spectrum of a digital signal suitable for a certain communication media. The line codes usually have no dc content (direct current, frequency component at 0 Hz). We want to get rid of the dc that does not transmit any information but wastes power.

Another reason for line encoding is to help to synchronize the receiver. In digital transmissions the receiver must be synchronized with the transmitter in order to receive the information when each new symbol arrives. For this the data should be transmitted in a form that contains synchronization information so that there is no need to transmit additional clock or timing signals.

The systems that use only line coding, but not modulation, are called baseband transmission systems. The spectrum of the line signal is still in the frequency range of the original message's "baseband." In radio systems both coding and modulation are used.

Line coding can be used to increase the data rate as shown in, for example, Figure 4.17, where each sequence of 2 data bits is encoded into four-level pulses for transmission. At the receiving end decoding is carried out and the original bits, 2 for each received symbol, are regenerated. Note that the symbol rate on the line is half of the bit rate seen by the data source and the destination and thus the required bandwidth of the channel is reduced to half compared to binary transmission. The line code in Figure 4.17 also cancels dc and similar code is used in ISDN subscriber lines. Note that Gray coding, in which neighbor symbols differ by only one bit, is used in Figure 4.17. The symbol in error is typically a neighboring symbol of the transmitted symbol and with the help of Gray coding only one information data bit in error is generated.

Baud rate on the line is half of the bit rate seen by the source and the destination
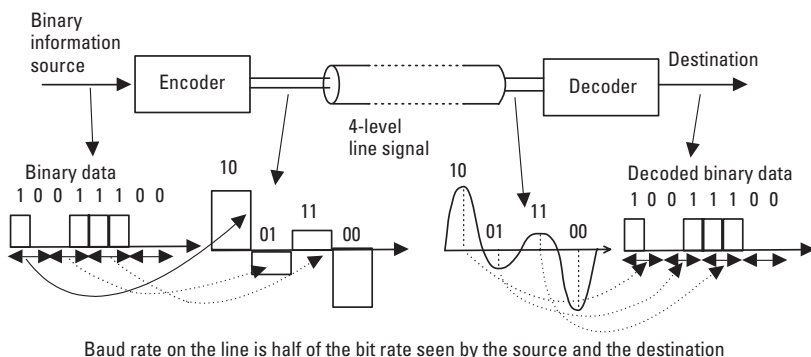
Figure 4.17 An example of the line coding.

We often combine coding and modulation and instead of four or more pulse amplitude values we may transmit four symbol values in carrier waveforms with, for example, four different phases. This so called QPSK was discussed in Section 4.2.4 and it can be seen to be a combination of four-level line coding followed by ordinary phase modulation.

4.4.2 Spectrum of Common Line Codes

To determine what kind of impact line encoding has on the spectrum, we look at the characteristics of some common line codes. Figure 4.18 presents their power density spectrums showing how the signal power of random data is distributed over the frequencies.

4.4.2.1 Nonreturn to Zero (NRZ)

NRZ is the most common form of digital signal used internally in digital systems. Each symbol has a constant value corresponding to binary symbol values 1 and 0. The spectrum has a high dc component, and there are no discrete spectral components at the harmonic frequencies of the data rate. The harmonic frequencies are multiples of the data rate. An external clock signal is always needed for the timing of the receiver.

4.4.2.2 Return to Zero (RZ)

RZ each symbol is cut into two parts. The first half of the symbol represents the binary value and the rest of the symbol is always set to zero. Because pulses are shorter than in the case of NRZ the spectrum is wider, as we saw in Figure 4.2, and the spectrum of a random data has strong discrete frequency components at the harmonic frequencies of the data rate. With the help of
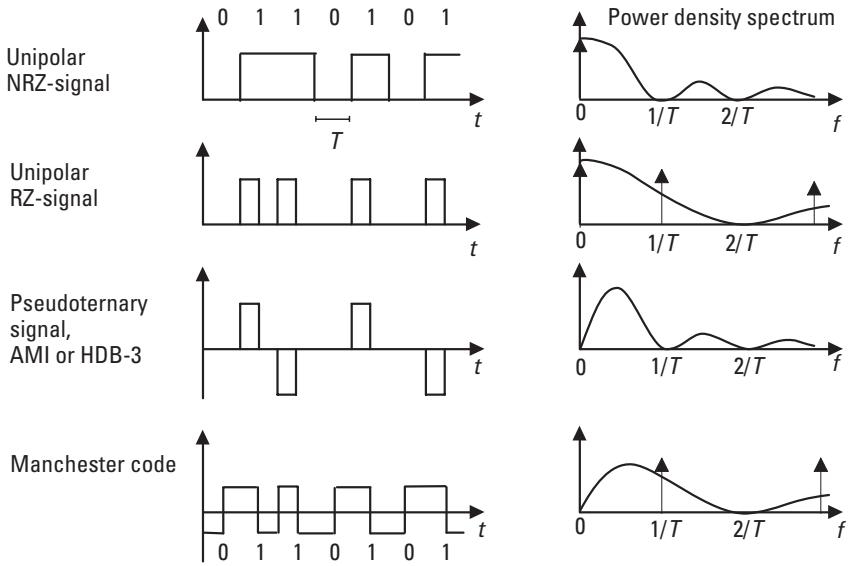
Figure 4.18 Common line codes and their power spectrums.

these components, timing information can be extracted from the signal spectrum and an external clock is not necessarily needed. However, because RZ code has high low-frequency content and a wide spectrum (see Figure 4.18), it is never used in long-distance transmission. Another problem is that synchronization is lost if the data content is all zero for a long period of time.

4.4.2.3 Alternate Mark Inversion (AMI)

If every other mark or 1 of the NRZ or RZ symbols is transmitted as an inverted voltage polarity, an AMI signal is produced. The advantage of this is that no dc component is present on the transmission line. The dc component is unwanted because it does not carry any information; it merely wastes power. With the help of this kind of code we can avoid the problem caused by transformers on the line. Transformers are needed on copper cable lines for matching impedance, for overvoltage or surge protection, and for other purposes. Direct current does not propagate through transformers.

AMI code is used in American telecommunications network in primary rate 1.5-Mbps transmission systems. We may extract the timing information by rectifying the AMI signal into an RZ signal in the receiver and then the discrete spectral components appear as in the spectrum of RZ code in Figure 4.18.

4.4.2.4   High-Density Bipolar 3

HDB-3 was developed from AMI and standardized for European primary rate 2-Mbps systems. HDB-3 overcomes the problem of the original AMI code that occurs in the timing when a data message contains long periods of subsequent zeroes. In this coding scheme, a pulse with the same polarity as the previous one is added in such a way that no more than three sequential zeroes are allowed. In the decoder these pulses are taken away according to the AMI coding rule that they violate.

4.4.2.5   Manchester Coding

Manchester coding is used in LANs. Binary digit 1 is coded as a "+ to –" transition and binary 0 as a "– to +" transition. The most important advantage of the Manchester code is that each symbol contains the timing information and the receiver needs only to detect the transition in the middle of each received symbol to extract the clock signal. Its main disadvantage is a wide spectrum because of short pulses and this is why it is suitable for LANs but not for long-distance transmission.

4.5   Regeneration

In long-haul transmissions the transmitted signal is attenuated and amplifiers or repeaters are needed. *Analog amplifiers* amplify the signal at the input, and the signal contains both the desired message and channel noise. In every amplifier and cable section some noise is added and the S/N decreases with distance.

Unlike analog amplifiers, digital repeaters are regenerative. A regenerative repeater station consists of an equalizing amplifier that compensates the distortion and filters out the out-of-band noise and a comparator as shown in Figure 4.19. Output of the comparator is high if the input signal is above the threshold voltage V_{ref}, and low if the input is below the threshold value. The regenerator also contains timing circuitry, which extracts the clock signal from the received data, and a D-type flip-flop which decides if a digit is high (1) or low (0) at the instant of the rising edge of the clock signal (see Figure 4.19). At the rising edge of the clock signal the input value is read into the output by the D-type flip-flop. The output value remains the same until the next rising edge of the clock signal. The operating principle of a binary regenerative repeater is presented in Figure 4.19. The regenerated digits that contain no noise are delivered to the destination or via a cable to the next repeater station (in the case of an intermediate repeater).

Original transmitted data

Amplified and equalized data with noise

Distorted and attenuated line signal

Regenerated data without noise

Regenerator

V_{ref}

1 0 1 1 0

D

1 0 1 1 0

V_{ref} Clk

D-type flip-flop

Clk

Time

Rising edges of the clock signal

Error rate versus S/N in each repeater

10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}

10 12 14 16 18 20   S/N dB
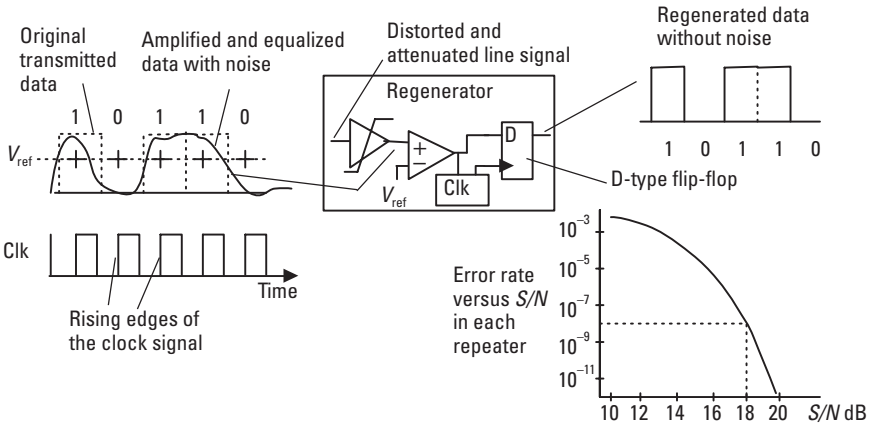
Figure 4.19 Operating principle of a regenerative repeater.

If the equalized signal is below threshold V_{ref} at the input of the comparator, the output is low and a zero is regenerated at the rising edge of the clock signal. If noise is too high, the input of the comparator may be above threshold even though a zero is transmitted. If this occurs at the rising edge of the clock signal, the value 1 is regenerated and an error has occurred. In the same way, high values may be in error if noise reduces the high-amplitude value below the threshold level at an instant of the rising edge of the clock signal. Then 0 is regenerated and an error has occurred.

How frequently errors occur depends on the noise level or S/N. If noise is assumed to have a *Gaussian amplitude distribution* (as thermal noise does), the error rate (bit error probability) follows the shape of the curve, error rate versus S/N, in Figure 4.19.

As an example let us assume that we have a channel, for instance a cable, that attenuates a signal so much that the resulting S/N in the repeater is 15 dB. The error rate would then be around 1×10^{-5} according to the curve in Figure 4.19. If we place a new repeater in the middle of the repeater section (in the middle of the cable), attenuation of the signal is 3 dB less, giving an S/N value of 18 dB in both repeaters and the error rate at both repeaters would be 1×1^{-8}. This means that one error occurs on average after 100,000,000 correct bits. Now we have two repeaters and we have an overall error rate of 2×10^{-8} because each of them creates on average one error in each sequence of 1×10^8 bits. We can see that the improvement of 3 dB in the S/N that we achieved with the help of the new repeater reduces the number of errors by a factor of 0.001.

In practice, the error rate of an operational transmission system is often much better and we have close to error-free transmission and the exact equivalent of the original signal is received in the end regardless of the distance (the number of repeaters).

The error rate decreases rapidly with noise as shown in Figure 4.19 because of the Gaussian nature of thermal noise. Not only thermal, but many other types of noise in real-life systems are assumed to follow a Gaussian distribution. With this model the reduction of noise by 1 dB improves the error rate by factor of 10 or more, as seen in Figure 4.19. The digital transmission systems installed in telecommunications networks are designed in such a way that noise is low enough in all regenerators and the error rate is extremely low. For example, optical line systems usually have a design practice of worst-case lifetime error rate of 1×10^{-10}. In normal operational conditions the error rate is several orders of magnitude better and they operate nearly error free.

From the error rate curve in Figure 4.19 we see how the error rate depends on the S/N. From the error rate we can easily calculate the mean time between errors when the data rate is known. Table 4.2 gives examples of error rates and mean times between errors for a 64-Kbps (ISDN B-channel) data channel.

Digital systems have a certain threshold value for the S/N. From the curve in Figure 4.19 and Table 4.2, we find that if the S/N is worse than 18 dB, errors occur quite frequently. At a few decibels better value for the S/N, the transmission is almost error free. The S/N values in the

Table 4.2
Examples of Error Rates and Mean Times Between Errors for a 64-Kbps Channel

S/N (dB)	Error Rate	Mean Time Between Errors
10.3	10^{-2}	1.5 ms
14.4	10^{-4}	150 ms
16.6	10^{-6}	15 seconds
18	10^{-8}	26 minutes
19	10^{-10}	2 days
20	10^{-12}	6 months
21	10^{-14}	50 years

Figure 4.19 curve and Table 4.2 are examples and are based on certain assumptions. The actual S/N value in decibels at a certain error rate of a specific system depends on the system characteristics and how the S/N is defined and measured. However, the shape of the error curve is the same as in Figure 4.19 and the threshold value is usually between 8 and 20 dB.

When the S/N of a digital system decreases, errors occur more and more frequently and when the error rate becomes too high, information is lost. An error rate of 1×10^{-3} is standardized to be the worst allowed communication quality for PCM speech in the telecommunications network. If the error rate becomes worse, ongoing calls are cut off. Data are transmitted in large packets and if a packet contains one or more errors it needs to be retransmitted. As a rule of thumb, we can say that data transmission requires an error rate of 1×10^{-5} or better, otherwise retransmissions slow down the end-to-end transmission data rate.

4.6   Multiplexing

Multiplexing is a process that combines several signals for simultaneous transmission on one transmission channel. Most of the transmission systems in the telecommunications network contain more capacity than is required by a single user. It is economically feasible to utilize the available bandwidth of optical fiber or coaxial cable or a radio system in a single high-capacity system shared by multiple users. The main principles of multiplexing are described in the following sections.

4.6.1   Frequency-Division Multiplexing (FDM) and TDM

FDM modulates each message to a different carrier frequency. The modulated messages are transmitted through the same channel and a bank of filters separates the messages at the destination (Figure 4.20). The frequency band of the system is divided into several narrowband channels, one for each user. Each narrowband channel is reserved for one user all the time. FDM has been used in analog carrier systems in the telephone network. The same principle is also used in analog cellular systems in which each user occupies one FDM channel for the duration of the call. In such a case, we call the process *frequency-division multiple access* (FDMA) because the frequency-division method is now used to allow multiple users to access the network at the same time.

A more modern method of multiplexing is TDM, which puts different messages, for example, PCM words from different users, in nonoverlapping
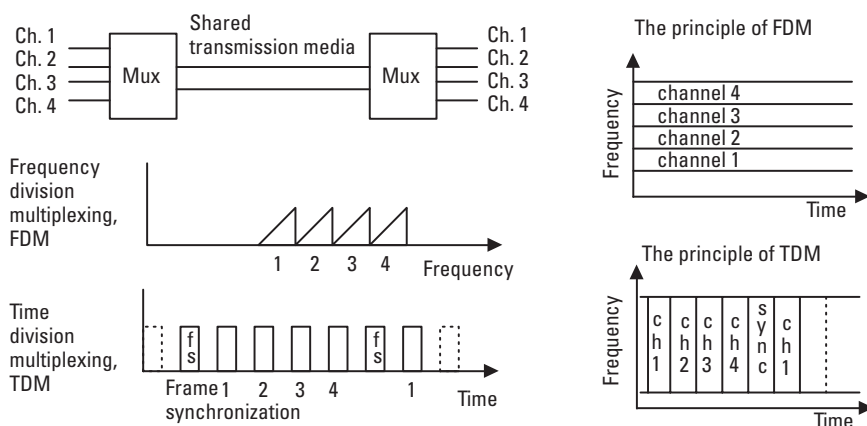
Figure 4.20 Multiplexing methods FDM and TDM.

time slots. Each user channel uses a wider frequency band but only a small fraction of time, one time slot in each frame as shown Figure 4.20. In addition to the user channels, framing information is needed for the switching circuit at the receiver that separates the user channels (time slots) in the demultiplexer. When the demultiplexer detects the frame synchronization word, it knows that this is the start of a new frame and the next time slot contains the information of user channel 1.

This method of TDM is used in high-capacity transmission systems such as optical line systems but also in digital cellular networks where we call it *time-division multiple access* (TDMA). One user occupies one time slot of a frame, and the time-division principle allows multiple users to access the network at the same time using the same carrier frequency.

4.6.2 PCM Frame Structure

We introduced the principle of TDM in the previous section. As an example of TDM and to get a clear view of TDM, we now look at the most common frame structure in telecommunications networks, namely, the primary rate 2,048-Kbps frame used in the European standard areas. This is the basic data stream that carries speech channels and ISDN-B channels through the network and it is called E-1. The corresponding North American primary rate is 1.544 Mbps, which carries 24 speech channels and it is known as DS1 or T1. It is also introduced in this section.

In the European scheme, the primary rate frame is built up in digital local exchanges that multiplex 30 speech or data channels at bit rate of

64 Kbps into the 2,048-Kbps data rate. ITU-T defines this frame structure in Recommendation G.704.

4.6.2.1   The 2-Mbps Frame Structure

PCM-coded speech is transmitted as 8-bit samples 8,000 times a second, which makes up a 64-Kbps data rate. These eight-bit words from different users are interleaved into a frame at a higher data rate.

The 2,048-Kbps frame in Figure 4.21 is used in the countries implementing European standards for telecommunications. It contains 32 time slots, and 30 of them are used for speech or 64-Kbps data. The frame is repeated 8,000 times a second, which is the same as the PCM sampling rate. Each time slot contains an eight-bit sample value and the data rate of each channel is 64 Kbps. These voice channels or data channels are synchronously multiplexed into a 2-Mbps data stream, which is often called E1 (first level in European hierarchy). For error-free operation the tributaries (64-Kbps data streams of the users) have to be synchronized with the clock signal of the 2-Mbps multiplexer. The data rate of 2,048 Kbps for the multiplexer is allowed to vary by 50 *parts per million* (ppm), and as a consequence each user of the network has to take timing from the multiplexer in the network and generate data exactly at the data rate of the multiplexer divided by 32.
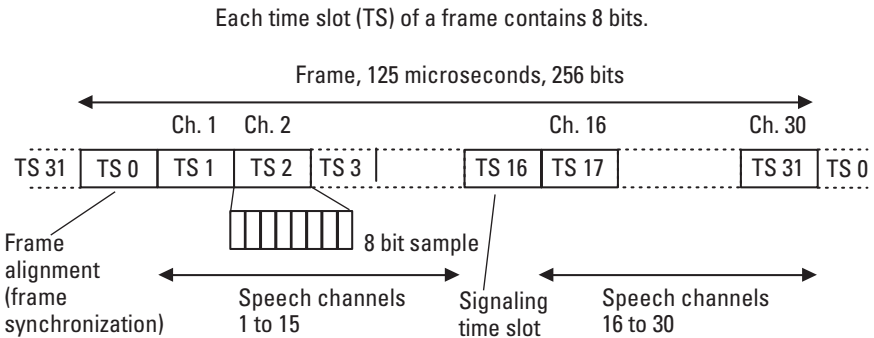
Each time slot (TS) of a frame contains 8 bits.

Frame is repeated 8,000 times in a second which is the same as PCM sampling rate.
Each frame contains one sample of 30 different speech signals.
One time slot (TS0) is used for frame synchronization and one (TS 16) for signaling.
Data rate: 8,000 1/s * 8 bits * 32 = 2,048 Kbps

Figure 4.21   The 2,048-Kbps frame structure from Recommendation G.704.

Frame Synchronization Time Slot

The frame alignment word is needed to inform the demultiplexer where the words of the channels are located in the received 2-Mbps data stream. The frame synchronization time slot (TS0) includes frame alignment information and it has two different contents that are alternated in subsequent frames (Figure 4.22). The demultiplexer looks for this time slot in the received data stream and, when it is found, locks onto it and starts picking up bytes from the time slots for each receiving user. Each user receives 8 bits in 125-μs periods, which makes 64 Kbps. A fixed alignment word is not reliable enough for frame synchronization because it may happen that a user's data from one channel simulates the synchronization word and the demultiplexer might lock to this user time slot instead of TS0. This is why there is one alternating bit (D2) in time slot 0 (see Figure 4.22) and due to this the demultiplexer is able to detect the situation where one channel constantly transmits a word that is equal to the *frame alignment word* (FAW).

To make frame alignment even more reliable, the *cyclic redundancy check 4* (CRC-4) procedure was added in the mid-1980s. C-bits are allocated to carry a four-bit error check code that is calculated over all bits of a few frames. The receiver performs error check calculations over all bits of the frames and it is able to detect false frame alignment even if the frame alignment word is simulated by one user that alters bit two.

A frame, 125 microseconds

| TS31 | TS0 | TS1 | TS2-30 | TS31 | TS0 | TS1 | | TS31 | TS0 | TS1 | | TS31 | TS0 | TS1 |

FAW ... FAW

Frame alignment word

Bit number	1	2	3	4	5	6	7	8
FAW	C	0	0	1	1	0	1	1

TS0 in every other frame that does not contain FAW.

Bit number	1	2	3	4	5	6	7	8
TS0 word	C	1	A	D4	D5	D6	D7	D8
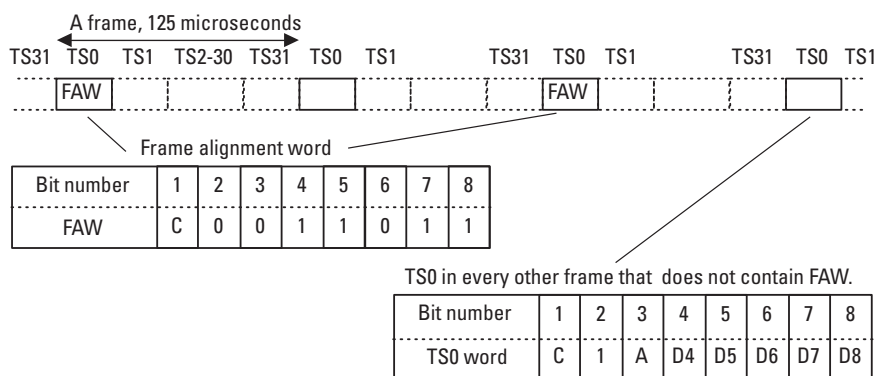
A = Far end alarm, alarm condition "1".
D = Spare bits that can be used for specific point-to-point low data rate applications (for example, network management information).
Bit 2 alternates from frame to frame to prevent accidental simulations of the frame alignment signal.
C = CRC-4 procedure for protection against simulation of frame alignment and enhanced error monitoring. If not in use, C-bit is set to "1".

Figure 4.22 The 2,048-Kbps frame alignment word in TS0.

Each receiver of the 2,048-Kbps data stream detects errors in order to monitor the quality of the received signal. Error monitoring is mainly based on the detection of errors in the frame alignment word. The receiver compares the received word in every other TS0 with the error-free frame alignment word. In addition to the frame alignment word, the CRC-4 code is used to detect low error rates. Errors in the frame alignment word do not give reliable results when the error rate is very low. It may take a long time before an error is detected in TS0 although many errors may have occurred in other time slots of the frame. The C-bit in Figure 4.22 is set to 1 if CRC is not used [5].

The TS0 in every other frame also contains a far-end alarm information bit A as shown in Figure 4.22. This is used (set to 1) to tell the transmitting multiplexer that there is a severe problem in the transmission connection and reception is not successful at the other end of the system. This can be caused by, for example, a high error rate, loss of frame alignment, or loss of signal. With the help of the far-end alarm, consequent actions can take place. These actions include rerouting user channels to another operational system.

D-bits can be used for transmission of network management information. At international borders they are usually set to 1.

Multiframe Structure of the Signaling Time Slot

Time slot 16 (TS16) is defined to be used for the channel associated signaling to carry separate signaling information to all user channels of the frame. TS16 is a transparent 64-Kbps data channel like any other time slot in the frame. Thirty channels share the signaling capacity of TS16. A frame structure is needed to allocate the bits of this time slot to each of the 30 speech channels. The information about the location of the signaling data of each speech channel is given to the signaling demultiplexer with the help of the multiframe structure containing a multiframe alignment word for multiframe synchronization. The data rate available for each speech channel is 2 Kbps. Because the CAS signaling systems are or will in the near future be replaced by common channel signaling we do not cover multiframe structures in detail here.

For CSS, multiframe is not needed and the signaling information of all users is carried in data packets and any time slot can be used for this. Each packet carries information about the call to which it is related and signaling information. CCS packets can in some cases, for example, in the short message service of GSM, also carry user data.

4.6.2.2   The 1.544-Mbps Frame Structure

The primary data rate in the United States and Japan is 1.544 Mbps instead of the 2.048 Mbps used in areas that go by European standards. Both

European PCM frame and the 1.544-Mbps frame are repeated at PCM sampling rate that is 8,000 times in a second. The frame structure shown in Figure 4.23 is used in North America and known as T1 or DS1 frame [5].

   The North American PCM system accomplishes frame alignment differently than does the European 2-Mbps system. Like its European counterpart, it uses eight-bit time slots, but each frame contains 24 channels. To each frame, one-bit frame, a frame alignment, or synchronization bit (S-bit) is added, and we get a 1.544-Mbps data rate as shown in Figure 4.23. A multiframe is constructed from 12 subsequent frames and their 12 S-bits make up the 6-bit frame and 6-bit multiframe synchronization words [5].

   In T1 there is no reserved time slot for CAS information as we have in the 2-Mbps frame structure. Instead of that, the least significant bit of each channel in every sixth frame is used for signaling. As a consequence, only seven bits in each time slot are transparently carried through the network and the basic user data rate is 56 Kbps instead of the 64 Kbps in the European systems.

   For frame synchronization and for demultiplexing of signaling information, frames make up a multiframe structure with two alternative lengths, a superframe containing 12 frames or an *extended superframe* (ESF) containing 24 frames. The framing bits of ESF, one in each frame, carry frame synchronization information including CRC code and data channel for network management messages. The detailed structure of the multiframe is explained, for example, in [5].
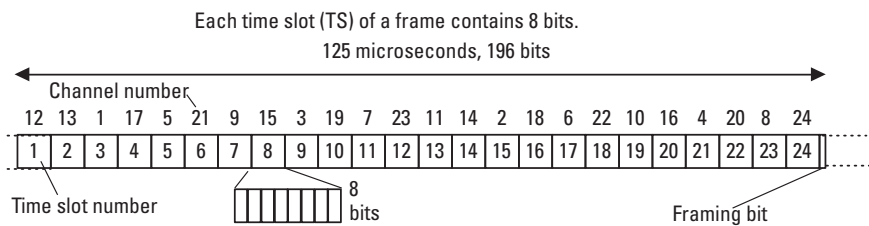
Each time slot (TS) of a frame contains 8 bits.
125 microseconds, 196 bits

Channel number

| 12 | 13 | 1 | 17 | 5 | 21 | 9 | 15 | 3 | 19 | 7 | 23 | 11 | 14 | 2 | 18 | 6 | 22 | 10 | 16 | 4 | 20 | 8 | 24 |

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |

Time slot number   8 bits   Framing bit

Frame is repeated 8,000 times in a second which is the same as PCM sampling rate.
Each frame contains one sample of 24 different speech signals.
To each frame 1 bit, called a framing bit, is added.
(24 time slots * 8 bits + 1 bit) *8,000 = 1,544 Kbps.

One bit in each slot in every sixth frame is replaced by signaling information.
As a consequence, only 7 out of 8 bits can be used transparently
through the network. Therefore, a basic channel capacity is 56 Kbps.

Figure 4.23 The 1.544-Mbps PCM frame.

In transatlantic connections, E1 frames are adapted to the T1 frame structure and transcoding between μ-law and A-law PCM is carried out. Each time slot in E1 is transmitted further in one time slot of T1.

4.6.3   Plesiochronous Transmission Hierarchy

A primary rate of 1.5 or 2 Mbps is usually too slow for transmission in trunk or even in local networks. This was noticed in the early 1970s and the ITU-T, at that time CCITT, standardized the higher data rate systems for transmission in the latter half of the 1970s. The digital systems of those days carried primarily analog information and end-to-end synchronization was rarely required. The first standardized digital higher-order transmission hierarchy is known as *plesiochronous digital hierarchy* (PDH). We review first the European hierarchy of higher-order multiplexing.

4.6.3.1   European PDH for Higher-Order Multiplexing

The higher-order multiplexers of PDH are allowed to operate according to their own independent clock frequencies. These standards are based on plesiochronous operation ("almost the same data rate"), which allows a small frequency difference between tributary signals that are multiplexed into a higher aggregate rate. For example, at 2,048 Kbps the frequency tolerance was standardized at 50 ppm, and at 8,448 Kbps the allowed tolerance is 20 ppm. This means that, for example, the data rate of a 2,048-Kbps system may deviate by 100 bps.

The basic principle of the European standard for higher-order multiplexers is that each multiplexer stage takes four signals of a lower data rate and packs them together into a signal at a data rate that is a little bit over four times as high, as shown in Figure 4.24. In addition to tributaries, aggregate frames contain frame alignment information and justification information.

The tributary frequencies may differ slightly and their frequencies must be justified to the higher-order frame. This process, called *justification* or *stuffing,* adds a number of justification bits to each tributary in order to make the average tributary data rates exactly the same. In the demultiplexer these justification bits are extracted and the original data rate for each tributary is generated.

At each hierarchy level the tributary signals are *bit interleaved* to the aggregate data stream, which means that the aggregate data stream contains one bit from tributary 1, one bit from tributaries 2, 3, and 4, and then again from tributary 1, and so on. Additional bits are needed in the frame for frame synchronization (frame alignment) and justification, and therefore the next
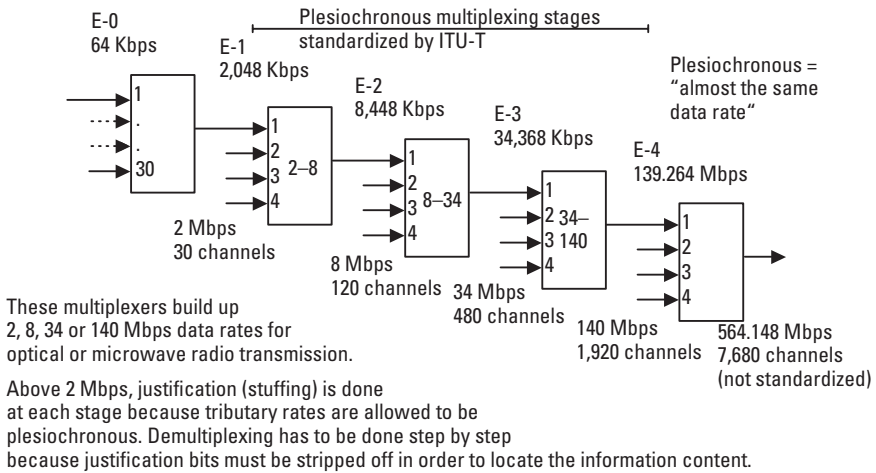
Figure 4.24 The PDH (European standard).

level has a slightly higher rate than four times the nominal tributary rate. Justification bits are added to tributaries so make their data rates equal for framing. The frame also contains some spare bits that can be used, for example, for management data transmission for a network management system. Bits for far-end alarms are included in the frames just as in the 2,048-Kbps frame discussed previously.

The standards for PDH ensure compatibility in multiplexing between systems from different manufacturers. The management functions are not standardized and they differ from manufacturer to manufacturer.

Only the local interfaces and the multiplexing scheme are standardized in PDH. The multiplexers are connected for transmission via standard interfaces at 2, 8, 34, or 140 Mbps to separate line terminal equipment or to a higher-order multiplexer as shown in Figure 4.24. The line interfaces of the line terminals for copper cable, optical fiber, and radio transmission are manufacturer specific so the vendor has to be the same at both ends.

4.6.3.2   North American PDH for Higher-Order Multiplexing

The North American PDH is shown in Figure 4.25. Higher-order rates are DS1C (3.152 Mbps), DS2 (6.132 Mbps), DS3 (44.736 Mbps), and DS4 (274.176 Mbps) [5]. The higher-level multiplexers are named in such a way that we know the DS levels, which are being combined. For example, M13 in Figure 4.25 has inputs from level DS1 and outputs at level DS3.
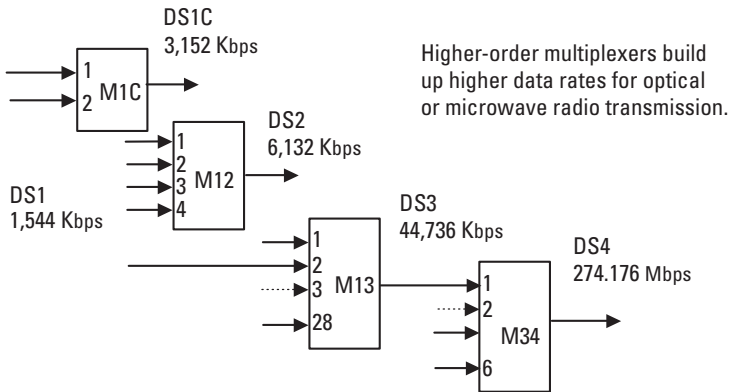
Figure 4.25 North American PDH.

As we can see in Figure 4.25 the higher-order bit rate for each multiplexer is a little bit higher than the sum of the tributary data rates. The aggregate data stream at each level contains, in addition to tributary signals, framing information and the stuffing bits that are used to justify tributary data rates, which may have slightly different data rates, into the higher-order frame. In the demultiplexer these stuffing bits are stripped off and the original tributary rate is produced.

4.6.4 SDH and SONET

The PDH higher-order systems were standardized more than 20 years ago. By the end of the 1980s, a lot of optical fiber cable had been installed and analog networks upgraded into digital networks. Then researchers realized that new standards were required to meet future requirements.

Problems with the PDH standards include the following:

• Access to a tributary rate requires step-by-step demultiplexing because of stuffing (justification).

• Optical interfaces are not standardized but vendor specific.

• To use optical cables, a separate multiplexer for each level (e.g., multiplexing from 2 to 140 Mbps in European PDH requires 21 pieces of multiplexing equipment) and separate line terminals are needed.

- American and European standards are not compatible.

- Network management features and interfaces are vendor dependent.

- High data rates (above 140 or 274 Mbps) are not standardized.

ANSI started to study a new transmission method in the middle of the 1970s to utilize optical networks and modern digital technology more efficiently. This system is called the *synchronous optical network* (SONET) and it is used in the United States.

ITU-T made its own worldwide standard, called SDH, by the end of the 1980s. SDH is actually an international extension of SONET and it was based on SONET but adapted to European networks. A subset of SDH recommendations from the ITU-T was selected as a standard for the European SDH by ETSI. You might say that there are two different synchronous optical systems: SONET in the United States and SDH in areas of Europe where European standards have been adapted. The operating principles of SONET and European SDH are quite similar and they use the same data rate at some levels, as shown in Table 4.3.

Figure 4.26 shows data rates for European SDH as well as an example of SDH equipment. SDH is a standardized multiplexing system for both plesiochronous tributaries, for example, 1.5, 2, or 34 Mbps, and synchronous tributaries.

The main advantages of SDH over PDH standards are as follows:

- The data rates for optical transmission are standardized (i.e., vendor independent).

Table 4.3

Data Rates of SONET (United States) and Corresponding SDH Data Streams (Europe)

OC-N Optical Carrier Level	STS-N Electrical Level	Data Rate (Mbps)	SDH STM-N
OC-1	STS-1	51.84	
OC-3	STS-3	155.52	STM-1
OC-12	STS-12	622.08	STM-4
OC-24	STS-24	1244.16	
OC-48	STS-48	2488.32	STM-16
OC-192	STS-192	9953.28	STM-64

To network management
computer

An example of SDH equipment

Optical interface:
STM-1   155.52 Mbps
STM-4   622.08 Mbps
STM-16 2.488 Gbps
STM-64 9.953 Gbps

Synchronous
add/drop
multiplexer

Optical interface:
STM-1   155.52 Mbps
STM-4   622.08 Mbps
STM-16 2.488 Gbps
STM-64 9.953 Gbps

SDH is used as a tranmission
layer in ATM networks

ATM interface cards of work
stations usually use the STM-1
electrical interface standard.

155.52 Mbps electrical
140 Mbps
34 Mbps
1.5 Mbps, 2 Mbps, and/or
64 Kbps interfaces

Figure 4.26   The synchronous digital hierarchy of ETSI.

- Different systems are included in standards, for example, terminal, add/drop, and cross-connection systems. These systems are discussed in Section 4.7 and they make SDH networks more flexible than PDH systems, which include only terminal multiplexer functionality.

- Access to the tributary data rates is efficient (no step-by-step multiplexing is required).

- The system is tolerant against synchronization and other system faults. Standardized redundancy functions allow operators to switch from a faulty line to an operational line.

- In the future, network management is slated to become vendor independent, with sophisticated management functions.

SDH is replacing PDH systems in the transport network. By *transport network* we mean the flexible high-capacity transmission network that is used to carry all types of information. By *flexible* we mean that telecommunications operators are able to easily modify the structure of the transport network from the centralized management system. This makes the delivery times for leased lines shorter. Leased lines are needed, for example, for LAN interconnections between the offices of a corporation.

4.6.4.1 Multiplexing Scheme in SDH

The transmission data streams of SDH are called *synchronous transport modules* (STMs) and they are exact multiples of STM-1 at the 155.52-Mbps data rate, as we can see in Table 4.3. STM-1 data are simply byte interleaved with other STM-1 data streams to make up a higher transmission data rate; no additional framing information is added. Byte interleaving means that, for example, an STM-4 signal contains a byte (8 bits) from the first STM-1 tributary, then from the second, third, and fourth tributaries, and then again from the first one. The demultiplexer receives all STM-1 frames independently.

The STM-1 frame is repeated 8,000 times a second, a rate equal to the PCM sampling rate. This makes each 8-bit speech sample visible in a 155.52-Mbps data stream. When PCM coding is synchronized to the same source as SDH systems, we can demultiplex one speech channel just by picking up 1 byte from each STM-1 frame. The frame contains frame alignment information and other information such as management data channels and pointers that tell the location of tributaries in the frame.

If tributaries are not synchronous with the STM-1 frame, a pointer (a binary number) in a fixed location in the STM-1 frame tells the location of each tributary. By looking at the value of this pointer, we can easily find the desired tributary signal. This is a great advantage over PDH systems, which require step-by-step demultiplexing (to separate information and stuffing bits) to the level of the tributary that we want to take out from the high-data-rate stream.

Multiplexing in SDH is quite a complicated matter because the multiplexing supports many different PDH and SDH streams to be multiplexed into an STM-1 stream. For example, a single STM-1 may carry 63 E-1 signals or alternatively one E-4 signal. The STM-1 frame structure and how ATM cells are inserted in it are demonstrated in Chapter 6 as an example of SDH framing. A more detailed treatment of the framing subject is not included here.

4.6.4.2 Data Rates of North American SONET

The *synchronous transport signal level 1* (STS-1) is the basic SONET module that corresponds to STM-1 of SDH. These modules have a bit rate of 51.840 Mbps and they are multiplexed synchronously into higher-order signals STS-N. Each STS-N signal has a corresponding optical signal called an *optical carrier* (OC-N) for optical transmission. Table 4.3 presents data rates for SONET and corresponding signal levels for European SDH.

An STS-1 signal consists of frames and the frame duration is 125 μs (8,000 times a second, that is, equal to the PCM sampling rate) just as in

SDH. Each frame contains 810 bytes that makes up a bit rate of 51.840 Mbps. Transport overhead information such as frame synchronization and pointers uses 27 bytes in each frame and the rest of it is used for payload; for example, for 1.544-Mbps signals that contain PCM speech channels. The detailed multiplexing scheme of either SONET or SDH is not presented here; for more detailed information the reader may refer to, for example, [5].

SONET and SDH were originally designed for transmission of 64-Kbps PCM channels. In Chapter 6 we will see how they are used when data consist of IP packets or ATM cells.

4.7 Transmission Media

Transmission systems may use copper cable, optical cable, or radio channels to interconnect far-end and near-end equipment. These channels and their characteristics are introduced next.

4.7.1 Copper Cables

Copper cable is the oldest and most common transmission media. Its main disadvantages are high attenuation and sensibility to electrical interference. Attenuation in copper cable increases with frequency approximately according to the following formula:

$$A_{dB} = k\sqrt{f}\ dB \qquad (4.13)$$

where A_{dB} is attenuation in decibels, f is the frequency, and k is a constant specific for each cable. This formula gives us approximate attenuation at other frequencies if the attenuation at one frequency is known. For example, if we measure that attenuation of a certain cable is 6 dB at 250 kHz, then at the four times higher frequency of 1 MHz it is approximately 12 dB. The speed of signal propagation in a copper cable is approximately 200,000 km/sec. The three main types of copper cables are shown in Figure 4.27.

4.7.1.1 Twisted Pair

A twisted pair consists of two insulated copper wires that are typically 0.4 to 0.6 mm thick or about 1 mm thick if insulation is included. These two wires are twisted together to reduce external electrical interference and interference from one pair to another in the same cable. The twisted pair is symmetrical and the difference in voltage (or to be more accurate, electromagnetic wave)

Examples of usage:

Twisted pair: Subscriber loops; telephone, ISDN; and ADSL and twisted pair LANs

Electrical signal

Open-wire lines: Telephone subscriber loops in rural areas; low capacity telephone FDM carrier systems

Electrical signal

A pole

Coaxial cable: Analog and digital high capacity systems in the telecommunications networks; broadcast radio and TV antenna systems; coaxial LANs

Electrical signal

Inner conductor

Outer conductor and sheath

Figure 4.27 Copper cable as a transmission medium.

between these two wires contains the transmitted signal. Twisted pair is easy to install, requires little space, and does not cost a lot. Twisted pairs are used in the telecommunications networks in subscriber lines, in 2-Mbps digital transmissions with distances up to 2 km between repeaters, in DSLs up to several megabits per second, and in short-haul data transmissions up to 100 Mbps in LANs.

Unshielded twisted pair (UTP) cables used in LANs are categorized as UTP Cat 3, 4, and 5. Cat 3 is a voice-grade cable designed for voice frequency applications, such as local loops. The characteristics of Cat 5 cable are specified up to a 100-MHz frequency and they are suitable for high-speed LANs operating at 100 Mbps or 1 Gbps (see Chapter 6).

4.7.1.2 Open-Wire Lines

The oldest and simplest form of a two-wire line uses bare conductors suspended at pole tops. The wires must not touch each other, otherwise short circuit occurs in the line and communication will be interrupted. New open-wire lines are rarely installed today but they are still in use in rural areas as subscriber lines or analog carrier systems with a small number of speech channels.

4.7.1.3 Coaxial Cable

In a coaxial cable, stiff copper wire makes up the core, which is surrounded by insulating material. The insulator is encased by a cylindrical conductor.

The outer conductor is covered in protective plastic sheath. The construction of the coaxial cable gives a good combination of high bandwidth and excellent noise immunity. Coaxial cables are used in LANs (original 10-Mbps Ethernet), in antenna systems for broadcast radio and TV, and in high-capacity analog and digital transmission systems in telecommunications networks and even in older generation submarine systems.

4.7.2   Optical Fiber Cables

Optical fiber is the most modern of the transmission media. It offers a wide bandwidth, low attenuation, and extremely high immunity to external electrical interference. The fiber optic links are used as the major media for long-distance transmission in all developed countries and high-capacity coaxial cable systems are gradually being replaced by fiber systems.

An optical fiber has a central core (with a diameter around 8 or 60 µm) of very pure glass surrounded by an outer layer of less dense glass. A light ray is refracted from the surface between these materials back to the core and it propagates in the core from end to end. The principle of optical cable transmission is presented in Figure 4.28. Compare the dimensions of optical fiber with the diameter of a human hair that is approximately 100 µm.

The principle of optical fiber transmission has been known for some decades. The breakthrough of optical fiber technology had been expected to occur ever since the first half of the 1970s. However, the development of fiber manufacturing technology and optical component technology was slower than expected, and the commercial breakthrough was delayed until the mid-1980s. Since that time all new high-capacity and long-distance cable systems, including submarine systems, have used optical fibers as a transmission medium. The advantages of optical fibers include these:
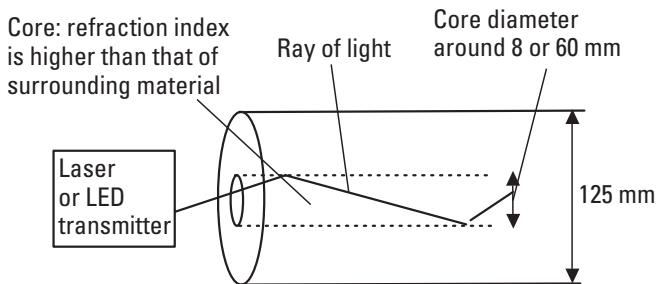
Figure 4.28   Optical fiber.

- *High transmission capacity:* Optical fibers have a very large bandwidth and they are able to carry very high data rates, up to 50 Gbps.

- *Low cost:* The cost of the fiber has decreased to the level of a twisted-pair cable; however, the coating and shielding of the cable increase the cost by a factor of two or more.

- *Tolerance against external interference:* Electromagnetic disturbances have no influence on the light signal inside the fiber.

- *Small size and low weight:* Fiber material weighs little and the fiber diameter is only of the order of a hundred micrometers instead of a millimeter or more for copper wire.

- *Unlimited material resource:* Quartz used in glass fibers is one of the most common materials on Earth.

- *Low attenuation:* Attenuation in modern fibers is less than half a decibel per kilometer and it is independent of the data rate.

One disadvantage of optical fibers is that they are more difficult to install than copper cables. Installation and maintenance, for example, repair of a broken fiber, require special equipment and well-trained personnel. Another disadvantage is that the radiation of light from a broken fiber may cause damage to the human eye. The safety standards set by IEC restrict the allowable maximum optical power that can be used and they also specify if equipment has to be able to switch off the transmitter in the case of a fiber fault. Note that visible light has a shorter wavelength (700–400 nm) than light used in optical systems.

Fibers are divided into two main categories, *multimode* and *single-mode* fibers. Multimode fibers, with diameters of 125/60 µm cladding/core are used in short-haul applications such as optical LANs. They use low-cost *light-emitting diode* (LED) transmitters at the 850-nm wavelength where attenuation of a multimode fiber is of the order of 2 dB/km. This was the first wavelength range, that is, the "first window," used for optical transmission. In multimode fibers several modes, reflected light rays, propagate through the fiber. Propagation delay is different for each ray, and the light energy of different rays is received with different delays, which causes *dispersion*, that is, spreading of light pulses as they travel through an optical fiber as shown in Figure 4.29.

The shorter the light pulses are, the higher the impact of this so-called "modal dispersion" and this makes multimode fibers suitable only for relatively low data rates. High attenuation makes them feasible only for short-haul systems.
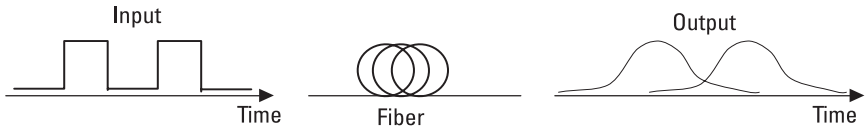
Figure 4.29   Principle of dispersion.

Single-mode fibers with approximate diameters of 125/5 μm are used in the telecommunications network in high-data-rate and long-distance applications. They allow only one mode to propagate through the fiber and modal dispersion is greatly reduced. Wavelengths of 1.3 or 1.55 μm in the second or third window in Figure 4.30 are used in single-mode fibers and then attenuation is of the order of 0.5 dB/km or even less. Semiconductor lasers are used as transmitting components and systems typically tolerate cable sections of tens of kilometers without intermediate repeaters. Long-haul, high-capacity coaxial cable systems required a repeater after every 1.5-km cable section! This partly explains the cost reduction of long-distance telecommunications during past few decades.
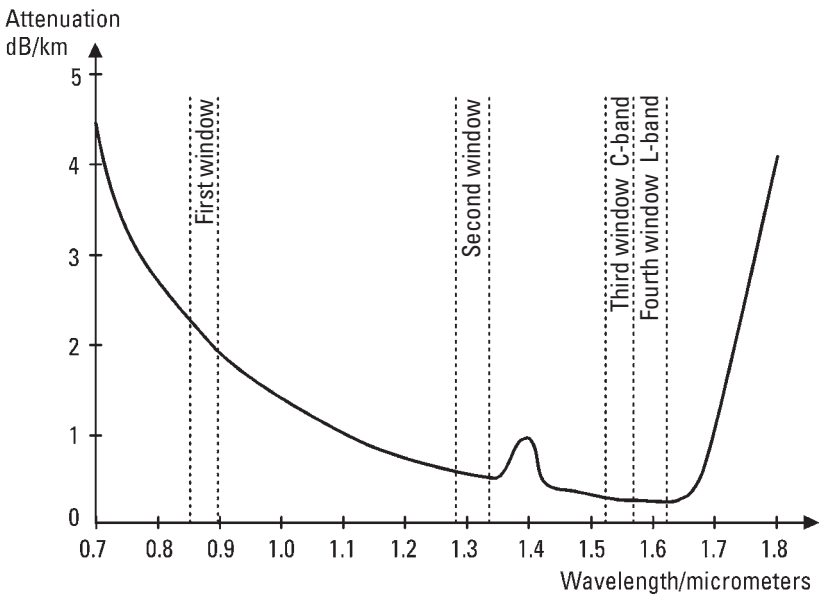
Figure 4.30   Attenuation of an optical fiber.

Note that the single-mode fibers require high-precision optical components and connectors because of the small core diameter and this makes their cost high compared with components used for multimode fibers. There are several types of single mode fibers but *nondispersion-shifted fiber* (NZ-DSF), which is optimized for 1.55-μm windows and DWDM, is the preferred type for new optic installations.

4.7.3   Radio Transmission

The most important advantage of radio transmission over cable transmission is that it does not require any physical medium. Radio systems are quick to install and because no digging of cable into the ground is required, the investment costs are much lower.

One important factor that restricts the use of radio transmissions is the shortage of frequency bands. The most suitable frequencies are already occupied and there are many systems with a growing demand for wider frequency bands. Examples of other systems using radio waves are public cellular systems, professional mobile radio systems, cordless telephones, broadcast radio and TV, satellite communications, and WLANs.

The use of radio frequencies is regulated by the ITU-R at the global level and, for example, by ETSI at the European level and the FCC in the United States. To implement a radio system, permission from a national telecommunications authority is required.

4.7.4   Satellite Transmission

In satellite communications a microwave repeater is located in a satellite. An Earth station transmits to the satellite at one frequency band and the satellite regenerates and transmits the signal back at another frequency band. The frequencies allocated by ITU for satellite communications are in the frequency range of 1 to 30 GHz. Figure 4.31 illustrates point-to-point transmission with the help of a geostationary or geosynchronous satellite using the 6/4-GHz satellite band.

The satellites used in the telecommunications network are usually located in a so-called "geostationary" orbit so that they seem to be in the same location all the time from the point of view of the Earth station, as shown in Figure 4.31. The distance of this orbit is around 36,000 km from the equator on the Earth's surface and this introduces a long transmission delay that is approximately 250 ms from the transmitting Earth station to the receiving Earth station. The speaker has to wait for a response for approximately 0.5 seconds and this disturbs an interactive communication.
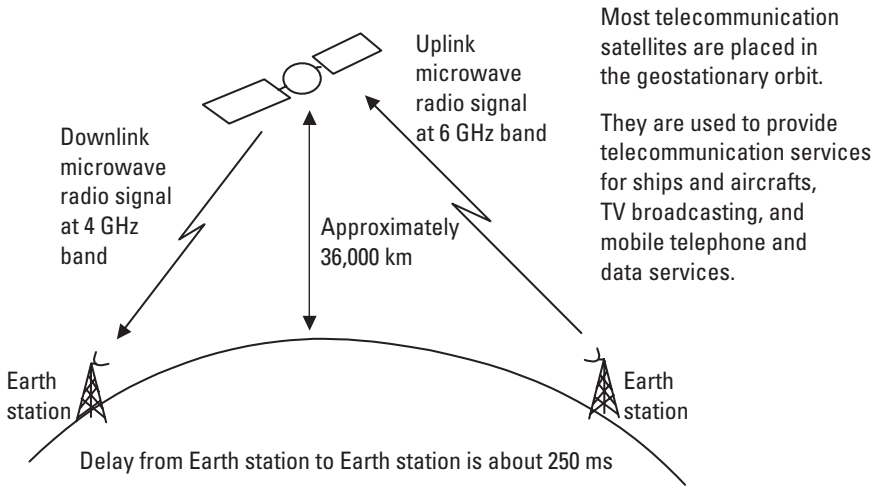
Figure 4.31   Satellite transmission.

However, satellite systems can provide telephone service to areas where no terrestrial infrastructure for telecommunications exists.

To provide wide coverage and smaller delay in mobile telephone service, many lower orbit satellite telephone systems have been developed and put into use. They have not been successful because *public land mobile systems* (PLMNs), such as GSM and CDMA, have grown rapidly and taken the major share of mobile telephone business.

One major application for satellite communications has been broadcast satellite TV. A TV program from a single satellite may be received in any part of a continent simultaneously making distribution cost per customer low. Satellite systems may also provide an attractive solution for data communications, for example, for a global hotel chain that needs a global data service to keep reservation databases synchronized.

4.8   Transmission Equipment in the Network

Many different systems are needed in the telecommunications network to transmit signals via various different channels. We review the most common transmission devices or systems in this section. Some of them were already discussed in the previous section, and some of them are also shown in Figure 4.32.
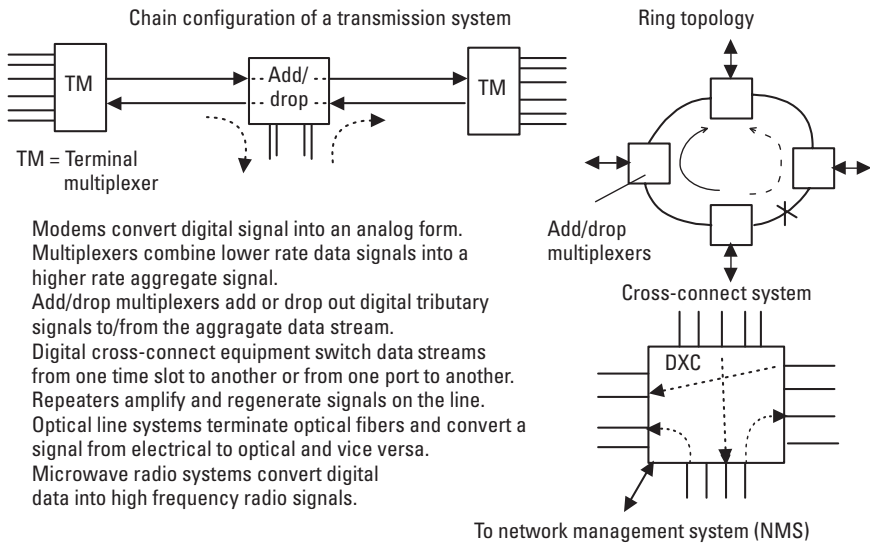
Figure 4.32 Transmission equipment and system topology.

4.8.1 Modems

A modem is a piece of equipment that includes a modulator and demodulator. Modems are used to transmit digital signals over an analog channel. Functionality of voice-band modems is described in Chapter 6 and they are used to transmit and receive data from a PC to/from an analog telephone channel. The microwave radio systems are sometimes also called modems because they send digital information over a microwave radio link, and in order to do this, they also carry out modulation and demodulation processes.

4.8.2 Terminal Multiplexers

Terminal multiplexers (TMs) or multiplexers combine digital signals to make up a higher bit rate for high-capacity transmission (Figure 4.32). The digital multiplexing hierarchies in use are PDH and SDH, which are replacing older generation PDH systems. These multiplexing schemes were described in Section 4.6.

4.8.3 Add/Drop Multiplexers

A transmission system in the network may be just a point-to-point system or it may be built as a chain or as a ring system as shown in Figure 4.32. These

configurations make efficient use of the high system capacity feasible when only a small fraction of the total transmission capacity is needed on each equipment site. The add/drop multiplexers are used in these configurations to take out (drop) some channels from the high-rate data stream and add or insert other channels into it.

4.8.4   Digital Cross-Connect Systems

The *digital cross-connect* (DXC) systems are network nodes that can rearrange channels in data streams (Figure 4.32). They make the network configuration of the transmission network flexible, because, with the help of these nodes, a network operator is able to control actual transmission paths in the network remotely from the network management center. The basic functionality of DXC is the same as the functionality of digital exchanges that establish speech or ISDN connections. However, DXC is controlled by the network operator, not by a subscriber, and its configuration is not changed as frequently.

Cross-connect systems are available that are able to switch high-order data rates, not just 64 Kbps as ordinary exchanges do. DXC may also contain redundancy functions that automatically change configurations so as to bypass a faulty transmission section.

SDH and SONET networks often use a ring topology like that shown in Figure 4.32 for higher reliability. These standards specify redundancy functions and a node in a ring may switch traffic from a faulty connection to the redundant path as shown in Figure 4.32.

4.8.5   Regenerators or Intermediate Repeaters

Intermediate repeaters are needed if the communication distance is very long. They amplify an attenuated signal and regenerate the digital signal into its original form and transmit it further. The operation priciple of a regenerator was described in Section 4.5.

4.8.6   Optical Line Systems

Optical line systems contain two terminal repeaters at each end of the fiber. They convert an electrical digital signal into an optical one and vice versa. These systems include, as most other transmission systems do, supervisory functions such as fault and performance monitoring. Note that SONET and SDH systems include multiplexing functions as well as the functions needed for optical transmission. In PDH multiplexers, optical line systems are

separate devices that are interconnected with standardized interfaces, which were discussed in Section 4.6.

As we discussed in Section 4.2.5, optical systems transmit light energy pulses to the fiber; they do not use light as a carrier the same way as in radio communications. In bidirectional systems two fibers, one for each transmission direction, are needed as shown in Figure 4.33. However, development of semiconductor laser technology has made narrow bandwidth lasers available and several parallel optical signals at different wavelengths can use the same fiber. This *wavelength-division multiplexing* (WDM) uses an optical coupler to combine optical signals (WDM multiplexer) and optical filters (WDM demultiplexer) to separate optical signals at the receiving end as shown in Figure 4.33.

4.8.7 WDM

Many single-mode fiber cables have been installed and technical solutions that increase fiber capacity without installation of new cable have become very attractive as the demand for transmission capacity increases. Particularly in long-distance systems, WDM has become popular and it can increase fiber capacity by a factor from 10 to 100.
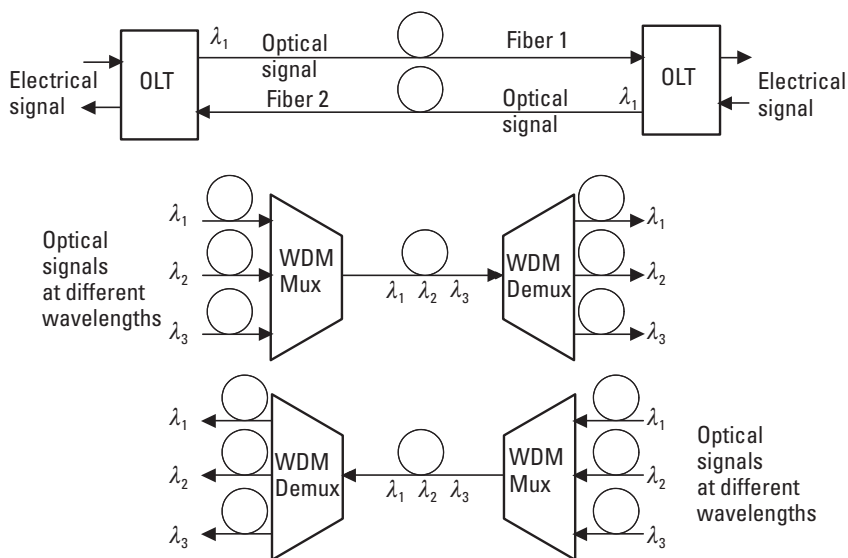
Figure 4.33 Optical fiber system and WDM.

Cooled *distributed feedback* (DFB) lasers are available in precisely selected wavelengths. The ITU has defined a laser grid for point-to-point DWDM systems based on 100-GHz wavelength spacing. There are 45 defined wavelengths in a range from 196.1 THz (1,528.77 μm) to 191.7 THz (1,563.86 μm) in the third window ("L" band in Figure 4.30), which is a compatible range for the EDFAs discussed later. Manufacturers can deviate from the grid by extending the upper or lower bounds or by spacing wavelengths more closely, typically at 50 GHz or even down to 25 GHz to double or triple the number of channels. Each optical channel can be used for transmission of light pulses at 10 Gbps, or an even higher data rate at 100-GHz spacing, and, with the help of DWDM technology, a pair of fibers can provide data capacity of several hundreds gigabits per second.

Most DWDM systems support standard SONET/SDH optical interfaces. Often short-haul STM-16 (2.4 Gbps) at the 1310-nm wavelength is used as an input signal for DWDM systems but also other interfaces, such as OC-192 for 10-Gb Ethernet, can be supported. The basic structure of a DWDM system is shown in Figure 4.34. Only one transmission direction is shown in the figure. Transponders in Figure 4.34 convert incoming optical signals into ITU-standard wavelengths. Each transponder is designed to support a certain interface, for example, STM-16, and it carries out optical-to-electrical conversion, signal regeneration, and electrical-to-optical conversion and transmits signals to the optical multiplexer at one wavelength specified by ITU.

DWDM technology has improved, and will continue to further improve, utilization of fiber bandwidth close to the huge capacity of optical fibers that will be achieved in the future by coherent radio-like optical technology.
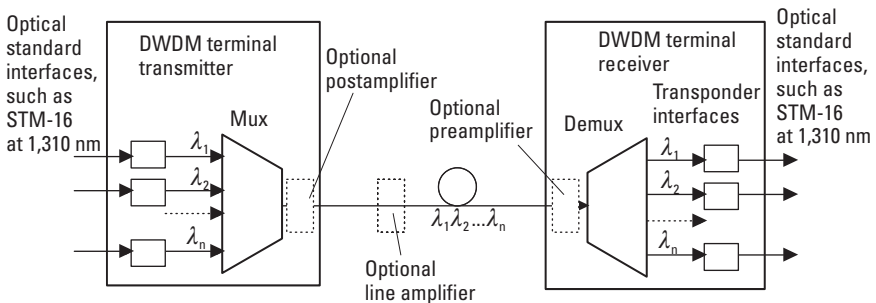
Figure 4.34   DWDM system (one transmission direction only).

4.8.8   Optical Amplifiers

The section length of a long-haul optical system from optical transmitter to receiver is limited to some tens of kilometers depending on the transmission data rate although attenuation of a fiber is quite low. In the case of longer systems, regeneration or amplification is required. Regeneration of optical DWDM signals is very expensive because it requires an optical demultiplexer and demultiplexer, regeneration of each signal in electrical form, and optical receivers and transmitters for each wavelength. Optical amplifiers offer a more attractive solution for implementation of long-haul DWDM systems and they can be used to boost the DWDM output signal, to amplify all wavelengths on the line, or to amplify the received signal before the optical demultiplexer as shown in Figure 4.34.

There are many different optical amplifiers, but *erbium-doped fiber amplifiers* (EDFAs) in particular have become popular in long-distance trans-missions of high-capacity DWDM signals. They operate in a low attenuation wavelength range from 1,520 to 1,565 nm (ITU grid) and gain signals at all wavelengths by typically 30 dB or more. Figure 4.35 shows a simplified dia-gram of an EDFA. A weak optical input signal containing many wavelengths in the 1.54-μm range enters the erbium-doped fiber, into which light at 980 or 1,480 nm is injected using pump lasers. This injected light stimulates the erbium atoms to release their stored energy as additional 1,540-nm light as the input signal is inserted [5]. As this process continues down the doped fiber, all optical signals in the 1,540-nm range grow stronger.

With pump power from a few milliwatts to 100 mW, EDFAs achieve gains from 20 to 50 dB extending fiber sections between amplifiers to 100 to 200 km. The spontaneous emission in the EDFA also adds noise to the sig-nal, which limits the number of concatenated optical amplifiers. At distances longer than 600 to 1,000 km, the signal must be regenerated, which requires
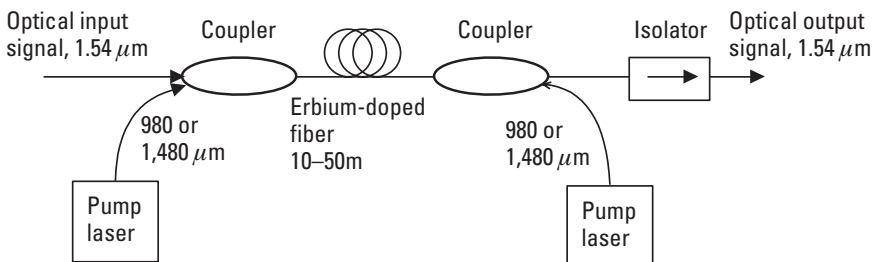
Figure 4.35   Optical EDFA amplifier for DWDM signals.

optical demultiplexing, optical-to-electrical conversion, electrical regeneration, electrical-to-optical conversion, and optical multiplexing.

4.8.9   Microwave Relay Systems

Microwave relay systems are radio systems that may be used for point-to-point transmission instead of copper or optical cable systems. They convert digital data into radio waves and vice versa. They also perform supervisory functions for remote performance and fault monitoring from the network management center. Figure 4.36 illustrates the structure of a point-to-point radio relay system used in the telecommunications network.

Microwave radio relay systems usually operate at radio frequencies in the range from 1 to 40 GHz. These frequencies are focused with parabolic dish antennas and applicable communication distances range from a few kilometers up to approximately 50 km depending on the frequency in use and the characteristics of the system. The radio waves at these frequencies travel along a straight line and therefore this kind of radio transmission is called *line-of-sight* transmission. The higher the frequency, the higher the propagation loss, as we saw in Section 4.2, and the shorter the communication distance. At very high frequencies, weather conditions influence attenuation and transmission quality, which restricts the available frequency
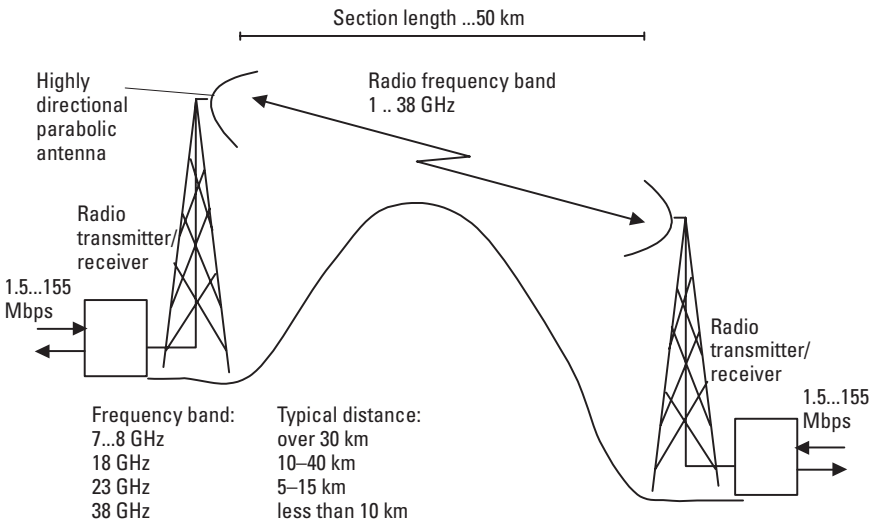
Figure 4.36   Microwave radio transmission.

band suitable for radio transmission and maximum transmission distance. Figure 4.36 shows examples of how communication distance depends on the radio frequency in use.

4.9 Problems and Review Questions

Problem 4.1

How wide a bandwidth does a pulse with duration of (a) 1 ms and (b) 1 μs require if only the strongest part of the spectrum needs to be transmitted? What is the bandwidth of a carrier wave with a duration of (a) 1 ms and (b) 1 μs?

Problem 4.2

What is continuous wave modulation and why is it often used in transmission systems?

Problem 4.3

(a) Draw the spectrum of a cosine wave at a frequency of 1 kHz. (b) Draw the spectrum of an AM signal when the carrier frequency is 100 kHz and the modulating message is a cosine wave at 1 kHz. (c) Draw the spectrum when the modulation method is SCDSB. (d) Draw the corresponding spectrum of SSB modulation.

Problem 4.4

(a) Draw the constellation diagram (or signal space diagram) for an 8-PSK signal so that the in-phase carrier waveform represents bit combination 000. Write in the diagram which bit combination each signal could represent. Take care that you minimize the bit error rate. (b) Draw the constellation diagram of a 16-QAM signal where a carrier with a 45° phase shift and a high amplitude corresponds to a bit combination of 1100. Write bit combinations for each signal so that the bit error rate is minimized. [*Hint:* Use Gray code for two bits for columns (I component) and two bits for rows (Q component) and combine them for each signal in the constellation.]

Problem 4.5

Explain how the radio wave propagation modes differ at (a) low-frequency, (b) medium frequency, and (c) and ultra high frequency bands.

Problem 4.6

Estimate the transmission capacity of an optical fiber that operates over the 0.9- to 1.6-m wavelength range if coherent optical transmission is used. Assume that the speed of light is the same as in space (300,000 km/sec) and the following modulation methods are in use: (a) Voice signal bandwidth is 4 kHz and it is SSB modulated into the fiber. (b) Voice signal is PCM coded and transmitted in a binary form through the cable. Assume that the modulation scheme in use is capable of transmitting 1 bps/Hz.

Problem 4.7

Derive on your own the formula, $L = [4\pi f l/c]^2$, step by step for the free-space loss (see Section 4.2.6). Use the formula for the effective aperture area of isotropic antenna, $A_{ei} = \lambda^2/(4\pi)$, and a spherical surface area $A = 4\pi l^2$ over which transmitted power is distributed.

Problem 4.8

Show that the equation for radio wave attenuation in decibels, $L_{dB} = 92.4 + 20 \log_{10} f/\text{GHz} + 20 \log_{10} l/\text{km}$ dB, follows from the equation of attenuation $L = [4f\pi l/c]^2$. Note that, for example, $f = f/\text{GHz} \times 10^9$.

Problem 4.9

The approximate distance between an Earth station and a geostationary satellite is 40,000 km. (a) What is the attenuation of the uplink radio section at the 6-GHz frequency? (b) What is the attenuation in the downlink direction at 4 GHz?

Problem 4.10

Consider a cell in a GSM cellular network operating at 900 MHz and a cell in a DCS-1800 network operating at 1.8 GHz. The DCS-1800 base station is installed in the same site as the GSM base station. Assume that all system parameters except frequency are equal and use the free-space loss formula. What would be the radius of the DCS-1800 cell if the radius of the GSM cell is 1 km?

Problem 4.11

How much higher transmission power is needed, according to the free-space loss formula, if the radio transmission distance is doubled (for the same performance)?

Problem 4.12

A telecommunications network operator is aiming to update a GSM network with DCS-1800 base stations. The cells of GSM (900 MHz) are designed for a maximum transmission power of 1W. What should be the maximum transmission power of DCS-1800 (1.8-GHz) base stations with the same cell structure? Assume here a free-space environment and that the only difference between systems is the frequency.

Problem 4.13

What is the approximate gain of the satellite TV antenna when the diameter of the dish is 0.6m and the frequency is 10 GHz? How much better is the S/N ratio if the antenna is changed to a larger one with diameter of 1m?

Problem 4.14

What is the received power level (dBm) and power (W) when transmitted power is 1W, frequency 1 GHz, distance 1 km, and transmitter and receiver antenna gains are 14 and 2 dB, respectively? Assume a free-space loss approximation for link loss.

Problem 4.15

What are the theoretical maximum symbol rate r and the maximum binary bit rate C through the following baseband channels: (a) bandwidth $B = 3$ kHz and S/N = 20 dB (degraded speech channel); and (b) bandwidth $B = 5$ MHz and S/N = 48 dB (typical video channel)?

Problem 4.16

How many bits can be encoded into each symbol in the case of baseband systems (a) and (b) in Problem 4.15? How much higher is the data rate in case (b) in Problem 4.15 because of the wider bandwidth and how much higher is the bit rate because of the improved S/N compared with the channel in case (a) of Problem 4.15?

Problem 4.17

Estimate how many symbol values (signals in the constellation diagram) there should be in the case of a 28.8-Kbps modem using QAM if the symbol rate is 3,200 bauds.

Problem 4.18

Why do we perform line encoding before data are transmitted to the transmission channel?

Problem 4.19

Explain how binary values 1 and 0 are represented in the following codes: (a) NRZ, (b) RZ, (c) AMI, and (d) Manchester.

Problem 4.20

Explain the operating principle of a regenerator (regenerative repeater).

Problem 4.21

What are the main two multiplexing methods and how do they operate?

Problem 4.22

Explain the structure of a 2-Mbps PCM frame.

Problem 4.23

Explain the structure of a 1.5-Mbps PCM frame.

Problem 4.24

Explain what is PDH?

Problem 4.25

What is SDH and what advantages does it provide over PDH?

Problem 4.26

The measured attenuation at 1 MHz of a 1-km copper cable pair is 18 dB. What is the approximate attenuation at (a) 250 kHz, (b) 500 kHz, (c) 2 MHz, and (d) 4 MHz?

Problem 4.27

What are the advantages of (a) optical transmission, (b) microwave radio transmission, and (c) satellite transmission? Compare their characteristics.

Problem 4.28

What do we mean by *dispersion* in optical fibers?

Problem 4.29

What do we mean by *dense wavelength-division multiplexing*?

Problem 4.30

Calculate the one-way delay and two-way delays of a transmitted signal from one Earth station to another Earth station via geostationary satellite. The distance between a satellite and each Earth station is assumed to be 40,000 km.

Problem 4.31

STM-1 contains 63 primary 2-Mbps data streams and each of them contains 30 time slots for speech. (a) How many simultaneous calls (64 Kbps) can be transmitted over a single fiber pair used by the STM-16 optical system? (b) What is the number of simultaneous calls if a DWDM system using a 100-GHz wavelength grid from 1,528.77 nm/196.1 THz to 1,563.86 nm/191.7 THz is implemented? (c) The STM-16 signal is transmitted through each optical channel. What will be the total data rate of the DWDM system from part (b)?

Problem 4.32

Why did optical amplifiers become so popular in long-distance networks after the introduction of DWDM technology?

References

[1]   Carlson, A. B., *Communication Systems:, An Introduction to Signals and Noise in Electrical Communication,* New York: McGraw-Hill, 1986.

[2]   Redl, M. S., K. M. Weber, and M. W. Oliphant, *An Introduction to GSM*, Norwood, MA: Artech House, 1995.

[3]   Walke, B. H., *Mobile Radio Networks,* Chichester, England: John Wiley & Sons, 1999.

[4]   Tabbane, S., *Handbook of Mobile Radio Networks,* Norwood, MA: Artech House, 2000.

[5]   Freeman, R. L., *Telecommunication System Engineering*, 3rd ed., New York: John Wiley & Sons, 1996.

5

Mobile Communications

The major application for wireless communications has been speech. Radio telephones have been around for many decades, but the capacity of these systems has been very limited. These radio telephone networks consisted of only a few *base stations* (BSs) with which mobile units communicate, and each BS covered a large geographical area. The number of simultaneous calls inside the area covered by one BS was restricted to the number of channels available for this BS. Therefore, the capacity of these systems was low and the radio telephone service was available only to professionals.

During the 1970s, the development of digital switching and information technologies made modern cellular telephone systems feasible. The cellular principle offered a solution to the capacity problem. Different analog cellular standards were developed in Nordic countries, the United States, and Japan at the end of 1970s.

In this chapter we introduce first the idea and operation of cellular radio systems in general. The common principles of cellular systems are valid for any public land mobile network. Then we will review other mobile systems such as paging systems, cordless telephones, and WLANs. In the last section of this chapter, we review the structure and operation of the GSM network. Our goal in this chapter is to provide the reader with an understanding of what is required of the network to enable someone to receive or initiate a call anywhere in the world. The natural requirement for this is that compatible service be available. We use GSM as an example of a digital cellular system because it is currently the dominant global digital technology.

5.1   Cellular Radio Principles

The main problem of conventional radio telephone networks was low capacity because of the limited frequency band available for this service. Cellular networks provide a solution for this by using the same frequencies in multiple areas inside the network. This principle of frequency reuse with the help of a cellular network structure was invented at Bell Laboratories during the 1960s. The technical development of radio-frequency control, the microprocessor, and software technologies made cellular networks feasible by the end of 1970s. Here is a list of the most important common characteristics of cellular systems:

- Frequency reuse provides a much larger number of communication channels than the number of channels allocated to the system.

- Automatic intercellular transfer, or a handover, ensures continuity of communication when there is a need to change BSs.

- Continuous monitoring of communication between the mobile and BS verifies the quality and detects the need for a cell transfer.

- Automatic location of mobile stations within the network ensures that calls can be routed to mobiles.

- Mobile stations continuously listen to a common channel of the network in order to receive a call.

Figure 5.1 presents the basic elements of a simplified cellular network. BSs are radio transmitter/receivers by which the *mobile stations* (MSs, such as telephones) are connected to the wire-line network. The BSs are connected to the *mobile switching center* (MSC) by primary rate digital connections. The MSC acts as a local exchange in the fixed network. In addition to the switching and other functions of an ordinary telephone exchange, the MSC also keeps track of the subscribers' locations with the help of location registers. We discuss this equipment in the following section.

Note that all cellular networks are designed to act as access networks. Their main purpose is to make mobile subscribers accessible from the global (fixed) telecommunications network. The mobile cellular networks always rely on a fixed network. They have no switching hierarchy similar to that of a fixed network (see Chapter 2) and international calls are connected via a fixed network.
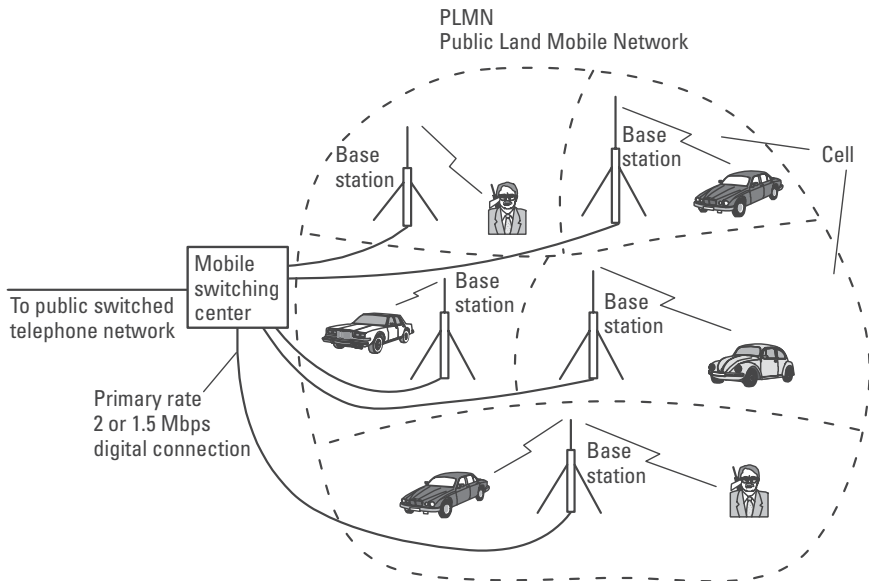
Figure 5.1 Basic structure of a cellular radio network.

5.2   Structure of a Cellular Network

This section reviews the structure of a general cellular network. The detailed structure of a cellular radio network, the terminology of network elements, and their detailed functions are dependent on the network technology in question.

5.2.1   Cellular Structure

Instead of covering an entire area with high-power fixed radio stations, the way older generation radio systems had to, the area of a cellular network is divided into small cells of only a few kilometers or less across as shown in Figure 5.2. Areas where subscriber density is high are covered by smaller cells than areas where subscriber density is low. The power BSs and MSs are automatically decreased with the decreased cell size.

The BSs and MSs (telephone) are controlled to keep their transmission power as low as possible. This low-power transmission does not interfere with other users of the same frequency (reuse of frequencies) some cells away from this cell. This is how each frequency channel can be used again and again and, in principle, a network operator can increase capacity without
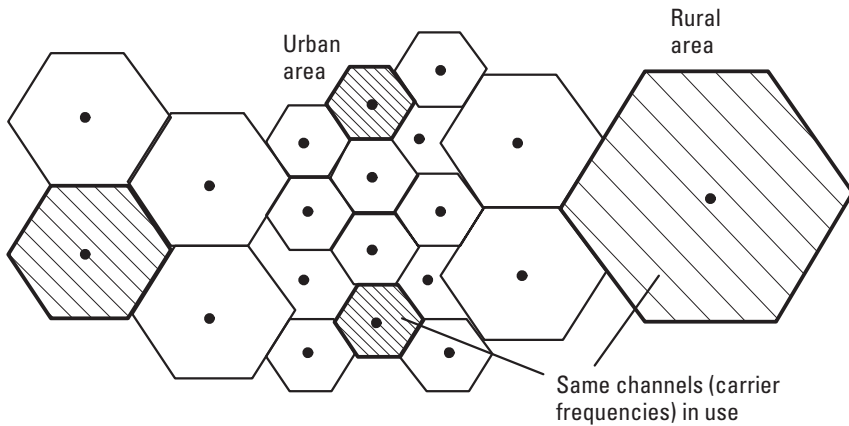
Figure 5.2 Cellular structure of a mobile radio network.

limit by reducing cell size. Naturally, this requires investment in additional BS sites. How often each carrier frequency is used is termed the *frequency reuse factor* and it depends on the system. Note that in the CDMA cellular system, which is introduced in Section 5.4.5, neighbor cells may use the same carrier frequency and there channeling is based on the spreading code instead of frequency (and time slot).

The consequences of reduced cell size are handier and less expensive telephones as well as longer operational life for the battery. Low transmission power also provides a safety improvement from the users' point of view. Because of public concern about handheld terminals and their adverse effects on health, low transmission power has become increasingly important.

In a conventional fixed network, telephone calls are always routed to one fixed telephone socket, as we saw in Chapter 2. In a cellular network a subscriber is located in one cell at a time. Now the network has to include additional intelligence to be able to connect a call to the cell where the called subscriber is available at that time. To succeed at this, the cellular networks have two databases or registers, a *home location register* (HLR) and a *visitors location register* (VLR), and with them the network is able to manage the mobility of its subscribers.

5.2.2 HLR and VLR

When subscribers purchase a mobile telephone, they are registered in the HLR of their own mobile telephone operator. The HLR stores their up-to-date subscriber information such as where (in the area of which VLR) they

are located presently, what services they have the right to use, and a number where she has transferred calls. The HLR is the global central point where their information is available wherever they are located. When a call is routed to them, the dialed subscriber's telephone number tells the network where their HLR can be found.

VLR stores information about every subscriber in its area. The VLR informs the HLR when a new subscriber arrives in its area. It also contains more accurate information of where (to which cell or group of cells) to connect incoming calls directed to a certain subscriber. The VLR is usually integrated into a mobile telephone exchange but the HLR is usually a physically separate efficient database system.

5.2.3 Radio Channels

Each BS provides two main types of channels, as shown in Figure 5.3: the common control channel and the dedicated channels. In the downlink or forward direction (from network to mobile stations) information such as network identification, location information, designated power level, and paging for incoming calls is sent on the common control channel of each cell. When MSs are in idle mode (no ongoing call) they are continuously listening to the common control channel of one cell. In the uplink or reverse direction

HLR, home location register, stores subscriber information and updated location information (VLR address). Each subscriber is registered in one fixed HLR. VLR, visitors location register, stores subscriber information of each MS located in its area.

Microwave link to mobile switching center

Lines to mobile switching center

Base station

Common control channel that all mobiles listen to when they are in idle mode. Call request messages are sent on this channel in the case of an outgoing call and paging messages are sent in the case of an incoming call.

Dedicated channels that are used for calls (speech or traffic channels) when a mobile is in dedicated mode. Each dedicated channel has its own control channel for performance monitoring and control.

Figure 5.3 The main types of radio channels.

of the common control channel the MSs send, for example, call-request messages in the case of outgoing calls and location update messages when they notice that they have arrived in a new location area.

One dedicated user channel or a traffic channel is allocated for each call. During an call, a MS is said to be in dedicated mode. Each dedicated channel requires the transmission of control information in addition to speech transmission. This is needed for transmission power control of mobile stations and for transmission of performance monitoring information from MSs to the network. When the call is cleared the dedicated channel is released and available for other users.

In Figure 5.3 we see that BSs are connected to the mobile switching center by a radio relay system or by a cable line (optical or copper cable). Especially in rural areas microwave links are attractive because cables are usually not available for BSs and they are very expensive to install. Microwave radio requires and antenna but this is not a problem—an antenna tower is always available because it is needed for the BS antennas.

5.3   Operating Principle of a Cellular Network

In the fixed telephone network each subscriber is identified by the number of a certain subscriber loop that is connected to a certain telephone socket. In the case of a cellular telephone the identification is in the telephone set (MS) itself. The cell structure of the network and the mobility of the user require the cellular network to keep track of the location of each MS in order to be able to route a call to the destination.

We now review the principles of how the cellular network manages the mobility of users and how calls are initiated and received. We introduce the operation of a cellular network in general; therefore, the terms and operation presented may not be consistent with the terms and detailed operation of a particular network technology.

5.3.1   MS in Idle Mode

The MS is preprogrammed to know the frequencies of the control channels. When it is switched on, the mobile scans these frequencies and selects the BS with the strongest common control channel. Then the MS transmits its unique identification code, which may be its telephone number (or other identification code depending on the system), over the control channel in order to inform the VLR. The VLR, with the help of the identification of the MS, determines the address of the subscriber's home country and the home network. Then the MSC/VLR transmits the signaling message toward the

home network. The message is then routed to the HLR, which is then informed that this specific subscriber is now located in the area of a certain VLR. The HLR stores this information. Now the HLR is able to route the calls to the right MSC/VLR, which routes it further to the mobile subscriber.

The MS then continuously listens to the common control channel and, if necessary, transfers to the control channel of another cell (Figure 5.4). Each network is divided into small location areas that contain a group of cells. All BSs inside a certain location area send the same global code dedicated for that location area on the common control channel. If the MS moves, changes the channel and the location information sent by the network changes; the MS notices it and informs the network, which then updates the location information stored in the VLR and HLR (if needed).

5.3.2 Outgoing Call

The number that a user wants to call is entered into the memory of the mobile telephone through its keypad. When the user presses the Call button, the mobile telephone sends a set of signaling messages to the BS via the common control channel, as shown in Figure 5.4. These messages contain the dialed digits, which the BS passes to the MSC for routing.
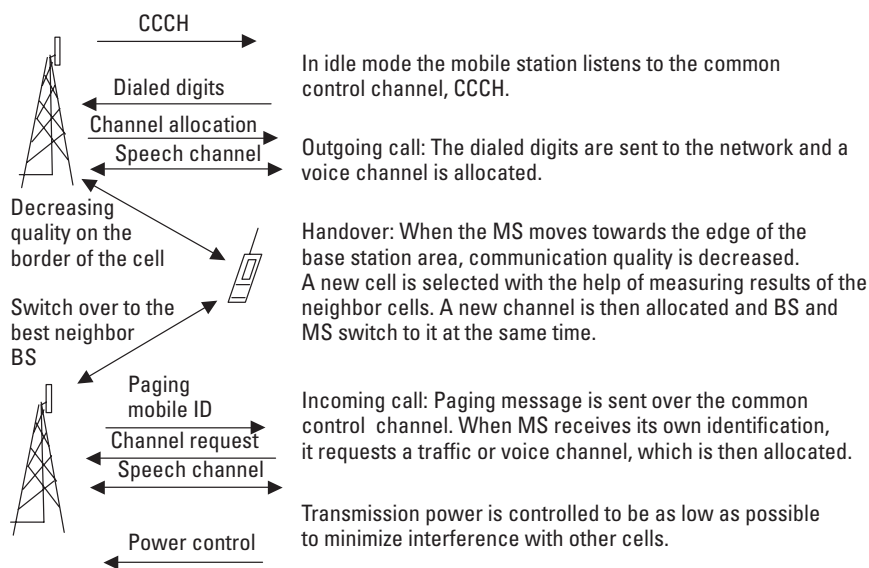
Figure 5.4 Basic operation of the cellular network.

The MSC analyzes the dialed number, passes the digits to the public telephone network for call establishment through the PSTN, and requests a BS to allocate a dedicated speech channel for the calling mobile. The MS and BS switch to this channel when the called party answers and the conversation is allowed to start (Figure 5.4).

5.3.3   Incoming Call

When a call is to be connected to the MS, the HLR determines to which VLR address the call should be routed. This address is global, containing the country and network codes according to international telephone numbering scheme. The call is then routed to the MSC/VLR, which knows the more exact location (the location area) of this specific subscriber inside its area. A paging message with MS identification is sent on the common control channel of all BSs in that area where the subscriber is currently located. The receiving MS continuously listens to this channel and when it receives the message containing its own identification it requests a speech channel and a channel is allocated for this call. The BS and MS switch to the allocated channel, the telephone rings, and when the subscriber presses the Call button, the call is connected.

5.3.4   Handover or Handoff

During a call the quality of the connection is continuously monitored and the transmission power of the MS and BS is adjusted to keep the quality at a sufficient level while at the same time keeping the transmission power as low as possible. When an MS moves close to the border of a cell, the transmission power is adjusted to the maximum allowed for that cell. As an MS moves further away from the BS, the S/N of the channel decreases and the error rate increases. If the quality falls below a predetermined level, a new channel is allocated in a neighboring cell and both the BS and the MS are requested to switch to the new channel at the same time instant. The cellular network has analyzed the measuring results before the switch and estimated the quality between the MS and neighbor cells. The best alternative is selected for a new cell.

5.3.5   MS Transmitting Power

During the planning phase of a cellular network, the maximum transmitting power for each cell is defined. This power is dependent on the desired cell size and on geographic conditions. The transmitting power of the common

control channel of the BS is adjusted to a level that is high enough to cover the cell area but not higher than necessary. During a call the network, to minimize interference between cells that use the same frequency, continuously controls the transmitting power of the MS and the BS. This also saves the battery of the MS.

5.4   Mobile Communication Systems

So far we have looked at the generic operation of cellular mobile radio systems because of the importance of these systems. However, there are many other important mobile communication systems, and we briefly introduce some of them in this section.

5.4.1   Cordless Telephones

Cordless phones were originally developed for the residential market and they were designed to cover only one local area such as a house and garden. They support only local mobility and should not be considered competitors for cellular mobile networks. We now look at the most important applications of cordless telephones.

5.4.1.1   Residential Use

The only advantage of cordless telephones over fixed telephones in ordinary residential use is a wireless handset that allows some mobility. The BS of a cordless telephone is connected to the fixed telephone socket and only one handset for each base station is typically in use (Figure 5.5). The BS unit contains a battery charger for the handset. Many systems in use are still analog *first generation cordless phones* (CT1).

Residential use | Telepoint or wireless local loop, WLL | Cordless corporate network
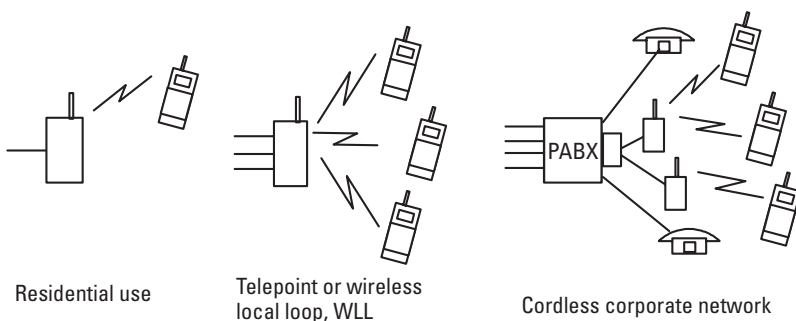
Figure 5.5   Cordless telephones and their applications.

5.4.1.2   Telepoint and WLL

Digital *second generation cordless telephone technology* (CT2) was developed for so-called "telepoint" use in addition to residential markets and offices. Telepoint was a service in which BSs were installed in key locations in a city such as railway stations and airports. A user of this service could take his or her digital cordless telephone from home or office (or rent a cordless telephone) and make a call outside via the telepoint BS. Subscribers were usually not able to receive a call. This service was not successful and most telecommunications network operators have abandoned it. The main reason for this was rapid expansion of cellular mobile service, which allows much better service and mobility.

The latest digital cordless technologies, such as *Digital European Telecommunications* (DECT), are also used in some areas to provide WLL service. With DECT technology a new operator that does not have its own cable network can provide telephone service. The WLL applications were seen to be important to generate competition in the area of traditional fixed telephone subscriber service provision. With the help of cordless technology, a new network operator can efficiently provide a service that is better, in terms of mobility, than the competing fixed telephone service by the operator who owns the cables of the fixed access network. However, the importance of WLL has decreased because of the reduced costs of cellular telephone service.

5.4.1.3   Cordless Corporate Network

In most companies internal wireless communications as well as external communications rely on the public cellular networks. The corporate telephone network is built on the fixed telephone service provided by the PABX/PBX of a company. One attractive application of modern digital cordless technologies, such as DECT, was considered to be cordless corporate networks where the PABX is upgraded to control wireless DECT telephones in addition to wire-line telephones. This technology supports handover and terminals can move freely inside the area of one PABX that controls multiple base stations. Internetwork mobility management functions make it possible to extend the mobility of DECT to other office sites of a corporation and probably even to the local public network if the local public network operator supports DECT technology. The corresponding American technology is called a *personal access communication system* (PACS).

5.4.2   Professional or Private Mobile Radio (PMR)

The PMR systems are dedicated and independent mobile radio systems. Some of them are just simple "walkie-talkie" type radios, others are complex

networks that use a technology similar to that of public cellular mobile radio systems.

One typical PMR is owned by a taxi operator. It supports telephone calls and some data communication between a control desk and a number of car telephones in the area. A small number of radio channels are allocated for each of these systems inside a geographic area.

Traditionally, each organization has built its own mobile radio system that is completely independent from others. The modern systems utilize a so-called "trunking" principle, which means that a group of radio channels is shared between several organizations. Radio channels are used on demand just as our call reserves one of the fixed channels, or "trunks," from one exchange to another. This improves the utilization of radio frequencies and is economically feasible because of reduced investments for network infrastructures. For each organization a closed user group is set up and this VPN operates in the same way as if it were physically separate. The systems are logically separate, but they use any free radio channel from a common channel pool.

These resource-sharing networks, also called trunked networks, are managed by a network operator. They are configured to provide a specialized service for each VPN of a corporate customer. The use of a frequency band is optimized by sharing it between multiple user organizations.

5.4.2.1 Operating Principle of the Trunked Networks

In the trunked network, a central equipment allocates a free channel from a common channel pool in real time to the user who requests it, and to him alone, for the duration of the communication. For each organization a VPN is defined in the system. Dynamic channel allocation uses the radio capacity efficiently and users still feel as if they had a separate dedicated system in use. Each user organization may have its own dispatch station just as in a separate conventional dispatch radio system.

Figure 5.6 illustrates the principle of channel allocation in the case of conventional PMR and of trunked PMR. There are three conventional dispatch radio networks in this simplified example with one radio channel allocated for each; that is, each organization has only one channel available whether others communicate through their channels or not. There is a demand for three simultaneous calls, two for organization 1 and one for organization 3, and one of the calls is blocked although one radio channel is free.

The lower part in Figure 5.6 presents the principle of a trunked radio system. Now a pool of all three radio channels is shared by all users. The

Conventional PMR (dispatch) network:
One channel for each organization

Trunked network:
Radio channels (spectrum) are shared by all users who may belong to separate networks
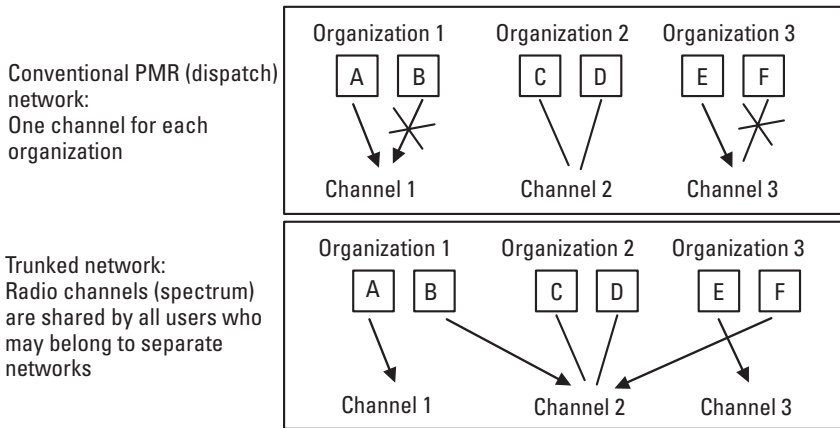
Figure 5.6 Operating principle of a resource-sharing network and a trunked network.

channels in this pool are allocated on demand and blocking occurs only when the total number of calls exceeds three in this simplified example.

 To further improve utilization of radio frequencies, the trunked networks utilize a cellular structure and technology that is similar to that used in public cellular networks.

5.4.2.2   Trunked Networks

Many analog trunked networks are in use. The frequency bands in use are different from those used in public cellular networks. Most of the current analog trunked networks provide enhanced voice services such as priority call in case of an emergency and group calling (group is defined by the network operator), and each network typically contains a terminal that has a dispatching role.

 The networks also support built-in data communication features such as predefined and user-defined text messages. They may also provide telemetry services such as remote control of unmanned stations; measurements of temperature, wind force, and water level; and alarms for buildings and remote alarm control (on/off machines or lights). For taxi companies they may provide an automatic response of the status at the moment and automatic vehicle location with the help of the *Global Positioning System* (GPS). These features may also be used by rescue services, transport companies, and the forestry industry.

 The analog networks in use are different from country to country and even within one country, many incompatible systems may be in use. New

digital systems are evolving that are aimed at supporting a wider area service. One of them is the terrestrial trunked radio system, which has the goal of providing a compatible service in all European countries.

5.4.2.3 Terrestrial Trunked Radio

A modern digital standard for a pan-European PMR system has been developed by ETSI and it is known as *Terrestrial Trunked Radio* (TETRA). This system was originally called Trans-European Trunked Radio and it is different from GSM but based on GSM's experiences. It uses different frequency bands and provides some services that are not available in GSM, for example, mobile-to-mobile communication. The TETRA networks are built for public safety organizations such as police, fire brigades, and border guards. These systems use the 380- to 400-MHz frequency band. Later the 410-, 450-, and 870-MHz frequency bands will be put into use by the commercial TETRA service for taxi, transport, railway, and other organizations.

Like all trunked systems, TETRA uses a cellular network structure and channel allocation on demand to improve spectral efficiency. It is a digital system, uses an efficient speech-coding method, and tolerates high interference, which further improves spectral efficiency.

Why do we need a separate network when the public cellular networks provide a service that can define a closed user group for an organization? One reason is because the operation of emergency services is so essential for a community that a separate network is required. The main reasons behind this are as follows:

- Availability of capacity is independent of the activity of ordinary subscribers to the public cellular networks. In an emergency situation public cellular networks may become overloaded.
- The structure and services of the network can be modified independently from the public service according to the users' needs.
- Some required features are not supported by the public cellular networks, for example, direct mobile-to-mobile communications and end-to-end encryption.

General features of TETRA systems are listed here:

- Efficient use of spectrum, cellular structure, trunked (shared) radio resources;
- Efficient use of investments; BSs and exchanges shared between several organizations;

- National or even international coverage;
- Standardized multivendor equipment;
- Support of the VPN for each user organization of the network, each of which can modify their resources, such as the usage of channels (mobile-to-mobile, mobile-to-base station) and priorities;
- Each user organization has its own "dispatcher station" from which an operator can communicate with all terminals;
- A number of channels that can be permanently or temporarily allocated for a certain organization (quarantined share of recourses);
- Open channel (mobile-to-mobile and point-to-multipoint) communication supported;
- Prioritization of organizations and user groups;

The standardization of TETRA took place after GSM was up and running, and the requirements were slightly different from those for public cellular networks. The standardization work was carried out by ETSI, which had also specified the GSM. For these reasons TETRA technology is closely related to that of the GSM but differs in details.

Some key technical specifications of the TETRA system are as follows:

- TDMA/FDMA channel access method;
- A 25-kHz carrier spacing (FDMA);
- Four user channels per carrier (TDMA);
- *Frequency-division duplex* (FDD) principle with 10-MHz duplex distance;
- A 28-Kbps maximum user bit rate (all four time slots of one carrier used by a single user), packet- or circuit-switched;
- Speech coding at the 4.8-Kbps data rate.

5.4.3   Radio Paging

Paging systems are simplex systems and they transmit short texts or simply generate an audible beep. Pagers are small and inexpensive wireless communication devices that are used by subscribers to receive messages without disturbing their current activities. There are two basic types of radio paging networks, on-site pagers and wide-area pagers. On-site pagers cover a local

area like a building or a hospital. Wide-area pagers may cover a whole country.

New paging technologies are available such as the *European radio messaging system* (ERMES). However, in many countries use of paging systems has decreased because many cellular systems provide a similar or even better bidirectional messaging service.

5.4.4   Analog Cellular Systems

In Section 5.3 we introduced the operation of cellular networks in general. The first cellular technologies were analog and they became available in the first half of the 1980s. These systems are often referred to as first generation cellular systems and these are the most important analog cellular systems:

- *Advanced Mobile Phone System* (AMPS) in the United States;
- *Nordic Mobile Telephone* (NMT) used in Nordic countries;
- *Total Access Communications System* (TACS) in the United Kingdom.

These systems are quite similar but incompatible. They use a frequency band in the range of 800 to 900 MHz (NMT uses 450 MHz as well) and frequency modulation. The frequency band is divided into channels and one of these is allocated for each call. We call this radio access principle frequency-division multiple access.

5.4.5   Digital Second Generation Cellular Systems

In this section we review the most important digital cellular networks that came into use in the first half of the 1990s. We often refer to these systems as second generation cellular systems.

5.4.5.1   GSM

GSM operates at the 900-MHz frequency band and it became the most widely used second generation cellular technology. The structure and operation of the GSM network are explained in Section 5.5. In GSM the subscription information is stored on a smart card and a subscriber can change his or her mobile telephone any time. When he or she inserts his or her card into the new telephone, he or she has access to exactly the same service as previously. The access method used in GSM is TDMA, in which each frequency channel is divided into time slots for multiple users.

5.4.5.2   Digital Cellular System at 1,800 MHz

Digital cellular system at 1,800 MHz (DCS-1800) is also known as GSM-1800. It is based on GSM technology but operates in the 1,800-MHz frequency band and provides much higher capacity than GSM in terms of the number of users. DCS-1800 is a technology for the European implementation of *personal communications network* (PCN), but it is in use in other parts of the world as well. The goal of PCN is to provide a mass mobile telecommunications service in urban areas.

5.4.5.3   Personal Communications Network and Service

Note here that the term *personal communications* refers to cellular mobile communications in which a call is routed to a person who carries a terminal instead of a fixed terminal location as in the conventional fixed telephone network. The PCN and *personal communications service* (PCS) simply refer to microcellular systems that emphasize low-cost and high-capacity cellular service and a hand-portable terminal with a long battery life. In Europe the DCS-1800 system is also called PCN because it is the implementation technology for PCN.

In the United States several digital technologies are used to implement the high-capacity cellular service that is known as PCS. These technologies are GSM-1900 (GSM at 1,900 MHz), NADC (known also as D-AMPS or US-TDMA), and CDMA. All of these network technologies are briefly introduced next. Note that all systems at higher frequency bands (1,800–1,900 MHz) are referred to as personal communications systems and systems below 1 GHz are referred to as cellular systems.

5.4.5.4   PCS-1900

As just mentioned, many technologies are specified for implementation of PCS in the United States and one of them is GSM-1900. GSM-1900 is based on GSM/DCS1800 technology but adapted to the frequency allocation of North America. These three GSM-based systems are so similar that with the help of a multimode mobile station a subscriber can use all of these networks with the same terminal and subscription (same subscriber card). We illustrate the structure and operation of this system in Section 5.5 as an example of a modern digital cellular system.

5.4.5.5   North American Digital Cellular

Both the United States and Canada have implemented digital techniques to increase the capacity and quality of the existing AMPS system. The *North American digital cellular* (NADC) system implements digital radio

communication in the frequency band of AMPS. It divides the channels of the analog AMPS into six time slots (TDMA). With the help of time division, three or six (half-rate speech mode) users share an analog 30-kHz AMPS channel. The terminals with dual-mode capability use a digital system when it is available; otherwise, the analog AMPS service is used [1]. Because of this principle the NADC system is also known as *dual-mode AMPS* (D-AMPS).

The NADC network system is able to provide service for even the oldest analog AMPS terminal. The common control channels of analog AMPS and NADC are compatible and a mobile station, analog or dual mode, first searches the *forward (downlink) control channel* (FOCC) that occupies one FDMA channel. Then the terminal informs the network with a signaling message that contains the information about its capabilities. There are three enhanced modes of operation in addition to the original analog AMPS: *narrowband AMPS* (NAMPS) (an analog enhancement designed by Motorola, which increases the capacity of the system), CDMA, and NADC.

Upon recognizing the mobile's enhanced digital capabilities, the network will assign a *digital traffic channel* (DTC) to the mobile for a call. If a DTC is not available, then an analog channel is assigned instead. When a channel is assigned a channel number (frequency), a time slot number, a timing advance value, and a mobile power setting are given to the MS. A timing advance value is needed in all TDMA systems and it defines the transmission time of the MS. A distant MS has to transmit earlier than an MS close to the BS, otherwise subsequent transmissions from mobile stations would overlap in the BS's receiver. The channel is maintained until disconnect time with the help of the continuous quality measurement of the communication and handoffs or handovers if required.

5.4.5.6 CDMA

CDMA was selected in the early 1990s to become the main digital cellular standard in the United States. The main difference between CDMA and other technologies discussed previously is that on the radio path it does not use either FDMA or TDMA. Instead, the mobiles use the wide frequency band all of the time with the help of a unique code for each user. This unique code is used to spread the signal over a wide frequency band and to detect the wanted signal at the receiving end. This American system is also referred to as *narrowband CDMA* (N-CDMA) or the *Interim Standard-95* (IS-95) system [1].

Operating Principle of CDMA

The operating principle of CDMA radio transmission is not as easy to understand as FDMA or TDMA. Figure 5.7 shows a very simplified diagram of a
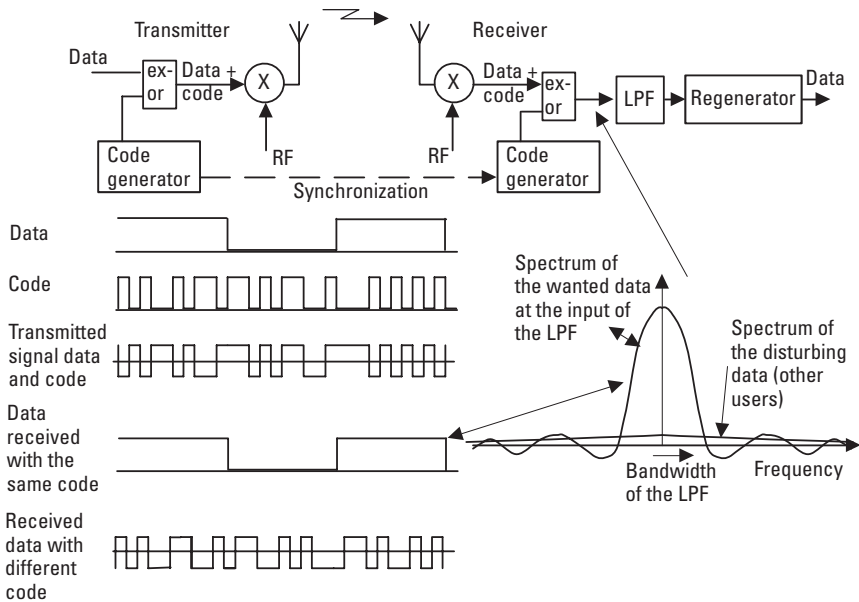
Figure 5.7 Operating principle of CDMA.

CDMA system. In Figure 5.7 the spreading code data rate is 10 times higher than the information data. In the actual IS-95 system the code has a more than 100 times higher data rate than the user data. The exclusive-or operation is performed in the transmitter with the user data and the spreading code. The exclusive-or operation gives a high state when the data and code have different states and a low state when they are equal. In our simplified example, the bits or chips of the code and data+code are then 10 times shorter than the bits of the original data. We saw in Chapter 4 that the shorter the pulses, the wider the spectrum they have. Thus the spectrum of each bit is now 10 times wider and the spectrum of the original data is spread over a 10 times wider frequency band. After modulation with the RF carrier, a wide frequency band is occupied with this CDMA radio signal.

In the receiver the received signal is first demodulated and then the same code is used to detect the wanted signal. The same exclusive-or operation is performed in the receiver and original 10 times longer data pulses are reproduced. The resulting data at the input of the lowpass filter in the receiver is the original *low-rate data*. We may imagine that the receiver, using the right code, has collected the signal energy from the wide frequency band to the baseband.

The other signals (of other users) on the channel were generated with different codes and they are received as a random high-rate signal with a wide spectrum, as shown in Figure 5.7. Most of these disturbing signals are filtered out in the *lowpass filter* (LPF) of the receiver, whereas most of the desired low-rate data gets through the LPF. At the output of the LPF other signals are seen as noise on top of the desired data. The regenerator detects the original data and this detection is error free if noise is not too high.

For proper operation the receiver has to be accurately synchronized with the transmitter and the simultaneously used codes have to be selected to minimize interference. The CDMA also requires accurate and frequent adjustment of the transmission power levels because the power of the users influences the S/N and error rate of the other simultaneous users. The CDMA principle provides many advantages compared with FDMA or TDMA systems. It utilizes radio resources more efficiently and it is not sensitive to multipath fading and narrowband radio disturbances. It transmits continuously with low transmission power so the safety risk for the users of handheld phones is reduced.

IS-95 CDMA System

The CDMA system supports dual-mode operation just like one of the other American systems, NADC. The CDMA resources exist in the same frequency band with the traditional AMPS system and it occupies 41 AMPS channels (1.23 MHz). The CDMA users use their unique codes to share this frequency band. The common control channels are also spread over the CDMA band by their own spreading codes, which are known by the MSs. When a MS is in the idle mode it uses the code of the FOCC to listen to the network and to be able, for example, to receive a paging message in the case of an incoming call. When a call is connected, a new code is allotted to the user for dedicated speech communication.

CDMA is an interesting technology and it provides many other features that we have not discussed, such as soft handover or handoff. To perform those tasks, a MS might use more than one BS (operating at the same carrier frequency) at the same time with the different codes. Multiple BSs receive a signal from the MS simultaneously and the MS combines signals received from different BSs. The MS does not need to switch over from one BS to another at a certain time instant as is done with the hard handover in FDMA/TDMA networks.

We restrict our discussion about CDMA and other cellular systems to brief introductions of the most important networks. For further information about CDMA, the reader should refer to, for example, [1].

5.4.5.7   Japanese Digital Cellular (JDC)

JDC system is also known as *personal digital cellular* (PDC). It is a separate system from the previous analog one but it utilizes dual-mode terminals that are able to use existing analog systems as well. The network technology is close to that of the European GSM.

5.4.6   Third Generation Cellular Systems

The main forces behind development of the third generation systems (3G) have been driven by the second generation systems' low performance data services, incompatible service in different parts of the world, and lack of capacity. In the 1990s, the ITU started a project to develop a future global 3G system, which is known today as *International Mobile Communications* (IMT)-2000.

5.4.6.1   IMT-2000

The IMT-2000 system was designed to be a global system for third generation mobile communications. It was developed by the ITU, which called it previously future public land mobile telecommunications system. Many problems have prevented the achievement of mutual understanding among countries regarding this system. Among the problems are frequency allocation in different continents, existing different second generation infrastructures, and different political interests. As a consequence a common understanding about detailed implementation technology was not achieved and IMT-2000 will not be a globally compatible technology; it will instead act as an umbrella for compatible services provided by different underlying technologies.

Even though third generation systems will use different technologies, but the development of mobile terminal technology will partly solve the incompatibility problem for users. With the same terminal we will be able use different networks and the services they provide. The most important network technology for 3G is UMTS.

5.4.6.2   UMTS

UMTS is a European concept for integrated mobile services and it is based on the GSM and GPRS. Its goal is to provide a wide range of mobile services wherever the user is located. For UMTS cordless (TDD), cellular and satellite interfaces are defined. It will provide multimedia service with data rates up to 2 Mbps for steady MSs and up to 384 Kbps for moving MSs.

The cellular radio access method for UMTS approved ETSI is *wideband CDMA* (WCDMA). The basic operating principle is the same as in CDMA, which was introduced previously. The new frequency band at the 2-GHz range is allocated for UMTS. The channel bandwidth is 5 MHz, and each channel is used by all cells.

The core network of UMTS is based on the core network of GSM and GPRS. The UMTS BSs can be added to the GSM/GPRS network to operate in parallel with GSM base stations. Even handovers between UMTS and GSM/GPRS are supported.

5.4.6.3   CDMA2000

The main 3G technology for the United States is based on second generation IS-95 CDMA. CDMA2000 is specified to use a sophisticated modulation scheme to increase the data rate over an ordinary 1.25-MHz CDMA channel. The problem with 3G systems in the United States is that a much smaller frequency band is available for 3G service than in areas following European frequency division.

5.4.7   Mobile Satellite Systems

One application of satellite communications is for point-to-point transmission, as presented in Section 4.7. Satellites also provide mobile communications services to ships and aircraft, and they are used in desert areas where other communications services are not available. In the systems that use geostationary satellites, mobile stations are expensive and the cost of service is quite high. Plans were made to implement lower cost satellite services that could be used with handy MSs. The multimode MSs could use satellite or, if available, lower cost PLMN, such as GSM or CDMA. Examples of these systems are Iridium and Globalstar.

These systems use a number of satellites that are in orbit at a 700- to 10,000-km distance from the Earth instead of in a geostationary orbit, which is at a distance o 36,000 km. The satellites are circling the Earth in such a way that some of them are visible all the time from any point on the Earth's surface (Figure 5.8). Each of the satellites performs base station functions and takes care of the large cell below it.

These systems need and use functions similar to those of cellular networks. Examples are the mobility management and handover functions, which are used to manage the movement of satellites (BSs) instead of subscribers. Earth stations of a satellite system control the operation of satellites and behave as connection points to the public land networks.

Mobile satellite systems use many
low or medium orbit satellites that
move around the Earth

Multimode terminals use satellite service
if land mobile network is not available

700...
10,000 km

Earth
station

PSTN

A satellite that
operates as a base
station of a public
land mobile
network, PLMN.
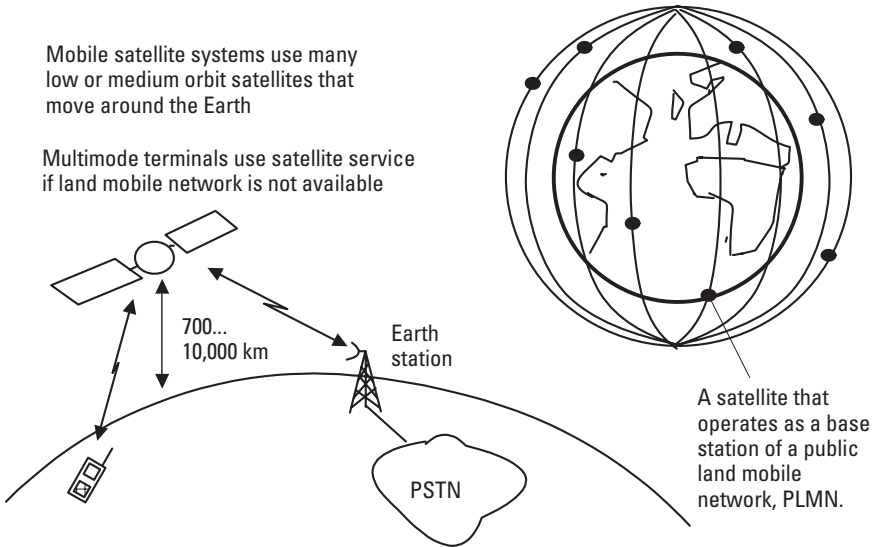
Figure 5.8 A mobile satellite system.

Most satellite projects have been financial catastrophes. Their business plans were done at a time when international mobile communication services were not available and when the expansion of digital cellular systems started. By the time satellite services became available, most business travelers already carried their own digital mobile telephones and the market for satellite service was much reduced.

5.4.8 WLANs

Many working environments would benefit from having available short-haul high-data-rate wireless data transmission. Examples include hospitals, factory floors, stores, and conference and exhibition centers. An approach similar to that of a wired private LAN is needed and that approach is a WLAN. A major step in the development of WLAN technology was Standard IEEE 802.11b, which was approved in 1999. Earlier standards had many implementation options and compatibility between different products was not good enough to make them popular.

Standard IEEE 802.11b uses a 2.4-GHz license free frequency band and its maximum data rate over the air interface is 11 Mbps. To be compatible with earlier 1- and 2-Mbps IEEE 802.11 standards, Standard IEEE 802.11b sends all frame header information at 1 Mbps, which reduces the

user data rate. Acknowledgments and the channel reservation mechanism handle a share of the air-interface capacity, and the actual higher protocol data rate is of the order of 6 Mbps, which is shared by all users and between the two transmission directions.

Standard IEEE 802.11b uses four different modulation schemes, one for each of four data rates: 1, 2, 5.5, and 11 Mbps. If the quality of the radio channel becomes worse, a more noise-tolerant modulation scheme is accessed and the data rate is reduced.

The base stations of WLAN systems are called *access points* (APs) and they are connected to wire-line Ethernet. WLANs are actually designed to operate as wireless extensions to wire-line backbone Ethernet.

The bandwidth of the IEEE 802.11b radio signal is 11 MHz, and there is only enough space for three nonoverlapping channels at the 2.4-GHz band. This places a severe limit on the data capacity when the number of users increases. Higher capacity WLAN technologies, such as IEEE 802.11a operating at the 5-GHz frequency band, have been developed to solve this problem.

WLAN networks are available in airports, hotels, and conference centers to provide Internet access to customers. WLAN technology is also becoming more popular in the educational and office environments. WLAN technologies may be a solution for high-data-rate short-haul data services when integrated with third generation systems.

5.4.9 Bluetooth

Bluetooth technology allows for the replacement of proprietary cables that connect one digital device to another with a universal short-haul radio link. Mobile computers, cellular handsets, printers, keyboards, and many other devices can be embedded with Bluetooth radios. Bluetooth was developed by the Bluetooth Special Interest Group (SIG, http://www.bluetooth.com), founded by Ericsson, IBM, Intel, Nokia, and Toshiba.

A small wireless Bluetooth network connecting, for example, a user's computer to its peripherals is called a *personal area network* (PAN). PAN contains one or more piconets. One Bluetooth piconet contains a single master and up to seven active slaves. The master polls slaves and orders each of them to transmit in turn. For voice applications Bluetooth specifies a synchronous channel that transmits at a bidirectional 64-Kbps constant bit rate between a master and a slave. This can be used to implement cordless telephones or hands-free sets for a cellular telephone.

Bluetooth systems use the same 2.4-GHz license free frequency band as WLANs and they can coexist in the same area. The wideband WLAN signals and narrowband Bluetooth signals do not interfere much.

Bluetooth uses *frequency hopping spread-spectrum* (FHSS) technology, in which data are transmitted in bursts and the carrier frequency is changed after each burst. There are 79 carrier frequencies with 1-MHz spacing over which the transmission frequency hops. Each piconet uses a different pseudorandom hopping sequence over the 79 carriers. Several piconets can operate in the same area simultaneously because their signals interfere only at a time when they happen to occupy the same frequency channel.

The modulation rate of Bluetooth is 1 Mbps, which all devices and both transmission directions in the piconet share. If we compare WLAN and Bluetooth technologies we see that WLAN is a system for a work group (LAN) and Bluetooth is for only a single user (PAN). The number of devices in the Bluetooth network is very limited and data rate available for each device is quite low.

5.5   GSM

As an example of a digital cellular network, we introduce the structure and operation of the GSM network. The European digital cellular system GSM was developed by CEPT during the 1980s, and this work was continued by ETSI. The acronym GSM came originally from the standardization working team, but GSM is presently understood to mean Global System for Mobile Communications. Two other cellular networks are based on GSM technology: the European DCS-1800, which operates in the 1.8-GHz band, and the American GSM-1900, which operates in the 1.9-GHz band. Our discussion in this section is valid for all of these networks.

In GSM, unlike in analog mobile networks, subscription and mobile equipment are separated. Subscriber data are stored and handled by a *subscriber identity module* (SIM), which is a smart card belonging to a subscriber. With this card the subscriber can use any mobile telephone equipment just if it were his or her own. The radio equipment is called *mobile equipment* (ME) and we can say that the mobile station consists of two parts, ME and SIM; that is: MS = SIM + ME.

5.5.1   Structure of the GSM Network

A simplified architecture for the GSM network is presented in Figure 5.9. For a more detailed look at the structure and functionality of the GSM network, the reader should refer to [1, 2].
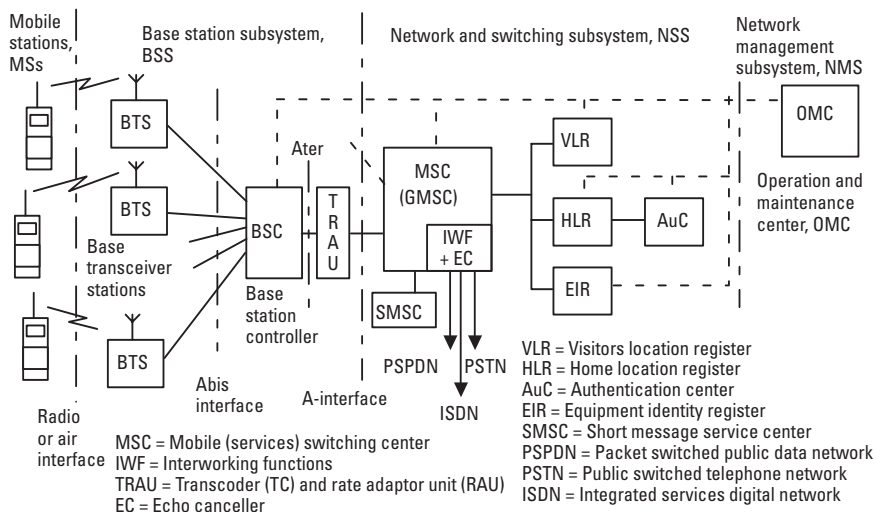
Figure 5.9 Structure of the GSM network.

5.5.1.1 Radio Network

MSs are connected to the mobile switching center (MSC), via a *base station subsystem* (BSS). The BSS consists of a *base station controller* (BSC) and many *base transceiver stations* (BTSs) that are controlled by one BSC. The roles of the network elements are introduced in the following sections.

5.5.1.2 MSC

Like any local exchange, the MSC establishes calls by switching the incoming channels into outgoing channels. It also controls the communications, releases connections, and collects charging information.

As a mobile switching system, the MSC together with the VLR performs additional functions such as location registration and paging. It also transfers encryption parameters, participates in the handover procedure when required, and supports *short message service* (SMS). The SMS is a service integrated into GSM that enables users to transmit and receive short text messages.

In each cellular network there is at least one *gateway MSC* (GMSC) that provides connections to other networks. The MSC in Figure 5.9 performs gateway functions in addition to other MSC functions. The GMSC works as an interface between the cellular network and the fixed networks and it must handle the signaling protocols between the fixed networks and

network elements of PLMN. The GMSC also controls echo cancellers, which are needed between the fixed and cellular network because of long speech-coding delays.

5.5.1.3   HLR

All subscriber parameters for each mobile user are permanently stored in one HLR. The HLR provides a well-known and fixed location for variable routing information. The main functions of the HLR are as follows:

- Storage of the subscriber data, for example, services available for this subscriber;
- Location registration and call handling, central store for subscriber location data;
- Support for encryption and authentication;
- Handling of supplementary services (e.g., barring or call transfer);
- Support for the short message service.

The HLR is implemented by an efficient real-time database system that may store the subscriber data of 1 million subscribers.

5.5.1.4   VLR

The VLR provides local storage for all of the variables and functions needed to handle calls to and from the mobile subscribers in the area related to that VLR. The information is stored in the VLR as long as the mobile station stays in that area. The VLR communicates with the HLR to inform it about the location of a subscriber and to obtain subscriber data that includes information about, for example, what services should be provided to this specific subscriber. The main functions of the VLR are as follows:

- Storage of data for subscribers located in its area;
- Management and allocation of the local identity codes to avoid frequent use of a global identity on the radio path for security reasons;
- Location registration and call handling;
- Authentication;
- Support of encryption;
- Support for handover;

- Handling of supplementary services;
- Support for SMS.

The VLR is a database system that is usually integrated in each mobile exchange MSC.

5.5.1.5   Authentication Center (AuC)

The security data of a subscriber are stored in the AuC that contains a subscriber-specific security key, encryption algorithms, and a random generator. The AuC produces subscriber-specific security data with defined algorithms and gives it to the HLR, which distributes them to the VLR. A PLMN may contain one or more AuCs, and they can be separate network elements or integrated to the HLR. The same subscriber-specific key and algorithms are also stored in SIM. There is no need to send them over the network and on the radio path.

5.5.1.6   Equipment Identity Register (EIR)

The EIR is a database that contains information about mobile terminal equipment. There is a white list for the terminals that are allowed to use the service, a gray list for terminals that need to be held under surveillance, and a black list for stolen mobile terminals. Those terminals whose serial numbers are found on the black list are not allowed to use the network.

5.5.1.7   Interworking Functions

The *interworking function* (IWF) is a functional entity associated with the gateway MSC. It enables interworking between a PLMN and a fixed network, for example, an ISDN, a PSTN, and a public switched data network. It is needed, for example, in the case of data transmission from GSM to PSTN. It converts digital transmissions used inside the GSM network to modem signals for PSTN. It has no functionality with the service that is directly compatible with that of the fixed network.

5.5.1.8   Transcoder and Rate Adapter Unit

A transcoder (TC) is needed to make conversions between GSM voice coding (13 or 7 Kbps) and PCM coding (64 Kbps), which is used in the fixed network. In the case of data transmission, transcoding is disabled. For data, a rate adapter unit (RAU) is needed to adapt SM data service to service provided by the external network. For example, if the GSM user has 14.4-Kbps data access to ISDN, RAU inserts its data into the 64-Kbps data stream of an

ISDN B-channel in a specified way so that the other end knows where the user data can be found. The functions of the TC and RAU are often combined into a single piece of equipment called a *transcoder and rate adapter unit* (TRAU).

5.5.1.9   Echo Canceler (EC)

The EC is needed at the interface between a GSM network and the PSTN. The efficient speech coding of GSM introduces such a long delay that echoes reflected by a hybrid circuit in the subscriber interface of the fixed network (see Chapter 2) of the fixed service would be disturbing. The echo canceler eliminates this echo.

5.5.1.10   Short Message Service Center (SMSC)

GSM provides a paging service that is called short message service. The point-to-point SMS provides a mean of sending messages of a limited size to and from MSs. An SMSC acts as a store-and-forward center for these short messages. A short message transmitted by a subscriber is first forwarded through the network to the SMSC of his or her home network operator. The SMSC stores it, extracts the destination telephone number from the message, and forwards the message to its destination. The service center is not standardized as a part of a PLMN, but the GSM network has to support the transfer of short messages between SMSCs and the MSs.

5.5.1.11   Operation and Maintenance Center (OMC)

The OMC is a network management system for the remote O&M of a GSM network. The alarms of GSM network elements and traffic measurement reports are collected there. The O&M system handles features related to system security, faults, and network configuration updates.

5.5.1.12   Interfaces Inside GSM Network

The interface between the MSC and BSC is called the *A-interface* as shown in Figure 5.9. It is standardized and BSSs and MSCs from different vendors at the opposite side of the interface are compatible. Speech is PCM coded (see Chapter 3) at this interface. Another important interface is the *Abis-interface* between the BTS and BSC. At this interface speech is GSM coded, which requires less transmission capacity than the PCM coding. The Abis-interface is not completely standardized and, as a consequence, both BTSs and BSCs have to be purchased from the same manufacturer. The Ater-interface is not standardized either but it is used for terrestrial connections between the BSC and MSC. Speech is GSM coded at the Ater-interface and the transmission

capacity needed at the Ater-interface is one-fourth of the capacity of the A-interface.

5.5.2 Physical Channels

The multiple-access scheme used in GSM utilizes two access methods, FDMA and TDMA. Up to eight users may share one of the 200-kHz frequency channels, which is divided into eight time slots.

5.5.2.1 FDMA and TDMA

A basic concept of GSM transmission on a radio path is that the unit of transmission is a series of about 100 modulated bits. This is called a *burst* and it is sent in time and frequency windows called a *slot* as shown in Figure 5.10. The central frequencies of the slots are positioned every 200 kHz (FDMA) within the system frequency band and they occur every 0.577 ms (TDMA). All time slots of different frequencies in a given cell are controlled by the synchronization broadcast from the BTS transmitted in the common control channel of that cell.

5.2.2.2 Separation of Transmission Directions in Time and in Frequency

For bidirectional user channels, the two directions are related by the fixed separation of frequencies and time instant. The fixed frequency gap between transmission directions is called the duplex distance and it is 45 MHz (in the

A transmission burst occupies a window in time and frequency called a slot.
There are eight time slots on each carrier frequency.
Eight simultaneous calls may use the same frequency.
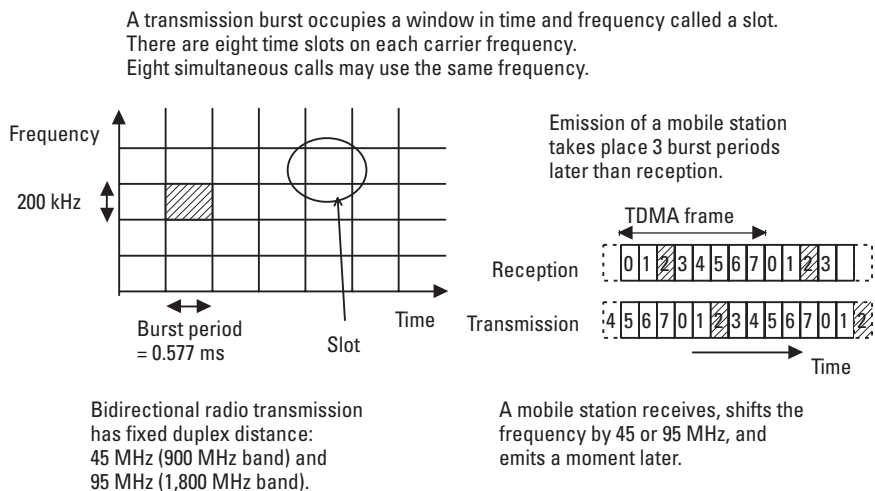
Figure 5.10 Multiple-access scheme of GSM.

900-MHz band) and of 75 MHz (in the 1,800-MHz band) and this duplex principle is called *frequency-division duplex* (FDD). The separation in time is three time slots, as shown in Figure 5.10. This principle makes the implementation of mobile equipment efficient because there is no need to transmit and receive simultaneously. Two bursts after reception on the downlink or forward frequency, the mobile equipment sends on the uplink or reverse frequency as shown in Figure 5.10.

One time slot in each eight-slot TDMA frame represents one physical channel. Each call typically occupies one of the eight physical channels at one carrier frequency.

5.5.3   Logical Channels

The physical channels at the GSM radio interface are divided into logical channels. They fall into two main categories, dedicated channels and common control channels. There are many different logical channels and the distinction between them is based on the purpose and the information transmitted via a channel. These logical channels are mapped onto one physical channel defined as one slot (usually TS0) in each TDMA frame and transmitted as regular radio bursts.

5.5.3.1   Traffic Channel and Associated Slow-Rate Control Channel

When the call is connected, two channels on the radio path are dedicated to it: the *traffic channel* (TCH) and the *slow associated control channel* (SACCH). The SACCH is used, for example, to transmit power control information to the MS and measurement results from the mobile stations to the network. These two channels belong to the dedicated channels because they are allocated for one user.

5.5.3.2   Common Control Channels

There are several logical common channels in each cell. These altogether typically occupy one fixed time slot (typically TS0) at a fixed frequency. The common control channel in the downlink direction transmits, for example, the following information from the network to the MSs:

- Synchronization information of frequency and time slots;
- Information about common channels that is used by neighboring cells;
- Location area and network identification;
- Paging messages for incoming calls and channel assignment for a new call.

In the uplink direction, from the MS to the network, the common control channel is used, for example, for call-request messages from the MSs.

5.6 Operation of the GSM Network

In this section we introduce the operating principles of a cellular network. To do this, we illustrate the GSM network with a few simplified examples. They show how location update is performed, how a mobile call is established, how handover is performed, and what the security functions of the GSM network are.

Each GSM subscriber is registered into one HLR of his or her home network. This HLR is the central point that provides subscriber information regardless of where he or she is presently located.

5.6.1 Location Update

The cellular mobile network has to be aware of the location of its subscribers at all times to be able to route incoming calls to them. The location update procedure takes place every time a MS moves to another location area or when a user switches her telephone on in a different location than where she was located previously.

The geographical position of a GSM mobile is known at the accuracy of a *location area* (LA), which typically consists of a number of cells. The BTSs of those cells need not be connected to the same BSC. When an incoming call to a mobile subscriber arrives, it is paged through all the cells belonging to the LA where this specific subscriber is known to be.

The MS is responsible for location updates and performs this updating in idle mode, that is, when a call is not connected.

The MS surveys the radio environment constantly and, when it detects that it could be served best in a new LA, performs a normal location update procedure to change the location information in its present VLR and in the HLR (if needed). We say that the mobile station has roamed to another LA. In dedicated mode, during a call, the procedure called handover, which we will discuss later, may be required. If the LA is changed during a call, the location update takes place after the call is cleared.

Location update may take place inside one network when the LA is changed or between different networks that may be located in different countries. The latter case requires a roaming agreement between network operators to allow a subscriber to use the other network in addition to her home network. Figure 5.11 illustrates the location update procedure that
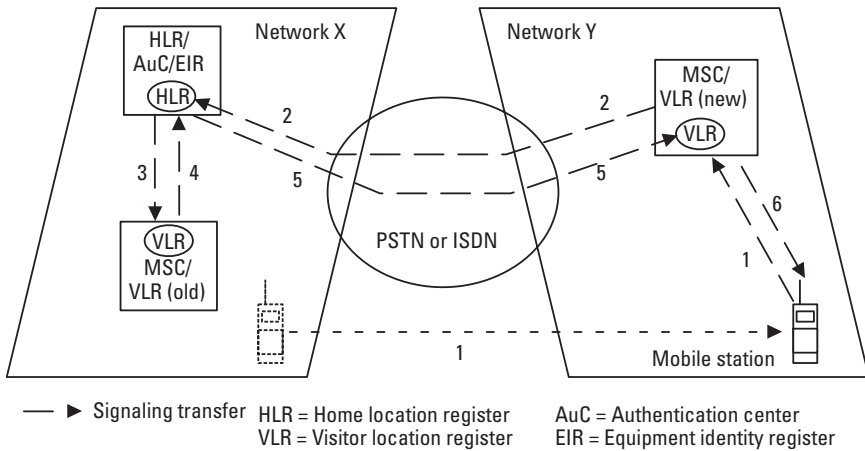
Figure 5.11 Location update in GSM network.

occurs when a mobile station is switched on in another network Y in another country. This example assumes that the mobile station has been switched off in the home network, network X, and that the network operators of networks Y and X have a roaming agreement that allows cellular subscribers to use the services of another network.

For location update the following main operations are carried out (see Figure 5.11):

1. When the MS has roamed to another LA, it scans the common control channels. When it finds a common control channel, it detects the LA code, which contains country and network identifications. If the MS cannot find the same LA code it has stored previously, it requests a location update from the network.

2. The MSC/VLR requests the global identity code of the mobile [*international mobile identity subscriber* (IMSI) stored into SIM, not the same as the telephone number]. With the help of this, the MSC/VLR knows in which country the home network of this mobile is found. The MSC/VLR sends a signaling message via the international CCS7 signaling network toward the home country of this cellular subscriber. The message includes country code, network code, and subscriber identity. The message also includes the address of this new VLR to inform the HLR about the new location of the MS.

3. When the HLR receives the message it requests the former "old" VLR, where this subscriber was previously located, to remove information about the subscriber.

4. The VLR (old) acknowledges and removes the subscriber information from its database.

5. The HLR updates the location information and sends the subscriber information, including security codes, to the new VLR.

6. The (new) MSC/VLR stores the subscriber information, performs authentication of the MS, and acknowledges location update. The MS will now show the name of network Y on its screen.

5.6.2 Mobile Call

Figure 5.12 illustrates how the GSM network routes a call to a subscriber who has roamed to another network. We assume here that both the calling and called subscribers are originally registered in the same home network, network X. Called subscriber B has traveled to another network Y and switched on her MS. Then the location update, which was illustrated in the previous section, takes place. Then mobile user A calls MS B from the home network.

We can identify the following main phases when the call is established from MS A in the home network to MS B located in another network (Figure 5.12).
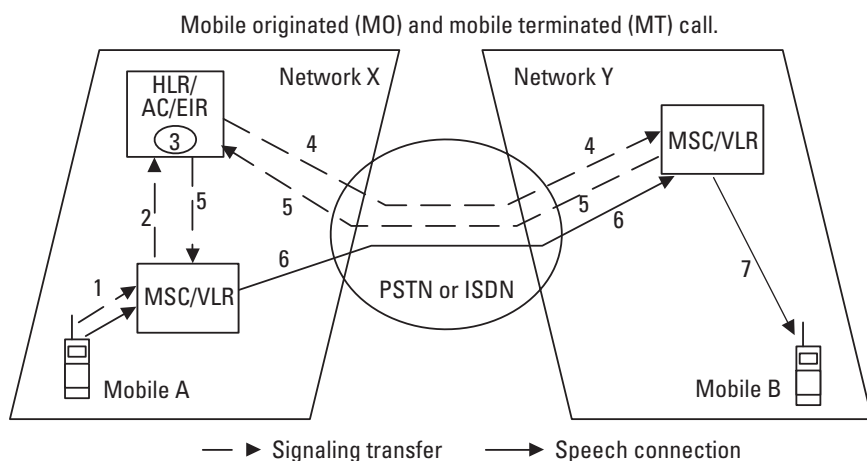
Figure 5.12 Mobile call in a cellular network.

1. MS A initiates a call to MS B, which is currently located in another network. The call connection request and other signaling information are transmitted via the radio path and BSS to the MSC. The telephone number of subscriber B is transmitted to the MSC/VLR.

2. The MSC recognizes mobile B (in this example) as a subscriber of its own network and requests the roaming number from the HLR of subscriber B. The roaming number is a temporary telephone number that is used for call establishment via a PSTN.

3. The HLR of subscriber B knows the identification of the "visited" VLR where mobile B is currently located. When mobile B was switched on, the MSC/VLR of network Y sent its address to the HLR (location update). The HLR builds up a signaling message that includes the identification of called subscriber B together with the address of the visited MSC/VLR.

4. The HLR requests the visited VLR to provide a roaming number.

5. The MSC/VLR of network Y has a pool of roaming numbers that look like the ordinary telephone numbers of that country. The visited MSC/VLR then allocates one roaming number to subscriber B, stores it in its database, and sends it to the HLR, which then forwards it to the MSC/VLR in network X.

6. The MSC/VLR of network X routes the call toward the MSC/VLR in network Y using the roaming number for dialing digits and the call is then routed the same way as any other telephone call.

7. When the MSC/VLR in network Y receives the call identified by the previously allocated roaming number, it associates this with subscriber B and initiates paging toward MS B. The roaming number is then released for reuse.

To keep Figure 5.12 simple, the GMSCs at the border of networks X and Y are not shown as separate network elements. There is always at least one GMSC in each individual GSM network. The GMSC is a signaling interface point to other networks and it is able, for example, to route signaling messages toward the right HLR inside its own network.

The telephone call to a roamed subscriber is currently always connected via the GMSC of the home network, and the roamed subscriber pays for the connection from the home network to his or her present location. Later it may become possible to connect calls directly.

5.6.3 Handover or Handoff

The main reason to perform handover is to maintain call connection regardless of the movement of the MS over cell boundaries. The structure of a GSM network requires the possibility to execute handovers at four levels, as shown in Figure 5.13.

The BSC is responsible for handover because it occurs most often between two cells under one BSC. The handover process should be as quick as possible so that communication is not disturbed. To perform handover quickly, the BSC collects measurement data from MSs and BTSs, processes it, and updates ordered candidate cell lists for handover for all the MSs that have an ongoing call.

Handover is most often necessary between BTSs of neighboring cells, case (b) in Figure 5.13, which are controlled by the same BSC. The BSC controls the handover and performs the channel switch from an "old" cell to a "new" cell. Sometimes there may be a need to switch communication from one channel to another in the same BTS [case (a) in Figure 5.13]. This may be necessary because of temporary interference. Also this handover is controlled and performed by the BSC.

The inter-BSC handover, case (c) in Figure 5.13, occurs if an MS moves to a neighboring cell that is controlled by a different BSC. Now the BSC cannot perform switching. Instead, it has to request the MSC to execute the handover switching to the target cell. When a new connection is
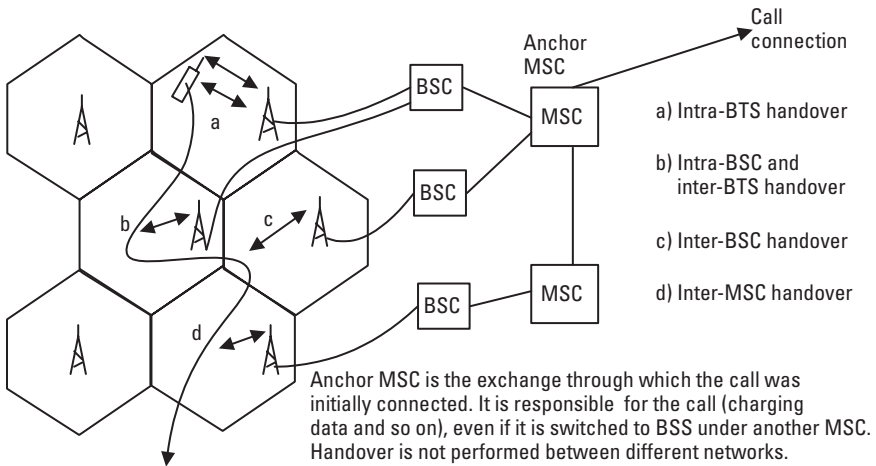
Figure 5.13 The four different cases of handover.

established from the MSC to the new BTS and the MS accesses the new channel, the MSC performs the switching.

When the neighbor cell is located under a different MSC, an inter-MSC handover may be required. Then the BSC requests the anchor MSC to establish the connection with the help of a new MSC to the new cell. The anchor MSC performs the switching. The anchor MSC is the exchange via which the call was originally connected. The anchor MSC controls the call until it is cleared even though it may use other MSCs to maintain the call through handovers.

Handover can also be executed to arrange traffic between cells to avoid unsuccessful calls due to geographically uneven load. In this case the calls of the MSs that are located close to the border of a loaded cell are switched to a neighboring cell that has more free capacity.

As an example we now look at how the most complicated handover, inter-MSC handover, is performed. The handover procedure between two MSCs, shown in Figure 5.14, includes the following main phases:

1. The mobile station moves across the cell border and the BSC (old) decides to initiate handover to another cell (new). The decision is based on the measurement information sent by the mobile and by
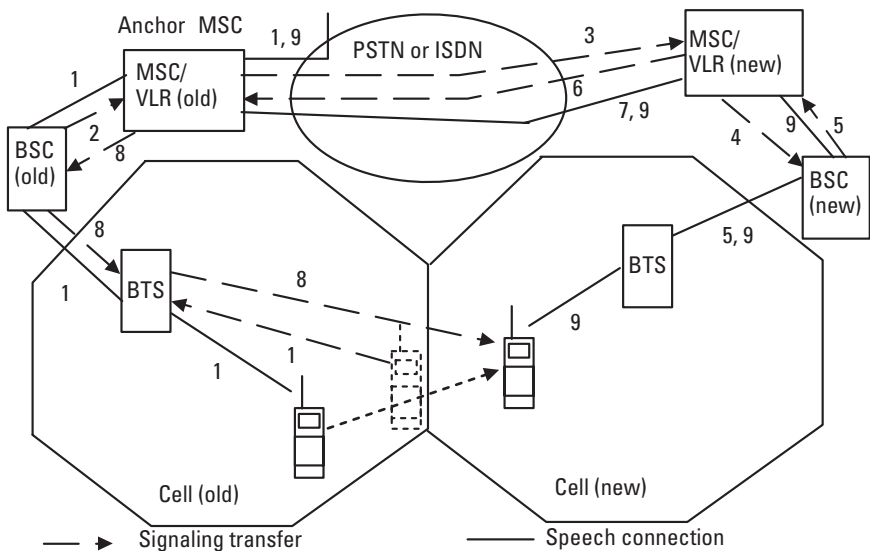
Figure 5.14 Handover or handoff between two MSCs.

the BTS. The measurement information includes, in addition to traffic channel measurement results, identifications of neighboring cells and the measurement results of them. The mobile station continuously measures the common channel of each neighbor cell in addition to the traffic channel in use for the call.

2. The BSC (old) notices that the best cell candidate is not under its control and requests the MSC (old) to begin handover preparation to the new cell. The MSC (old) recognizes that the proposed cell (new) is connected to another MSC.

3. The MSC (old) requests a handover number from MSC (new). The handover number is a temporary telephone number that is used to establish a connection via PLMN, ISDN, or PSTN for the handover.

4. The new MSC requests allocation of a traffic channel from the BSC (new).

5. The BSC (new) allocates the channel and informs the MSC (new).

6. The MSC (new) allocates a handover number and sends it to the MSC (old).

7. The MSC (old) routes a call toward the MSC (new) using the handover number as dialed digits.

8. When the routing is complete and channel is established from the anchor MSC to the new cell, the new MSC/VLR commands the mobile station, via the MSC (old), to switch to the new traffic channel (frequency and time slot of the new cell) and MSC (old) to perform switching.

9. The old MSC switches to the new channel and the new MSC and BSC connect the speech path through the reserved channels in the new cell. Notice that the call is still controlled by the old MSC, which has the role of anchor MSC and it, for example, produces charging records. Finally, the channel of the old cell is released.

5.6.4   GSM Security Functions

In the GSM special attention is paid to security aspects, such as security against forgery and theft, security of speech and data transmission, and security of the subscriber's identity. Use of a radio transmission makes the PLMNs particularly sensitive to the misuse of resources by unauthorized persons and the eavesdropping of information exchanged on the radio path.

For security functions the AuC delivers random numbers and precalculated keys for authentication and ciphering to the HLR. It then sends them with other subscriber information to the VLR, where location update is performed.

We now review the four most important security functions of GSM network, which are shown in Figure 5.15. In addition to these functions, the GSM SIM card is protected by a *personal identity number* (PIN), similar to a credit card "password." The ME may also provide additional security features.

5.6.4.1 Authentication

For authentication AuC provides authentication triplets to the VLR via the HLR. These include a signed response, random number, and ciphering key. Each triplet is used only once and AuC delivers new triplets on demand.

The principle of authentication is to compare the subscriber authentication key Ki in the authentication center and in the SIM without ever sending the Ki on the radio path. For authentication the network sends a random number to the mobile at the beginning of each call. The SIM then uses an algorithm, A3, to process a response that is dependent on the random number as well as on the secret subscriber specific key stored in the SIM. The
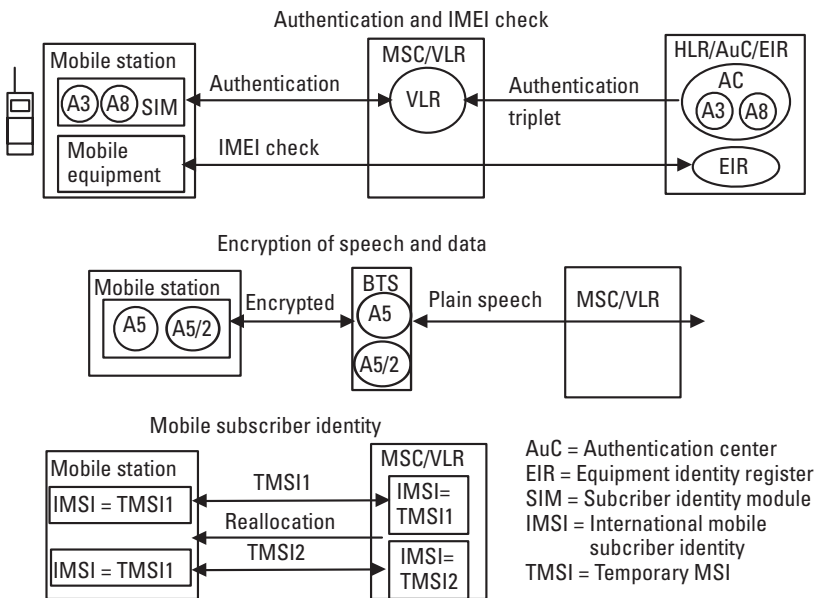
Figure 5.15 The security functions of GSM.

AuC has also computed this response, called *signed response* (SRES), and delivered it in the authentication triplet to the VLR. The VLR performs a comparison and if they match, the MS is allowed to use the network.

5.6.4.2   IMEI Check

The *international mobile equipment identity* (IMEI) check procedure is used to ensure that the mobile equipment does not belong to the black list where the EIR stores the serial numbers of stolen mobiles. If an IMEI is found on the black list, a connection cannot be established. The IMEI is a manufacturer-specific code that is stored in each piece of mobile equipment when manufactured.

5.6.4.3   Encryption of Speech and Data

Speech and data are encrypted before forwarding the radio or air interface. The main algorithm for ciphering is A5, which defines how the ciphering sequence is generated. For encryption an exclusive-or operation is performed with data and the ciphering sequence. An encryption algorithm uses the ciphering key that is calculated by the authentication center and by the SIM. The ciphering key depends on the subscriber-specific key together with the random number that is given to the mobile station at the beginning of each call.

5.6.4.4   Mobile Subscriber Identity

The MS is normally addressed over the air interface by using a *temporary mobile subscriber identity* (TMSI), which is allocated for each mobile located inside an LA. The global identity of the mobile, *international mobile subscriber identity* (IMSI), which is stored into SIM, is very seldom sent over the air interface to prevent eavesdropping devices from using it as trigger information. A new TMSI is allocated for the next call when communication is in ciphered mode.

IMSI is a global subscriber identification but it is not the same as the telephone number. A subscriber may have several telephone numbers, for example, one for telephone and one for fax, connected to one IMSI in the HLR. He or she may also change SIM (IMSI is changed) but keep the same telephone number.

5.6.5   GSM Enhanced Data Services

The business of mobile data services is expected to grow fast when mobile telephone business becomes mature. The original second generation cellular systems provided quite a low data rate (9.6 Kbps) and they utilized a

symmetrical circuit-switched operation principle that is not optimum for data applications. Most data services have the traffic characteristics of asymmetric high-data-rate "bursty" data transfer. To meet increased demand for better data services, standards have been developed for a new channel coding (14.4 Kbps), *high-speed circuit-switched data* (HSCSD), GPRS, and *enhanced data rate in GSM evolution* (EDGE).

One physical channel of GSM carries a 22.8-Kbps data rate and transmission of user data at only 9.6 Kbps seems wasteful. Most of the overhead is used for error control. For those users who are close to the base station, channel coding can be reduced to improve the user throughput from 9.6 to 14.4 Kbps while keeping the same bit error rate. If the user moves and the quality of the channel decreases, channel coding is changed back to the original rate of 9.6 Kbps, which tolerates higher interference.

HSCSD is sometimes called *multislot service* and it increases data throughput by combining one to four time slots on one carrier frequency into a single data channel. The maximum user data rate of HSCSD is 4 × 14.4 Kbps = 57.6 Kbps. HSCSD is a circuit-switched service and its user fee is based on the connection duration and the number of time slots used.

EDGE implements a new modulation scheme in the GSM air interface. Originally GSM used binary modulation that was very noise tolerant. In good propagation conditions we can use a less noise tolerant nonbinary modulation scheme and increase the user data rate via the same channel. EDGE triples the user data rate with the help of 8-PSK modulation, which transmits three bits in each symbol instead of one. EDGE defines several coding schemes and by selecting a suitable modulation and coding scheme the system can adapt its operation to channel conditions. EDGE will increase the user data rate from the original 9.6/14.4 Kbps per time slot to 59.2 Kbps (no error control) per time slot. The most important application for EDGE will be GPRS, which is then called EGPRS (GPRS with EDGE).

We have used GSM here as an example of a digital cellular radio system and illustrated the functionality of the network with a few simplified examples. A more detailed study of cellular network functionality is beyond the scope of this book. More comprehensive descriptions about the operation of GSM are given in [1, 2].

5.7 GPRS

GPRS will surpass HSCSD because it provides optimum service for data users. It is a genuine packet-switched system in which the radio channel is

reserved only for the time during which data transmission takes place. It supports asymmetrical transmission and radio resources in uplink and downlink directions are assigned independently. The radio channel is reserved only for the time of transmission although a virtual connection (see Section 6.2) exists at all times for each GPRS subscriber.

GPRS users share physical channels allocated for packet-switched service. It offers a real packet access method and supports charging based on the amount of transferred data.

5.7.1 GPRS Network Structure

Because GSM was originally designed for circuit-switched service, the introduction of packet-switched transmission requires some significant functional and operational changes. GPRS introduces two new network nodes, *GPRS support nodes* (GSN) to support end-to-end packet transfer. They are *serving GPRS support node* (SGSN) and *gateway GPRS support node* (GGSN), which are shown in the simplified network architecture in Figure 5.16. To keep the figure simple, many GSM network elements, such as EIR and SMSC, have been excluded. Circuit-switched calls are routed from the BSC via the MSC to PSTN.
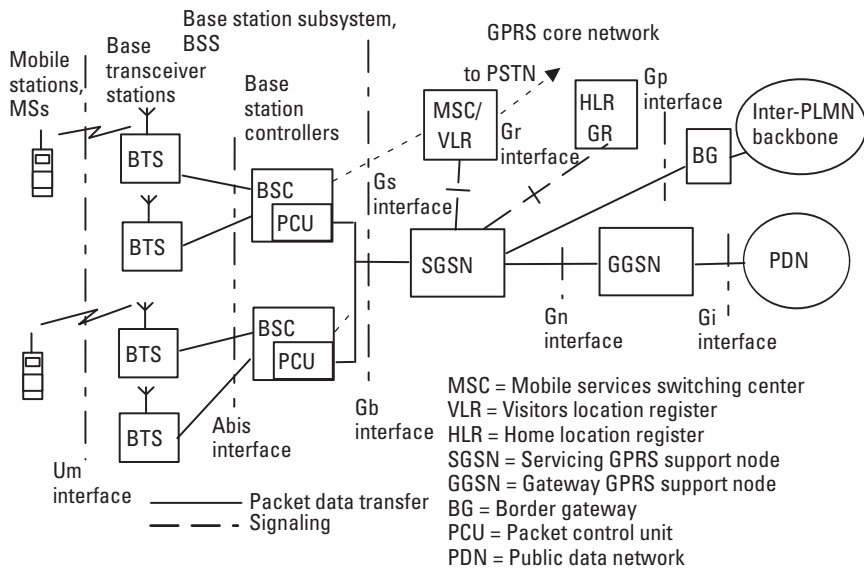
Figure 5.16 GPRS system architecture.

For GPRS operation the HLR is enhanced with GPRS subscriber data and routing information. The HLR is to be updated to include *GPRS register* (GR), which stores packet user related data, such as IP address of the present SGSN. The GR stores routing information (SGSN address) and maps IMSI to one or more *Packet Data Protocol* (PDP) addresses if addresses are permanently assigned to subscribers. Typically, an *Internet Protocol* (IP) address is assigned for a subscriber on demand, that is, when she attaches GPRS. Dynamic address is released in GPRS detach, when the MS is disconnected from the GPRS network. The major upgrades in the BS subsystem are new channel coders in BTSs and *packet control units* (PCUs) in BSCs. PCUs take care of the packet transmission between MSs and SGSN.

5.7.2 GPRS Network Elements

GPRS is designed to leave BSSs almost unchanged. A PCU is added to the BSS and it routes packet-switched data to a separate GPRS core network. The roles of the new network elements are introduced below.

5.7.2.1 SGSN

The SGSN support node is responsible for the delivery of packets to all MSs within its service area. SGSN plays the same role in GPRS as MSC/VLR in the circuit-switched GSM network. It detects new MSs in its area, performs authentication, ciphering, and IMEI check, and it sends and receives packets to and from the MS. It also collects *charging data records* (CDRs), performs session and mobility management, and supports SMS. Mobility management contains GPRS attach/detach, routing area update, location area update, cell change (in ready mode), and paging. Cell change corresponds to handover and for that SGSN takes care that unacknowledged packets are routed to a new cell and possibly to the new SGSN if the new cell is under different SGSN. Session management contains PDP context activation, deactivation, and modification. PDP context activation means establishment of a "virtual circuit" between the MS and GGSN for IP packet transfer. The SGSN handles protocol conversion between the IP protocol and the LLC protocol that is used between SGSN and MS. The SGSN performs TCP/IP compression according to RFC 1144 to save radio capacity [3].

5.7.2.2 GGSN

The GGSN support node acts as a logical interface to external packet data networks. GGSN acts as a router and hides the GPRS network infrastructure from the external networks. GGSN remains an anchor point when SGSN is

changed due to a cell change. When the GGSN receives a packet addressed to a specific user, it checks its database to determine if the address is active. If it is, GGSN uses its PDP context (containing SGSN address for tunneling) to forward the packet to the relevant SGSN. If the address is not active, the data are discarded. GGSN collects charging information based on usage of network resources.

The GGSN corresponds to the GMSC in circuit-switched operation. The main function of GGSN is to handle interactions with external data network. It acts as a router hiding the GPRS network from the external network, typically the Internet. GGSN updates the location of the MS according to the information from SGSNs and routes packets to and from the SGSN, which serves the destination MS.

Within the GPRS network, PDUs or packets are encapsulated at the originating GSN (either SGSN or GGSN) and decapsulated at the destination GSN. In between the GSNs, IP tunneling is used to transfer PDUs. This means that a user data packet is inserted into an IP packet, which contains the IP address of the destination GSN (see Section 6.6.4). The GGSN maintains routing information used to tunnel packets to the SGSN that is currently serving the destination MS. All GPRS user-related data, needed by the SGSN to perform the routing and data transfer functionality, are stored within the GR/HLR.

5.7.2.3   PCU

The BSC is upgraded with a PCU, which supports all GPRS protocols and controls and manages most of the radio-related functions of GPRS. It splits up long LLC frames into short RLC frames for ratio transmission. The PCU's function is to set up, supervise, and disconnect packet-switched calls. It also supports cell change, radio resource configuration, and channel assignment. The BTS is merely relay equipment without protocol functions and it performs the modulation, demodulation, and channel measurements. The PCU may be located anywhere between the SGSN and the BTS [3].

5.7.2.4   Border Gateway (BG)

The BG is not specified by GPRS specifications and PLMN operators have to define its functionality in their roaming agreements. It may contain firewall functions to ensure secure connections over the inter-PLMN backbone network. The BG may be integrated into the GGSN.

The GPRS network contains some network elements that are not shown in Figure 5.16. Two important examples of these are the *charging gateway* (CG), which collects charging information from SGSN and GGSN

and sends it to the billing system, and the *domain name server* (DNS), which maps logical domain names to IP address numbers the same way as in any IP network (see Section 6.6.10).

5.7.2.5 GPRS Network Interfaces

Several interfaces are defined in GPRS standards, the most important of which are as follows:

- Gb, between BSS and SGSN;
- Gn, between SGSN and GGSN;
- Gi, between GGSN and external network;
- Gs, between SGSN and VLR;
- Gr, between SGSN and HLR.

Interface specifications define protocols needed for packet-switched operation. Typically IP packets are transmitted from the MS to the external PDN and additional IP tunneling is used in the GPRS core network.

5.7.2.6 MS

GPRS requires completely new terminals, which can be regular mobile telephones, PC cards, or specific modules built in to a machine. These terminals are divided into three classes:

1. *Class A terminals* can handle both circuit-switched and packet-switched services simultaneously and independently.
2. *Class B terminals* can handle either circuit- or packet-switched service at one time. It can automatically switch between these two modes. For example, in the case of an incoming circuit-switched call, it may suspend packet transfer and resume it afterward.
3. *Class C terminals* must be manually set into one of the modes. In the circuit-switched mode, it is cannot be accessed for packet-switched traffic and vice versa. A special case of class C mobile is a packet-only terminal integrated into laptop.

5.7.3 Operation of GPRS

GPRS provides genuine packet-switched radio access and packet service users share the radio channels allocated for GPRS. Information about whether the

network provides GPRS service and which channels (frequencies and time slots) are allocated for packet users is broadcast on the cell broadcast channel.

HLR/GR stores information about the services and present location of its MSs (SGSN). SGSN, which corresponds to MSC/VLR, stores more accurate location of MSs and data related to their current service, such as the current IP address. It also performs security functions, such as authentication.

When a GPRS user wants to access a packet-switched service, for example, the Internet, he or she performs GPRS attach. Then IP address is allocated for the MS and the user sees the Web page of his or her ISP's browser. The URL (see Section 6.6.11) of this default page is stored in his or her subscriber information in HLR/GR and transferred to SGSN at the time of GPRS attach. The GGSN in Figure 5.16 acts as a border router between GPRS and the Internet; the GPRS network looks the same as any other IP network from outside the Internet. The GGSN stores the routing table for all active IP addresses to be able to route packets to the correct SGSN. The SGSN then routes packets further to the cell where the destination MS is currently located.

Physical radio channel, one time slot in each TDMA frame, transmits data blocks that occupy one time slot in four subsequent TDMA frames. Multiple users share each physical channel and each downlink packet contains identification of the destination MS. In the uplink direction, some data blocks are reserved for uplink channel requests from MSs. When an MS wants to transmit packet data it requests an uplink packet channel. According to these requests, SGSN assigns uplink capacity to MSs. Each downlink data block contains MS identification that is allowed to transmit the next block in the uplink direction.

We see that a mobile station can be connected to GPRS service continuously, but it reserves capacity only if it needs to transmit or receive. Charging can be based on a low fixed monthly fee and the fee based on the amount of transmitted and received data. This is makes GPRS superior to earlier circuit-switched alternatives, such as HSCSD.

5.8 Problems and Review Questions

Problem 5.1

What are the main advantages of cellular systems compared with the old generation radio telephone systems that did not utilize a cellular network structure?

Problem 5.2

An analog radio telephone network has a frequency band of 100 (bidirectional) FDMA channels. The network covers a 50 × 50–km urban area. Give the maximum number of simultaneous calls in the network if (a) only one base station is in use; (b) the network is upgraded to a cellular network with a cell size of 10 × 10 km and the frequency reuse ratio is 1:9 (each channel is used again in every ninth cell); (c) cell size is reduced to 1 × 1 km; and (d) cell size is reduced further to 0.35 × 0.35 km (that is, equal to the minimum size of cells in early GSM). For simplicity, assume here that the cells are rectangular and all channels are used as traffic channels. [*Hint:* Divide all channels of the network between a cell cluster (group) of nine cells. Then repeat this cluster to cover the geographical area of the network.]

Problem 5.3

What are the two main types of channels used in each cell of a cellular mobile system?

Problem 5.4

What is handover? Explain the handover principle, that is, how it is carried out in a cellular network.

Problem 5.5

How does the cellular network route an incoming call to a subscriber located anywhere in the network or even in a different country? What are the roles of the HLR and VLR in the routing of an incoming call?

Problem 5.6

Explain the main phases that occur in the radio interface of a cell when an outgoing call is requested. Explain also what happens when an MS in a cellular network receives an incoming call.

Problem 5.7

Explain the applications of cordless telephones. How do cordless systems basically differ from cellular systems?

Problem 5.8

Explain the structure of a GSM network. What are the main network elements and what are their roles in the operation of GSM?

Problem 5.9

Explain the multiple-access method of GSM.

Problem 5.10

Explain how location update is performed in GSM. What triggers it and what happens after that?

Problem 5.11

Explain how a call is routed from a GSM MS to another MS of the same home network. Assume that (a) both are located in home network and (b) both have roamed to another country.

Problem 5.12

Explain how handover is performed in the GSM network.

Problem 5.13

Explain the security functions implemented in the GSM network.

Problem 5.14

What are the main new network elements that are needed in GPRS network? What are their roles in GPRS operation?

Problem 5.15

What are the basic operating differences between GPRS and circuit-switched GSM?

References

[1] Redl, M. S., M. K. Weber, and M. W. Oliphant, *An Introduction to GSM,* Norwood, MA: Artech House, 1995.

[2] Mouly, M., and M. B. Pautet, *The GSM System for Mobile Communications,* Paris, France: Michel Mouly and Marie-Bernadette Pautet, 1992.

[3] Heine, G.A., and Inacon GmbH, *GPRS from A–Z,* Norwood, MA: Artech House, 2000.

6

Data Communications

In the beginning of this chapter we clarify some key terms that we need to describe a certain data communications principle or a system. Then we introduce the concept of data communication protocols, trying to convey a concrete feel for the layered data communication protocol stacks and the reason why we define data communication architectures with the help of the protocol layers. Then we describe various data communications systems used in access networks and for local- and wide-area data communications. In the latter half of this chapter, we concentrate on the Internet and describe its structure, operation, and services.

6.1   Principles of Data Communications

The first data communications system was the telegraph. It was invented more than 100 years ago. The letters to be transmitted were converted into a code called Morse code. The codes were transmitted as pulses along a wire or as radio-frequency bursts in the case of wireless telegraph. Then the development of data communications slowed, but during the last few decades data communications have expanded rapidly as computers have become tools for everyone in both business and residential environments.

6.1.1   Computer Communications

Modern computers manipulate bits, binary symbols, of electrical energy. When a computer communicates with another computer it sends these bits along a cable between them. This is relatively easy if the computers are within the same room or a building. If the distance is longer, a telecommunications network is required that provides an end-to-end communications channel. Data communications can be accomplished by means of many various alternatives, some of which we discuss in the following sections.

6.1.2   Serial and Parallel Data Communications

In a transmission network only one channel is usually allocated for one end-to-end connection in each direction. Let us use as an example source of data a simple *American Standard Code for Information Interchange* (ASCII) terminal. We press keys on the keyboard and each keystroke generates a 7-bit binary word (8 bits with parity) corresponding to the letter or number of the key pressed. For example character *a* corresponds to the binary sequence 1000011 (the first bit on the left) [1]. If we have only one channel available, we have to send bits of this word in turn (first bit on the left) to the channel; such a case represents *serial* data transmission (see Figure 6.1).

When serial transmission is used between a computer and its peripheral device, a parallel clock signal may be used for timing. In serial transmission

Figure 6.1 Serial and parallel transmission.

over longer distances we want to manage with one channel and we have to use a line code to insert timing information into the data stream. This synchronization information enables the receiver to determine when it has to detect each individual received bit. How we implement this depends on whether we use an asynchronous or synchronous transmission mode, as described in Section 6.1.3.

If a computer needs to communicate with, for example, a printer in the same room, parallel communication is often used. A special cable with several wires is provided between the computer and the printer and all 8 bits of a data word, corresponding to one character, are transferred at the same time in parallel over the cable. Parallel data transmission is much quicker than serial, but we can typically use it only over short distances. The maximum is usually of the order of 10m.

Communicating terminal devices in data communications are called *data terminal equipment* (DTE) and the equipment that terminates the transmission channel that goes through the network is called *data circuit-terminating equipment* (DCE). A modem that we use for data transmission over a telephone network is a typical example of DCE. Many different interface specifications exist for DTE and DCE, and the most common standards are defined by the ITU-T and the *Electronic Industries Association* (EIA). One of the most common data interfaces is ITU-T's V.24/V.28, which corresponds to EIA Standard RS-232-C.

6.1.3 Asynchronous and Synchronous Data Transmission

Over longer distances we use serial transmission either in an asynchronous or synchronous transmission mode. Serial transmission over long distance requires that the timing information for the receiver be transmitted together with the data so that a separate clock signal is not required.

In asynchronous transmission only a small number of bits are transmitted at a time, usually 8 bits that correspond to one ASCII character. In the beginning of each block of 8 bits of data, a *start bit* is sent to indicate to the receiver that it should prepare to receive 8 bits of data (see Figure 6.2). For synchronization the receiver has to know the data rate, which has to be set in advance, so that when it detects the start bit it is able to receive the few following bits. After these bits a *stop bit* is sent that terminates the 8-bit data block. The next block of data is synchronized independently with the help of a new start bit preceding the data bits.

In asynchronous transmission, a simple error-detecting scheme called *parity* can be used. We may use even or odd parity error checking. If even
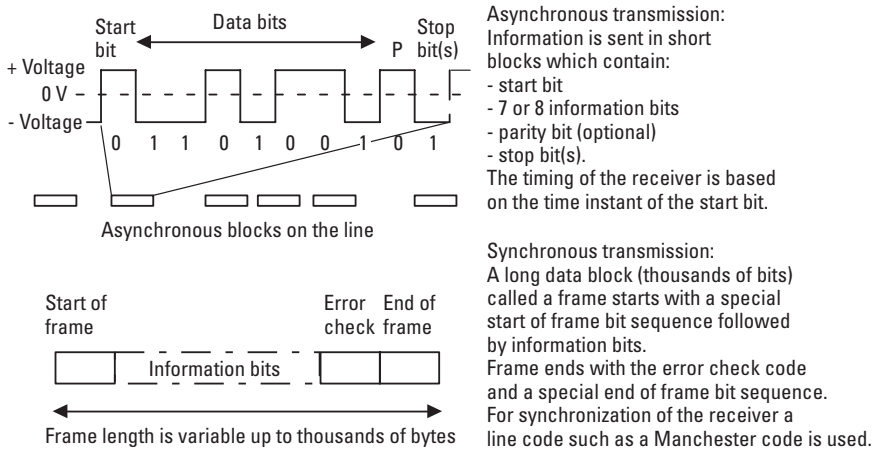
Figure 6.2 Asynchronous and synchronous transmission.

parity is used, the total number of "1" bits in the block, including data bits and the parity bit, is set to be even with the help of the parity bit. In the case of odd parity, the parity bit is set to "1" or "0" so that the total number of "1" bits in the block is odd. To detect possible transmission errors, the receiver determines whether the received number of "1" bits is even or odd depending on the parity agreed. We will see later that this parity check method is a simple example of a data link layer protocol.

Asynchronous transmission is used for the transmission of ASCII characters in conventional terminal–mainframe computer communications. For larger information blocks it is used in some file transfer protocols such as KERMIT and X-LINK. In these protocols special "start of block" characters are sent at the beginning. Then information follows as asynchronous words and at the end special "end of block" characters are sent.

Synchronous transmission is a more modern principle for transmitting a large amount of information in a frame (see Figure 6.2). Each frame starts with a special start-of-frame bit sequence and the frame may contain more than 1,000 bytes of information. Each frame also contains error control words and an end-of-frame sequence. The receiver uses the error control section of the frame to detect if errors have occurred in transmission. The most common detection method for error detection is a *cyclic redundancy check* (CRC). It is much more reliable than the parity check method discussed previously. In the case of errors the transmitter retransmits the frame in error. In the most common protocols the receiver sends an acknowledgment to the

transmitter in the other transmission direction for received error-free frame or frames. If errors have occurred, the frame is not acknowledged in a predefined period of time and the transmitter sends it again.

In asynchronous transmission the start bit provided the required timing information for each byte of data. Most synchronous transmission methods are so-called "bit-oriented" protocols in which data blocks are not divided into separate bytes because many types of information, such as graphics, is not presented as a set of bytes. Unique start-of-frame and end-of-frame sequences or flags are used to provide frame synchronization. These flags should be unique and actual data must not include similar data sequences. One common method used to avoid frame misalignment is to use bit stuffing or zero insertion, as shown in Figure 6.3. Consider a flag (01111110) used in the popular *high-level data link control* (HDLC) protocol. After the start-of-frame flag the sequence of six subsequent 1's is not allowed in the data section of the frame. To avoid that, a 0 is inserted in the end of each sequence of five subsequent 1's. In the receiver each 0 following five subsequent 1's is discarded. If binary 1 follows five subsequent 1's, the frame is declared to be finished (end-of-frame flag) [1].

Synchronous transmission requires that the bit timing information be inserted into the data stream itself with the help of line coding because frames are very long. As an example, many LANs use the Manchester line code that we described in Chapter 4.

The principles we have discussed above are contained in physical layer and data link layer definitions in the data communication architecture
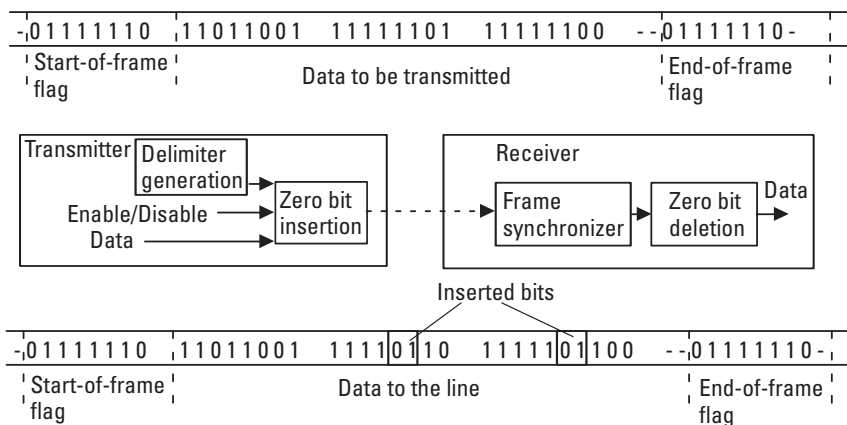
Figure 6.3 Bit stuffing or zero insertion.

described in Section 6.3. As we will see later, these two protocol layers deal with aspects of how data communications over a physical connection between two machines are arranged.

Connection from a computer to another through the data communications network requires a switching function, which routes data frames or packets from the source host to the destination host. We introduce next the basic alternatives for routing and, as we will see in Section 6.3, these functions are implemented into the data link layer or network layer in the protocol hierarchy.

6.2   Circuit and Packet Switching

We can divide data connections through a telecommunications network into different categories based on the principle of how the communications circuit is built between the communicating devices. Data communications through the telecommunications network may use three basic different types of circuits:

1. *Leased or dedicated:* The cost of a leased line is fixed per month and depends on the capacity and length of the connection.
2. *Circuit switched or dial-up:* The cost of switched service depends on the time the service is used, the data rate, and the distance.
3. *Packet switched:* The cost is often fixed and depends on the interface data rate. In some packet-switched networks cost may depend on the amount of transferred data. Agreements with the service provider may specify other parameters that influence the cost, such as the maximum data rate or average data rate.

For corporate data networks, the leased-line solution is often attractive when the LANs of offices in a region need to be interconnected. The network operator provides a permanent circuit and the monthly cost is fixed and depends only on the agreed-on data rate. Over long distances, however, leased lines become expensive and switched service is often preferred. In such a service, several corporate networks share transmission capacity and the cost of the backbone of the telecommunications network operator.

Within the switched category there are two subcategories, circuit- and packet-switched networks as shown in Figure 6.4, both of which are used for data transmission. Figure 6.4 also shows some sample networks and what switching principles they use.
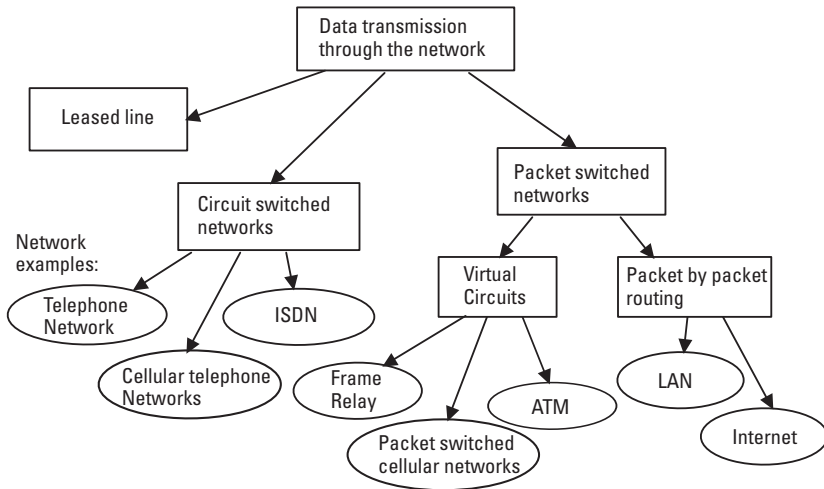
Figure 6.4 Leased lines and circuit- and packet-switched networks.

6.2.1 Circuit Switching

Circuit-switched networks provide fixed bandwidth and very short and fixed delay. It is the primary technology for voice telephone, video telephone and video conferencing. The disadvantage is that it is inflexible for data communications where the demand for transmission data rate is far from constant but varies extensively over short time scales.

Some older generation data networks used the circuit switching principle. In the beginning a circuit-switched connection is dialed up by the data source. The routing is based on the destination subscriber number given when the circuit is established. The connection is released after the communication is over (see Figure 6.5). During a conversation, the data capacity of the connection is fixed and it is reserved only for this conversation regardless of whether the data capacity is used or not. At the end of the call, the circuit is released. ISDN as well as the telephone network use the circuit-switching principle.

6.2.2 Packet Switching

Packet-switched networks are specially designed for data communication. The source data are split into packets containing route or destination identifications. The packets are routed toward the destination by packet-switching nodes on the path through the network. The major drawback of the packet-

switched technology is that it usually cannot provide a service for applications that require constant and low delay. There are two basic types of packet-switched networks as illustrated in Figure 6.5: virtual circuits and datagram transmission.

In the case of virtual circuits, the virtual connection is established at the beginning of each conversation or it is permanently set up and every packet belonging to a certain connection is transmitted via the same established route. The main difference between circuit-switched physical circuits and virtual circuits is that many users share the capacity of the transmission lines and channels between network nodes if virtual instead of physical circuits are used. At a certain moment active users may use all the available capacity if other users are not transmitting anything. The complete address information is not needed in the packets when the connection is established. Only a short connection identifier is included in each packet to define the virtual circuit to which the packet belongs. The operation of switched virtual circuits is explained in more detail in Section 6.2.4.

Circuit switched data transfer

Physical circuit

Signaling
Data transfer
Signaling

The circuit is first established, then data is transferred and in the end the circuit is released. The capacity of the circuit is not available for other users. Examples: telephone network, ISDN

Packet switched data transfer

Optional routes for packets

Address   Data

In true packet switched data communication there is no dedicated connection between communicating devices. Each packet includes complete destination address and is sent and routed independently. One example is the Internet.

Packet switched data transfer with virtual circuits

Virtual circuit

(Signaling
Data transfer
Signaling)

Permanent virtual circuit exists (or virtual circuit is established for conversation). All data is transferred via the same path. (In the end the virtual circuit is released.) Each packet includes circuit identification. Capacity between nodes is shared by all users. Examples: frame relay, ATM

Figure 6.5 Circuit- and packet-switched data transfer.

Another method for packet-switched data communications is connectionless datagram transmission in which routing devices perform routing procedures, and each packet contains a full destination address. We discuss this layer 3 (network layer) routing principle next.

6.2.3   Layer 3 Routing and Routers

In the case of layer 3 routing every data packet carries the complete global destination information (network layer address of the destination) and all packets are routed independently. As a consequence, each packet may use a different route and arrive out of sequence. The operating principle of the Internet belongs to this category.

The routing procedure is performed at the network layer, layer 3, and it requires analysis of each packet and the routing decision based on the destination address on each router. Packets are stored and, when the route and the corresponding port of the router are defined, the packets are forwarded to the next router on the path to the destination. This operating principle make routers slower than the switching devices that operate at the data link layer, layer 2, and use virtual circuits. The operating principle of the data link layer switches is discussed next.

6.2.4   Switching and Routing Through Virtual Circuits

The routing of packets is based on the virtual circuits in many public data networks, such as frame relay or ATM networks. Each frame or cell on a virtual circuit contains identifying information about the circuit to which it belongs. This identification has a different name in different networks but we call this identification the *virtual circuit identifier* (VCI). During the circuit establishment phase, signaling messages are exchanged between user equipment and a network and end-to-end virtual circuits are set up in each node on the way through the network. Often a permanent virtual circuit is set up by the network operator when agreement is made about a data connection between corporate sites.

In the network each circuit established between nodes has a certain identification number and there is no global identification that could be used on all links through the network. Instead, one of the free circuit identifications on each intermediate link is allocated for each virtual circuit being established and the connection tables of switching nodes in the network are updated to contain all established circuits, as shown in Figure 6.6.

VCIs have only local significance on a specific network link and, therefore, virtual circuit identifiers are changed as a frame traverses the virtual path through the network.
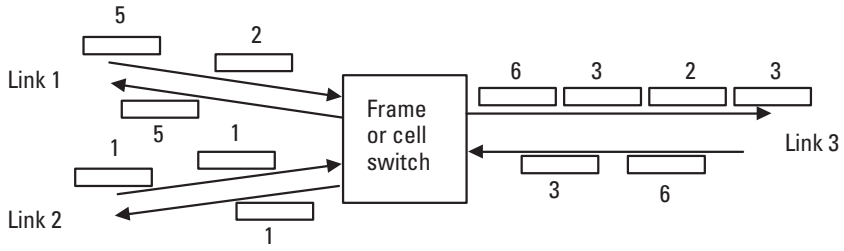
VCI = Virtual circuit identifier

Link 1 routing table		
VCI-in	Link-out	VCI-out
2	3	2
5	3	6

Link 2 routing table		
VCI-in	Link-out	VCI-out
1	3	3

Link 3 routing table		
VCI-in	Link-out	VCI-out
3	2	1
6	1	5

Figure 6.6 Switching of frames on virtual circuits.

When a frame is received from a certain link, the frame switch simply reads the VCI and combines the incoming link number to determine the corresponding outgoing link and VCI. The new VCI is then written into the frame header and the frame is queued for forwarding on the appropriate link. The order of frames is preserved and routing them is very fast because the routing process does not require analysis of a global address.

In Figure 6.6 a frame switch has connection tables for each incoming data link. Let us assume that a frame with identification 3 is received from link 3. The switch looks up the link 3 routing table and finds out that this frame should be transmitted to link 2, so identification 3 is replaced by 1.

This process is fast because it does not require any network layer routing with a global address. Instead switching is done in the data link layer. The VCI is also very short and the utilization of data capacity in this kind of network is more efficient than if the global address were included in each frame or packet.

6.2.5 Polling

In local data networks, such as shared media LANs described in Section 6.5, all devices are usually connected to shared transmission media and they

broadcast their data frames to the network, and only the destination machine, indicated by destination address in the frame, receives it. Figure 6.7 shows two physical topologies for a broadcast network, a bus and a star. Logically both networks have a structure of a bus and each frame, transmitted by one station, reaches all other stations in the network.

This type of broadcast channel is dynamically (i.e., on demand) allocated for hosts in one of two main ways. Devices using network may be autonomous or a master or central control computer may give permission to one device to transmit at a time. Computers in an Ethernet LAN are autonomous; there is no central control computer that allocates network for users and anyone can use a free channel when needed, as we will see in Section 6.5. In this kind of decentralized channel allocation method, each computer has to decide itself whether or not to transmit.

The traditional way to allocate a single channel between a mainframe computer and its terminals is for the main computer (master) to poll the terminals (slaves), each one in turn. The master sends regularly poll frames that contain identification of a slave and possibly data from master to slave. The slave responds to the poll frame with the data frame.

The main advantage of this old principle is that operation of slaves is very straightforward and it is also used in some modern systems. Examples of

Broadcast network, decentralized channel allocation
autonomous computers, each has equal right to transmit on demand.

Physical bus, logical bus

Physical star, logical bus

Broadcast network, centralized channel allocation,
master polls each slave in turn.

Poll for station D

Response to poll frame

Master

Master

Response to
poll frame

Poll for
station D
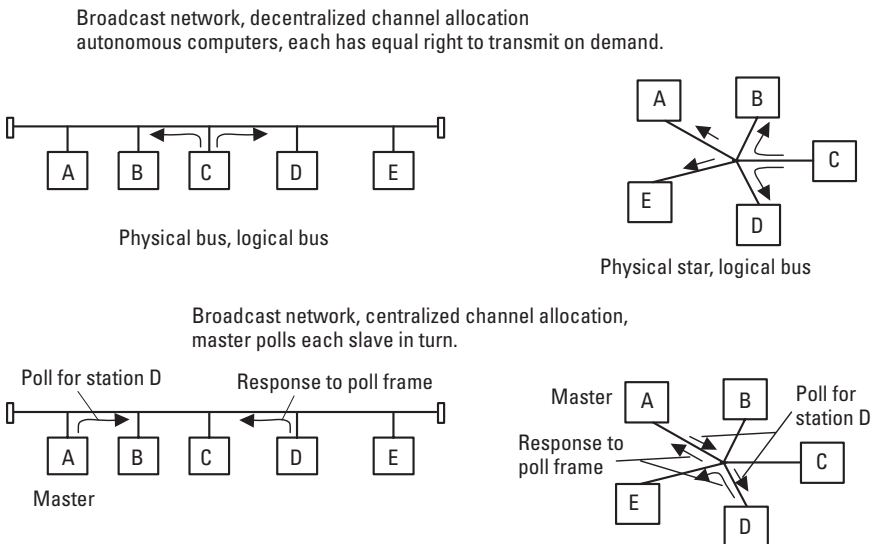
Figure 6.7 Dynamic broadcast channel allocation methods.

these systems include *universal synchronous bus* (USB), a high-speed interface between a PC and its peripherals, and Bluetooth, a wireless connection between a PC or cellular phone and its peripherals. Another advantage of polling is that no contention situations arise and each slave is guaranteed to get the transmission capacity that the master allocates to it. A disadvantage is the capacity that is wasted by the process of polling slaves that have nothing to transmit.

6.3   Data Communication Protocols

The computers that communicate have to understand each other. They have to speak the same "language." This common language is defined as a data communication protocol. A detailed protocol specification that enables two different systems to communicate includes many communications rules such as how the letter "A" is presented as a binary word and what the voltage is of bit "1."

As we will see, many specifications are needed to enable data communications between systems. An ISO standard *Open Systems Interconnection* (OSI) gives guidelines about how this complicated set of communications rules is divided into smaller groups of rules and functions that are called *layers.* This helps us concentrate on one group of functions (= protocol of a certain layer) at a time and we do not have to think about the functions for which other layers are responsible. For example, if we are specifying the error detection code that belongs to the data link layer of OSI, we need not worry about the power levels of optical transmission or the shape of electrical pulses in the cable. These things are the problems of the lowest layer, called the physical layer in OSI.

In the next section we review the OSI reference model that was standardized by the ISO and try to clarify the importance of the principle of the layered structure in data communications.

6.3.1   Protocol Hierarchies

To reduce the design complexity of computer communications hardware and software, the needed functionality is organized as a series of layers, each built on its predecessor (see Figure 6.8). Many proprietary protocols are in wide use in addition to the available international standards. All of them use some form of layering. The number of layers, the name of each layer, the contents of each layer, and the function of each layer may differ from network to network.
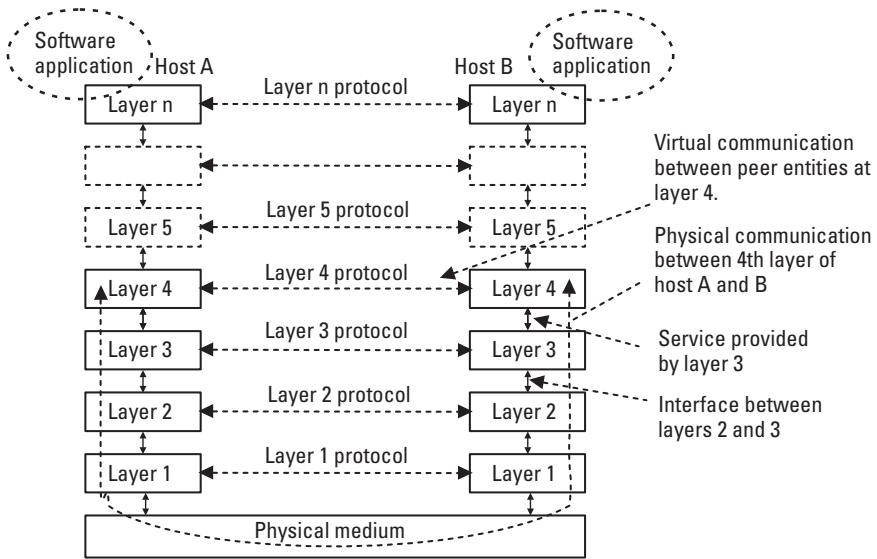
Figure 6.8 Protocol hierarchy.

In all networks, the purpose of each layer is to offer certain services to the higher layers, shielding those layers from the details of how the provided services are actually implemented.

6.3.1.1 Protocol

Each layer in one machine carries on a conversation with the corresponding layer in another machine as shown Figure 6.8. The rules and conventions used in this conversation are collectively known as the *protocol* of this layer. We can say that the protocol specifies the format and meaning of the information that a layer sends to the layer below. This information is received and understood by the corresponding layer at the other end if exactly the same detailed protocol specification is implemented there.

With the help of its protocol each layer below provides services to the layer above it. The provided service is often specified separately from the protocol specification. We could say that service specifies what the layer looks like from the point of view of the next layer above. For example, if a layer provides data transmission with or without error detection, the layer above may select which one it wants to use. How they are implemented, that is, how layers at opposite ends communicate to provide the service, in the layer is specified in the protocol specification.

The interfaces between layers are defined to be as simple and clear as possible and each layer performs a specific collection of well-understood functions.

6.3.1.2   Protocol Stack

The set of layers and their specified protocols are known as a *protocol stack*. For successful communications both computers have to use exactly the same protocol stack where each layer at both ends complies with the same detailed standard.

6.3.2   Purpose and Value of Layering

The purpose of each layer is to provide certain services to the higher layers, shielding those layers from the details of how the provided services are actually implemented. Without this abstraction technique it would be difficult to partition the design of communications hardware and software into smaller manageable design problems, namely, the design of individual layers.

This also makes it possible to replace one layer with a new implementation without affecting other layers. Consider, for example, a LAN in which the same software applications may use both token ring and different Ethernet LAN technologies. We illustrate the fundamental idea of the layered protocol structure next with an analogy.

Imagine that two philosophers, one in Egypt and one in the Philippines, want to communicate remotely (at layer 3 in Figure 6.9). The philosophers have a jargon specific to their profession and only another philosopher is able to understand it completely. This corresponds to the protocol of the layer 3.
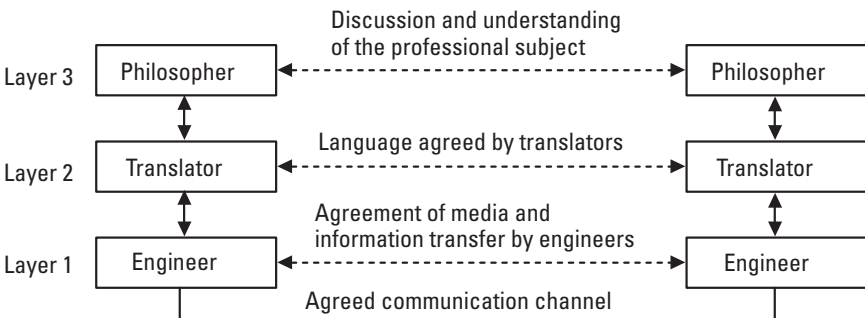
Figure 6.9   Purpose and value of layering.

Because these philosophers have no common language, they each need a translator (at layer 2). To establish a communications channel, each translator contacts an engineer (at layer 1). When the Egyptian philosopher wishes to discuss something with another philosopher, he passes the message across the 3–2 interface to his translator at layer 2 who uses the language that he has agreed to previously with the layer 2 translator at the other end. The translators use their best common language, which may be English, and this agreed common language of the translators corresponds to the layer 2 protocol.

The translator then gives the message to layer 1 for transmission. Engineers at layer 1 may use any channel they have agreed on in advance. This physical communications may use a telephone network, a computer network, or some other means. This engineer and the communications channel arranged by him correspond to the layer 1 protocol.

When a message arrives in the Philippines, it is received from layer 1, translated by the translator (layer 2) at that end, and passed to the receiving philosopher. Let us now imagine that these English-speaking translators are replaced by others, for example, because of a lunch break. These new translators notice that French is a better common language for them and they agree to use that. The service provided to layer 3 by layer 2 remains the same and the philosophers do not notice that the protocol of one lower layer is completely changed.

In the same way, engineers can change the communications channel in use and upper layers may not notice and do not even care how the communication is arranged as long as the quality of service is acceptable. Note that each protocol layer is completely independent of the other layers, and higher layers do not have to concern themselves about how communications is actually arranged by the lower layers, that is, what protocol they use as long as service provided remains the same.

6.3.3 Open Systems Interconnection (OSI)

In the late 1970s the ISO began to work on a framework for a computer network architecture that is known as the OSI reference model. The purpose of this model was to eliminate incompatibilities among computer systems.

In 1982, ISO published ISO 7498 as a draft international standard. This document is just a framework about how communications protocols should be designed, not a detailed specification needed for compatibility. CCITT/ITU-T published it as Recommendation X.200.

OSI was originally designed for computer communications. Today data and voice are not necessarily separated into different networks. Many

times the network does not know and is not interested in what information the transmitted data contain. ISO and ITU-T specify all new networks and systems according to the layering principle of OSI and several detailed protocol specifications for OSI layers have been published for various purposes. However, some worldwide systems are not designed according to OSI and the most important of them is the Internet. The Internet is based on standards that are openly available but not approved by ISO or ITU-T.

The name OSI comes from the goal to make systems open for communications with other systems. Any manufacturer is free to use these "open" specifications. Anyhow, many data communications systems are still proprietary systems and their specifications are the property of one vendor, so they are not available to others.

6.3.3.1 OSI Reference Model

In the OSI model, communications is divided into the seven layers shown Figure 6.10. The OSI reference model lists what each layer should contain, but it does not specify the exact services and protocols. The detailed specification of each layer is published as a separate international standard.

Note that the layers below the transport layer care about the data transmission through the network from host A to host B. The transport layer is
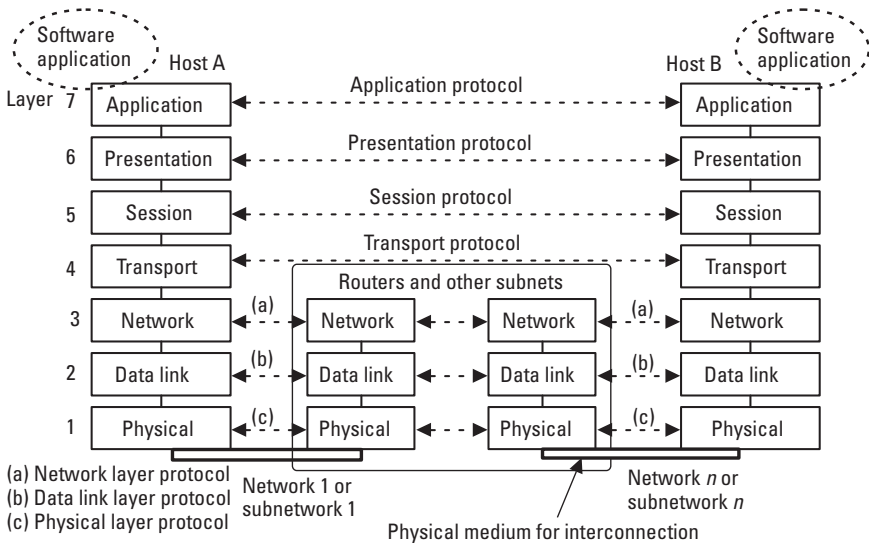
Figure 6.10 The OSI reference model.

the lowest end-to-end layer and it uses the network to implement the service for the session layer.

When we look at what kind of functions each layer performs, we notice that the lower the layer we look at, the more functions we see that are related to the network technology used for the actual data transmission. The more we look at the upper layer, the more we see common functions available for software applications running in hosts. As we see in Figure 6.10, layers 4 through 7 are all implemented only in the communicating end machines. We do not need layers 4 through 7 at all for actual end-to-end data transmission; this is accomplished by layers 1, 2, and 3. The only purpose of the uppermost layers is to help software applications, and to do this they provide more sophisticated services than just a stream of data. As an example, this stream of data from the network layer may contain some errors. Each application software designer should design an error recovery scheme in his application if this service is not provided by the transport layer protocol.

Note that data link layer and physical layer may be completely different in each network or subnetwork according to OSI terminology [2]. For example, in Figure 6.10, host A could use ISDN for connection of network 1 and host B could use Ethernet technology in network *n.*

6.3.3.2   Physical Layer

The physical layer is concerned with transmitting bits over one hop in a communications channel. The main design issue is to make sure that when one side sends a 1 bit, the other side receives is as a 1, rather than as a 0 bit. Typical specifications of the physical layer are the duration of a bit in microseconds, the number of volts used to represent a 1 and a 0, a number of pins, and the connector type used. Physical layers of the systems are designed to operate practically error free. If the physical layer cannot deliver error-free data to the upper layer, it does not perform retransmission for error recovery, but leaves the consequent actions to the upper layers.

The specifications of the physical layer deal with mechanical, electrical, and procedural interfaces and the physical transmission medium. The transmission medium is understood to be below the physical layer but the specifications include the characteristics required by it.

Let us look at some examples of the physical layer protocols:

- *V.24, RS-232-C, EIA-232D (latest version):* electrical characteristics of the asymmetrical data signals and their usage;
- *IS 2110:* specification of a physical data interface connector;

- *I.430 (IS 8877):* ISDN basic rate user interface;
- *ANSI 9314:* specification of optical interface for wideband data network called *fiber distributed data interface* (FDDI);
- *IEEE 802 and ISO 8802 series:* physical interfaces of Ethernet-based LANs and WLANs.

The last two examples contain data link layer specifications as well.

6.3.3.3   Data Link Layer

The data link layer builds the frames and sends them to the following node on the line via the physical layer. It receives frames, checks if these frames are error free, and delivers error-free frames to the network layer. The data link layer at the receiver may send acknowledgment of error-free frames to the transmitting end. The transmitter may retransmit the frame if no acknowledgment is received within a certain time period. Note that this procedure takes place between each pair of nodes on the way.

The ISO has specified the data link layer for LANs and divides the specifications into two sublayers: (1) the medium access (MAC) sublayer and (2) the *logical link control* (LLC) sublayer. This division is necessary for LANs because of the complexity of the data link layer in this kind of application. In LANs computers are connected to the same network and they share the transmission capacity of a broadcast channel. The MAC sublayer cares about the functions dependent on the network hardware. The most popular LAN technology is carrier sense multiple access with collision detection (CSMA/CD), or the "Ethernet," which is available at many data rates (see Section 6.5). If we upgrade our network to a higher data rate LAN, we change only the MAC sublayer. The LLC considers most of the data integrity aspects, such as retransmission and acknowledgments, and it remains the same. In the case of a simpler point-to-point link there is no need for a separate MAC layer and one data link layer protocol specification may cover the whole data link layer.

In a LAN each computer has its own MAC address (hardware address). This address is used to identify the source and destination of each frame in the broadcast channel. With the help of this address, computers can have a point-to-point connection via a broadcast channel that is shared by many other point-to-point connections. Note that this hardware or MAC address is used only inside a LAN, it is not transmitted to other networks (see Section 6.5.4).

Some examples of data link layer protocols are as follows:

- *IS 3309:* HDLC; variants are used in most modern networks, such as GSM and ISDN;

- *Q.921, LAP-D, ISDN layer 2:* HDLC-based data link layer protocol;

- *IEEE 802.X = IS 8802-X:* MAC layer "Ethernet"-type LANs and WLANs;

- *IEEE 802.2 = IS 8802-2:* LLC of LANs (when OSI stack is in use); for a complete LAN data link layer both 8802-2 (LLC) and 8802-X (MAC) are needed.

6.3.3.4 Network Layer

The layers below the network layer are only interested in the point-to-point connections between two nodes. The network layer has some knowledge about the structure of the network and, together with the network layers of the other nodes it services, packets are routed through the network to the destination. Each node has its own (network layer) global address.

A key issue is to determine how packets are routed from the source to the destination. Routes can be based on static tables at the network layer that are rarely changed, or they can be dynamic to reflect the current network structure and operational conditions, such as load.

The hosts connected to the network are autonomously sending packets when they wish. They usually are not informed about the traffic density of other hosts or network connections. If many hosts happen to be active at the same time, too many packets are transmitted and, hence, have the potential to get in the way of each other and form bottlenecks inside the network. The control of such congestion also belongs to the network layer.

In public data networks, an accounting function (if applied) is often built into the network layer. The software in the network layer must count how many packets or bytes are sent by each customer in order to produce the charging information.

In isolated small broadcast networks (such as Ethernet), routing is so simple that the network layer is not needed at all. MAC or hardware addresses identify the hosts inside the LAN. However, if and when these networks are connected to other networks, network addresses are mandatory. Note that the MAC addresses used in the data link layer have no importance outside one LAN.

Here are some examples of network layer protocols:

- *X.121:* addressing of digital networks;

- *Q.931, I.451:* ISDN D-channel, layer 3;

- *Internet Protocol of the Internet:* not approved by ISO but it performs basically the same functions as the network layer protocols of OSI.

Figure 6.11 shows the relationship between the OSI reference model and the more popular TCP/IP protocol stack, which is explained further in Section 6.6.

6.3.3.5 Transport Layer

The transport layer is the first true end-to-end layer. The protocols of hosts from the transport layer upward use the network as an end-to-end connection for communication. The source message may be split into shorter segments or packets, and the destination transport layer may be the first point where pieces belonging to the same message meet again. The destination transport layer then reproduces the original message from the received data segments.

The transport layer acts as an interface layer between network connection-oriented lower layers and application service-oriented upper layers. Its responsibility is (typically) to check that end-to-end transmission is error free, that packets are not lost on the way, and that data were delivered in their original order to the upper layer. For this it may include end-to-end acknowledgment and retransmission procedures.
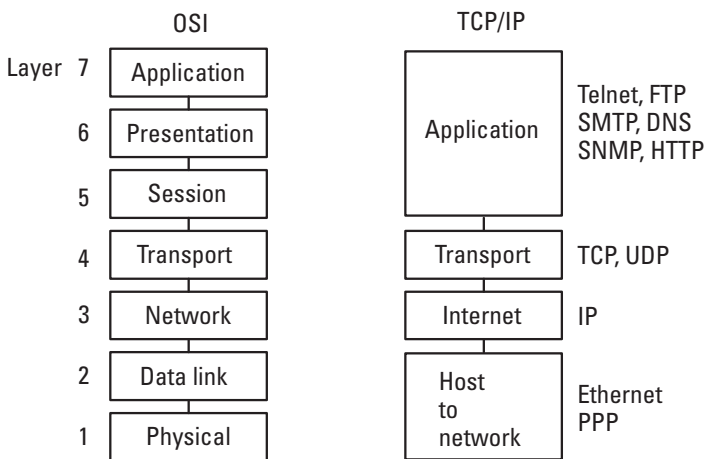
Figure 6.11 The TCP/IP stack and OSI reference model.

The transport layer usually provides two basic service classes to the session layer:

1. *Transport of isolated datagrams through the network:* Transmitted messages may arrive in different order and errors may occur. Examples: UDP of the Internet (actually this does not belong to OSI protocols), and the Transport Protocol, class one (TP1), of OSI (IS 9072).

2. *Error-free point-to-point channel:* Such a channel delivers messages in the same order in which they were sent. Examples of these are the Transmission Control Protocol (TCP) of the Internet (not included in OSI protocol standards) and TP4 of OSI (IS 8072/8073).

UDP and TCP protocols are explained in Section 6.6.

6.3.3.6 Session Layer

The transport layer ensures that end-to-end transmission between computers is successful. Actually, the task of communications is accomplished by the four layers below the session layer. The three uppermost layers are not needed for data transmission but they help make applications compatible so that the application programs running in computers understand each other.

The session layer allows users on different machines to establish sessions between them. It can be used, for example, to allow a user to log in to a remote time-sharing system or to transfer a file between two computers.

A session layer allows ordinary data transport, as does the transport layer, but it also provides some enhanced services useful for some applications. One of these services is to manage dialogue control. Sessions can allow traffic in both directions at the same time or in only one direction at a time. If traffic is allowed only one way at a time, the session layer can help by keeping track of whose turn it is. The session layer also provides a token management function and, with the help of this, only the host holding a token may perform critical operations.

Another service by the session layer is to support successful transmission of large files. Without this service a single error might destroy the whole file, which would then have to be retransmitted. To eliminate this problem, the session layer provides a way to insert checkpoints into the data stream so that after a crash only the data after the last checkpoint have to be repeated.

An example of the session layer standards is the International Standard IS 8326/8327 (X.215/225 of ITU-T) that defines the connection-oriented session layer service and protocol. In TCP/IP a separate session layer does not

exist and there all application support functions are integrated into the application layer, as shown in Figure 6.11.

6.3.3.7   Presentation Layer

As we saw, the lower layers deal primarily with the orderly transfer of bits or data from source to destination. The presentation layer is instead concerned with the format of the transmitted information. Each computer may have its own way of representing data internally, so agreements and conversions are needed to ensure that different computers can understand each other.

The job of the presentation layer is to encode the structured data from the computer's internal format into a bit stream suitable for transmission. This may require compression, for example. The presentation layer at the other end decodes the compressed data to the required representation at the destination. The presentation layer helps both computers understand the meaning of the received bit stream the same way.

Different computers have different internal representations of data. All IBM mainframes use *extended binary coded decimal interchange code* (EBCDIC) 8-bit codes as character code; whereas practically all others use ASCII 7- or 8-bit options. The Intel chips number their bytes from right to left, whereas Motorola chips number theirs from left to right. Because computer manufacturers rarely change these conventions, it is unlikely that any universal standards for internal data representation will ever be adopted.

One solution to ensure compatibility is to define a presentation layer standard for the "network representation" of data and any computer may communicate with another if each of them converts its internal representation to this standardized network format. When this is implemented into each computer, all can communicate with all others and there is no need for data conversion between each pair of computers. Other tasks for the presentation layer are data compression and encryption.

Some examples of presentation layer protocols are IS 8824-1 through -4, which are standards for the representation of data structures, and Abstract Syntax Notation 1 (ASN.1; abstract because it is just a representation). ASN.1 descriptions are quite similar to any high-level programming language and include definitions of data structures such as integer and floating-point number. Another example includes IS 8825-1 and -2, which are encoding rules for ASN.1 defining how representations are encoded into a bit stream for transmission.

In the TCP/IP protocol stack a separate presentation layer does not exist and its functions are integrated into the highest layer, the application layer, as shown in Figure 6.11.

6.3.3.8   Application Layer

The application layer contains the application-specific services that use the services of lower layers. User applications that perform the tasks that computers are purchased for are not included in the application layer, but they communicate with the help of the application layer protocol. An example of a user application is a word processing program.

Often needed communications applications, such as file transfer or an ASCII terminal, have been defined as the application layer protocols to serve any user application that needs their functions. Communications applications provide a common vendor-independent service for user applications of any vendor. The application layer services are usually available for the programmer as other services of the operating system in use. With the help of these services software application programmers (designing, e.g., word processing software) do not have to worry about actual data transmission at all. They may use all of the services of the protocol stack in their development environment.

One example of application protocols is electronic mail. In addition to a service similar to file transfer, it provides ready-made functions for deleting, sending, and reading mail. The specifications of the application layer protocol define, for example, the format of addresses and message fields.

To distinguish between application programs and the application layer protocol, let us look at an example of electronic mail. We may have an application running on top of the application layer in our computer. This program may provide nice colors, a user-friendly editor, and separate windows for addresses and messages. It may also provide a user-friendly addressing method, that is, we can give a destination address such as "John" that is converted by the software to the format that the application layer understands. Note that the application layer service provides the communications services required, but we may enhance them with application software for local purposes.

Some examples of application protocols are as follows:

- *X.400, the message handling system (MHS) of ITU-T:* electronic mail;

- *IS 8571, file transfer access and management (FTAM) of ISO:* file transfer protocol of OSI;

- *FTP and other application layer protocols of the Internet:* see Section 6.6. FTP is not defined strictly according to OSI.

The importance of the OSI protocols just discussed is decreasing as the Internet expands. OSI protocols are official standards; to meet all needs they

are very complex, and their usage is restricted to public telecommunications networks. However, their design principles are valid for all protocols and that is why we have spent some time with OSI. OSI is also valuable model for analysis and comparison of different protocols.

6.3.4   TCP/IP Protocol Stack

Instead of OSI protocols, a major share of data communications use TCP/IP, which is used in the global Internet. The relationship of TCP/IP to OSI layers is shown in Figure 6.11 [3]. The Internet as a network, its services, and its operation is introduced in Section 6.6. Section 6.6 also illustrates operation of the most important protocols belonging to the TCP/IP protocol stack.

Those readers who are not familiar with any data communication protocol may have found our discussion quite abstract. To make the operation of protocol layers more concrete, we illustrate in the next section how actual data packets are handled when they are transferred down and up through the protocol stack. Further clarification is given in Section 6.6, in which the TCP/IP stack is described from the bottom up.

6.3.5   Data Flow Through a Protocol Stack

Let us assume that the user of the source machine performs an action that creates the message, M(A), which is produced by a process running in the application layer (OSI layer 7) in the source machine (see Figure 6.12). This message could be an e-mail that we transmit to the other computer through the network. The message is passed from the application layer directly to the transport layer in the TCP/IP protocol stack. In the OSI stack the presentation layer transforms the message in a certain way (e.g., text compression) and then passes the new message to the session layer (5). The session layer, in this example, does not modify the message but regulates the data flow to prevent an incoming message from being handed over to the presentation layer while it is busy. Data units given to the lower layers are called protocol data units, for example, an *application protocol data unit* (APDU).

In most networks a data packet has a certain maximum length, but usually there is no limit to the size of messages accepted by the transport layer. If the message is very long, the transport layer must break it up into smaller units, adding a header to each unit. The header includes control information such as a sequence number. In many networks, such as the Internet, transmitted units may arrive in a different order than they were transmitted. With the sequence number the transport layer at the destination machine is able to build the original message by placing the transmitted pieces into the correct order.
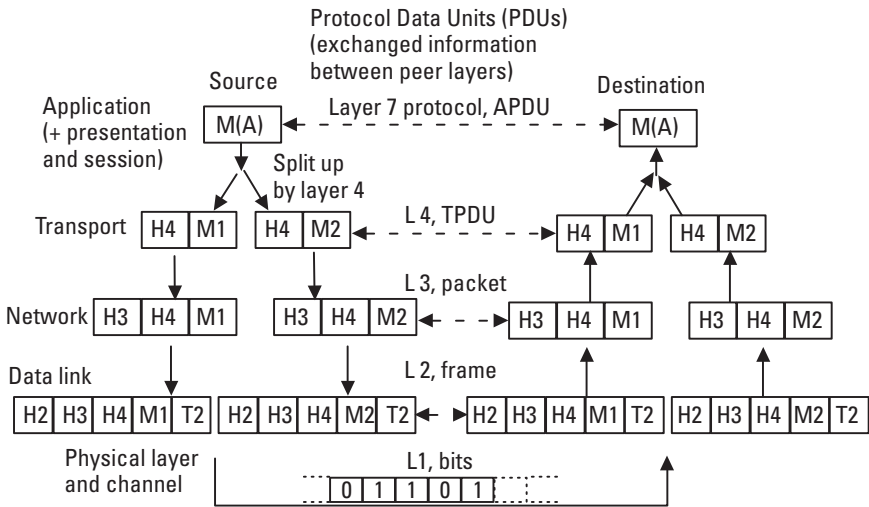
Figure 6.12 Data flow through a protocol stack.

The network layer (3) looks up the routing table and decides which of the outgoing lines to use. It attaches its own headers such as the address of the destination network layer and passes the data to data link layer (2). The network layer message is often called a *packet.*

The data link layer adds a header and also a trailer and gives the resulting unit to the physical layer for transmission. The header may include a start-of-frame flag and physical addresses in an LAN. The trailer is needed for end-of-frame flag and error detection. The message of the data link layer is often called a *frame.*

The physical layer transmits the bits given by the data link layer to the physical media, such as an LAN cable. It may, for example, convert bits into light pulses for optical fiber cable transmission.

In the receiving computer the message moves layer by layer upward. A corresponding layer at the other end as shown in Figure 6.12 strips off the header of a layer. None of the headers for layers below a certain layer *n* are passed up to the layer *n.* Thus, each layer receives the message as it was transmitted by a corresponding layer in the source machine. They act as if they were connected directly, not through the lower layers. For example, when the data link layer of the destination machine has checked through the error detection field in the trailer (T2) to determine that the frame is error free, the error check bytes are removed before the data are given to the network layer.

If the reader still feels that the preceding illustration was too abstract and wants to understand thoroughly the principle of protocols and layers, she may study the operation of one protocol stack, for example, TCP/IP, layer by layer. This is the most efficient way get a concrete grip on protocols; and when one protocol is understood, new ones are easy to learn. We will describe TCP/IP in more detail later in Section 6.6 and a comprehensive description is also given, for example, in [4].

6.4   Access Methods

To use data services, a user's computer has to access the data network. Various access technologies are available for business and residential needs. We review the most important access systems in this section.

6.4.1   Voice-Band Modems

The word *modem* comes from the combination of the two devices, modulator and demodulator. Modulation converts a digital signal into an analog signal for transmission through a channel, and demodulation performs the conversion back to the original digital baseband data signal. Voice-band modems are needed when an analog voice channel of the telephone network is used for data transmission.

The frequency band of the voice channel is 300 to 3,400 Hz and the baseband digital information is transferred to this band through CW modulation. The CW modulation methods used in voice-band modems are exactly the same as those used for radio transmission (see Chapter 4).

As we know, CW modulation may vary three characteristics of a carrier: amplitude, frequency, or phase. The corresponding basic modulation methods are AM, FM, and PM. All these basic modulation methods are used in the voice-band modems.

As we see in Figure 6.13, the only analog section in the connection through a modern telecommunications network is the subscriber line of the local-access network. The fastest standardized voice-band modems can support data rates up to 33.6 Kbps. The maximum user data rate is of the order of 30 Kbps even though the transmission rate inside the PSTN is 64 or 56 Kbps (data rate of PCM coded voice channel). Half of the end-to-end data capacity is wasted because of analog subscriber lines that perform A/D and D/A conversions at both ends.

New modems with essentially higher capacity will not be standardized because voice-band modems are already quite close to the theoretical
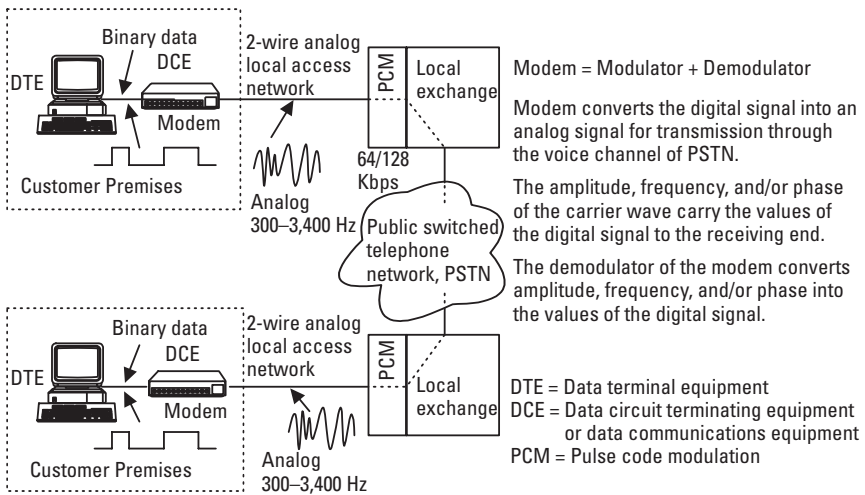
Figure 6.13 Modem link over the PSTN.

maximum capacity of the voice channel and many higher data rate access technologies have become available. If an analog subscriber line is replaced by an ISDN line, the full capacity of the allocated channel in the network can be utilized and end-to-end data rate will then be 64 Kbps (B-channel) or 128 Kbps (two B-channels).

6.4.1.1 V Series Recommendations of ITU-T

The ITU-T (CCITT) has defined many standards for voice-band modems with a variety of speeds and these recommendations are identified by the letter V and a number attached to it. Modems of different manufacturers work together if they support the compatible V standard. Many modern modems support previous lower speed standards as well and they are able to adapt their speed to the level supported by the other end. To illustrate development of voice-band modems and modulation methods, some examples of modem standards are described briefly next.

- *V.21:* 300-bps full-duplex (bidirectional transmission). One of the first modems that used carrier frequencies at different transmission directions (1,080 and 1,750 Hz). FSK is used so that binary 1 corresponds to the carrier frequency of the direction in question (1,080 or 1,750 Hz) minus 100 Hz and binary 0 corresponds to the carrier frequency plus 100 Hz.

- *V.22:* 600/1,200-bps full-duplex. This standard provided an acceptable dial-up data connection for the transfer of text messages in both directions. The transmission directions use different carrier frequencies. One example of user applications is a remote text mode terminal. The modulation scheme is PSK with two or four carrier phases and a modulation rate of 600 bauds.

- *V.22bis:* 2,400-bps full-duplex. This modem was designed to update the V.22 modem at the end of the 1980s. The data rate was doubled with 16 amplitude-phase combinations phases (16-QAM) of the carrier. The modulation rate is 600 bauds.

- *V.23:* 1,200/600-bps modem that transmits 1,200 or 600 bps and 75 Kbps in the reverse direction. This asymmetrical transmission provides enough capacity to send keystrokes from the terminal while transmitting larger amounts of data in the other direction. FSK is used in both directions and 1,300 Hz corresponds to 1 and 2,100 Hz corresponds to 0 in the 1,200-bps direction. In the 75-bps direction frequencies are 390 Hz as 1 and 450 Hz as 0.

- *V.32:* 9,600-bps full-duplex. The modulation method is QAM, a combination of amplitude and phase modulation. The modulation rate is 2400 bauds and 16 combinations of carrier amplitudes and phases are used.

- *V.32bis:* This modem is an enhancement of V.32 with a new modulation scheme. It transmits data at 14.4 Kbps. The modulation method is QAM with 128 different combinations of amplitude and phase of the carrier. The modulation rate is 2,400 bauds.

- *V.34:* This standard supports data rates up to 28.8 Kbps full duplex over dial-up telephone lines and uses QAM with a modulation rate of 3,200 bauds. Error-free operation at this high data rate requires a very clean speech channel. If errors occur too frequently, this modem falls back to lower speed in steps of 2,400 bps to reduce the number of errors.

- *V.34+:* Enhancement to V.34 with a data rate of 33.6 Kbps. The modulation method is QAM and the modulation rate is 3,200 bauds as in V.34.

- *V.90:* This standard supports 33.6-Kbps upstream and 56-Kbps downstream data rates. Note that the downstream 56-Kbps rate requires that the source computer have digital access to PSTN and A/D conversion is not performed in the transmitting end.

The highest data rate modems use so many constellation points that errors occur frequently. To reduce the error rate, they add error control bits to correct most errors and this method where modulation and error control coding are combined is known as *trellis-coded modulation* (TCM).

It is likely that essentially higher data rate voice-band modems will not be standardized. Essentially higher data rate service requires end-to-end digital connections provided by, for example, ISDN instead of speech channels.

Note that data transmission with a voice-band modem does not require anything other than just a modem at the end of the subscriber line and an analog voice-band circuit through the network. The voice-band modems that we have discussed in this section use the telephone network exactly the same way as ordinary telephones.

A V.90 modem provides a 56-Kbps data rate to subscriber premises (downstream) and a lower 33.6-Kbps data rate transmission in the opposite direction. This device is not actually a voice-band modem because the downstream data are not modulated to the speech channel and carried through analog telephone channels the same way as speech is. The source machine transmits data over a digital connection to the network and at the other end a PCM encoder converts the digital data stream into analog signals for subscriber loop from which the receiving device reproduces the data stream. Modulation to the analog speech channel is not carried out in this direction. In upstream transmission a voice channel is used and the data rate is restricted to 33.6 Kbps.

The interfaces of a voice-band modem are shown in Figure 6.14. External modems support a standardized physical interface (usually RS-232C, EIA-232D, or V.24) via which data are exchanged usually as asynchronous frames, as we saw in Section 6.1.3. A binary 0 corresponds to a voltage of +3V to +15V and binary 1 to –3V to –15V.

Separate wires are dedicated to the control signals that are used to control data flow between devices. The two example control signals in Figure 6.14 are used for handshaking between modem (DCE) and terminal (DTE) in the following manner. When a terminal wants to send data, it indicates that by setting the request to send (RTS) signal on (+3V...+15V) and modem responses by setting the clear to send (CTS) signal on. If a terminal transmits data too fast, the modem sets CTS off and while it is off the terminal does not transmit. Many other control signals have been defined and reader may refer their functionality in, for example, [3].

In addition to the basic modem functionality that allows a user to transmit data over an ordinary telephone channel, most modern modems include additional functions as introduced in the following sections.
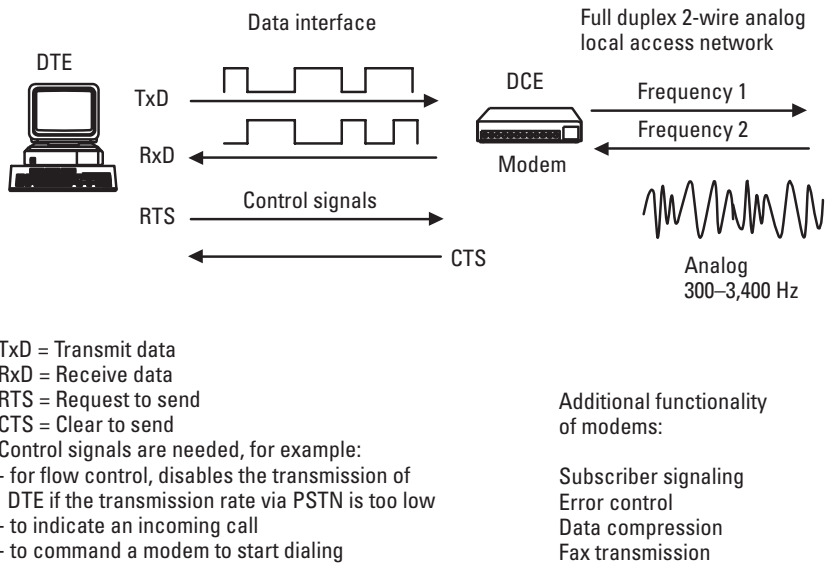
Figure 6.14 Interfaces and operation of a voice-band modem.

6.4.1.2   Error Control

Modems implement the physical layer channel from a terminal to the host at the other end. Errors may occur in the transmission channel between modems, for example, because of the noise on the subscriber line. Many modems send, in addition to the data, error check information and with the help of these they are able to detect and probably correct some bit errors. Both ends have to support the same error control protocol. One international standard for error correction in modems is Recommendation V.42 of the ITU-T.

In addition to error detection and correction in modems, most communications software packages include error recovery functions implemented at higher layers end to end. For example, if TCP of TCP/IP is used and a received TCP data segment contains errors, errors are detected, and retransmission is requested by the far-end communications software where the TCP layer is implemented.

6.4.1.3   Data Compression

Data compression makes it possible for the transmission rate at the interface between a computer and a modem to be much higher than the actual transmission rate through the network. For example, text can sometimes be

compressed by a factor of four or even more. Several methods of data compression are available. As a simple example of the compression of text information, we can imagine that the most common characters are not transmitted in ASCII form but with very short codes; less frequently needed characters would use longer bit strings. This principle would save some transmission capacity. One international standard for data compression is the V.42bis recommendation. Many proprietary standards are widely in use as well.

6.4.1.4   Facsimile Transmission

Many modern modems include facsimile functionality that enables a user to send and receive faxes without printouts. These modems comply with the Group 3 fax recommendations of the ITU-T and transmit facsimile information in digital format at 9,600 or 14,400 bps. We can envision facsimile equipment as a combined scanner and a modem. Faxes and fax modems also perform compression since runs of 0's (blank paper) are very common.

The Group 4 fax is designed to use a 64-Kbps B-channel but it has not become popular because of low penetration of ISDN service. It can communicate with a Group 3 fax, in which case the bit rate for a Group 3 fax will be used [5].

6.4.1.5   Dial-Up Modems

All modern modems are able to transmit multifrequency signaling tones to the telephone network to establish a connection. Voice-band modems include signaling functionality similar to that of a telephone and an external telephone is not needed for dial-up.

6.4.1.6   Operation of a Voice-Band Modem Connection

Modems operate at various data rates over a voice-band telephone channel. Modern modems support many different data rates and they can adapt the transmission data rate to the quality of the channel. In Chapter 4 we saw that the maximum transmission data rate depends on the bandwidth and noise of the channel. If the S/N is degraded (noise level increases), the data rate has to be decreased to keep the transmission error rate low enough. Modems are also able to adapt their data rate and error control scheme to the capability of the other end. To do this, they exchange control data sequences during connection establishment.

The analog signal from a modem is PCM coded into a 64-Kbps data stream at the subscriber interface of a local telephone exchange. The absolute maximum capacity of the transmission channel through the telephone

network can never exceed 64 Kbps. Some quantizing noise is introduced in the quantizing process of PCM, as we learned in Chapter 3, and it reduces the end-to-end data rate from the maximum of 64 Kbps. The present highest rate modems operating at 33.6 Kbps are quite close to the theoretical maximum when we consider quantizing noise, and we can never develop essentially higher rate voice-band modems. The next step after voice-band modems is ISDN, which provides a two or four times higher data rate and speeds the call establishment process. Other options for even higher data rate access are DSL and cable modems, which we discuss later in this chapter.

6.4.2   ISDN

We introduced ISDN in Chapter 2 as a new generation telephone network. Now we look at it again from the data service point of view. As we saw, the full capacity of digital telecommunications network is not utilized by voice-band modems. If we rarely need higher data rate service, it may be attractive to use a circuit-switched telecommunications service to provide the connection only when it is needed. The ISDN provides switched end-to-end digital $n \times 64$-Kbps circuits that we can use for voice or data. Figure 6.15 presents an example of an interconnection when ISDN *basic rate interfaces* (BRIs), 2×64 Kbps, are available at both ends of the circuit.
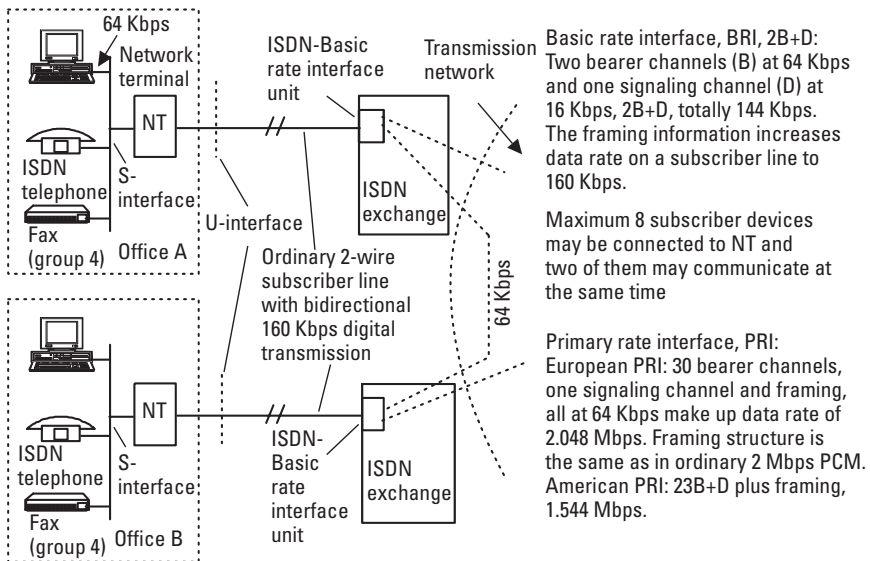
Figure 6.15 Basic rate interface and ISDN connection.

The basic rate interface provides two independent 64-Kbps circuits, and the routing of one B-channel is independent of the routing of the other channel. This allows residential users to have two independent telephone connections via one two-wire subscriber line, or alternatively one line for telephone and the other for a simultaneous connection to the Internet. Network terminals provided by network operators contain one analog interface and a PCM codec for ordinary analog telephone. The provision of a 64-Kbps end-to-end digital connection by ISDN also allows quicker and better quality Group 4 facsimile transmission. The BRI of ISDN 2B+D (2 <insert> 64 Kbps + 16 Kbps) is designed to replace the present analog subscriber telephone interface in the future. However, ISDN is a circuit-switched technology, in which the user fee is based on the call duration and its data rate is quite moderate for Internet access. When new access technologies have evolved, many residential customers prefer to keep ordinary analog telephone and order higher data rate DSL or cable modem access for data services instead of ISDN.

In corporate networks, many B-channels are required and these are provided by the *primary rate interface* (PRI) that has the structure of 30B+D (30 × 64 Kbps + 64 Kbps) or 23B+D. The PRI utilizes 2.048- or 1.544-Mbps transmission in the local-access network and it is able to support many simultaneous (ISDN) telephone calls. This interface is used for PABX connections to the public network and rarely for data connections. The frame structures at 1.544 and 2 Mbps were explained in Chapter 4.

6.4.3   DSL

The access technologies discussed earlier do not utilize all of the potential capacity of the symmetrical twisted cable pair of a subscriber loop. A family of technologies, known as DSL, or digital subscriber line, has been developed to increase the data transmission rate over ordinary local loops to the order of a few megabits per second and it is simultaneously available for ordinary telephone service. This is far beyond the capacity of ISDN subscriber lines. The ISDN data channels are expensive dial-up circuits that are switched by ISDN exchanges and each connection minute increases the subscriber's telephone bill. In the case of DSL, data and speech are separated at the local exchange site. Then the data portion is connected to the data network for Internet access. Customers pay a fixed monthly fee for a high-data-rate connection that is always on. We review now a few DSL techniques and their applications.

6.4.3.1   Applications of DSL

The carriers or network operators are aiming their DSL services mainly at residential users. For them it provides high-data-rate access to the Internet

and at the same time an ordinary telephone connection over a local loop. In these applications, ADSL, which transmits at a higher data rate downstream than upstream, and its variants are preferred.

Corporate network managers can also take an advantage of the benefits that DSL offers. For the interconnection of LANs between offices in the same region, DSL offers a low-cost, high-data-rate access option. In these applications symmetrical DSL or HDSL is often preferred. Figure 6.16 illustrates some applications of DSL: remote access to a data center, Internet access, and interconnection of LANs.

DSL replaces the ordinary local loop, and DSL modems are needed at both ends of the line. If an ordinary telephone connection is to be available simultaneously, the lowpass filter, splitter, at the carrier's central office, splits off the voice channel and routes it to the PSTN. A *DSL access multiplexer* (DSLAM) terminates the data channel at the other end of the subscriber loop and sends traffic onto the carrier's backbone data network, implemented by IP, ATM, frame relay technology, or fixed data circuits, where it heads to a remote data center or the Internet.

DSL is mainly designed to improve the utilization of subscriber cables in the access network. However, because it requires fewer intermediate repeaters, system cost is reduced and DSL will replace conventional primary rate, 1.5- or 2-Mbps, copper cable transmission systems inside the core network as well.
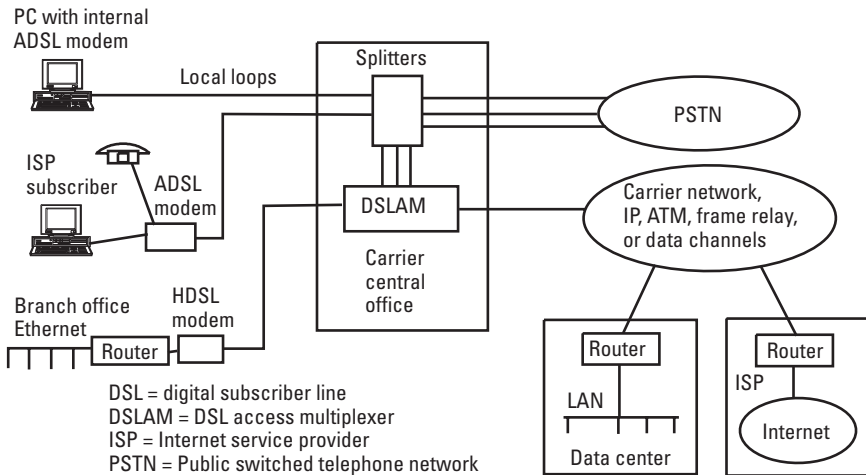
Figure 6.16 DSL in the local loop.

6.4.3.2   DSL Techniques

DSL technologies are still evolving and many alternative technologies are available today and new ones are under standardization. The most important technologies, their transmission distances, and data rates are presented in Table 6.1. We introduce these technologies here. Note that these technologies are evolving and the characteristics given in the table are not final. The data rates and distances in Table 6.1 are approximate maximum figures and they are given just for the comparison of the different DSL technologies. We may expect that some of the technologies introduced here will disappear and some of them will receive wide acceptance in a few years time.

Which technology a network operator chooses for its service depends on many things, for example, the subscriber loop lengths and cable network characteristics in the operator's network. In Europe more than 90% of all subscriber loops are less than 3 km and most technologies in Table 6.1 seem to be feasible. However, the higher data rate and/or the longer the distance, the more effort is required for installation and troubleshooting, which may make DSL less attractive, especially when competition has pressed service fees low.

6.4.3.3   ISDN DSL and Consumer DSL

For residential markets, some carriers in the United States offer low-speed *ISDN DSL* (IDSL) access. IDSL is essentially ISDN without the ISDN

Table 6.1
DSL Technologies, Access Distances, and Service Rates

DSL Technology	Reach (km)	Downstream Data Rate	Upstream Data Rate	Analog Phone	Market
IDSL	8	144 Kbps	144 Kbps	No	Residential
G.lite ADSL	5	1.5 Mbps	640 Kbps	Yes	Residential
HDSL	4	2/1.5 Mbps	2/1.5 Mbps	No	SME*
SDSL, G.shdsl	5–6	2.3 Mbps	2.3 Mbps	No	SME
G.dmt ADSL	3	...8 Mbps	...1.5 Mbps	Yes	Residential SME
VDSL	0.1–2	...52 Mbps (34 Mbps)	6 Mbps (34 Mbps)	Yes	Residential SME

*SME = small and medium size enterprises.

switch. The two B-channels of ISDN BRI are multiplexed to offer a dedicated 128 Kbps of bandwidth for data only. This technology does not provide a simultaneous voice channel as do other DSL technologies, but it operates over longer distances than higher speed technologies.

6.4.3.4   High-Bit-Rate DSL

A conventional primary rate transmission PCM system operating at a 1.544- or 2.048-Mbps data rate over twisted-pair copper cable uses two cable pairs, one for each transmission direction. In a typical cable, signal attenuation together with crosstalk (interference from other systems in the cable) restricts the transmission distance and a regenerator is required after about each 1.5-km cable section. These conventional 1.544- and 2048-Mbps systems use AMI and HDB-3 encoding.

The *high-bit-rate DSL* (HDSL) increases the section length and thus reduces the need for intermediate repeaters. This technology uses 2B1Q (two bits are transmitted in each four-level symbol) encoding that has superior spectral and distance characteristics. HDSL uses two (or sometimes three cable pairs) and thus it is not a consumer access technology. It provides the same data rate for E1 or T1 in both directions and is suitable for *small and medium size enterprises* (SMEs) where upstream traffic has equal volume.

HDSL systems use two cable pairs for full-duplex transmission. The data rate is divided between pairs. In one pair, to one direction, it is only half of the data rate of conventional systems that use different cable pairs for each transmission direction. Further improvement is achieved with the help of an efficient line code. The line code in use is 2B1Q, which means that each pair of bits is coded into one quaternary symbol with four values to the line. This is the same line code that is used in ISDN basic rate subscriber lines for 160-Kbps bidirectional transmission and each symbol carries two bits of information. That reduces the symbol rate on the line to half of the binary rate and the lower transmission rate decreases attenuation and crosstalk. Taken together, these developments double the transmission distance compared to the distance of conventional systems.

The HDSL system transmits the same data rate to both directions just as conventional 1.5/2-Mbps copper cable transmission systems. It will replace them in other applications in the telecommunications network, such as in ISDN PRI connections, because it requires fewer intermediate repeaters, which reduces costs. HDSL is not a consumer access technology because it is symmetrical, uses two pairs, and does not allow a voice-band telephone connection to coexist in the same subscriber loop.

6.4.3.5 Asymmetrical DSL and G.Lite

A symmetrical connection has the same capacity in both transmission directions. The conventional T1 and E1 (1.5- and 2-Mbps) transmission systems and HDSL systems are symmetrical in this sense. However, many applications do not require as much capacity from a subscriber to the network as from the network to a subscriber. One example of this type of application is *video-on-demand* (VoD), which transmits one video program to a subscriber via an ordinary telephone subscriber pair. A subscriber needs only a narrowband channel to the network that enables her to select and control the video program. ADSL was originally developed for VoD. This service has not been successful, but ADSL has ideal characteristics for residential Internet users.

ADSL uses a single pair and transmits downstream at a high data rate and at a lower data rate in the upstream direction. Figure 6.17 shows how the ADSL technique is used for VoD service. In this application the downstream video channel capacity is 1.5 or 2 Mbps, the upstream control channel is 16 or 64 Kbps, and an ordinary telephone call is possible over the same subscriber line simultaneously. A downstream data rate of 1.5 or 2 Mbps can be used over 6-km-long subscriber loops. ADSL terminals modulate the video signal and control signal to a higher frequency band that the telephone or ISDN basic rate signal does not use as shown in Figure 6.18. As an example the frequency band up to 410 kHz is in use and the transmission distance is restricted to approximately 5 or 6 km in the case of a 1.5- or 2-Mbps data rate. Figure 6.18 shows the spectrum allocation used in some VoD field trials.

A standardized ADSL G.dmt technology supports downstream data rates up to 8.1 Mbps at a 3-km distance as shown in Table 6.1. G.dmt ADSL

ADSL, Asymmetrical digital subscriber line
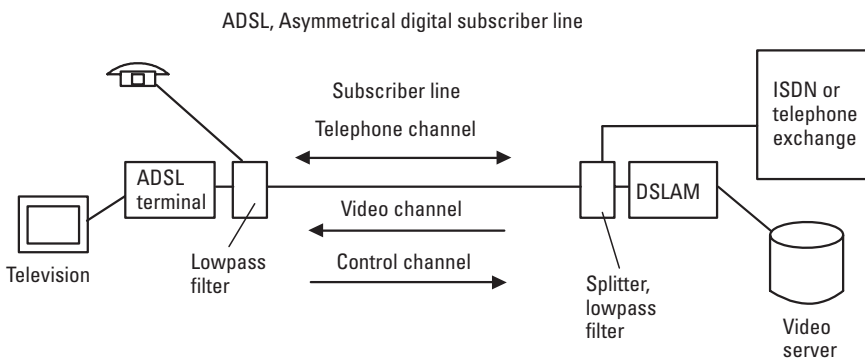
Figure 6.17 Video-on-demand and ADSL.
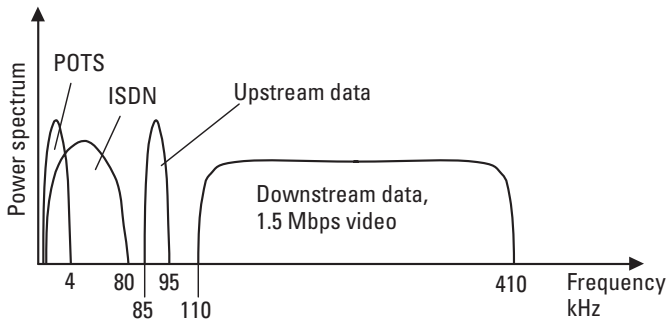
Figure 6.18 Spectrum of 1.5-Mbps ADSL use in VoD.

uses *discrete multitone* (DMT) modulation in which the entire frequency band is divided into 4.3125-kHz-wide subbands, called bins. Bins are numbered from 0 to 256, and the upper cutoff frequency of each bin is given as $k \times 4.3125$ kHz, where k is the bin index. Then the upper cutoff frequency of the ADSL band is 256×4.3125 kHz = 1.104 MHz. Figure 6.19 shows allocation of the bins when an ordinary telephone is used over the same subscriber loop simultaneously. If ISDN is used over the same subscriber loop, the lowest bins of upstream data are not used.

DMT ADSL uses a fixed symbol rate, that is, each bin transmits a symbol for a fixed length and then all of them simultaneously change to the next symbol. The symbol rate is 4,000 bauds and each symbol carries 0 to 11 bits of information using QAM (actual symbol rate on line is slightly higher
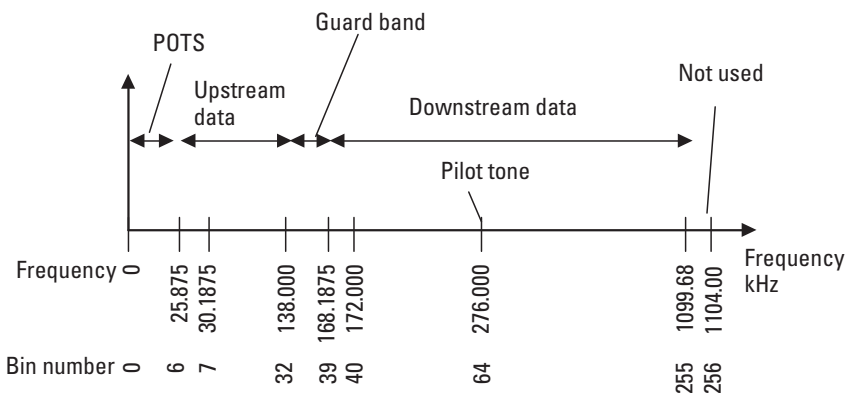
Figure 6.19 Bins and their usage in ADSL.

because every sixty-ninth symbol carries synchronization information instead of user data). The DMT equipment determines the S/N for each of the bins separately and, based on the results, allocates the information bits to be sent to each tone or bin. Then the bins with better S/N transmit more bits per symbol than bins with worse S/N, as shown in Figure 6.20.

The ADSL access system is set to operate at a data rate that the customer has ordered by setting bins to be used and the average number of bits transmitted in each symbol (in each bin). The system may automatically, according to Figure 6.20, adapt its operation into line conditions.

At the time ADSL was specified, asynchronous transfer mode (ATM), which is introduced in the end of this chapter, was expected to be the major backbone network technology. To transmit data efficiently end-to-end ADSL was defined to split data into ATM cells for transmission over the subscriber line. Now when most of the traffic is IP packets, the IP packets are rebuilt at the other end from ATM cells transmitted over ADSL.

One of the major problems with ADSL is that installation to the customer premises often requires the network operator's maintenance personnel to visit the site. To make installation so easy that a customer can manage it himself, a light version of ADSL, known as G.Lite or ADSL Lite, was developed by the ITU. It is does not require a filter in the customer premises, its maximum downstream data rate is 1.5 Mbps, and its maximum upstream direction is 640 Kbps.

6.4.3.6 Symmetric DSL

The SDSL system transmits the data in both directions just as HDSL but it uses a single pair. Because both transmission directions operate at a high data rate, the near-end crosstalk is higher and the data rate lower than in ADSL
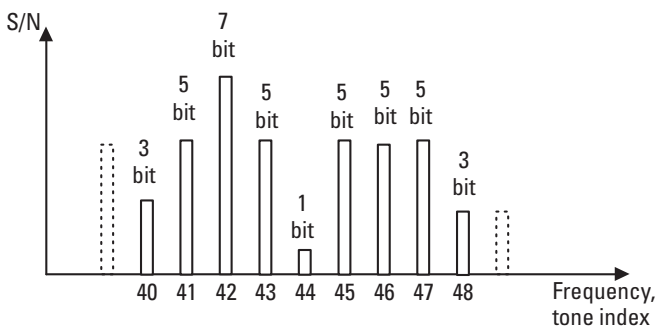
Figure 6.20 Bit allocation to tones or bins.

(see Table 6.1). ITU's standard G.shdsl contains an integrated 64- Kbps voice channel providing *voice over DSL* (VoDSL) service.

6.4.3.7 Rate-Adaptive DSL (RADSL)

An often-used term, RADSL refers to modern DSL technologies, such as ADSL.dmt, SDSL, and VDSL, that can adapt their operation to maximize transmission rates over a cable pair. To achieve this, it adapts loading of each bin to its S/N as explained earlier. However, the DSL access data rate is often set to be fixed and then RADSL technology can ensure that the defined data rate is achieved in various loop conditions.

6.4.3.8 Very-High-Bit-Rate DSL (VDSL)

VDSL is an evolving technology that aims to provide access to wider band services via ordinary telephone subscriber pairs. The transmission data rate from the network to the subscriber's premises is up to 52 Mbps and up to 6 Mbps in the opposite direction over a single pair (see Figure 6.21) [5]. Its symmetrical configuration allows an up to 34-Mbps data rate in both directions. The distance over an ordinary cable pair without intermediate repeaters is quite short, between 0.1 and 2 km depending on the data rate and loop conditions.

Subscriber loops from exchange site are usually longer than VDSL can tolerate and the network-side VDSL equipment has to be installed close to the customer. Then a copper wire DSL part of the circuit might only include the drop line to a residence or business.

6.4.3.9 Summary of the DSL Technologies and Markets

As we have seen, many DSL technical alternatives are available and which technology operators choose for their service depends on many things, such
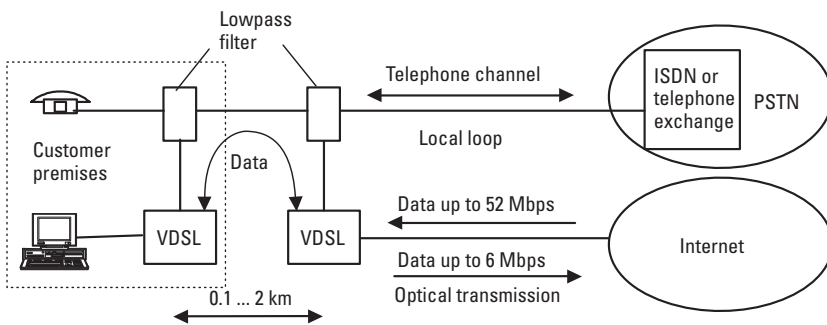
Figure 6.21 VDSL.

as access network length and quality statistics, competition, and their business strategies (e.g., residential or business). However, these technologies have important advantages over the competing technologies for high-speed Internet access such as the cable modems of the cable TV networks and ISDN. A point-to-point local loop is available to most homes and DSL can utilize it to provide access to a residence. It is straightforward to implement because each user has a dedicated point-to-point line. In the cable TV networks, we have to combine and split data to/from many users. ISDN has a low data rate and it requires network operator investments to the infrastructure to manage the increased load of the exchanges. DSL removes traffic from the switched network and reduces the congestion that Internet users might cause.

Since expansion of cellular networks, the importance of subscriber loops for ordinary telephone service has decreased, whereas the demand for wideband Internet access has increased. Subscriber loops provide a high penetration media for wideband Internet access and DSL is a key access technology in the evolution where speech is going more and more wireless, releasing the cable network for wideband data services.

6.4.4 Cable TV Networks

Another media that is widely available for residential Internet access is a cable TV network. Traditionally it has been one-way broadcast media providing a set of broadcast TV channels to the home. The structure of a traditional cable TV network is shown in Figure 6.22. International and national TV programs are received from a geostationary satellite at a central distribution point, known as the *head end* (HE). Local programs are added and the set of TV channels is directed to various neighborhoods by fiber optic cables, which terminate into various fiber nodes. Some hundreds of homes nearest to each fiber node receive their programs in analog form from the coaxial cables [6]. The hybrid fiber coaxial cable infrastructure seen in Figure 6.22 was originally designed for unidirectional TV broadcast distribution only.

High-speed interactive communications across a cable TV access network are made possible by the combination of an upgraded two-way *hybrid fiber coaxial cable* (HFCC) infrastructure, with a cable modem installed in the home and a cable modem termination system, installed at the HEs (see Figure 6.23).

The 54- to 550-MHz frequency band is allocated for broadcast TV channels of 6 MHz each [6]. One or more of these 6-MHz channels is reserved for downstream data and voice. Upstream data carrying data or
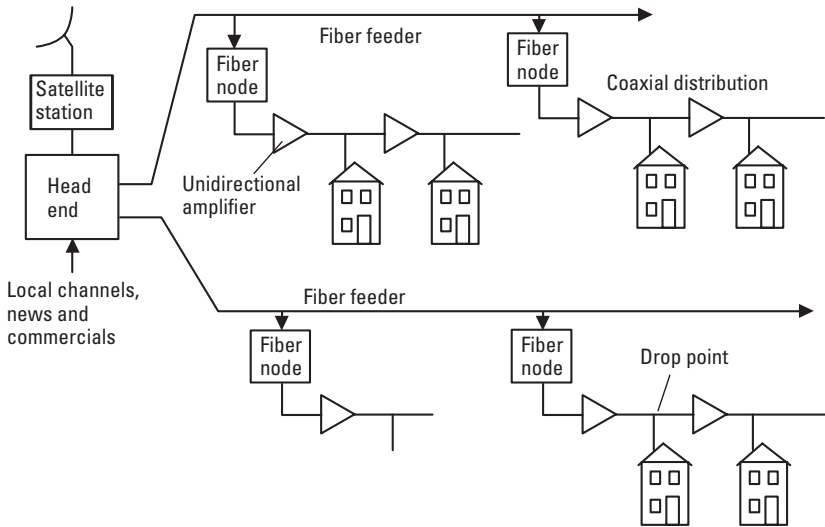
Figure 6.22 Traditional cable TV plant.
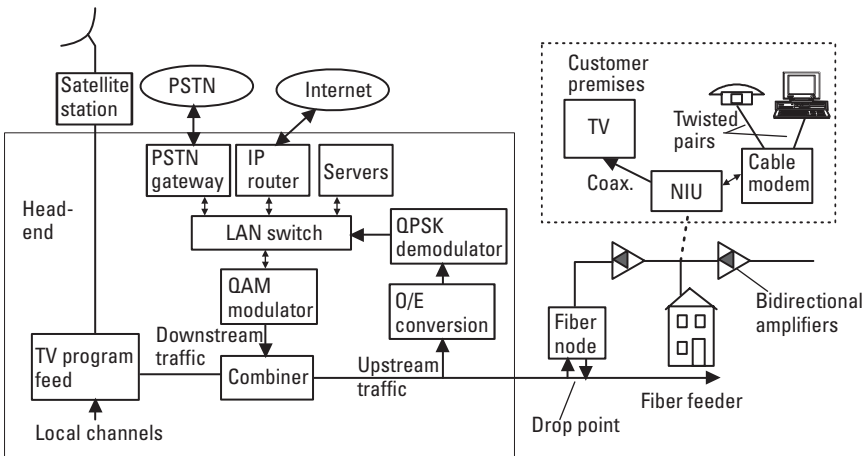
Figure 6.23 Cable TV plant modified for cable modem service.

voice use 6-MHz channels in the 5- to 42-MHz frequency range. Major modifications are required in the network to carry upstream traffic. First, strands of optical fiber must be allocated for upstream signals. The HE has to be equipped with a modulator and combiner for downstream and receiver

and demodulator for upstream signals as shown in Figure 6.23. Second, fiber nodes and coaxial cable amplifiers have to be changed to bidirectional devices. A customer premises *network interface unit* (NIU) splits up voice/data signals and TV broadcast channels. Data between the LAN switch at the HE and the cable modem at the customer premises are transmitted in standard 10BaseT/Ethernet frames.

In the downlink direction 64 QAM or 256 QAM with 6-bit or 8-bit symbols, respectively, is used and data rates around 30 or 40 Mbps are achieved through each 6-MHz downlink channel. Note that all users of the channel share this capacity.

Uplink frequency band is noisier because of the branching structure, which adds noise from all branches, when you approach the HE. In the uplink direction robust modulation scheme QPSK is used, restricting the total data rate via one 6-MHz channel to a few megabits per second. Another problem in the uplink direction is congestion when many users share the same channel. A cable modem may jump to another channel when severe congestion occurs. Uplink congestion can also be solved by assigning time slots at the HE. In this case the cable modem termination at the HE divides uplink channels into TDM slots and assigns those slots to end points that want to send data.

As we saw earlier, the cable TV network provides existing media for other services such as data and voice. It is an attractive alternative for high-speed data access and many cable TV operators offer it with better terms than telecommunications operators can provide DSL access. The major difference between the cable modem and DSL offerings is that users of a cable TV network share the data capacity and performance depends on the activity of other users. Another major concern is security because every user of a cable modem system may receive data directed to other users in the same fiber feed.

6.4.5 Wireless Access

DSL and cable TV access rely on existing cable networks and they are very cost-effective solutions for the operators that own the access network. They are not willing to lease their cable network on reasonable terms to their competitors although there is a lot of political pressure to open access network competition. Wireless technology for fixed access provides cost-effective broadband access alternative for new service providers with much lower initial investments.

Some operators use WLAN technologies with directional antennas to provide fixed wireless broadband access. In some countries special frequency

bands are allocated for this application. A basic wireless access system consists of a LAN at the customer premises and a radio relay system connecting the LAN via radio waves to a service provider's router that is connected to Internet.

6.4.6   Fiber Cable Access

Access via fiber optic cable is superior in terms of quality and bandwidth. Where deployment cost is justified by service opportunity, fiber optic cable is being deployed in the last mile from a telephone central office to the subscriber. This approach is known as *fiber-to-the-home* (FTTH) or *fiber-to-the-office* (FTTO).

To divide fiber cable investments between multiple customers, a fiber connection can be built from the central office to a multiplexing point from which copper cable access is provided to multiple customers. This access method is known as *fiber-to-the-building* or *basement* (FTTB) [5].

6.4.7   Leased Lines and WANs

An enterprise consisting of multiple offices in an area usually needs continuous information access among sites. For this purpose a public network operator leases cable pairs or optical fibers for the connection between offices (see Figure 6.24). This is often the most economical way to interconnect LANs when the distance is of the order of a few kilometers. The line terminals shown in Figure 6.24 may be HDSL terminals for copper cable or optical terminals for optical fiber depending on the required data rate and distance.

In the case of a long-distance connection, it is not economically feasible to build for each customer to build its own dedicated physical connection. This would require repeaters and separate cable pairs or fibers throughout the country. Instead the required end-to-end transmission capacity is leased from the core network of the long-distance network operator. For long-distance connections the operator uses the same high-capacity optical transmission systems that are used for the interconnections of public exchanges in the network (see Figure 6.24). The basic data rate unit of the provided transmission rate through the network is 64 Kbps corresponding to the capacity of one time slot in the PCM frame (see Section 4.5). This is why the telecommunications carriers provide leased-line services with data rates in multiples of 64 Kbps.

The four-wire baseband connection and leased-line long-distance connections just explained are common examples of leased-line connections. The leased line is connected all the time, but dial-up or switched lines are
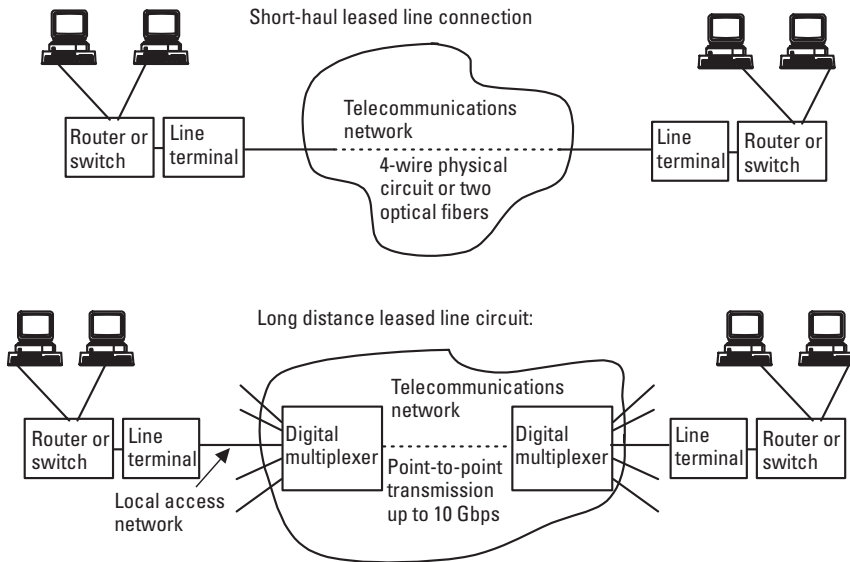
Figure 6.24 Regional and long-distance leased lines.

connected only on demand. However, the switched line requires higher investments in the network equipment and the fee is higher if the circuit is connected most of the time. In LAN interconnections the required capacity is often high and the connection is needed so frequently that the leased line often provides better service with a lower service cost in a regional corporate network. Another advantage of leased lines is high security because eavesdropping requires physical access to the dedicated channel.

Packet-switched alternatives are also available for long-distance interconnections. These are WANs and they use frame relay, ATM, or IP technology (see Sections 6.6, 6.7, and 6.8). They are often economically attractive because the core network costs are shared among many customers of the network operator.

6.5 LANs

The data communications systems that we described previously rely on the services provided by a public telecommunications network operator. However, there is a need for high-data-rate communications inside a building or a company; to satisfy these needs, local data communications networks, called LANs are built.

6.5.1   LAN Technologies and Network Topologies

LANs provide high-data-rate communications between computers, for example, inside one building. Because of the high transmission capacity (10 Mbps or higher) only short distances are allowed. The typical maximum transmission distance is a few hundred meters.

With help of the switching devices (switches or bridges) or routers, LANs can be interconnected to make up a wide-area corporate network. The bridges or switches interconnect separate LAN segments and switch frames from one segment to another with the help of a local hardware address that is stored in the interface unit of each computer. Routers are devices that use network layer addresses for the routing of packets and they are used to connect LANs to other networks, for example, to the Internet. Routers can also be used to interconnect LANs that use different technologies.

The basic structures of the two most common LANs, Ethernet and token ring, are presented in Figure 6.25. The original principle of all LAN networks is that all computers are connected to the same physical cable and they use it in turn. Information is sent in long frames that include the hardware addresses of both the source and the destination. These addresses are unique to each interface card plugged into a computer. Each computer listens to the cable and receives the frames that contain its own identification as a destination address.

Special protocols are standardized to make sure that only one computer transmits at a time. The complex standards of LANs specify OSI layer 1, the physical layer, and the so-called medium access sublayer (MAC) of layer 2 (the data link layer). The basic task of these protocols is to connect a
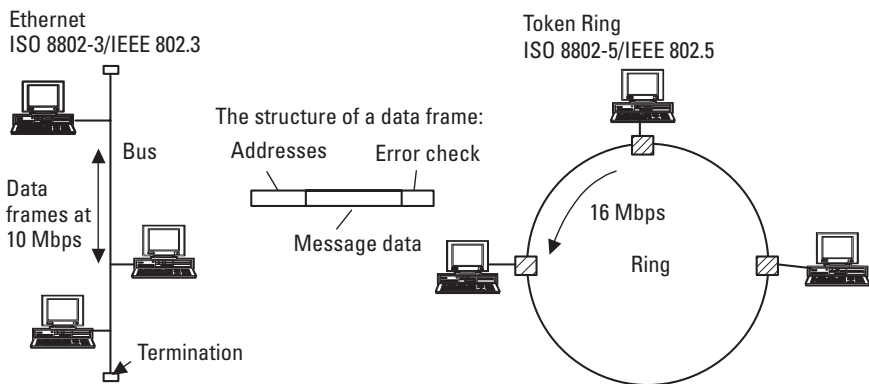
Figure 6.25 LAN structure.

computer to another via a shared medium as if they were connected by a point-to-point cable.

The most common LAN is the *Ethernet*, which has been standardized as ISO 8802-3 or ANSI/IEEE 802-3. Its original principle was invented by Metcalfe and Broggs and developed by Digital, Intel, and Xerox. It was called (DIX) Ethernet, and it became the de facto standard for LANs. The standardized protocols are not exactly equal to the original Ethernet but they can operate in the same LAN. An Ethernet LAN is logically a bus although its physical structure is often a star where all stations are connected to wiring center called a *hub*. We discuss Ethernet in more detail later in this section.

Another common LAN is the token ring, developed by IBM, and it is standardized as ISO 8802-5 or IEEE 802-5. The typical data rate of this LAN is 16 Mbps. In a token ring network, only a computer holding a special short frame called a *token* is able to transmit to the ring. The transmitted frame propagates via all computers in the ring and the station with the destination address reads it. The sending computer takes the frame from the ring and passes the token to the next station in the ring, which is then able to transmit. Physically the token ring is always built as a star although logically it still makes up a ring as shown in Figure 6.25. All computers are connected to a wire center that bypasses the workstations in the power off condition. When the power is switched on, the frames propagate from a workstation via a wire center to the next workstation in a logical ring. The token ring has some technical advantages over the Ethernet (no collisions, better bandwidth utilization, and deterministic operation) but it is much more complicated because of the token management and thus more expensive.

One important high-speed LAN is the *fiber distributed digital interface* (FDDI). Its operating principle is quite similar to that of a token ring but the data rate is higher, 100 Mbps. FDDI also allows longer distances and the maximum length of the ring is 100 km. The original transmission media of FDDI was optical fiber, but currently copper cables are also used for the connections between computers and a station attachment unit that connects workstations to the ring. The FDDI has been around since the 1980s and for many years it was the only technology that provided bandwidth higher than 10 or 16 Mbps. It was used as a backbone network to interconnect Ethernet or token ring LANs. Now that simpler high-speed technologies have become available the importance of FDDI has decreased.

There are many other standards for LANs but the vast majority of LANs in use utilize Ethernet technology because it is simple and inexpensive. In the following sections we concentrate on Ethernet networks.

6.5.2 Multiple-Access Scheme of the Ethernet

The MAC layer in the Ethernet is defined in ISO 8802-3/IEEE 802.3 and this access method is called CSMA/CD. This abbreviation stands for the following:

- *Carrier sense* (CD) means that a workstation senses the channel and does not transmit if it is not free.
- *Multiple access* (MA) means that many workstations share the same channel.
- *Collision detection* (CD) means that each station is capable of detecting a collision that occurs if more than one station transmits at the same time. In the case of a collision, the workstation that detects it immediately stops transmitting and transmits a burst of random data to ensure that all other stations detect the collision as well.

The original standard defined thick and thin coaxial cable networks operating at 10 Mbps. Many physical cabling alternatives have been added to the standard and the twisted-pair network 10BaseT has replaced most coaxial networks. In response to the increasing need for higher data rates in today's LANs, 100/1,000-Mbps Ethernet networks are released. The Ethernet offers a seamless path for the development of LANs into higher speeds while the present infrastructure of the network remains unchanged. To support this smooth development of LANs, the latest high-rate networks still use the same frame structure and the same managed object specifications for network management.

We now explain the operation of the CSMA/CD multiple-access scheme and the network structure of the original IEEE 802.3. The multiple-access method is most easy to understand with the help of bus-type coaxial cable network structure. Later in this section, we review the structure and operation of the twisted-pair and higher-data-rate variations of Ethernet.

6.5.3 CSMA/CD Network Structure

For collision detection it is essential to define the maximum delay of the network so that a station can be sure that transmission has been successful or collision has occurred (during transmission). In the case of a coaxial network, each cable segment is terminated by a 50-Ω resistor at both ends to avoid reflections. The maximum length of the cable segments and number of workstations (or transceivers) connected to each segment are specified. Thick coaxial cable (10Base5) specifications allow for a maximum section length of

500m and the maximum number of workstations in one segment is 100. A thin coaxial cable (10Base2) network allows a maximum section length of 185m and the maximum number of workstations in one segment is 30.

Thick coaxial cable was typically used in a backbone network that interconnects thin coaxial cable segments into which workstations are connected.

If the network is longer than one cable segment, repeaters may be used to regenerate attenuated signals. Repeaters are physical layer devices that retransmit signals in both directions. Logically the network remains a single physical network in which all frames are transmitted to every cable segment (see Figure 6.26).

Collision detection requires that the maximum delay not exceed a certain value and this restricts how many cable segments can be connected with repeaters. The definition states that the maximum number of repeaters in a 10-Mbps network between workstations is four and two of the segments between have to be link segments, which have no workstations.

If further extension to the network is needed, bridges or switches can be used. The physical size is then no longer a limitation because physical networks are now isolated from each other by a MAC layer device. It stores and forwards frames according their MAC layer addresses and acts as a separate workstation interface at each segment.

6.5.4 Frame Structure of the Ethernet

The MAC frame structure of IEEE 802.3/ISO 8802-3 is shown in Figure 6.27. Another popular frame structure that can be used is Ethernet II, which is also known as Ethernet V2.0 or DIX Ethernet or Ethertype or
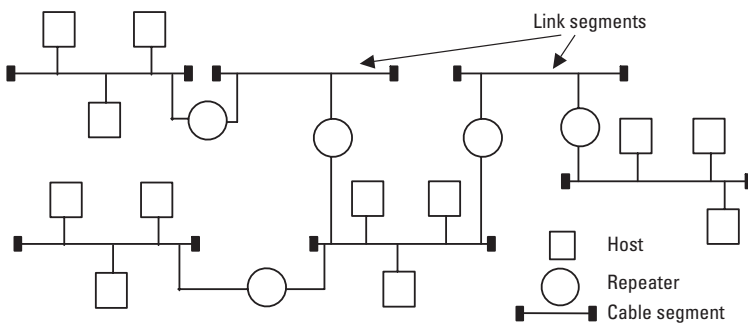
Figure 6.26 Example of a coaxial Ethernet network.
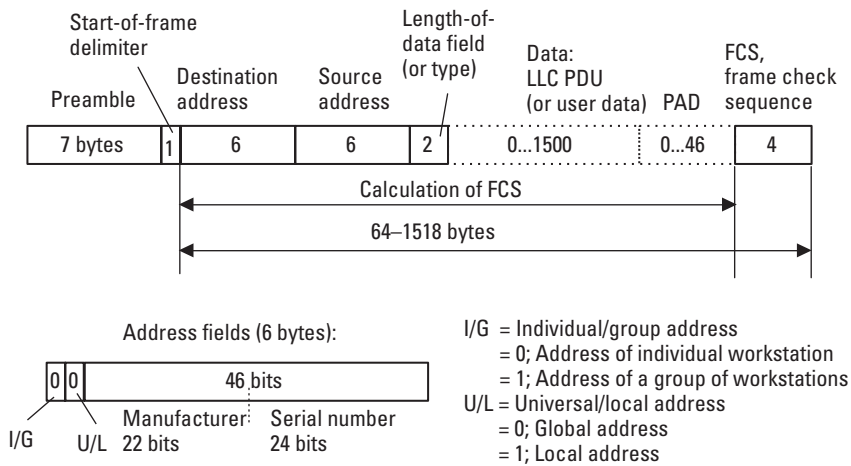
Figure 6.27 Frame structure of the Ethernet (MAC).

ARPA. They can coexist in the same LAN but communicating computers have to use the same frame structure. Now we explain the purpose and structure of the fields in the both popular frame types.

Each frame starts with the *preamble* of 7 bytes, each containing the bit pattern 10101010. The Manchester encoding produces a 10-MHz square wave that helps the receivers to synchronize with the sender.

The *start-of-frame delimiter* contains the bit sequence of 10101011 and indicates the start of the frame.

Both *addresses* contain 6 bytes, with the first bit indicating if it is the address of an individual workstation or a group address. Group addresses may be used for multicast where all stations belonging to the same group receive the frame. The second bit indicates whether the address if defined locally or if it is a unique global address. Normally global addresses are used and they are unique for each network card in any computer. The IEEE allocates an address range for each LAN card manufacturer [3]. When a card is manufactured, the manufacturer and serial number are programmed into it (see Figure 6.27). This ensures that no two cards will be using the same address in any network. Note that although these addresses are globally unique, they have only local importance. They are never transmitted to other networks.

If all stations in a LAN should receive the same message, all destination address bits are set to one. This is called a *broadcast address* and used, for example, by the address resolution protocol discussed in Section 6.6.

The *length-of-data field* indicates how many bytes there are in the data field, from 0 to the maximum of 1,500 (Hex 0000–05DC). If this number is higher than 1,500 in a frame, it cannot be an 802.3 frame. In this case the frame is a DIX Ethernet frame and a receiver interprets these two bytes as a protocol type information that defines a higher layer protocol. Here are some examples of type field hexadecimal values and corresponding higher layer protocols:

- *0800:* the *Internet Protocol* (IP) packet;
- *0806: Address Resolution Protocol* (ARP);
- *8137:* Novell IPX;
- *0000–05DC:* LLC, that is, the 802.3 frame.

The *data field* is where the PDU of the upper LLC sublayer of the data link layer is carried. In the case of DIX Ethernet (type field higher than 05DC hex.), the data field contains user data for the protocol identified by the type number.

For collision detection the minimum length of the frame is defined to be 64 bytes from the destination address to the checksum. If the data field is very short, the PAD field contains random data to extend the frame length to the minimum of 64 bytes.

The *frame check sequence* (FCS) is added to the end of the frame and with the help of this 32-bit check code the receiver is able to determine if errors have occurred. The 32-bit cyclic redundancy check (CRC-32) code is used for error detection. [The FCS is actually the remainder of the division when a binary number from the destination address to PAD (included) is divided by the specified binary number (in hex.: 10411DB7). The receiver divides the whole frame including the FCS by the same number and if the remainder is nonzero errors have occurred.] If errors are detected, the frame is discarded by the MAC layer and it is left up to the upper protocol layers to recover this situation. Note that if a frame is in error we cannot be sure that the destination address is correct and we may have received a frame that does not belong to us.

An IEEE 802.3 MAC frame (type/length 0000–05DC) data field does not give any information about the network layer protocol. However, it indicates that the data link layer contains an upper sublayer, LLC, on the top of MAC, carried in the data field. Network layer protocol is identified in the LLC PDU (in the MAC data field), which contains *destination service access*

point (DSAP) and *source service access point* (SSAP) numbers that define network layer protocol in the source and the destination machine.

6.5.5   CSMA/CD Collision Detection

Suppose that two stations both begin to transmit at the same time to the same cable. The minimum time needed to detect collision is the signal propagation time from one station to the other. However, in the worst-case scenario, the station cannot be sure that it has seized the cable until after two times the propagation delay. This is the case because the far-end station may transmit just before receiving the signal from the distant station. Then it takes another end-to-end propagation delay until this transmission is detected at the distant transmitting station. As a conclusion, a station can be sure that it has seized the cable and transmitted successfully after two times the worst-case propagation delay. As a consequence, to detect collision (before the transmission is finalized), the shortest frame has to be longer than two times the worst-case propagation delay. In the case of 10-Mbps coaxial network, the minimum frame length is 64 bytes and correspondingly the maximum length of the network is 2.5 km. The propagation speed in coaxial cable is approximately 70% of the speed of light and repeaters cause some additional delay.

6.5.5.1   Operation of Collision Detection

The Ethernet transmitter operates as a current generator (see Figure 6.28). When the pulse is transmitted, the current of –82 mA is driven to the cable and the pulse amplitude is –2V (25-Ω impedance). The Manchester line code used (see Chapter 4) gives the average current of about –41 mA when the transmission is on. The average voltage of the cable is monitored by an integrator (lowpass filter) and a comparator that compares average voltage in the cable with the threshold level, which is set to approximately 1.5V, as shown in Figure 6.28.

If two transmitters are active at the same time, each generates –41 mA on average and, with no attenuation taken into account, the average voltage is –2V with two active stations at a time. When three stations are active, the average voltage is –3V. If the average voltage goes below –1.5V, the output of the comparator changes state and the collision is detected (multiple stations are transmitting at the same time).

The principle just described is specified in the CDMA/CD standard (IEEE 802-3/ISO 8802-3). However, actual implementations may perform the collision detection in a different way. They may read signals back from

Operation principle of the transmitter:

Current of the transmitter:

Collision detection circuitry:

Voltage with one transmitter:
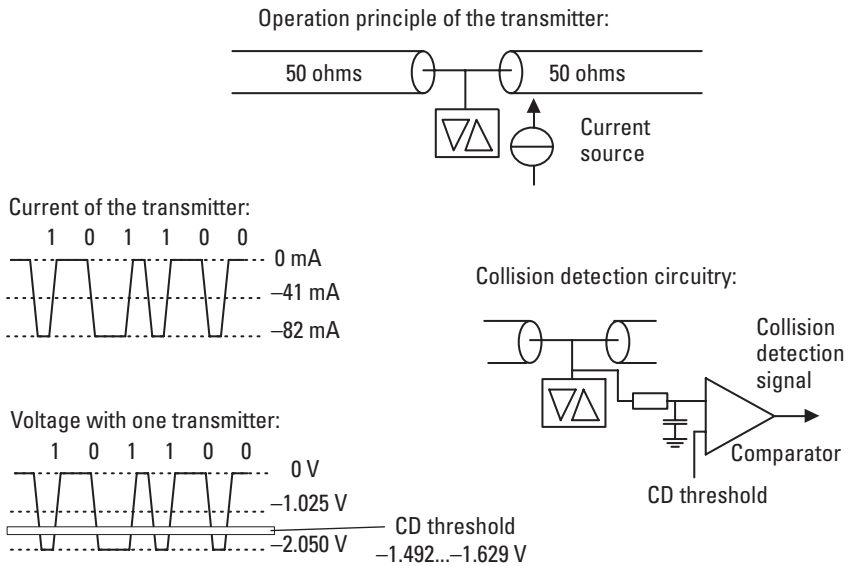
Figure 6.28 Collision detection in Ethernet.

the cable and if they are different from the original ones, collision is detected. They may measure the timing jitter of the pulse edges and detect collision if the edge locations in time do not occur at regular time instants. It is up to the manufacturer of the LAN cards to design the implementation as long as it is equal to or better than the one defined in the standard.

6.5.5.2   Contention Algorithm of CDMA/CD

Any station that has a frame to send may transmit at any time if the medium is free or at a transmission instant (see Figure 6.29). If more than one station decides to transmit simultaneously, a collision will occur. Each station that transmits detects the collision, aborts its transmission, waits for a random period of time, and then tries again (if no other station has started to transmit in the meantime and occupied the channel). Therefore, there will be alternating contention and transmission periods, with idle periods occurring when all stations are quiet.

6.5.5.3   Binary Exponential Backoff Algorithm

After a collision, time is divided into discrete slots with a length equal to the worst-case round-trip propagation time on the network. To accommodate the longest path allowed (2.5 km and four repeaters in coaxial network), the
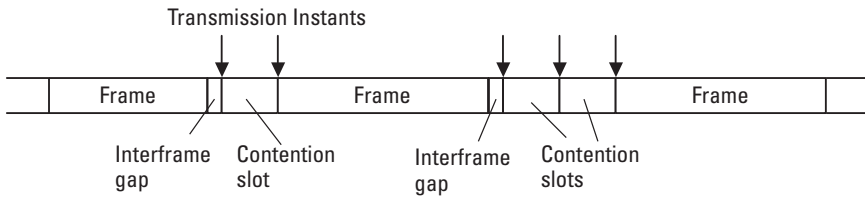
If more than one station transmits simultaneously, contention is detected
and both stations select a random number 0 or 1 and transmit again immediately or
wait for one contention slot (51.2 microseconds).
Depending on how many collisions have occurred, random number is selected from
the set $0.....(2^i-1)$, where i is the number of detected collisions
 If 10 or more collisions have occurred, selection range is 0...1,023.
When 16 collisions have occurred, the problem is reported to higher layers.

Figure 6.29   Contention algorithm of CDMA/CD.

slot time is set to be 512 bit times ($51.2\,\mu s$), the time that the transmission of
a minimum size frame (64 bytes) takes at the data rate of 10 Mbps.

After the first collision, each station waits randomly either 0 or 1 contention slot times before trying again. If two stations collide and each one picks the same random number, they will collide again. After the second collision, each station picks 0, 1, 2, or 3 at random and waits that number of contention slots. If a third collision occurs, then the next time, the number of slots to wait is chosen at random from the interval of 0 to $2^3 - 1$.

In general, after i number of collisions, a random number between 0 and $2^i - 1$ is chosen, and that number of slots is skipped. The probability of the next collision decreases with the number of previous collisions. After 10 collisions have been reached, the randomization interval is frozen at the maximum of 1,023 slots. After 16 collisions, the controller reports failure back to the computer. Further recovery is up to the higher layers and typically an error message is prompted. The probability of this situation is so small that it does not occur in normal operation but it may happen, for example, if the coaxial cable is cut off. Then each transmitted frame is reflected from the broken end of the cable and collision is detected for each transmission.

This described algorithm, called binary exponential backoff [3], was chosen to dynamically adapt to the number of stations trying to send. If the number of stations trying to send is high, a significant delay will result. However, if the stations had only options 0 or 1 from which to choose, and if there were 100 workstations, it would take years to have a successful transmission.

No simple mathematical solution is available to estimate CDMA/CD delays accurately. Practical experience has proved that to have reasonable performance out of 802.3 the loading has to be kept to the order of 40% or less on average of the maximum physical data rate.

The CSMA/CD as a MAC sublayer operation provides no acknowledgments and garbled frames are just discarded. If higher protocol layers use acknowledgments, they appear just like other frames in the network. Figure 6.30 shows an example in which there are three active stations in the CDMA/CD network. At time instant 0, both stations A and B transmit simultaneously and collision is detected. Then station A decides to transmit again but station B decides to wait for one contention slot time. Station C transmits at the same time as station A and a second collision occurs.

Now both A and C decide to skip one slot and station B seizes the network. Both A and C transmit when the network is free again and a third collision occurs. Now station A has suffered from three collisions and its range for the second transmission is zero to seven slots. Station A has now a wide range and it selects most probably a higher number than station C, which has had only two collisions. In the example shown in Figure 6.30, station C picks 1, waits one contention slot, and transmits. Station A picks 2 and transmits immediately when C has finalized its transmission.

Instant	Station A			Station B			Station C		
		Selection range	Selected number		Selection range	Selected number		Selection range	Selected number
0	T+C	0...1	0	T+C	0...1	1	L		
1	T+C	0...3	1	W			T+C	0...1	1
2	W			T			W		
...									
3	T+C	0...7	2	L			T+C	0...3	1
4	W			L			W		
5	W			L			T		
...									
6	T			L			L		

T = Transmits
C = Collision
W = Wait
L = Listen
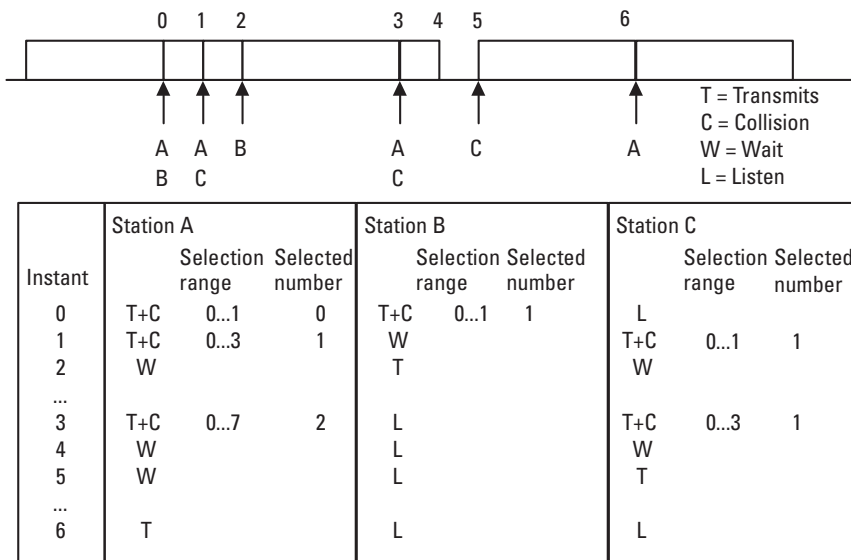
Figure 6.30 Contention example.

6.5.6   Twisted-Pair Ethernet

Ethernet today is always wired radially from a hub device or switch as illustrated in Figure 6.31. This is done for two basic reasons:

1. Bus cabled networks are difficult to manage and maintain. Faults in a segment are not easy to locate because a cable break anywhere in the segment prevents all communication. Also, addition of a new workstation or relocation of old ones is difficult.

2. Use of UTP copper cable is preferred because it is low cost, easy to install, and in most buildings spare twisted pairs are already in place. Attenuation of twisted pair is high and it cannot be used in a bus topology network.

The twisted pair CDMA/CD networks 10BaseT, 100BaseT, and 1000BaseT use twisted pairs to connect workstations to the wire concentrator, a hub. Twisted pair is easier and more flexible to install than coaxial cable and this has made 10BaseT very popular. In the simplest structure the concentrator or hub, which acts as a repeater, transmits frames from one workstation to all others as shown in Figure 6.31.

The 10BaseT system operates over two pairs of wires, one pair for receiving signals and one pair for transmitting signals. Each pair is terminated at the receiver input by matched impedance so that signal reflections
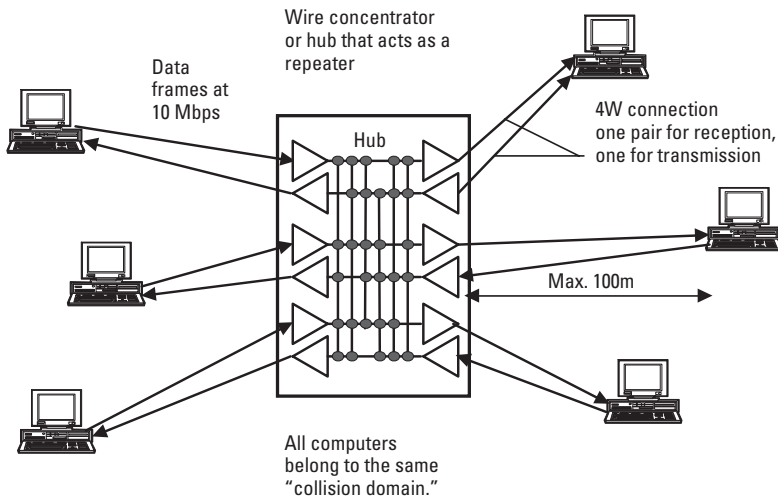
Figure 6.31   Twisted-pair shared media CDMA/CD.

are avoided. The maximum distance from workstation to hub is 100m for typical voice-grade twisted-pair cable. The hub contains electronics for signal reception, regeneration, and transmission. Note that logically the network in Figure 6.31 is still a bus in which all transmitted signals propagate to all other workstations. However, the major disadvantage of a physical bus is avoided because each workstation is separated from the bus by electronics and a break in one workstation's cabling does not disturb the operation of others.

In Figure 6.31 the signal from one workstation is forwarded to the reception pair of all other workstations, but not to its own reception pair. For collision detection the workstation merely needs to detect a signal on the reception pair. If a signal is received before its own transmission is terminated, someone else has transmitted at the same time and a collision has occurred. When a workstation is connected to a hub it operates in half-duplex mode, that is, it can either transmit or receive at a time.

The 10BaseT uses Manchester coding similar to that used by a coaxial network but symmetrically so that the signal voltage varies between +1V and –1V instead of 0V and –2V. Bit values are encoded into transitions ± or ± as shown in Figure 6.28 and care must be taken not to invert two conductors of a pair.

The requirements for a 10BaseT network collision detection feature to operate properly are exactly the same as those for a coaxial network. We have to take care that the worst-case propagation delay from one station to the most distant station does not exceed half of the transmission time of the shortest frame. At 10 Mbps this requirement is the same as in the coaxial networks, that is, there may be five cable segments and four repeaters (hubs) on the transmission path between two workstations.

The twisted pair restricts the transmission distance between a concentrator and a workstation to approximately 100m. Note that when the data rate is increased, the duration of the shortest frame is decreased and the maximum distance is correspondingly decreased.

A 10BaseT network can be extended by connecting hubs together with a twisted pair or fiber segment. In this structure one workstation in Figure 6.31 is simply replaced by another hub as shown in Figure 6.32.

In the network that we just described, all frames are transmitted to every segment in the LAN. We call this *shared media CDMA/CD* because all computers share the transmission capacity. The shared media networks shown in Figures 6.26 and 6.31 make up a single "collision domain"; that is, collision occurs if two or more computers anywhere in the network transmit so that two or more frames overlap.
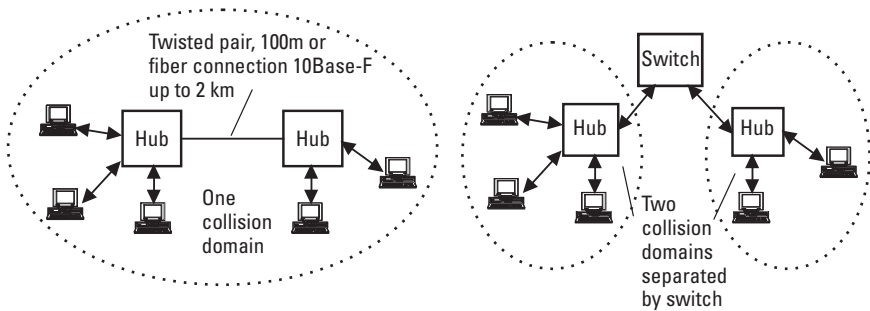
Figure 6.32   Network extension with and without a switch.

Bandwidth utilization of the shared networks (i.e., inside one collision domain) is poor because one transmitting workstation seizes the whole network although only the segments to the source and destination computers are needed for communication. As a rule of thumb, approximately 40 % of the network data capacity can be utilized on average in one collision domain; that is, all computers share the 4-Mbps transmission capacity of a 10-Mbps Ethernet network. If higher capacity is needed by workstations, the number of collisions increases, delays increase, and the average capacity used by successful transmissions decreases because frequent collisions occupy the channel. The switched LAN that is discussed next divides a LAN into multiple collision domains and the bandwidth utilization is much improved. However, broadcast frames propagate to all segments in switched Ethernet and the router is needed at the border of the broadcast domain.

6.5.7   Switched Ethernet Switches and Bridges

We can improve the performance of a CDMA/CD network by using switches (or switching hubs or bridges) instead of repeaters or hubs. Switches have replaced hubs in most CDMA/CD networks including coaxial ones where we called them *bridges*.

We may connect many repeater hubs as shown in Figure 6.32 to each other by a switch or a switching hub. The switches create separate collision domains because they do not forward collision signals from one port to another.

Switches do not transmit all received frames to all ports like repeaters. Instead they use MAC addresses and transmit frames to the direction where the destination is known to be located. Bridges are able to learn by listening

to the traffic. They read the source address in each frame and build up an address table containing all stations that have transmitted a frame as shown in Figure 6.33. If the location is not yet known, frames are transmitted to all ports. The address table is updated continuously to allow a workstation to move from one port to another.

In the network topology shown in Figure 6.31, we can change the repeater to a switch that is actually a fast multiport bridge. Now the frames from one computer to another are transmitted only from the source port to the destination port, and two other computers connected to different ports of the switch may transmit to each other at the same time. To allow this, the internal capacity of a switch must be much higher than the data rate at one port. Connections in 10BaseT use two pairs, one for reception and the other for transmission, and switches also allow full-duplex transmission between a workstation and the switch. Then, for example, two computers can transmit and receive simultaneously to each other and the maximum transmission capacity is increased from 10 to 20 Mbps. Note that collisions never occur in sections connected to ports 2 and 3 in Figure 6.33 because there is only one transmitter connected to each pair and the switch transmits only error-free frames further.
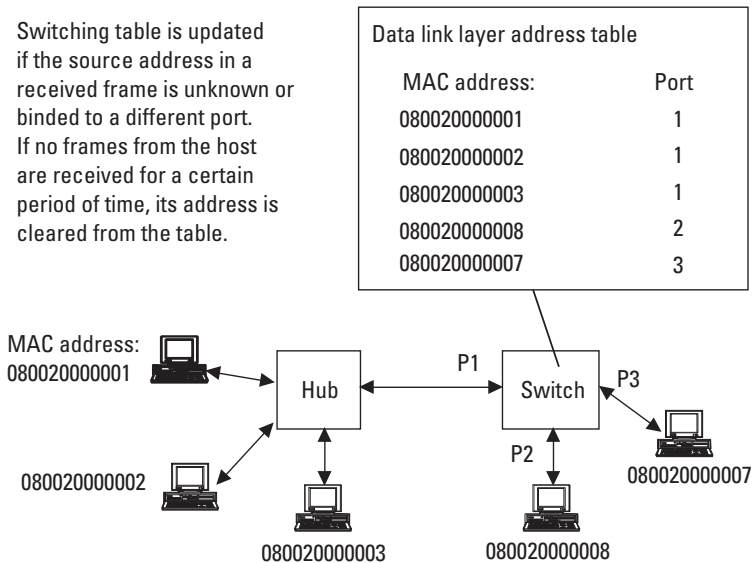
Figure 6.33 Address or switching table of an autolearning switch.

6.5.8   Fast Ethernet

The fast Ethernet standard is 100BaseT and carries data frames at 100 Mbps. This results in the reduction by a factor of 10 in the bit time, which is the amount of time it takes to transmit a bit on the Ethernet channel. Because 100BaseT operates at 10 times the speed of 10-Mbps Ethernet, all timing factors are reduced by the factor of 10. For example, the slot time is 5.12 μs rather than 51.2 μs. The maximum length of the network is shorter because of the shorter frame transmission time during which possible collisions must be detected.

The data rate is increased by a factor of 10 but the frame format and media access control mechanism remain the same as in coaxial Ethernet and 10BaseT. Only a 1-byte *start-of-stream delimiter* (SSD) and a 1-byte *end-of-stream delimiter* (ESD) are added in the beginning and end of the frame in Figure 6.27.

The topology of the 100BaseT network is equal to the 10BaseT shown in Figures 6.31 and 6.32. Connections between the workstations and a repeater are twisted pairs and their maximum length is 100m. The fast Ethernet standards include both full-duplex and half-duplex connections and operation over two pairs or four unshielded twisted pairs. Table 6.2 shows Ethernet technologies and their main characteristics. Media types show the required twisted-pair quality, where UTP category 3 means ordinary voice-grade twisted pair. The highest quality twisted pair is category 5 and its characteristics are specified up to a 100-MHz frequency.

Table 6.2
Preferential Order of Ethernet Technologies on Twisted Pair

Technology	Mode	Throughput/ Connection	Media
1000BaseTX	Full duplex	2 × 1 Gbps	4p UTP 5
1000BaseTX	Half duplex	1 Gbps	4p UTP 5
100BaseTX	Full duplex	2 × 100 Mbps	2p UTP 5/STP
100BaseT2	Half duplex	100 Mbps	2p UTP 3/4/5
100BaseT4	Half duplex	100 Mbps	4p UTP 3/4/5
100BaseTX	Half duplex	100 Mbps	2p UTP 5/STP
10BaseT	Full duplex	2 × 10 Mbps	2p UTP 3/4/5
10BaseT	Half duplex	10 Mbps	2p UTP 3/4/5

The Manchester coding used in 10-Mbps Ethernet is not suitable for higher data rates because it has very wide spectrum as explained in Chapter 4. Figure 4.18 shows that at 100 Mbps it has a strong spectrum at frequencies up to 200 MHz, which is too high for attenuation and the crosstalk characteristics of twisted pairs. To make the signal spectrum suitable for different quality cables, the following coding schemes are specified:

- The 100BaseTX uses 4B5B line coding in which four bits are encoded into a five-bit symbol to the line and the spectrum lies below 125 MHz, requiring category five cable pairs. For timing 5-bit symbols are defined in such a way that there are always transitions on the line signal for receiver synchronization.
- The 100BaseT4 uses 8B6T line coding (8 bits encoded into a symbol containing 6 three-level pulses) and it divides data between three pairs in each direction to manage with voice-grade pairs.
- The 100BaseT2 uses PAM5 encoding (five-level pulses) to reduce the spectral width and make it suitable for voice-grade cable pairs.

The segment length in all 100-Mbps networks is limited to a maximum of 100m to ensure that round-trip timing specifications are met. The fast Ethernet standard also specifies optical fiber connections that allow longer distances than a twisted pair.

Just as in the case of the 10BaseT system, we can extend or improve the performance of the 100BaseT network by using switches instead of repeaters (or hubs). Switches also allow full-duplex transmission in ports that use one of the full-duplex technologies shown in Table 6.2.

The fast Ethernet specifications include a mechanism for *autonegotiation* of the medium speed. This makes it possible for vendors to provide multiple-speed Ethernet interfaces that can be installed and run 10 Mbps, 100 Mbps, or 1 Gbps automatically. With the help of switches that support multiple data rates, we can gradually update the network and increase the data rate only where it is required. As a next step, the Gigabit Ethernet that we introduce later in this chapter further increases the capacity of the Ethernet networks. It provides a smooth path to gradually increasing the performance of Ethernet LANs.

6.5.9   Autonegotiation

The Ethernet specifications include mechanism for autonegotiation of the media speed as illustrated in Figure 6.34. Ethernet adapters can autosense

10-Mbps, 100-Mbps, and 1,000-Mbps operations, and with the help of this standardized feature it is possible to establish Ethernet networks that support all three speeds. Autonegotiation also detects whether a full-duplex (to switch) or half-duplex (to hub) operating mode can be used.

Figure 6.34 illustrates the autonegotiation process between the switch and *network interface cards* (NICs) of workstations.

Table 6.2 shows the order of priority of Ethernet technologies. When a workstation is connected to the network, an autonegotiation process takes place and the highest technology in Table 6.2 supported by both ends is used.

6.5.10 Gigabit Ethernet

The Gigabit Ethernet provides a 1-Gbps bandwidth with the simplicity of Ethernet at a lower cost than other technologies of comparable speeds. It will offer a natural upgrade path for current Ethernet installations, leveraging existing workstations, management tools, and training.

Gigabit Ethernet employs the same CSMA/CD protocol and the same frame format (with carrier extension) as its predecessors. Because Ethernet is the dominant technology for LANs, the vast majority of users can extend their network to gigabit speeds at a reasonable initial cost. They need not reeducate their staff and users and they need not invest in additional protocol stacks.

The Gigabit Ethernet is an efficient technology for backbone networks of Ethernet LANs because of the similarity of the technologies. As an example, for an ATM backbone network the frames of the Ethernet must be
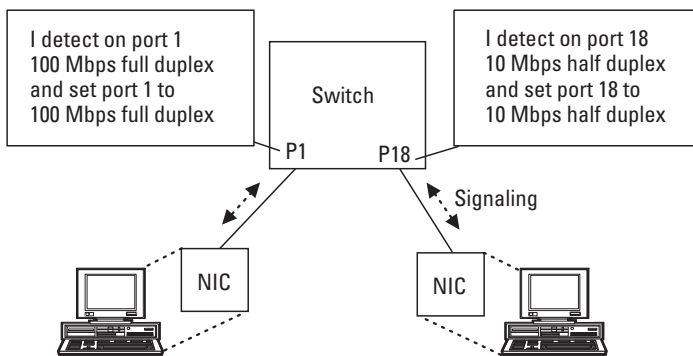
Figure 6.34 Autonegotiation between switch and NICs of workstations.

translated into short ATM cells and vice versa. The Gigabit Ethernet back-bone transmits Ethernet frames just as they are but at higher data rate.

The Gigabit Ethernet may operate in full-duplex mode, that is, two nodes connected via a switch can simultaneously receive and transmit data at 1 Gbps. In half-duplex mode it uses the same CSMA/CD access method principle as the lower rate networks.

The Gigabit Ethernet CSMA/CD method has been enhanced in order to maintain a 200-m collision diameter at gigabit speeds. Without this enhancement, minimum-size Ethernet frames could complete transmission before the transmitting station senses the collision, thereby violating the CSMA/CD method. Note that the duration of a frame is now only 1% of that at the 10-Mbps data rate.

To resolve this issue, both minimum CDMA/CD carrier time and the Ethernet slot time have been extended from 64 to 512 bytes. The minimum frame length, 64 bytes, is not affected but frames shorter than 512 bytes have an extra carrier extension. This so-called packet bursting affects small-packet performance but it allows servers, switches, and other devices to send bursts of small packets or frames to fully utilize available bandwidth. Devices that operate in full-duplex mode are not subject to the carrier extension, slot time extension, or packet bursting changes because there are no collisions.

Many transmission media have been considered for Gigabit Ethernet including multimode and single-mode optical fiber and category 5 twisted pair.

6.5.11 Upgrade Path of the Ethernet Network

As we have seen, the Gigabit Ethernet offers a smooth transition for an LAN to higher bit rates where they are needed. Often a greater bandwidth is first needed between routers, switches, hubs, repeaters, and servers. Figure 6.35 shows an example of how different Ethernet technologies could be used in the same network. In the example, we originally had a 10BaseT network in which the workstations were connected to repeaters or hubs. As capacity demand increased, repeaters were upgraded to switches at the same data rate or repeaters or switches at 100 Mbps. If 10 Mbps is high enough for an individual workstation, there is no need to update the network interface card of that computer as long as the port of the switch supports that data rate. The two 10/100-Mbps switches are 10-Mbps switches that were upgraded with a network interface card that connects them at 100 Mbps to the higher-level switches in the network. In the same way, the highest-level switches were upgraded with 1,000-Mbps cards for their interconnections and for the

To other
networks

100 Mbps

100/1,000 Mbps
switch

1,000 Mbps

100/1,000 Mbps
switch

Servers

Router

100/1,000 Mbps

100 Mbps

100 Mbps

100 Mbps
hub

10/100 Mbps
switch

100 Mbps
switch

10/100 Mbps
switch

100 Mbps

10 Mbps
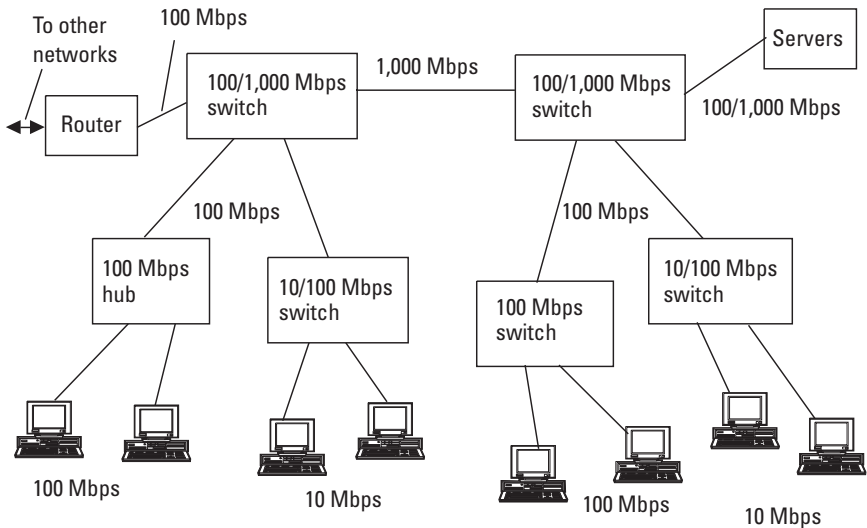
100 Mbps

10 Mbps

Figure 6.35 Ethernet network operating at 10, 100, and 1,000 Mbps.

connections to the servers. In the next upgrade we would probably replace the highest-level switches with genuine Gigabit Ethernet switches and use the old ones to replace lower-level switches or repeaters.

Ethernet technology is still evolving and it will soon support data rates of 10 Gbps.

6.5.12 Virtual LAN

A large physical LAN can be divided into many logical LANs. This improves the performance and security of the LAN because the traffic of, for example, a marketing department is separated from other traffic because it is on a dedicated logical LAN. A straightforward way to define a *virtual LAN* (VLAN) is to say that all computers connected to a certain ports of a switch make up one logical LAN and traffic is switched between these ports only.

Another more flexible way to configure a VLAN is to define the MAC addresses of all computers that belong to a certain network. This principle is more difficult to manage because a network manager has to define a VLAN for each MAC address. On the other hand, switches may dynamically configure themselves when a computer is transferred to a new physical location. This principle also allows a computer to belong to many VLANs. The third way to configure a VLAN is to define all computers using a certain network layer protocol to make up their own VLAN. The VLAN of a certain protocol

can be further divided into smaller VLANs by defining a certain set of the network layer addresses that make up a VLAN.

6.6   The Internet

The Internet has developed into the major information network in the world and this development will continue. We review here its development and the most important protocols on which its operation is based.

6.6.1   Development of the Internet

The worldwide Internet network developed from experimental computer networks in the 1960s to a worldwide university network in the 1970s and 1980s. The Department of Defense of the United States supported the original technical development. The aim was to design a fault-tolerant data network that would stay operational in crisis situations.

Internet technology is not as formally standardized as other public telecommunications networks. There is no globally authorized standardization body such as ITU-T where all nations together participate in the development of the network. However, some centralized control is required and there is an organization that manages the development of the Internet. The main institution in the *Internet Society* (ISOC) responsible for technical development is the *Internet Engineering Task Force* (IETF) in the United States. IETF updates Internet standards. Internet addresses, network numbers, are assigned by the Network Information Center (NIC) to avoid conflicts.

Technical specifications of the Internet are called *Requests for Comments* (RFCs) instead of standards. This gives the reader an idea about how official they are. This freedom in development of the Internet has speeded the growth of the network. Some RFC documents are approved and published as Internet standards, STD documents, by the *Internet Activities Board* (IAB). All organizations mentioned here belong to ISOC, which is a nonprofit international organization for global cooperation and coordination of Internet-related activities. Its headquarters is located in the state of Virginia in the United States [3, 5].

The Internet has been used by academics for more than 20 years. It used to be difficult to use, only some organizations had access to it, and the only users were academic specialists who were familiar with it. Because there were no commercial applications, usage charges were not considered at all for Internet technology. The academic information exchanged over the Internet

was public by nature and neither security nor charging functions were considered in the development of the Internet.

The development of a graphical user interface exploded onto the scene in the mid-1990s and the use of the Internet grew exponentially. This new graphical user interface is called the *World Wide Web* (WWW) browser and it has made the Internet easy to use for anyone. Nowadays many commercial *Internet service providers* (ISPs), who have access to the worldwide Internet service, provide Internet access for ordinary telephone and ISDN subscribers. Anyone who has a personal computer can access the Internet via the telephone or ISDN network. New higher data rate access systems, such as cable modems and ADSL discussed earlier, have become popular because they essentially improve customer access to Internet.

The Internet was originally designed for data applications only and it uses the genuine packet-switched transmission principle explained in Section 6.2. This is a very efficient method because the transmission connections in the network are used on demand. There is no circuit and fixed share of capacity for each user as is the case, for example, in ISDN. Because of this efficiency, the Internet will be used more and more for voice communications instead of PSTN. The usage of the Internet for international calls is very attractive because the international section of the call is often free of charge. No method exists to charge inside the Internet a certain type of usage and the user's fee is typically fixed or based on the time they are connected to the network of their ISP. However, because of the variable delay of packets, the quality of speech is not as good as in PSTN.

The Internet has turned into the major information network in the world, but problems that restrict its usage remain with this technology. The major problems are the inability to charge for services, lack of security, quality of the interactive real-time information, such as voice, capacity of the network when usage increases, and shortage of Internet addresses. However, the technical solutions to these problems are under development or implementation and the rapid growth of the Internet is expected to continue as more and more commercial applications become available. The Internet is expected to take a growing share of the telecommunications for which we presently use PSTN or ISDN.

6.6.2 Protocols Used in the Internet

Protocols used in the Internet are usually referred to as the TCP/IP protocol suite, which was introduced in Section 6.3 and is shown in Figure 6.11. As we saw, TCP/IP does not follow the OSI model exactly but it does follow a

layered structure. It was developed in the 1970s for fault-tolerant data communications, while the OSI model was designed to serve as a reference model for future protocol development.

TCP/IP is a collective term including all protocols in Figure 6.36, not only IP and TCP. The corresponding names of the OSI layers are given in Figure 6.36 although protocols in the figure do not exactly follow OSI specifications. Protocols in Figure 6.36 are, from the bottom up, as follows:

- Serial Line Internet Protocol (SLIP);
- Point-to-Point Protocol (PPP);
- Address Resolution Protocol (ARP);
- Reverse Address Resolution Protocol (RARP);
- Internet Protocol (IP);
- Open Shortest Path First (OSPF) Protocol;
- Routing Information Protocol (RIP)
- Border Gateway Protocol (BGP)
- Internet Control Message Protocol (ICMP);
- Transmission Control Protocol (TCP);
- User Datagram Protocol (UDP);
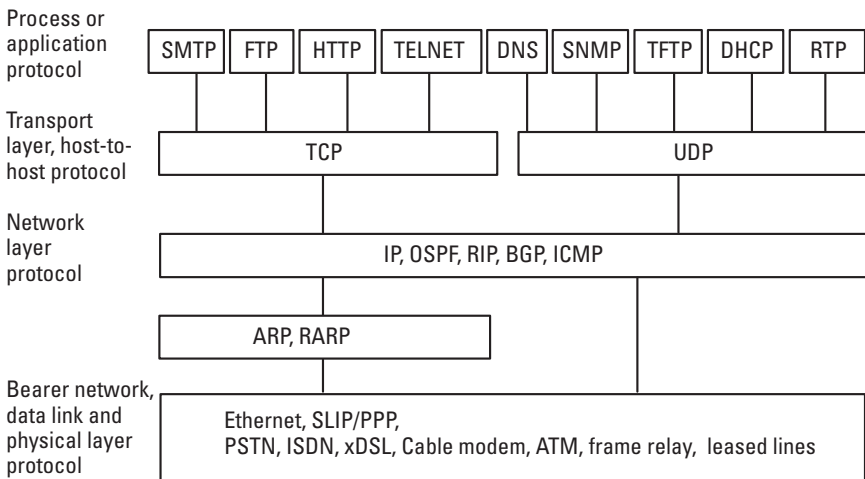
Figure 6.36 TCP/IP protocols and their interrelationship with the OSI reference model.

- Simple Mail Transport Protocol (SMTP);
- File Transfer Protocol (FTP);
- Hypertext Transfer Protocol (HTTP);
- Telnet, Virtual Terminal Protocol;
- Domain Name System (DNS);
- Simple Network Management Protocol (SNMP);
- Trivial File Transfer Protocol (TFTP);
- Dynamic Host Configuration Protocol (DHCP);
- Real-Time Transport Protocol (RTP).

We introduce these protocols later in this section. Figure 6.37 shows an overview of an Internet connection in which messages are routed independently by a connectionless IP layer from host to host with the help of the IP address. Each router has interfaces to many networks and based on the IP address it decides which way to choose. In the receiving hosts a message is given to either TCP or UDP depending on the protocol identification in the IP packet header. If connection-oriented TCP is used, a virtual connection is first established between hosts end to end. Then IP packets are transmitted, retransmitted if required, and reordered by the TCP layer so that the application feels as if there were a wire connection to the far-end hosts. The port number in the TCP or UDP packet identifies to which process or application the message is to be directed from TCP or UDP.
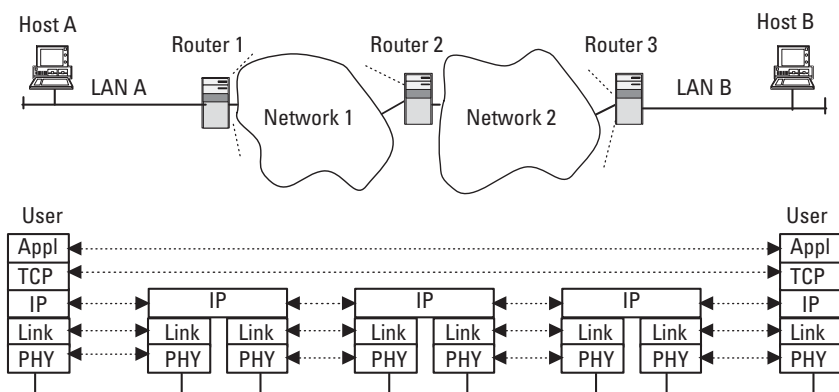
Figure 6.37 Example Internet connection.

Host A in LAN A in Figure 6.37 uses MAC addresses to communicate with router 1 as discussed in Section 6.5. At router 1, the data link layer protocol (MAC and LLC) is disassembled and only network layer data are forwarded to network 1, which may use point-to-point channels (without any MAC protocol) to router 2 at the other edge of the network. Router 2 knows from stored routing information that the destination IP address is located in the direction of router 3 and forwards the IP packet to that direction through network 2, which may use frame-relay technology below the network layer. Router 3 receives the IP packet, detects that it belongs to host B whose MAC address it knows, and attaches it to Ethernet frame together with the MAC addresses for itself and host B. The goal of this example is to make clear the relationship between MAC and IP addresses and to illustrate data flow through protocol stacks, which was discussed in generic terms in Section 6.3. In later sections we describe these TCP/IP processes in detail.

6.6.3   Bearer Network Protocols for IP

Most of the bearer networks mentioned in Figure 6.36 were discussed in Sections 6.4 and 6.5. Both physical and data link layer protocol functions are needed to transmit IP packets from a host to a router and from a router to another router, as shown in Figure 6.37. The Ethernet frame carries hardware or MAC addresses that make up a point-to-point connection over shared media between two machines as explained in Section 6.5. The Ethernet frame also contains error check and network layer protocol information as explained in Section 6.5.4.

Other bearer alternatives in Figure 6.36, such as PSTN and ISDN, carry data point to point and perform physical layer tasks. They do not contain data link layer functionality, and a protocol, such as SLIP or PPP, is needed to frame IP packets for physical transmission over serial lines.

SLIP is a simple framing protocol used to send IP packets across a telephone line. The problems with SLIP are that many incompatible versions of SLIP are in use, it does not do any error control, and it is not able to assign IP addresses dynamically. PPP is a more modern protocol, which solves all the problems of SLIP and can also send other protocols in addition to IP. PPP is used in many applications from residential Internet users to high-data-rate IP over SONET/SDH core network connections.

Typical application of these protocols is by a residential dial-up user of Internet service. To get access to the Internet, a user (or a modem) dials the telephone number of his ISP. The call is connected to a modem in the access server at the ISP's *point of presence* (PoP), that is, the point where access to

the ISP's service is provided. Now a point-to-point physical connection is established but to transmit independent IP packets necessary data link layer functions have to be implemented.

IP packets are transmitted every now and then and in the meantime errors on the line could be interpreted as IP packets if no framing and error control is implemented. To solve this problem, PPP uses the data link layer protocol, called *Link Control Protocol* (LCP), which performs framing of IP packets and error control. LCP also defines a negotiation mechanism, which is used in the beginning and end of the data link layer connection. End systems may, for example, agree to use frame numbering, acknowledgments and retransmission for error recovery. They are needed in wireless connections but not typically used in PSTN or ISDN connections.

Another problem is that the user's computer has no permanent IP address and before any communications an address has to be assigned. Typically each ISP has a much smaller number of IP addresses than customers. For dynamic IP address assignment, a Network Control Protocol (NCP) is used after data link layer connection is established by LCP. NCP assigns one of the ISP's IP addresses for the customer at the beginning of the connection and releases it at the end [3].

6.6.4   Internet Protocol

The IP is the core protocol of the Internet. It provides a service for the transfer of data units, datagrams, between the host computer and the router as well as between routers. At the IP level, each datagram is handled as a separate transfer and not as part of a larger data set.

The main task of the IP layer is addressing, which requires global Internet addresses, and routing of the IP packets from the source computer to their destination via a number of interconnected networks. The basic network elements in the IP network are routers and permanently connected computers (hosts) with different application protocols that provide services for Internet users. Each such element has at least one Internet address. They are different from the addresses used in the PSTN. The Internet addresses are global and their usage is internationally controlled by the NIC.

6.6.4.1   IP Addresses

Every host and router in the Internet has a unique fixed-length IP address, that defines the network and the host. No two machines connected to the global Internet have the same IP address. All IP addresses are 32 bits long and are used in source and destination address fields of IP packets. Figure 6.38 shows the format of an IP address.
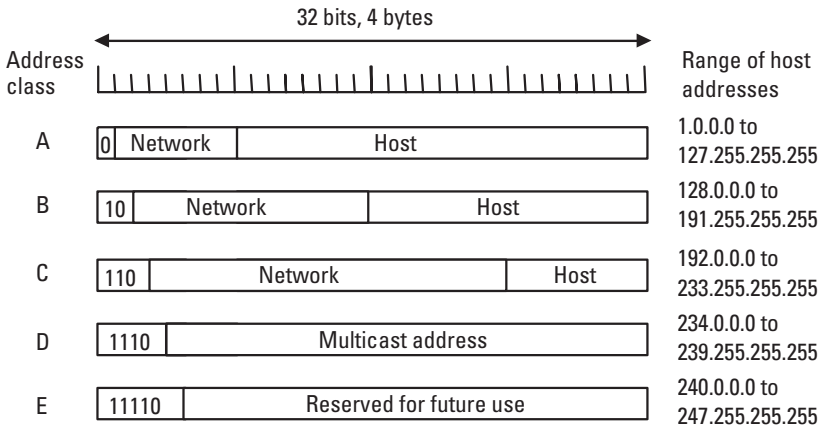
Figure 6.38 IP address format (IPv4).

Machines connected to multiple networks have different addresses on each network [3]. NIC assigns the network part, and the administrator of each network assigns the host part of addresses. IP addresses, which are 32-bit numbers, are usually written in dotted decimal notation as shown in Figure 6.38. For example, class C binary address 11000000 00101001 00000111 00110100, which is in hexadecimal form C0290734, is written as 192.41.7.52. Some addresses, such as lowest 0.0.0.0 and highest 255.255.255.255, have special uses, as shown in Figure 6.39 [3, 4]. Because
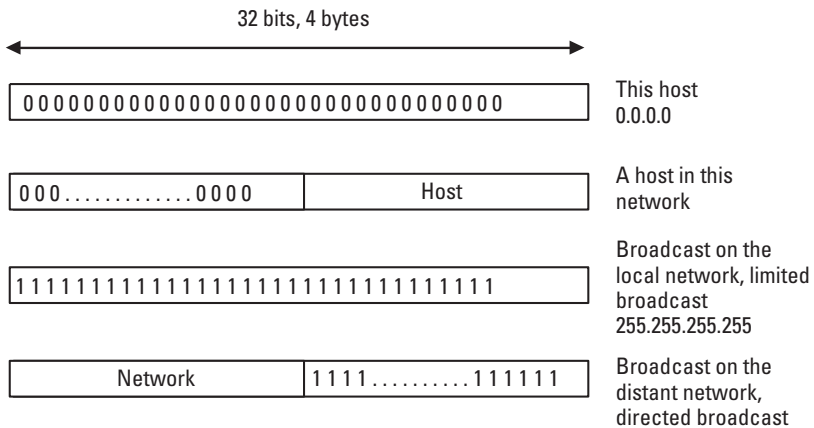
Figure 6.39 Special IP addresses.

of special use of all zero and all ones addresses a host address, where all bits in the host part are either zero or one, must not be used.

The IP address 0.0.0.0 is used by hosts being booted, but not afterward. IP addresses for later use may be assigned by the network for the host while it is booted. An address where all bits in the network part are zero refers to the current network, typically a LAN. All hosts in the network receive the IP packet with address consisting of all 1's. If only the host part is all 1's the packet is received by all hosts in the network identified by the network address.

6.6.4.2   Subnetworks

As seen earlier, all hosts in the network must have the same network number. A company that has one class C can have up to 254 hosts in its network and the use of these addresses have to be controlled over whole network, which may consist of multiple LANs. This could become a serious headache for network managers as the network grows and hosts are added and relocated. For easier management, a network can be divided into subnets so that a company's network still acts like a single network to the outside world. The network manager can decide to use, for example, two first digits in the host address section as a subnet address, as shown in Figure 6.40.

Now he or she may divide his or her network into four subnets, each containing up to 62 (0 and 63 are not used) hosts. If, for example, the class C network address is 221.109.65.0, the hosts are numbered from 1 to 254 (excluding 0 and 255). The 2 bits in Figure 6.40 identify four subnets and their host address ranges are 1 to 63. With subnet digits a whole 8-bit host part has values 1 to 63 (subnet 0), 65 to 127 (subnet 1), 129 to 191 (subnet 2), and
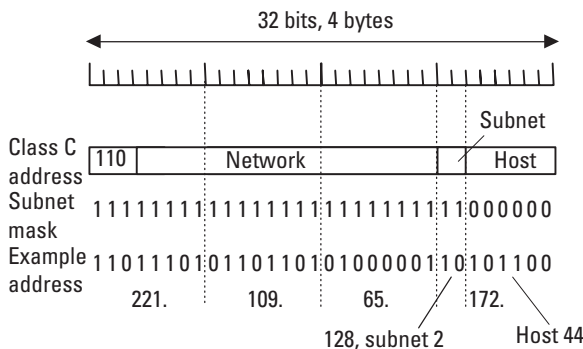
Figure 6.40   Example of subnet and subnet mask.

193 to 254 (subnet 3). Actually a few more addresses could be used without including hosts with all host part bits 0's or 1's, for example, 64 in subnet 1. However, in practice, it is probably better to follow the same addressing principles in all subnetworks. From the outside world, IP packets are routed with the help of a network number section and it is a matter for the company's internal staff to determine how host numbers are divided into subnets.

When an IP packet arrives at the router, it detects the address class from the first digits to see what section in the address represents the network. If it identifies its own network, the packet is forwarded to the host identified by the host part; otherwise, it is forwarded to the next router according to the stored routing table. If subnets are implemented, a subnet mask (shown in Figure 6.40) is defined and stored in the router. Now the received packet contains the router's network address and it performs a Boolean AND operation with the subnet mask to get rid of the host section. In our example, the result of this operation could give subnet address 221.109.65.0, 221.109.65.64, 221.109.65.128, or 221.109.65.192. This address is then looked up in the routing tables to find out how to get to hosts in a given subnet. The example address in Figure 6.40 is 221.109.65.172, which gives subnet address 112.109.65.128 as a result of an AND operation with a subnet mask and this indicates that the destination host is located in subnet 2.

Classless interdomain routing (CIDR) is a technique that divides network addresses into smaller address ranges in the same way subnets are divided as explained earlier. With the help of CIDR, for example, an ISP can split up its address range for assignment to its customers. For example, CIDR notation 128.211.176.112/30 defines a subnet mask of 30 bits and a four-address block with highest address 128.211.176.115.

6.6.4.3   IP Header

Each IP packet contains a header, as shown in Figure 6.41. *Version* specifies the IP protocol version being used, in this case, version 4. The *Internet header length* (IHL) specifies header length as a number of 32-bit words. The minimum value of IHL is 5 and the maximum is 15. With the *type of service* field the host may specify the datagram priority. It also contains flag bits D (Delay), T (Throughput), and R (Reliability), which the host can set to 1 to indicate about which feature it cares most. In practice, most routers ignore the type of service field altogether.

The *total length* field tells the length of the IP packet including the header and user data. It gives the total number of bytes or octets and its maximum value is 65,535 bytes. The IP layer may divide long datagrams into shorter fragments, which is necessary, for example, when the data are
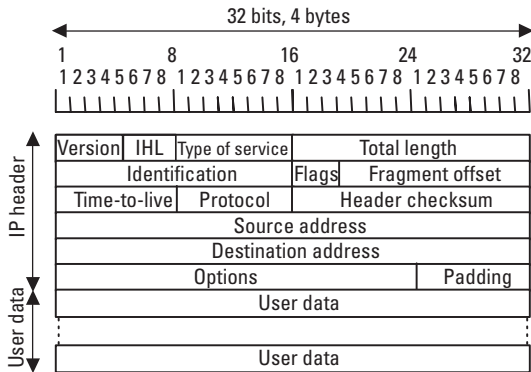
Figure 6.41 IP packet and header (IPv4).

transmitted over Ethernet, where the maximum size of a data field is 1,500 bytes. An Ethernet section almost always exists between end hosts, so long datagrams need to be split into segments with a maximum size of 1,500 bytes. Then we can say that the typical size of an IP datagram is not longer than around 1,500 bytes. The *identification* field has the same value for all fragments that belong to the same original IP datagram.

Flags contain one unused bit, a *don't fragment* (DF) bit, and a *more fragments* (MF) bit. By setting the DF bit the host can request the network to use a route where a datagram need not to be fragmented. When a datagram is fragmented all of its fragments, except the last one, have the MF bit set. All machines are required to accept fragments of length 576 bytes or less.

The *fragment offset* tells where in the current datagram this fragment belongs. The length of all fragments, except the last one, is a multiple of eight bytes. The fragment offset value tells where, in multiples of eight bytes, in original datagram this fragment starts. In the first fragment the offset is zero. The *time-to-live* field is supposed to count time in seconds and its purpose is to prevent datagrams for wandering around forever, which might happen if routing tables became corrupted. The maximum lifetime would then be 255 seconds. In practice, it is decremented by one on each hop. When it hits zero, the packet is discarded and a warning packet (ICMP message) is sent to the source host.

When the destination host has received all fragments, it assembles the complete datagram to be given to the higher layer protocol. The *protocol field* defines the higher layer and it is, for example, 6 for TCP and 17 for UDP. *Header checksum* verifies the header only. Every router must recompute it, because the time-to-live field changes at each hop. Source

and destination address fields contain the IP address described earlier. The *options* field needs not to be used but it may be useful for debugging routing problems. A network manager may, for example, set the options field to indicate that each router must insert its address to the options field. Padding fills up the IP header so that it contains complete 32-bit words.

6.6.4.4 IP Version 6

The main problems of IPv4 have been address shortages, poor security, and poor handling of real-time services. The new version, IPv6, was specified by IETF and its major goals [3] are listed next:

- Increase address space to support billions of hosts.
- Reduce the size of routing tables.
- Simplify the protocol to make the routing process faster.
- Provide better security.
- Improve quality of service, particularly for real-time services.
- Make roaming possible for hosts without changing its address.
- Permit the old and new protocols to coexist.

IPv6 specifications have been available for many years but version four is still the only one in wide use. However, it is clear that IPv6 will replace version 4 but it may take many more years, perhaps a decade. IPv6 header is shown in Figure 6.42 and it has a fixed length of 40 bytes.

The *version* field in Figure 6.42 has value 9 for IPv6 to tell the network layer how other fields should be interpreted. The *priority* field contains a higher value for higher priority. Values 0 to 7 are for typical data applications where transmissions can slow down in the case of congestion. For example, the recommended value for FTP is 4 and for Telnet 6 because a few seconds delay of one packet in the FTP stream is not noticeable but in the case of Telnet it is frustrating. Values 8 through 15 are for real-time traffic, such as audio or video, whose sending rate is constant.

The *flow label* field is zero for low-priority traffic and other values may be used to indicate to which data flow between source and destination hosts this packet belongs. Many flows may be active between a pair of hosts. Then routers may, for example, be set up to handle real-time traffic so that constant bandwidth is reserved for a certain flow.

The *payload length* field tells how many bytes follow the 40-byte header in Figure 6.42. The *next header* field allows extension to the header following
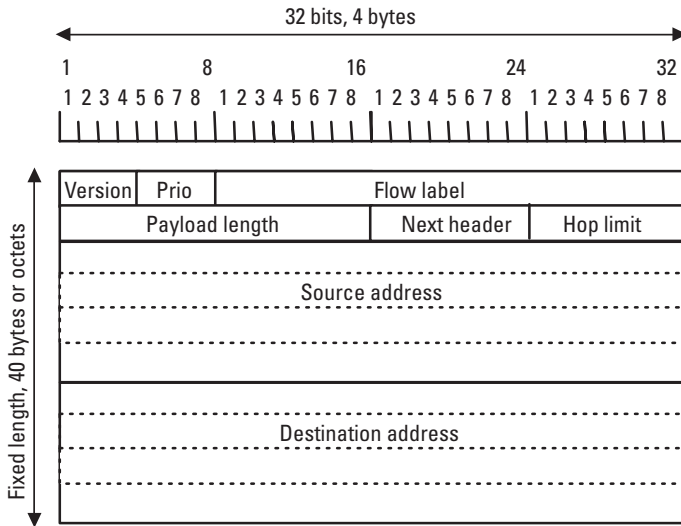
Figure 6.42 IPv6 packet header.

the fixed header in the figure. They are optional and, for example, the extension header for information exchange between routers for authentication and encryption is specified [3]. If there are no extension headers, the Next header field specifies a higher layer protocol such as TCP or UDP. The *hop limit* field prevents packets from living forever and it is used the same way as the time-to-live field in IPv4 header.

If we compare IPv4 in Figure 6.41 and IPv6 header in Figure 6.42, the IHL field is not needed in IPv6 because the header length is fixed. Protocol field is not needed because Next header (in the fixed header or in last extension header) determines the higher layer protocol, for example, TCP or UDP. Fragmentation field is not present in IPv6 header. A router that has received a longer packet than it can handle sends back an error message. Then the originating host splits up data into smaller packets. In this way, later transmission is much more efficient because routers need not fragment packets on the fly. Checksum is not present in IPv6 header in Figure 6.42 because its recalculation in every router reduces performance. The network is currently quite reliable and error checking is done at each data link layer of each connection and also end-to-end at the transport layer.

6.6.4.5 IPv6 Addresses

The most important problem with IPv4 is the shortage of addresses. IPv6 extends the address range so that approximately 3×10^{38} addresses are

available. If the entire Earth, land and water, were covered with computers, IPv6 would allow 7×10^{23} IP addresses per every square meter [3]. Addresses need not to be used efficiently and even in the most pessimistic scenario it is estimated that more than 1,000 addresses will be available for each square meter of earth surface. This seems to be far more than enough.

A notation that is used to write 16-byte addresses contains eight groups of four hexadecimal digits with colons between groups. An example address could look like this:

8000:0000:0000:0000:0ABC:DEF1:2345:789A

Because addresses have a lot of zeros, leading zeros in a group can be omitted and, for example, 0ABC can be written as ABC. If one or more groups are zero, they can be replaced by a pair of colons. The address above may then be written as follows:

8000::ABC:DEF1:2345:789A

An IPv4 address can be used in an IPv6 address field, and then the first 80 bits are zeros. This form of address can be written as a pair of colons followed by ordinary dotted decimal notation; for example: ::221.109.65.192.

IPv6 addresses start with a prefix, which defines which kind of address this is. For example, provider-based addresses have the prefix 010 and the following structure:

- Starts with 010 to indicate that this is a provider-based address.
- The following 5 bits define the registry where the provider can be found. There are operating registers for North America, Europe, and Asia.
- The next 3 bytes define the provider number.

A prefix is defined for local use addresses that have only local significance and each organization can use these addresses freely without conflict. They are not propagated outside organizational boundaries, and suit well those who have isolated their network by firewalls from the global Internet [3]. If messages with local destination addresses are transmitted by accident to the public Internet, routers see that this packet does not belong there and discard it.

6.6.4.6   IP Tunneling

As we have seen, an IP address consists of network and host sections and each host accessible from the Internet must have the IP address of the network where it is located. There are several reasons why routing should sometimes be based on another address, not the original destination host address. When we want to use the public Internet for VPN connections we might want to hide host addresses in IP packets for security reasons.

Another reason is to allow mobile terminals to roam to other networks and continue to use their home network address. *Tunneling* encapsulates an IP packet into a new IP packet, which has a new destination address of the current destination network to make roaming possible. Now routing is based on the outer IP header toward the network, where the mobile terminal is currently located. The original IP packet is de-encapsulated by a gateway in the destination network, which knows the route to the foreign destination terminal.

Tunneling is also used to improve the security of VPN connections. IP addresses of the gateways at the border of private and public networks are used for routing. The whole packet payload, including IP addresses of the destination and source hosts, can be ciphered and both data and the private network structure are hidden from outside world.

6.6.4.7   Mobile IP

The framework for IP mobility in the IETF is the *Mobile IP* (MIP). Its architecture and message flow are shown in Figure 6.43 [7]. In the MIP model, a mobile terminal has two addresses: the *home address* (HAddr) and the *care-of address* (CoA). The HAddr is the address that the terminal retains independent of its location. This address belongs to the home network of the terminal, which is the IP subnetwork to which the terminal primarily belongs. The CoA is the temporary address assigned to the terminal within a foreign network.

When the mobile terminal is located within its home network, it receives data addressed to the HAddr through the *home agent* (HA). When the mobile terminal moves to a foreign network, it obtains a CoA broadcast by the *foreign agent* (FA) in a *router advertisement* message as defined in RFC 1256. This CoA is then registered with the HA with a *registration request* message. Whenever a packet arrives at the HA addressed to the HAddr of the mobile terminal (1), the HA checks to see if the MT is currently located in the foreign network. In this case, the HA tunnels the packet within an IP packet addressed to FA (2). When the FA receives the packet it de-encapsulates it and forwards it to the mobile terminal (3). IP packets from the MT in Figure 6.43 are routed in the ordinary way and tunneling in that direction is not needed.
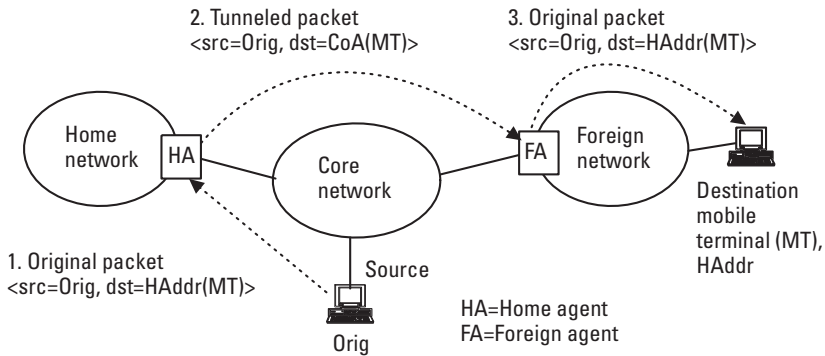
Figure 6.43 MIP architecture and message flow.

The advantage of MIP is that it relies on ordinary IP routing and the IP used by both source and destination needs no modifications. However, it introduces additional delay that is caused by tunneling of packets via HA.

6.6.5 Address Resolution Protocol

When a router is connected to other routers by point-to-point links, and it receives a packet to be routed, it just has to find out to which port it forwards the IP packet. However, usually a router connecting a corporate network to the Internet is connected to a LAN to which a set of hosts and other routers are also connected. To transmit an IP packet via a LAN to a certain host, the MAC address of the destination host must be known. Configuration files updated by a network manager could define a connection between MAC and IP addresses but this would be a permanent headache in large networks.

The Address Resolution Protocol maps an IP address to a MAC address as illustrated in Figure 6.44. A router of class C network 221.109.65.0 is connected to the external Internet and the network is divided into subnetworks 1 and 2 with subnet mask 255.255.255.192. When the packet from external network arrives (1), the router uses a subnet mask to find out the subnetwork of the destination host. In our example in Figure 6.44, the subnet mask gives 221.109.65.64 (2), which is subnet 1 connected to port 1. If the router does not know the MAC address of the destination, it sends an ARP request (3) carrying its own IP and MAC addresses in addition to the destination IP address. Broadcast MAC address (111...11) is used as a destination hardware address and the type field in the Ethernet frame is set to hex 0805 (2,054 decimal) to indicate that this is an ARP frame. All hosts in subnetwork 1 receive the frame and check if it contains its own IP address.

Hosts in subnet 1 are set to use default router 221.109.65.65
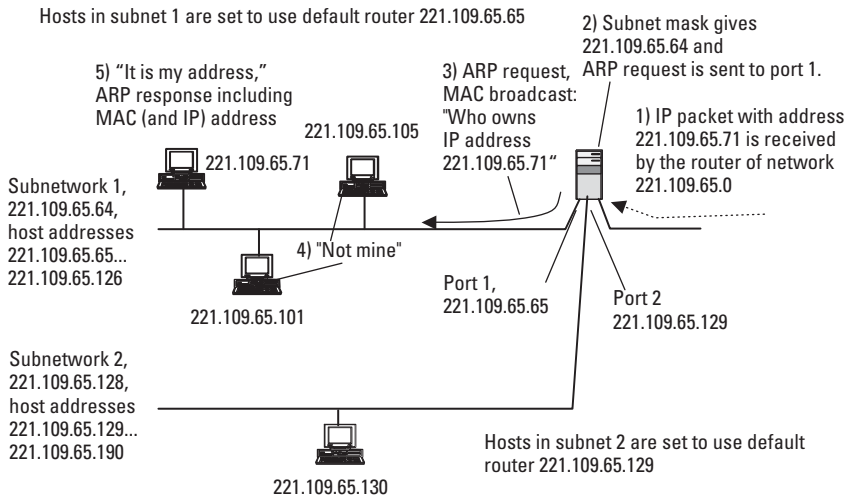
Figure 6.44 Operation of ARP.

The host having the IP address given in the ARP request frame responses with an ARP response frame containing its MAC and IP addresses. All hosts in the network may update their ARP caches by detecting ARP response frames and storing them for a few minutes. If one host in network 1 wants to transmit a packet to an IP address, it checks with the subnet mask to see if the destination belongs to my own network. If this is the case it uses ARP to find out the destination MAC address. If the IP address does not belong to the same network, a packet is sent to the default router (usually the first address of the subnet). The IP address of the default router, subnet mask, and own IP address are configured to all hosts.

Sometimes, for example in the case of diskless workstations, we might need to find out an IP address when the MAC address is known. The Reverse Address Resolution Protocol solves this problem. A newly booted workstation sends an Ethernet frame, where type field is hex. The 8035 (32,821 decimal) and destination address is the broadcast address. It actually asks: "My Ethernet address is given here; does anybody know my IP address?" The RARP server that takes care of IP addresses for diskless workstations responds with a RARP response containing the IP address allocated to that host.

6.6.6   Routing Protocols

As we can see in Figure 6.36, the network layer contains other protocols that control network layer routing. *Interior Gateway Protocols* (IGPs) are used

inside one *autonomous system* (AS) such as a LAN, and an *Exterior Gateway Protocol* (EGP) is used for communications between the exterior router and the other system [4].

One IGP protocol is OSPF, which is capable of dynamic updating of routing information. Routers exchange network topology and link status information with their neighbors and use that to derive the best, shortest path to the destinations [5]. Link metrics used in route computing include, for example, delay and throughput of the link. Another simple and popular IGP protocol is the *Routing Information Protocol* (RIP). A router running RIP broadcasts a routing update message every 30 seconds, which contains IP addresses and distances (hop counts) to those networks. All stations and routers running RIP update their tables accordingly [4].

One EGP protocol is the *Border Gateway Protocol* (BGP). The BGP router computes paths and tells its neighbors which routes it is using toward destinations [3]. Based on this information, neighbors select their own routes. For example, if router A tries to derive a path to a certain network C, and its neighbor B has told that it will use A as a second step toward network C, router A knows that it should use other neighbor routers instead of B.

Exterior gateway routing protocols, such as BGP, are designed to allow many kinds of routing policies, which are considered in routing decisions. Policies are manually configured into each EGP router. Typical policies involve political, security, or economic considerations. For example, a telecommunications network operator is happy to act as carrier of the traffic from its own customers, but not the others.

6.6.7 ICMP

The ICMP in Figure 6.36 is another protocol for network layer control. It is used especially for communications between the router and the sending host computer as shown in Figure 6.45.

ICMP must be implemented into every network element equipped with IP and it provides a means to communicate between the IP software of a host and a router. The value 1 (decimal) in the protocol field of the IP datagram in Figure 6.41 identifies the ICMP.

Communication problems may lead to discarding of the IP datagrams as shown in Figure 6.45 where router 3 is not able to transmit the datagram further. It then uses ICMP to inform the source host about the reason for the problem. The ICMP message contains a type field indicating the type of the problem and a set of parameters that may help the host to decide how to resolve the problem. Types of problems include, for example, the following:
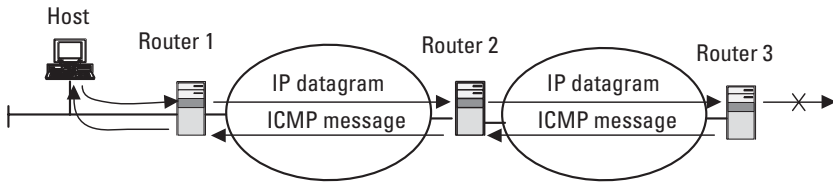
Figure 6.45 Operation of ICMP.

- *Destination unreachable:* The destination host is, for example, in the power-off condition.

- *Time exceeded:* Time-to-live field is subtracted to zero.

- *Source quench:* The source should reduce its transmission speed because the destination is not able to process incoming data or a router on the way does not have enough buffer space for IP packets.

- *Redirect:* A router finds that the source and next router are connected to the same network (according to the routing table) and requests the source to transmit the packet directly to the other router.

For a more detailed description of ICMP the reader should refer to [4].

6.6.8   Structure of Internet and IP Routing

The Internet can be seen as a collection of subnetworks or autonomous systems that are connected as shown in Figure 6.46. An IP Ethernet LAN of a company providing Web service and an ISP's network in Figure 6.46 are two examples of autonomous systems connected by regional and backbone networks. There is no real fixed hierarchy but there are several backbones, which consist of high-data-rate lines and fast routers [3]. Regional networks are attached to the backbones and LANs at universities, companies, and ISPs.

Figure 6.46 shows an example in which one European residential customer uses his ISP to access the Internet to view the Web page of a company in the United States. The glue that holds the Internet together is the Internet Protocol. For communications, the IP address from the ISP's address range is provided to the customer who wants to access the service. If the customer knows the uniform service locator of the Web page he wants to view, it is translated into the destination IP address and the exchange of IP packets can start. In the following sections, we introduce higher layer protocols that are needed to interpret information in the payload of IP packets properly.
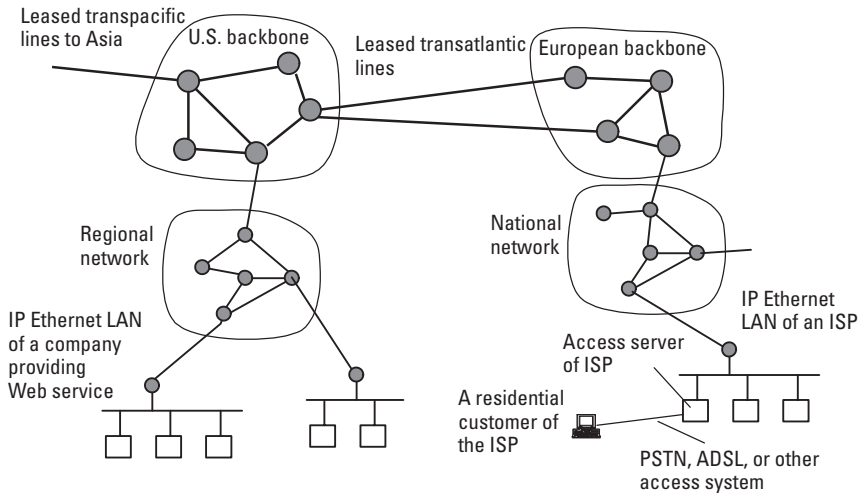
Figure 6.46 Interconnection of IP networks via IP switching network.

6.6.9 Host-to-Host Protocols

As shown in Figure 6.37, the protocols on the top of the network layer oper-
ate end to end. The two available options for end-to-end transmission service
are TCP for reliable connection-oriented communications and UDP for
connectionless datagram communication. Which one is used depends on the
needs of the application protocol on the top of them as shown in
Figure 6.36.

6.6.9.1 TCP

IP provides connectionless datagram transmission through the network. This
means that it routes each packet of data independently by using the IP
address in each packet. Most applications, such as file transfer, require that
the packets arrive in the original order and if one of them is in error it must
be transmitted again. The procedures required for these functions are imple-
mented in the TCP. The TCP that runs only in the data source and the des-
tination machines provides connection-oriented reliable communications
over connectionless IP network. To do this, it establishes a logical connec-
tion, determines if errors have occurred in packets, retransmits packets in
error, and rearranges the packets if they arrive out of order.

Figure 6.47 shows an overview of multiplexing and addressing in
TCP/IP protocol stack. The host is identified by the IP address and if the
protocol field in the IP packet has the value 6 (decimal), the payload of the
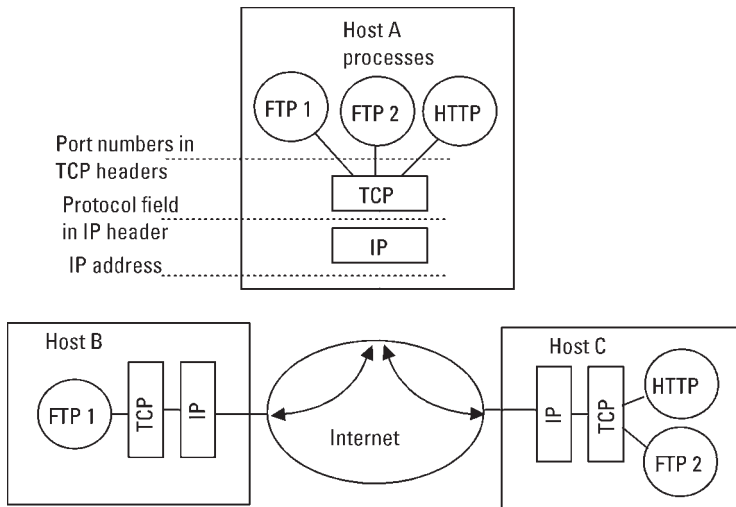
Figure 6.47 Addressing and multiplexing of TCP/IP.

IP packet is given to the TCP above. The TCP header contains *source* and *destination* ports, which define the processes that exchange information.

In the example in Figure 6.47, two FTP and one HTTP application processes run simultaneously on host A and they are distinguished with the help of the port numbers. IP packets carrying FTP 1 data contain IP addresses of hosts A and B for routing through the Internet while others carry IP addresses of hosts A and C.

Figure 6.48 shows TCP header fields that follow the IP address in the IP packet. The source and destination ports define the source and destination processes, respectively. Some standard server port numbers are defined, such as 25 (decimal) for SMTP and 23 for Telnet, and a set of them are available for use on demand.

The *sequence number* is the number of the first byte in the data segment carried in the TCP message. The *acknowledgment number* specifies the number of the first byte in the next segment expected to be received. The *data offset* reveals the number of 32-bit words in the TCP header; that is, it tells where the data section starts. All 6 bits in the reserved field are set to 0. The following 6 control bits are used as follows:

- URG is set if urgent pointer is in use.
- ACK is set if acknowledgment field is in use (always when the connection is set up).
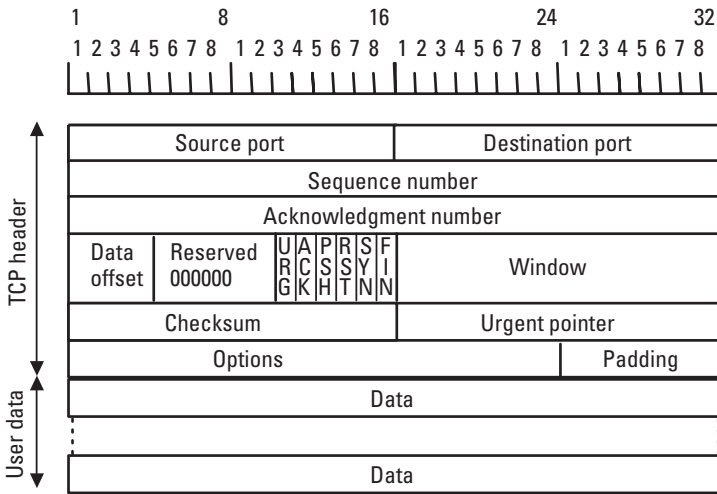
Figure 6.48 TCP header.

- PSH is set if data should be forwarded immediately to application process, no more segment is waiting.
- RST is set when the TCP connection is not accepted or to terminate a connection when a problem has occurred.
- SYN is used in connection establishment for synchronization of sequence numbering
- FIN indicates that this is the last data segment and there is no more to transmit.

The *window* indicates the number of bytes allowed to be sent by the other party without acknowledgment starting from the value given in the acknowledgment number field. The *checksum* is calculated over the entire TCP segment, header and data, and it is used for error detection. If the URG control bit is set *urgent pointer* tells the offset of urgent data relative to the current sequence number. The most important use for the *options* field is to indicate at the far end the maximum segment length that the host can handle. All Internet hosts are required to accept TCP segments of 536 + 20 = 556 bytes. Maximum segment size can be different in different directions.

6.6.9.2 TCP Connection Management

For establishment of a TCP connection, three-way handshaking, as shown in Figure 6.49, is carried out. A server that accepts TCP connection requests has
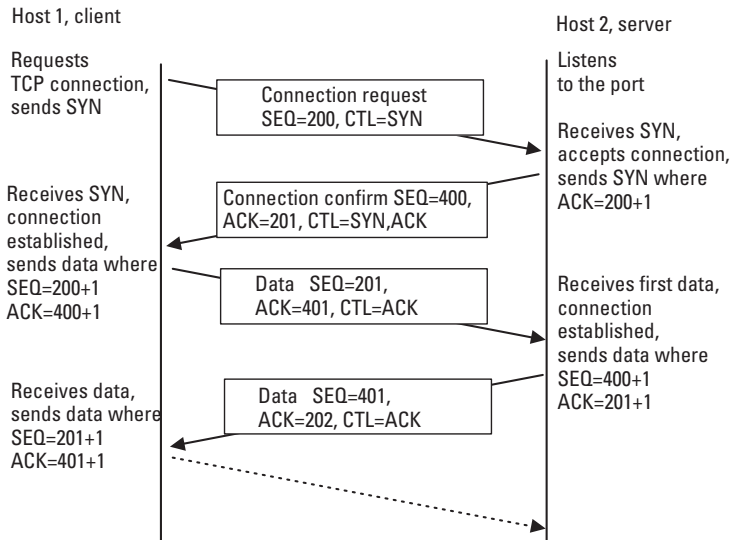
Figure 6.49 TCP connection establishment.

TCP software running and waiting connection requests to its port. For example, Web server waits connection requests to port 80 and DNS server to port 53. The combination of IP address and port is called a *socket* and it defines where the source or destination process is found.

A client uses one of its free ports and inserts its port number and the destination port number into the TCP header in Figure 6.48. It sets SYN bit on, ACK bit off, and writes the sequence number (which can be any number), for example, 200, to the TCP header. With the help of the options field it also attaches the maximum segment size it is able to handle. Optionally some user data could also be attached, for example, a password. Then the TCP segment is attached to the IP packet including source and destination IP addresses and sent to the server or first router on the way to the destination identified by the destination IP address. A retransmission timer is started at the time of transmission and if the response does not arrive before the timer expires, the connection request is retransmitted.

The packet is routed to the destination host, server, and to the identified port of which server process is listening. If the server accepts the connection, it replies with a TCP segment including its own sequence number, for example, 400, control bits SYN and ACK on, and acknowledgment number one higher that the received sequence number, in this case, 201. The server starts the retransmission timer when it transmits the connection confirm

message. When the client receives the segment it knows that the far-end computer accepts the connection and it has also understood the attached sequence numbering. If the server does not accept the connection, the RST bit is set in its TCP header to reject the connection request.

The server cannot yet be sure that the client has received its message properly and a third message is required until the two-way connection is established. The third message in Figure 6.49 may contain data, at the time of transmission the retransmission timer is started, and if the response from the server is not received in time, the client starts the entire process again. If the server receives a message properly it knows that the client follows its sequence numbering and a two-way connection is established.

When the connection is established, data transmission proceeds as shown in Figure 6.49. At each transmission an instant retransmission timer is started. If acknowledgment does not arrive in time, the timer expires and retransmission takes place. If there is no need for data transmission in one direction, TCP segments are sent without data to acknowledge the received segments before the retransmission timer expires. If the transmission delay is very long, for example, in satellite channels, TCP acknowledgment segments must be send by the transmitting Earth station, not by the destination host. Longer timers are then used over satellite hop. The receiving Earth station acts as a TCP source machine when it delivers data to the destination.

For efficient transmission, the data source should be allowed to send many TCP segments without waiting for separate acknowledgments for each of them. Otherwise, especially if the transmission delay is long, the source machine would spend most of the time waiting for acknowledgments. TCP uses the sliding window principle shown in Figure 6.50 to make transmission efficient and uses it for flow control as well.

The window field in each received TCP header (see Figure 6.48) tells the receiving party what transmission window size it should use. It defines the number of bytes that can be transmitted without acknowledgment as shown in Figure 6.50. Each time when acknowledgment is received, the pointers in Figure 6.50 are shifted to the right and new bytes can be transmitted. The two transmission directions are independent and both machines manage both transmission and reception windows separately.

Figure 6.51 shows an example of how the transmission window is managed. The reception window size of host 2 is 512 and it has told host 1 to use this transmission window size. Host 1 transmitted bytes 5 to 11 earlier and it has received acknowledgment up to byte 4. Now host 1 is allowed to send all 512 bytes inside its transmission window without waiting for further acknowledgments.
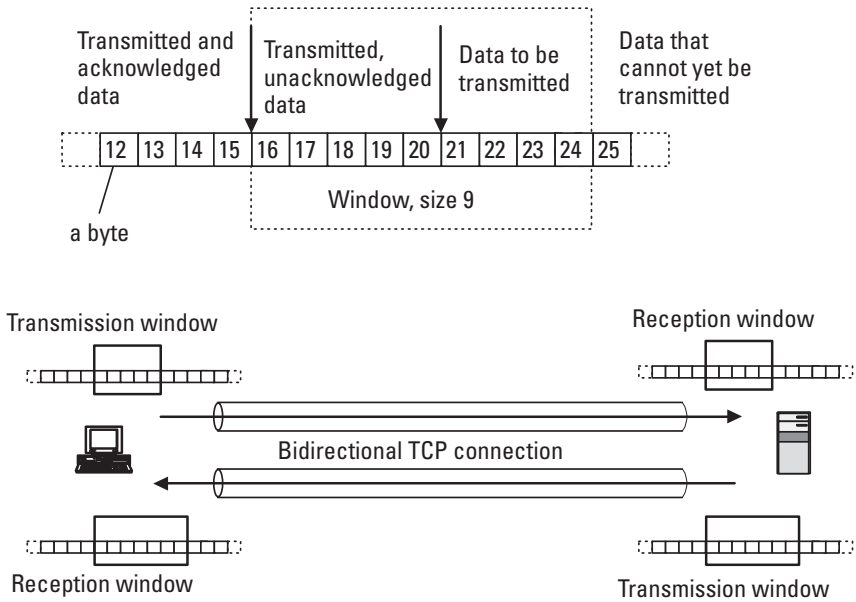
Figure 6.50 TCP window management.
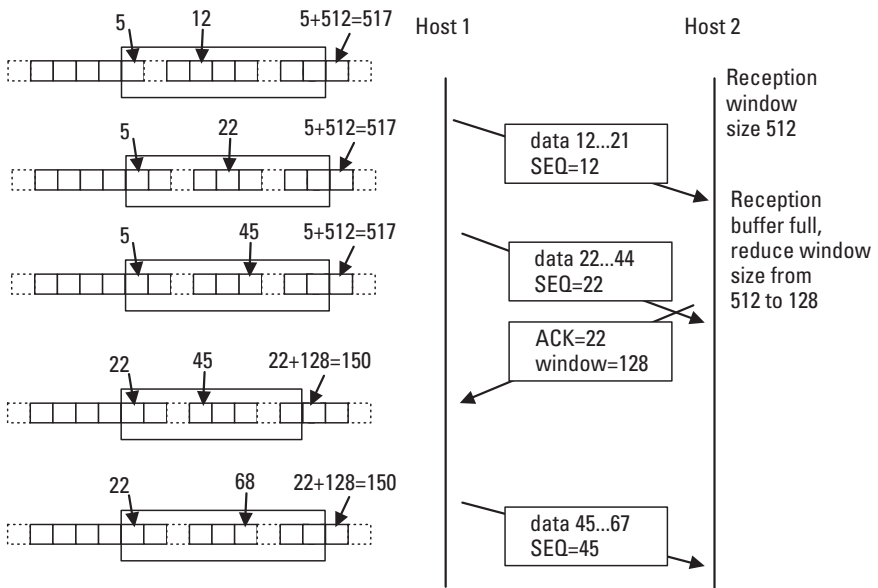
Figure 6.51 TCP window management example.

When acknowledgment is received the window is transferred so that it contains unacknowledged bytes only. In the example in Figure 6.51, free space in the receiver's buffer is reduced and host 2 requests host 1 to reduce its transmission window size to avoid buffer overflow. If the buffer becomes full, the receiver may command the transmitter to stop data transmission by setting the window size to zero.

The TCP connection is closed as shown in Figure 6.52. When an application program tells TCP that it has no more data to send, TCP will close connection in one direction. The sending TCP transmits the remaining data, waits until they are acknowledged, and then sends the TCP segment with control bit FIN in the Figure 6.48 set. The receiving end acknowledges the FIN segment, informs the application, and one transmission direction is closed [4].

When all data in the other direction have been transmitted, TCP in host 2 transmits the FIN segment. When acknowledgment of that is received by host 2, the TCP connection is closed in both directions.

The destination and source hosts use a window for flow control as explained above. The problem that cannot be managed by the end-to-end windowing mechanism is congestion in the interconnecting network. If the
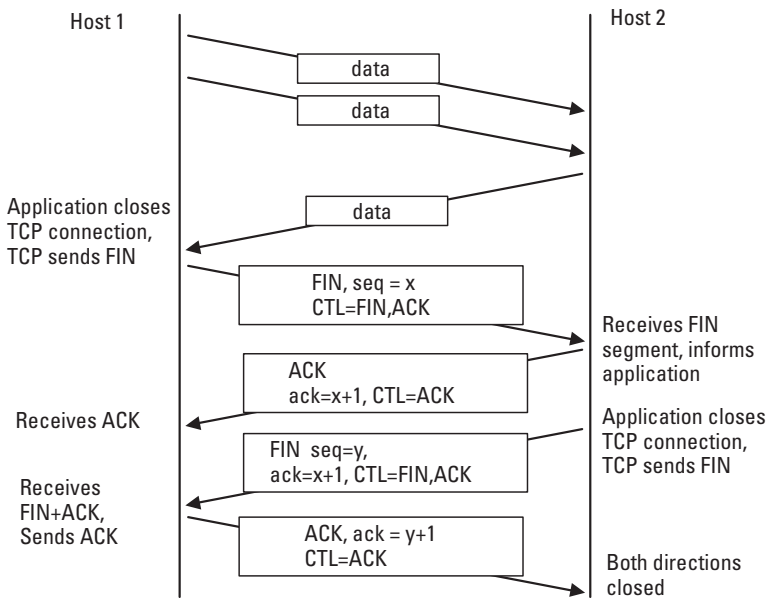
Figure 6.52 Closing the TCP connection.

destination host is able to handle arriving data, it does not reduce its window size although buffers in routers on the way may overflow. The alarm about this kind of situation may arrive to the sending host as an ICMP message or its transmission timer expires because of delayed acknowledgment. If this occurs the source host reduces its transmission rate so that it transmits only one TCP segment and waits for its acknowledgment (this is known as a "slow start"). If the acknowledgment arrives in time the source host starts to increase its transmission rate and transmits two segments before stopping to wait for acknowledgments. The number of segments before waiting acknowledgments is increased until the maximum window size is reached or congestion occurs again. Naturally, if there is no congestion, the transmission window (controlled by the receiver) is the only one that sets limits to the transmission rate.

6.6.9.3   UDP

The UDP is an alternative to TCP, which is used if reliable connection-oriented service is not required. UDP offers transmission with a minimum of protocol handling. No connection is established between end points; the source host just sends separate datagrams toward the destination. The UDP header is very short, 8 bytes instead of 20 (or more) bytes of TCP header, and it is shown in Figure 6.53.

Port numbers indicate the application or process for which the message is intended. Length is the number of octets in the packet including header and data. All Internet hosts have to accept UDP datagrams of length 512 bytes or less [5]. Checksum is calculated over both the header and data field. In the event of an error, the checksums derived by the receiving party do not match and the datagram is discarded. No further actions are taken.
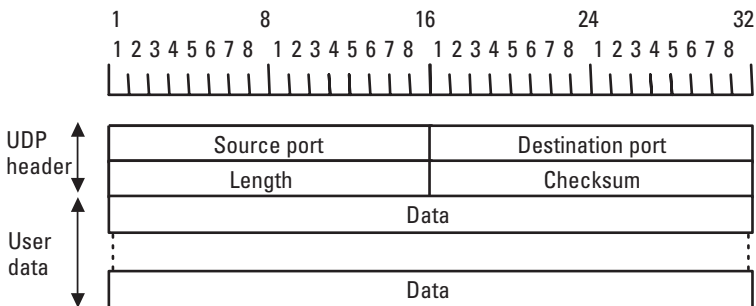
Figure 6.53 UDP header.

As shown in Figure 6.36, UDP is used by the Domain Name System and SNMP. It is also used by real-time applications that do not use retransmissions. They tolerate better lost datagrams than datagrams that arrive much delayed because of retransmissions.

6.6.10 Application Layer Protocols

The transport layer of TCP/IP provides reliable data transfer (TCP) or *unreliable datagram transfer* (UDP) over the Internet. All tasks for data transmission end to end are actually done at the transport layer and below. The application protocols use end-to-end communications provided by either TCP or UDP and provide specific services for software applications running in hosts. The most important application layer protocols are introduced next.

6.6.10.1 SMTP

E-mail is one of the most widely used applications on the Internet. Each e-mail user has his or her own mailbox in a computer and e-mail address. For example, the address tarmo.anttalainen@evitech.fi refers to computer evitech.fi where my mailbox is stored and which acts as my mail server.

The letter to be sent is first written in an e-mail application program. When it is sent SMTP in the sender's computer sets up a TCP connection to her mail server to port 25 defined for SMTP. SMTP defines message formats transmitted through an established TCP connection, that is, for example, how destination mail address, actual text, and possible file attachments are distinguished from each other. The TCP connection is terminated when the local mail server receives whole mail.

The local mail server then repeats the same procedure with the recipient's mail server or intermediate nodes, called *message transfer agents* (MTAs). MTAs are used in large companies to serve multiple local servers for external e-mail exchange. When mail enters the mailbox identified by the destination address, the destination user may access and extract it from the mailbox. For that she may use *Post Office Protocol* (POP) or *Internet Message Access Protocol* (IMAP) to get e-mail to her terminal for local processing.

6.6.10.2 FTP

FTP is an application protocol for the transfer of files among different computers. The objective is to provide users access to files in other computers. The user is able to view a remote computer's file catalog and request the transfer of any file of interest [5].

To copy a file, an FTP user establishes a TCP connection to the FTP server (port 21) for file transfer control. This is actually a Telnet connection via which commands for establishment of another TCP connection, used for actual data transfer, are given. Now a command, such as get <file>, may be given to copy a file from the FTP server to the user's computer. When using a browser's graphical user interface, we just click a button and Telnet and FTP do the actual work.

6.6.10.3 HTTP

HTTP is a standard protocol used on the Web. It is based on client–server communications between a Web browser (client) and a Web server. Its operation and how it is used in Web surfing are introduced in Section 6.6.11.

6.6.10.4 Telnet

The Telnet protocol provides a standard method for communications between terminal and terminal-oriented processes on a host computer. With the help of Telnet a user can log on to a remote host computer instead of his desktop PC.

Telnet is based on the *network virtual terminal* (NVT) concept, which represents a standard terminal and a set of services needed in the terminal session. Several application protocols, such as SMTP, are based on Telnet.

6.6.10.5 DNS

The Internet's IP addresses in binary, dotted decimal, or hexadecimal format are not especially user friendly. Instead of numbers domain names are used to identify hosts and a domain name system is needed to convert domain names (ASCII strings) into network addresses.

The DNS is a hierarchical database distributed to servers all over the Internet. The root is located at the top of the hierarchy followed by the upper domain layer as shown in Figure 6.54. The database root is implemented in a limited number of root servers, with the majority of them located in the United States. The upper domain layer is normally divided either by county or by organization type. Examples include these:

- *com:* commercial companies;
- *edu:* educational institutions;
- *gov:* government agencies (United States only);
- *net:* organizations that support Internet operators;
- *org:* other nonprofit organizations;

- *int:* international organizations;

- *se:* Sverige.

Figure 6.54 shows an example of the DNS. A company that has an assigned domain name, such as companyb.fr, is free to organize a hierarchy of domains beneath its assigned domain name. The addresses are stored in the DNS servers that make up a distributed network in which each server is responsible for one or more domains. If the local DNS server does not know the requested address, the query is sent higher up the hierarchy. The address query results in the IP address used for actual routing. Addresses are saved in the originating server for a period of time to avoid unnecessary querying. Due to network modifications they cannot be stored permanently.

The DNS is implemented as a large-scale client-server system in which clients and domain name servers exchange domain name server messages. Request or query messages contain the domain name and response messages contain the IP address or address of another server whom to contact.

How does the client host know where to begin the search for an address and how do the servers know how to find other servers? The client must know at least one name server and it is configured to every host. In our example each host in Figure 6.54 knows the IP address of the name server maintained by the network manager of company B. To ensure that domain name servers can reach each other, all name servers in the tree must know the

Figure 6.54 Hierarchy of the DNS.

address of at least one root server. Then all servers are accessible at least via the root server. The IP address is enough for DNS message exchange because all DNS servers use UDP and port 53 for communication.

6.6.10.6 SNMP

SNMP is a protocol for the transfer of network management messages between network elements, such as routers and switches, and network management center computers. SNMP defines messages to be exchanged; it has no functionality for actual management actions.

Most routers, LAN switches, and other LAN devices support SNMP, and many network management software packages are available for network control and monitoring with the help of SNMP.

6.6.10.7 DHCP

Each computer attached to the Internet needs to know its IP address before it can send or receive datagrams [4]. In addition the computer needs other information such as address of a default router, the subnet mask to use, and the address of the name server. DHCP can provide all the needed information.

DHCP uses IP and UDP protocols, and the payload of a UDP datagram contains a DHCP message. With the help of DHCP IP addresses need not be permanently assigned. Each time when the host joins the network, or is powered up, it requests an IP address and other needed information from the DHCP server. To access the server whose address it does not know, the client uses IP broadcast address 255.255.255.255 (see Figure 6.39). The client's hardware or MAC address is attached to the DHCP message. Also the server has to use an IP broadcast address in the response but by knowing the hardware address the response is received only by the client that sent the request.

6.6.10.8 TFTP

FTP is the most general file transfer protocol in TCP/IP protocol suite. Many applications do not need the full functionality that FTP offers, nor can they afford its complexity [4]. The TFTP is a simple file transfer protocol that uses UDP, as shown in Figure 6.36, and transmits files in fixed 512-byte blocks. It waits for an acknowledgment for each block before sending the next. UDP is an unreliable packet delivery system and TFTP uses time-outs and retransmissions to ensure that the entire file is received properly.

6.6.10.9 RTP

The RTP is designed to improve real-time services, such as digitized audio and video, over the Internet. It is designed to be independent from

underlying protocols and it cannot guarantee a specific level of service, for example, a certain constant data rate and delay. It uses UDP as a transport layer and cannot ensure timely delivery; such guarantees must be made by the underlying system. However, RTP provides sequence numbering for detection of out-of-order delivery and a timestamp that allows the receiver to control playback [4]. A *timestamp* defines the exact time at which the first octet of digitized data in the packet was sampled [4]. Protocol information of RTP is inserted into the UDP payload after the UDP header.

If some datagrams travel a much longer route and are much more delayed than others, an application layer protocol, such as RTP, cannot help it. For high-quality real-time transmission, additional control of lower layers is required. One proposed protocol for network layer control is the *Resource Reservation Protocol* (RSVP). The end points send an RSVP message to request resources and all routers on the way have to accept the request. If not, end-to-end QoS is not guaranteed.

6.6.11 WWW

The WWW is an architectural framework for accessing linked documents spread out all over the Internet. Its enormous popularity is the result of a colorful and easy-to-use graphical interface. The Web began in 1989 at CERN, the European Center for Nuclear Research. In 1994 the World Wide Web Consortium was founded for developing the Web and its protocols. In the mid-1990s, Netscape Communications Corporation launched a browser available to anybody free of charge and use of the Internet exploded.

With a WWW browser (client program) Internet users can download pages containing various types of information including text, graphics, photos, video, and audio. The three main questions are: (1) how does the user locate a piece of information in which he or she is interested? (2) how is information requested and delivered? and (3) how is the format of a Web page created? We attempt to address these questions next.

6.6.11.1 Uniform Resource Locators (URLs)

To access information or a resource, it must have a unique identification. A URL is used to indicate where the specific resource is found and which protocol should be used to fetch it. The structure of a URL is illustrated in Figure 6.55.

When the user provides a URL, the Web server transfers the requested file to the browser for display. The user may then click an item on screen, which the page designer has linked to another URL that may identify

http://www.w3.org/public/courses/internet.html

Protocol   Host   Search path   File
            computer
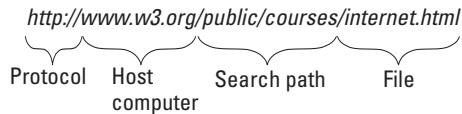
Figure 6.55 Structure of a URL.

a page that is then fetched from the other side of the world. The host part in Figure 6.55 is translated by the DNS to an IP address, which is in this case 18.23.0.24, and the protocol section defines HTTP and the TCP connection is established to port 80, the default port for HTTP, of the destination host. Search path defines where in the file hierarchy the file of interest is found. The file type *.html* in the example defines a file type that the browser shows on screen.

6.6.11.2 HTTP

HTTP is an ASCII-type application protocol primarily intended for client–server communication. A *client* is an application program, or browser, which sets up a connection to sends queries to a Web server. A *server* is an application program that accepts connection requests and responds to queries.

An HTTP request identifies the resource (URL) that the client is interested in and tells the server what to do with it. HTTP enables users to fetch different kinds of resources, such as text, picture, and audio. For HTTP a TCP connection is first established and requests and replies are sent through this connection. The URL identifies, with the help of DNS, the destination host, and the server port used for HTTP is 80.

HTTP requests contain information about what a server should do with the resource, the identification of the actual resource or page, and the HTTP version in use. HTTP replies consist of a header containing information about the resource, such as file type (how the browser should handle file), and actual resource or document. The standard known as *Multipurpose Internet Mail Extensions* (MIMEs) specifies content types included in responses to tell to browser how to deal with it. Figure 6.56 illustrates the procedure.

The steps that occur between the user's click and the page being displayed are as follows:

1. The browser determines the URL *http://www.w3.org/public/courses/internet.html* that is written or pointed out with a mouse.
2. The browser asks the DNS for the IP address of *www.w3.org*.
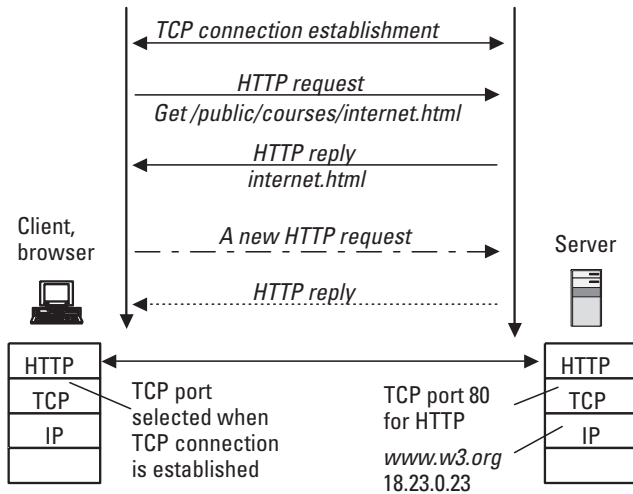
Figure 6.56 Procedure for HTTP requests and replies.

3. DNS replies with 18.23.0.23.

4. The browser establishes a TCP connection to port 80 on 18.23.0.23.

5. The browser sends a *GET/public/courses/internet.html HTTP/1.0* command.

6. The server *www.w3.org* replies with the file *internet.html.*

7. The TCP connection is released.

8. The browser displays all the text in *internet.html.*

9. The browser fetches and displays all images and audio files in *internet.html.*

The HTTP request contains simply an ASCII string *GET/public/courses/internet.html HTTP/1.0* where the first word defines the method (command) to be executed. The second section defines the path and the file, which is followed by HTTP version. The following methods are used in requests:

- *GET:* Indicates that the client wants to fetch the specific resource, in the example in Figure 6.56, the *internet.html* file. That is, GET indicates that the user wants to read the Web page defined by the file *internet.html.*

- *HEAD:* Request to read a Web page's header.
- *POST:* Adds information to an identified resource, for example, a Web page.
- *PUT:* Request to store a Web page.
- *DELETE:* Removes a Web page.
- *LINK:* Connects two existing resources.
- *UNLINK:* Breaks an existing connection between two resources.

The GET method or command, which we use most often, requests server to send the page. In response the server describes MIME content type for decoding. If the GET is followed by If-Modified-Since, the server sends data only if they have been modified after the date given in the request. The browser caches a set of pages in history and if the page is already stored, and not updated after that, there is no need to transfer it again. This is why clicking on the Back button on a browser often gives a very quick response.

The HEAD method asks for the page header only. It can be used to check when a page was last modified or to collect information for indexing purposes. The POST command is used to add information to a bulletin board system or to send information filled in a Web form.

The following methods allow a Web page manager to update information at the remote server. These requests usually contain content type and authentication information to prove that the user has permission to perform the requested action. The PUT method is the opposite of GET and it offers the ability to update Web pages on a remote server. The DELETE method is used to remove the page and LINK/UNLINK methods are for attaching new links or removing links between two Web pages.

6.6.11.3 Hypertext Markup Language (HTML)

HTML is a page-descriptive language that allows the user to navigate through a text with the help of links. The HTML document is created by providing different parts of text with markers, for instance <TITLE>. Pictures and links to other documents are also marked and included in a HTML document. Some word processors are able to produce an HTML document from any written text document.

The term *hypertext* refers to a capability to provide new information about a word or phrase by clicking it. The term *markup* comes from the old days when copyeditors actually marked up a document to tell to the printer which font to use.

A proper Web page consists of a head and body enclosed by <HTML> and </HTML> tags (formatting commands) [3]. The head starts with <HEAD> and ends with </HEAD>. The main item in the header is the title, delimited by <TITLE> and </TITLE> and is usually shown at the upper part of the browser screen. It may be the official name of the owner of the (home) page, for example, the name of the person or company that created the page. As we saw, HTTP provides a command that fetches only the heads of Web pages, so the head should briefly mention what this page is for. The body of the page starts with tag <BODY> and ends with </BODY>.

Figure 6.57 shows an example of the .html (or .htm) text file and how it appears on the browser's screen. <A> and </A> create the *anchor,* which defines a link. A URL is placed into the first tag and text to be displayed between tags. On the screen display, we see that the text is then underlined, and probably in color, to indicate that the user may click it to move to a new page.

Table 6.3 shows a selection of common HTML tags. There are many others and many tags have additional parameters that are not discussed here. Headings are generated by <Hn>, </Hn> and H1 represents the highest-level header and H6 the lowest-level header. Each item of the list starts with <LI> tag (there is no </LI>). The tag <UL> before the first <LI> tag and </UL>
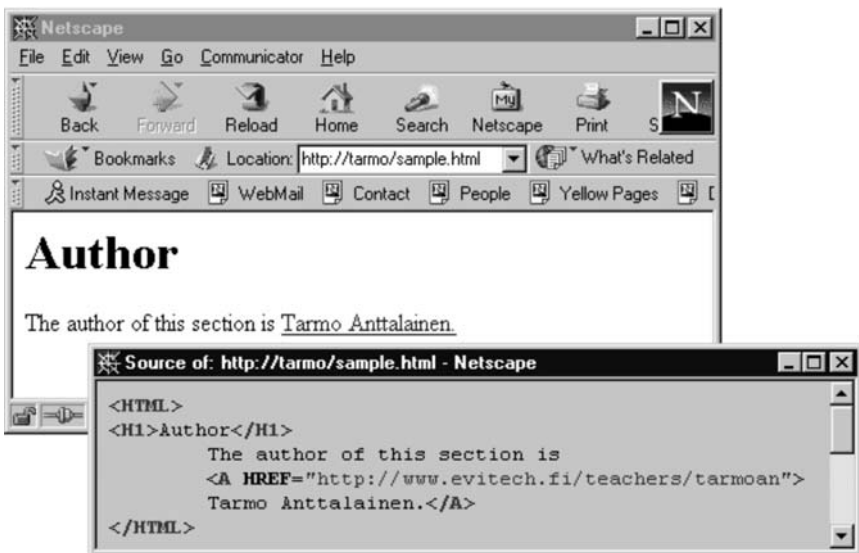
Figure 6.57 Example HTML document and its source file.

Table 6.3
Some Common HTML Tags

Tag	Description
<HTML>...</HTML>	Declares that Web page is written in HTML
<HEAD>...</HEAD>	Delimits page's head
<TITLE>...</TITLE>	The title not shown on the page
<BODY>...</BODY>	Delimits the page's body
<Hn>...</Hn>	Delimits level *n* heading
<B>...</B>	Shows text ... in boldface
<I>...</I>	Shows text ... in italics
<LI>	Start of each item on the list
<UL>...</UL>	Delimits unordered (bulleted) list of LI items
<OL>...</OL>	Delimits ordered or numbered list of LI items
<MENU>...</MENU>	Delimits a compact list of LI items
<P>	Start of paragraph
<HR>	Horizontal rule
<CENTER>...</CENTER>	Sets the item ... to the middle of the window
<IMG SRC="...">	Load an image from URL ...
<A HREF="...">...</A>	Define a hyperlink
<APPLET>...</APPLET>	A Java script to be downloaded

after the last one command the browser to add bullets in the front of each list item. If the list starts with <OL> followed by the list items, indicated by <LI>, numbers are shown instead of bullets. Numbered list section ends to </OL>.

We can insert into an HTML description images, video, and audio. MIME was defined to allow transmission of non-ASCII data through e-mail. It allows arbitrary data to be encoded in ASCII for byte-oriented transmission the same way as HTML text [4]. Table 6.3 shows how images are inserted. We simply refer to the URL defining the location and the file itself. The file extension tells to the browser how to deal with it. Image formats that are supported by practically all browsers are *Graphics Interchange Format* (GIF) and *Joint Picture Encoding Group* (JPEG). Also video and audio files can be inserted in the same manner.

Many comprehensive books are available about HTML for a reader who wants to see what it provides in detail and probably design her own home page. HTML, HTTP, and browsers are continuously being developed and new features implemented. However, Web page designers should take care that attached video and audio files are in formats that most browsers support. They should also note that all users might not have the latest browser version. A user probably will not want to visit a Web page again if it merely gives blank video windows and error messages.

Another thing that is very frustrating for Web users is that the information provided by WWW is often out of date. It would probably be better for the image of a company to provide no information rather than the wrong information via the Web.

6.6.11.4 Java

HTML is designed to show static pages on the screen of the browser. Every change on screen requires a client-server interaction, which may slow the response. In many cases, the response could be given locally by the browser if software that produces it were available. Examples are background music played locally while surfing, a game loaded to the browser, clicking a cat to make it meow, and complex forms (such as spreadsheets). These can be implemented with the language called Java.

An interactive Web page can point to a small Java program, called an applet, which the client downloads and runs locally. Now only information that cannot be produced locally needs to be transmitted and this approach may improve the performance of Web service. In particular when a user has wireless access to the Internet, it is essential to avoid unnecessary transmission and instead perform processing locally. Mobile terminals supporting Java are an important step toward wireless Internet.

When a Web page containing an applet is fetched, it is automatically executed in the client machine. This creates security risks. There will always be people around who enjoy designing applets that cause damage by, for example, reformatting a hard disk or searching confidential information and transmitting it to the Internet. In the design of Java security risks are considered but they are not entirely solved yet [3].

6.6.12 Voice over IP (VoIP)

The vast majority of information exchanged over the public telecommunications networks has been voice. The present voice communications networks, public telephone and ISDN, use the circuit-switching principle. Circuit switching provides good quality service and it does not require a complicated

encoding algorithm. A simple waveform coding scheme such as PCM, as discussed in Chapter 3, is sufficient for a circuit-switched connection that provides constant-bit-rate service. Charging for voice services has been straightforward—we simply pay for the duration of a call. This has been a relevant approach because each call reserves a certain data capacity whether there is speech on the line or not.

6.6.12.1 Voice Communications over Circuit- and Packet-Switched Networks

The characteristics of data transmission are different from waveform-coded speech, and the data networks that were developed to provide data services utilize packet-switched technology. These technologies include LANs, Internet, frame relay, and ATM. Packet-switched networks utilize network resources more efficiently than circuit-switched networks because the capacity in the network is dynamically shared among all users. If there are no data to be transmitted between two users, their share of the data capacity is available for other users. This difference in the operating principle makes a packet-switched network superior to a circuit-switched network when the data rate per user is not constant.

When all transmitted information including speech became digital, we saw that an integrated network providing all kinds of services was needed. Narrowband ISDN was developed but it is still circuit switched and not the optimum choice for data services. Then ATM was developed (for broadband ISDN) to support the constant-bit-rate service required by traditional speech and video encoders in addition to packet-switched data service. The use of the Internet has expanded rapidly and the majority of data transmission uses IP packets. When destination and source networks, usually LANs, and hosts use IP it is not efficient to split IP packets into ATM cells just for transmission. The speed of IP routers has improved, which has reduced importance of ATM technology.

The major packet-switched technology is the TCP/IP. The Internet has become very popular and access to it is available for every home with a telephone, a personal computer, and a modem or an ISDN network terminal. DSL technologies and cable modems, discussed in Section 6.4, provide wideband access to the Internet for residential customers. The ISPs provide access to the global Internet and charging for this service is based on the time of usage of the service or simply on a fixed monthly fee. This allows subscribers to utilize international data communications networks at a cost that may be lower than that of a local telephone call. The Internet can provide voice service, in addition to the Web, and allow subscribers to make international telephone calls via it instead of the telephone network.

6.6.12.2 Applications for VoIP

The implementation of VoIP service is attractive for subscribers because it reduces the cost of international and long-distance calls and it is also attractive to ISPs because it would increase the usage of Internet services. The technology for VoIP does not yet provide voice quality that is as good as a circuit-switched telephone network, but a lot of activity is being aimed at developing protocols for the implementation of high-quality voice service. The problem is that IP was designed for data communications and the packets suffer a long and variable delay, which decreases voice quality. In principle, samples of speech are transmitted in an IP packet payload and they do not arrive in regular intervals as they do in the case of circuit-switched service. To overcome this problem, the protocols of the Internet are being developed to provide a fixed share of network resources for each voice call through the network.

Figure 6.58 shows three possible ways to make telephone call over the Internet. In the first application example, a telephone subscriber dials the telephone number of the local gateway for an IP telephone service provider. The call travels over the PSTN to the nearest gateway that acts as an access point to the Internet. The service providers have their own telephone number prefix that connects a customer to the right gateway. Then the caller enters the destination telephone number and the gateway in the local office
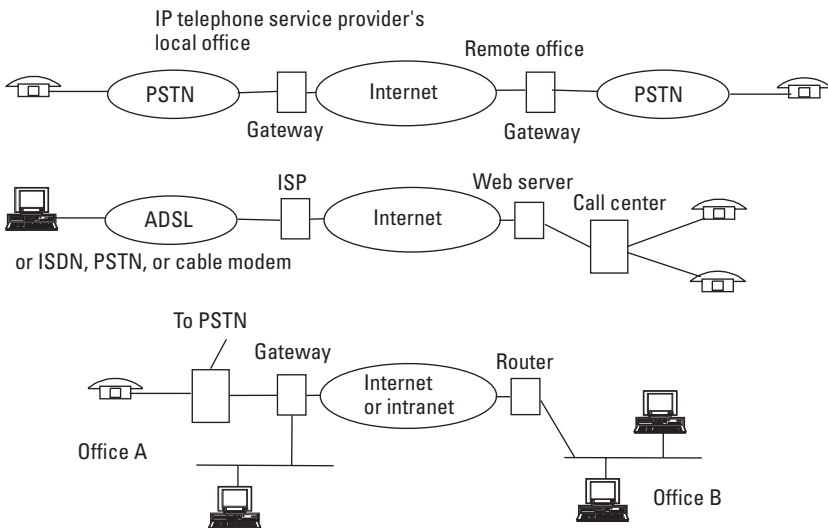
Figure 6.58 Voice over Internet applications.

establishes a connection over the Internet to the gateway in his remote office closest to the destination. Then the gateway in the remote office calls the destination subscriber via the local PSTN. Internet routing and speech processing is performed by the gateways and ordinary telephones can be used for the call. Now the Internet, instead of PSTN, carries a long-distance section of the call. In this way, international calls in particular can be provided at very attractive fees because only the local part of a telephone network is involved in the call.

The second application example in Figure 6.58 illustrates a customer surfing the Internet and a Web service provider with enhanced WWW service using VoIP. People surfing the Web can connect to a company's call center by clicking a Call button located on the company's Web page. Users can communicate with a customer service group, ordering department, or help desk by using their Web browser and a personal computer equipped with a compatible speech encoder. This is an important new feature as commercial use of the Internet expands.

The third example shows a company with locations in multiple sites and an intranet. Intranet connections between office A and B use the IP network. The IP network may carry secure VPN connections where IP packets are tunneled and ciphered, although firewalls are not shown in the figure to keep it simple. Now transmission capacity is available between offices and in addition to data VoIP can carry voice via the same VPN connections. There is no need to lease separate channels from a PSTN service provider for speech only. In office B no PABX equipment is required if the PCs contain suitable sound cards, headsets, and software for VoIP. External calls can be done via PABX in office A.

Another application that PSTN operators use is to replace their conventional trunk network with an IP network. Modern telephone exchanges contain VoIP functions and are able to establish calls alternatively via the IP network instead of conventional 64-Kbps trunk network channels. However, emergency calls are usually routed via a circuit-switched trunk network because it has much higher reliability.

6.6.12.3 VoIP Protocols

As mentioned earlier the quality of VoIP service is worse than that of PSTN service. The two main reasons are *delay* and *jitter* (variable delay). Typically 20 ms of speech is encoded into one IP packet and encoding, packetization, packet handling, and buffering will often lead to an overall delay of well over 200 ms, which disturbs our internal schema of interactive communication. To improve quality, several protocols are developed and evolving.

One protocol introduced in Section 6.6.10 is the Real-Time Transport Protocol, which operates on the top of UDP. It cannot guarantee high QoS because it has no control on the lower layer protocols where, for example, network layer congestion can occur.

Another protocol designed for control of lower layers is the Resource Reservation Protocol. An endpoint uses RSVP to request a simplex flow through an IP network with specified QoS bounds, for example, delay and throughput. If routers along the path agree to honor the request, they approve it; otherwise they deny it. Every router on the way has to support RSVP and the QoS requirements given. Both ends have to use RSVP to request QoS if it is needed in both directions [3].

Another issue is signaling, that is, how a telephone call is established over an IP network. Quite similar signaling phases that we illustrated in Chapter 2 are needed between endpoints or gateways. The main two protocols are the H.323 recommendation of ITU-T and the *Session Initiation Protocol* (SIP) of IETF. Both support signaling for multimedia sessions including video in addition to ordinary telephone calls. They define signaling messages, which are exchanged for call establishment, maintenance, and clearing.

IP networks can carry voice and video already and we may expect that these applications will expand as standards become mature and widely implemented. The Internet will also be used for facsimile calls and videoconferencing as standards evolve. Internet and private intranets will carry a larger and larger part of PSTN traffic but a lot of work remains to be done before IP technology can replace PSTN the infrastructure. The main issues to be solved are QoS and reliability.

6.6.13 Summary

Our main intention in this Internet section was to get a clear view of how most popular Internet services, such as Web and e-mail, are really implemented from the display of a Web page to actual data communications over LAN and other bearer networks.

The aim was also to clarify how layered data communications introduced in Section 6.3 operate in practice. Many important aspects were not covered but this introduction should give readers a basic understanding about communications methods from the physical layer to the application layer and software applications. This helps readers extend their knowledge and use another information source for more detail. Some of these sources are listed the end of this chapter.

6.7   Frame Relay

Frame-relay technology is widely used by network operators that provide long-distance data communications service to companies. It is designed for ordinary data applications and transmits data frames with variable length. The old packet-switched networks, such as X.25, were originally designed for a low-quality physical network and included data integrity checking at many protocol layers. With the present high-quality physical network, this is usually unnecessary. Frame relay leaves data checking and acknowledgment procedures to the network users and the protocols in use are much simpler and can support a much higher data rate. Frame relay supports data rates up to 50 Mbps.

Frame-relay technology is usually used to provide semipermanent connections for LAN interconnections. The network operator sets up a virtual connection between endpoints and frames with circuit identifiers are routed through the network as explained in Section 6.2.4. The network capacity is shared between users and the cost for long-distance connections is much lower than cost of leased-line connections.

Frame relay is a technology for data transmission and it does not support isochronous transmission, such as voice or video, which requires low and constant delay. A network technology that was designed to support isochronous services as well is known as asynchronous transfer mode, as discussed next.

6.8   ATM

Most packet-switched techniques make use of variable-sized packets and this leads to significant variations in the arrival times of the packets of a particular data stream. Because each physical connection may carry traffic from many individual data streams, it occurs every now and then that a specific packet is queued behind a number of large packets from other data streams that are waiting to be sent out on the physical connection. A further consequence is that switching is carried out by software that will eventually constrain the speed and performance of the network.

We saw in the 1980s that sooner or later all services would be integrated into a common network, which was called *broadband ISDN* (B-ISDN). However, the two main network technologies, packet-switched for data and circuit-switched for voice services, could not support the other main service type and the ITU decided to develop ATM.

ATM is a cell-relay technology, which uses small fixed-size frames called *cells*. Cell relay transmits frames with constant length, 53 octets, and provides both *variable-bit-rate* (VBR) service that is optimum for data transmission and *constant-bit-rate* (CBR) service for voice and video applications. CBR is not available in frame-relay technology.

ATM defines the structure of cells, continuous transfer of cells, and cell switching. Isochronous service is available by reserving certain fixed capacity of ATM cells from the network. ATM cells are packed into an SDH frame, STM-1, or into a SONET frame and then the physical data rate may reach 155 Mbps or higher. Significant advantages of cell-relay technology follow from the use of fixed-size small packets or cells instead of packets with variable lengths. The consequences of this principle are as follows:

- Delays in the network are much lower and more predictable. By ensuring that the cells from a specific data stream occur at regular intervals in the cell stream, it is possible to provide guaranteed bandwidth with low delay and jitter just as in circuit-switched networks.

- The fixed size of cells allows the switching function to be removed from software into hardware with a dramatic increase in switching speed.

ATM thus provides the benefits of circuit- and packet-switched networks, hence allowing all types of traffic to be integrated onto a single network. Many network operators use ATM technology in their core network. In ATM networks the switches are usually configured to provide semipermanent data connections. By *semipermanent*, we mean that these connections are not dialed up by users, but controlled from the network management center by a network operator.

6.8.1 Protocol Layers of ATM

ATM networks can be considered as a number of layers providing different functions. The ATM stack consists of a physical layer, ATM cell layer, and ATM adaptation layer, as shown in Figure 6.59. They do not correspond to the three lowest OSI layers.

ATM networks are connection oriented, which means that there is a connection establishment phase followed by a data transfer phase. During the connection establishment phase, a path (virtual circuit) through the network is built up and all cells of this call then use this path. This principle was
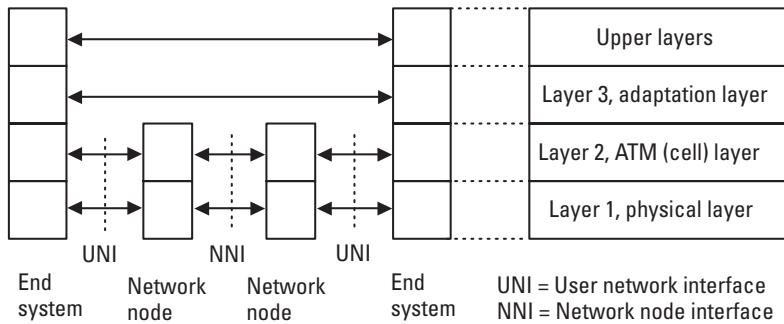
Figure 6.59   The protocol layers of ATM.

explained in Section 6.2.4. ATM thus provides guaranteed cell sequencing but some cells in a data sequence may be lost. The cells with errors are discarded by the network, and it is up to the end systems to detect and recover from a cell loss. If the network supports dial-up connections, the control of virtual paths and circuits is carried out by signaling on the subscriber interface called the *user network interface* (UNI). If dial-up connections are not supported, virtual paths are set up to each network node by the network operator. The interfaces between nodes in the network are called *network node interfaces* (NNIs).

6.8.2   Cell Structure of ATM

The ATM cell is 53 bytes long with 48 bytes reserved for carrying the payload and 5 bytes for the header (see Figure 6.60). This size was a compromise between the 64 bytes favored by the data community and the 32 bytes preferred by the voice community. The data fields of a cell are shown in Figure 6.60 are explained in the following sections.

6.8.2.1   Generic Flow Control (GFC)

The GFC field is used at a user interface only to control the data flow between the first ATM switch and the user node. Inside the network (i.e., in the NNI), this field is used for virtual path identification together with the other VPI fields. This is the only difference in the cell structure between UNI and NNI.

6.8.2.2   VPI and VCI

The majority of the header is taken up by VPI and VCI. Together they identify an individual circuit. They have only local significance and they change
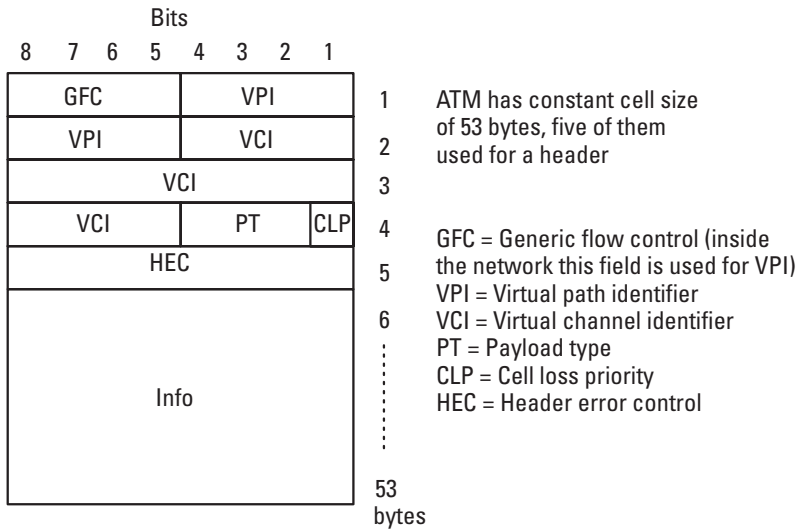
Figure 6.60 Structure of an ATM cell (UNI).

on the way through the network as explained in Section 6.2.4. They are used in the same way as the logical channel number in X.25 or the data link connection identifier of frame-relay technology.

6.8.2.3 Payload Type

Payload type specifies whether the cell contains user information or information to be used by the network itself, for example, for O&M. The network can use these maintenance cells between nodes to perform operations in, for example, a congestion situation.

6.8.2.4 Cell Loss Priority

The cell loss priority bit carries information between an ATM user system and the network. For example, in a congestion situation the network may use this field to define the priority of cells in the queues or to decide which cells are discarded first in the case of overload.

6.8.2.5 Header Error Control (HEC)

HEC is a checksum for the first 4 bytes. It makes possible the detection of multiple errors and the correction of a single error. ATM cells with more than one error will be discarded by the network. It is up to the end systems to detect and recover from such losses. The end systems also have to detect

errors in the user data. When an ATM switch updates the virtual circuit and path identifications of a cell, it calculates a new HEC for the following hop to the next switch.

6.8.3 Physical Layer of ATM

ATM cells can, potentially, be carried over most physical layer media but the ITU-T has defined SDH to carry ATM cells with speeds of 155.52 and 622.08 Mbps. Figure 6.61 illustrates how ATM cells are inserted into the payload of an SDH frame STM-1. The frame includes 9 × 270 bytes and it is transmitted 8,000 times per second. We review here only the STM-1 frame of SDH as an example, although this principle is valid for SONET as well but the detailed structure of the frame is different.

The SDH frame includes *section overhead* (SOH) that contains framing information, network management data, and other overhead information needed by optical SDH transmission systems. ATM cells are merged together with *path overhead* (POH) to the payload of an SDH frame. POH and data (ATM cells) together make up the *virtual container* (VC), which can start at any point inside the payload of the frame. The *administrative unit* (AU) pointer tells where the frame (VC) containing path overhead and ATM cells starts inside the payload [1]. In Figure 6.61 this starting point is assumed to be at the beginning of the payload.

At an SDH interface the concept of a continuous stream is used and dummy cells that are called idle cells are inserted if there is no traffic. In

ATM cell mapping onto an STM-1 frame:
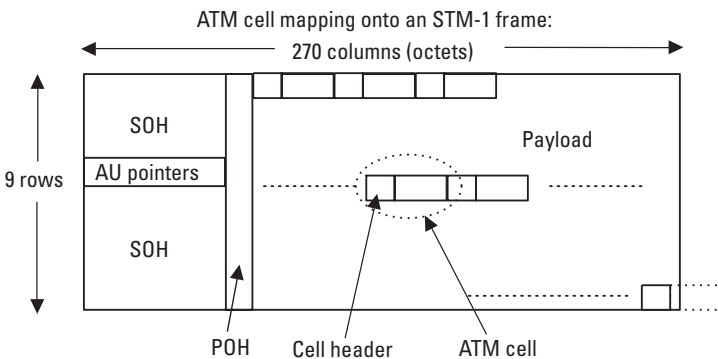
Continuous stream of cells; dummy cells are inserted if there is no traffic. Cell synchronization is based on an error check code in the header.

Figure 6.61 Physical layer of ATM.

addition to SDH the ATM cells can be carried in many other transmission systems such as DS-1 (1.544 Mbps), E-1 (2.048 Mbps), or SONET.

Because the cell stream is continuous, there are no explicit indicators for the start and end of a cell. The synchronizing scheme relies on the check code (HEC) in the frame header. The receiving equipment calculates the code over four subsequent bytes as each byte arrives and checks if the next byte corresponds to the calculated result. If they are the same, it is likely that a cell header has just been received. If the codes of a few cells match, the synchronization is accepted, and the equipment (ATM system) becomes operational.

If more than a defined number of cells arrive with a wrong code, the equipment assumes that synchronization is lost and attempts to reestablish it.

6.8.4 Switching of ATM Cells

When an ATM connection is established through the network, the cells of this connection carry a certain VPI that is changed for each link between switching nodes as explained in Section 6.2.4. The circuit identification is divided into two parts and ATM switches can operate at two levels, the virtual path level (a group of virtual channels) or virtual circuit level (see Figure 6.62). Virtual paths act as pipes for a collection of VCs. ATM switches may act at the VP level, the VC level, or both. A VP switch does not look at the VCs within a path, and end systems can freely establish and remove VCs without the network carrying VPs being involved.
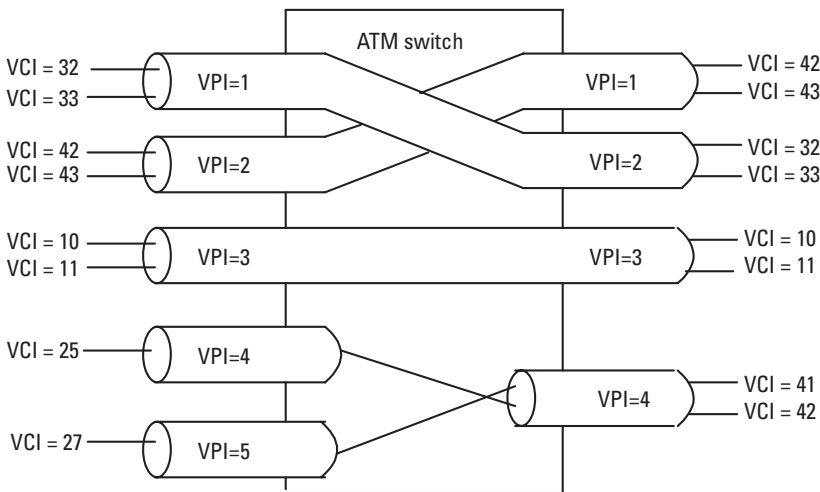
Figure 6.62 Virtual paths and channels of ATM.

The routing at the VP level allows the network operator to provide a VPN for a corporation just by cross-connecting virtual paths between offices. These paths provide the permanent connections between offices. The user may then configure his private network, with connections provided by the virtual channels, inside the virtual paths he has leased from a network operator.

To establish an individual call through a dial-up ATM network, the signaling cells with a specific VPI and VCI (reserved for signaling purposes at a UNI) are exchanged between the network and the user. Then the network defines a route with a certain VPI and VCI for this call in each link. For the routing of cells along an established virtual connection, the ATM switches look at both the VPI and VCI in each cell.

In Figure 6.62 virtual paths 1, 2, and 3 are cross-connected at the path level and virtual channel identifications remain unchanged. The figure also shows an example where virtual channels 41 and 42 of virtual path 4 are switched to the virtual channels 25 and 27 of virtual paths 4 and 5. In this case the switch updates both the VPI and VCI fields in the header of ATM cells before it transmits them to the next switch.

6.8.5 Service Classes and Adaptation Layer

ATM is designed to support different services and the purpose of the *adaptation layer* (AAL) is to make the cell transport of ATM suitable for different applications. The service classes are defined to support various types of applications. The corresponding adaptation layer protocols define how each class of service is implemented. The role of the AALs is to provide the mapping of particular types of traffic onto the underlying ATM cell layer. This requires that the AAL header containing the protocol information be added to the user data before transmission in the payload of ATM cells as shown in Figure 6.60. Depending on the adaptation layer in use, this reduces the user information to 48...44 octets in a cell [1]. We can divide services into four basic classes depending on whether the required service is constant or variable bit rate, isochronous or synchronous, or connection oriented or connectionless.

Figure 6.63 shows the service classes, their basic characteristics, and corresponding AAL protocol. The timing relation characteristic tells if the information about timing has to be available for the receiver of ATM cells. This may be required in the case of the reconstruction of PCM-coded speech, which requires that samples arrive at regular intervals. The timing information is used to control playback. Four service classes support different applications:

- *Class A:* Constant-bit-rate service for voice and video applications;

- *Class B:* Variable-bit-rate service with timing information for variable-bit-rate voice and video applications;

- *Class C:* Variable-bit-rate service for ordinary data applications;

- *Class D:* Variable-bit-rate service for connectionless transmission of very short data messages (no connection establishment).

Figure 6.63 shows also which AAL is used to implement each service class. It also makes a conclusion about their main characteristics.

6.8.5.1   Class A/AAL1

AAL1 provides the support for traffic, which requires CBR service, and it is mainly used for voice and video applications. This AAL is quite simple because there is no requirement for error detection and recovery for this type of traffic. The transfer of timing information over a call is the major function that AAL1 has to provide to ensure high-quality playback of the data. This service simulates leased-line data or voice circuits and one important application could be PABX voice channels in integrated corporate networks that utilize ATM technology.

Service classes A, B, C and D provide communication services for different applications. Adaptation layer defines the protocols for the mapping of particular types of traffic onto the cell layer. Protocol control information is transmitted in the information field of the cells.

Service Class	A	B	C	D
Timing relation	Isochronous		Asynchronous	
Bit rate	Const.	Variable		
Connection oriented, CO, or connectionless, CL	Connection oriented			CL
AAL protocol	AAL1	AAL2	AAL5 AAL3/4	AAL3/4

Cell header 5 octets

AAL 1-5

Information field for user data 44 .. 47 octets

Service classes and examples:
Class A: Constant bit rate service, ordinary video or voice
Class B: Variable bit rate service with timing information, for variable bit rate video or voice
Class C: Variable bit rate service, ordinary data transmission
Class D: Variable bit rate service with connectionless service, data transmission of very short messages (no connection establishment)

Figure 6.63 Service classes and adaptation layer of ATM.

6.8.5.2   Class B/AAL2

AAL2 provides the support for VBR traffic that requires maintenance of timing information during the call. Timing information is transmitted in the adaptation layer header. Examples of this type of traffic are variable-bit-rate voice and video applications in a LAN environment.

6.8.5.3   Class C/AAL5, AAL3/4

AAL3/4 is a complicated protocol and it provides both connection-oriented service of class C and connectionless service of class D. AAL3/4 was found to be too complex and inefficient for ordinary LAN traffic.

AAL5 evolved after AAL3/4 and it was designed to be simple and efficient. It supports only variable-bit-rate traffic, like burst data of LANs, with no timing relationships. AAL5 does not provide enhanced services and thus consequently it does not require much overhead for protocol information. It is the primary AAL used to provide LAN interconnections over ATM networks.

6.8.5.4   Class D/AAL3/4

This class supports variable-bit-rate traffic that requires no timing information. It supports connectionless service that does not require connection establishment. Class D is suitable for datagram transmission in which only a small amount of data is transmitted during one connection.

6.8.6   Applications and Future of ATM

ATM is used as a technology for high-data-rate backbone networks of some telecommunications network operators. ATM was expected to be a major backbone technology and some access technologies, such as ADSL, were specified to transmit ATM cells. However, because LAN and IP switching technology has developed to manage higher data rates, the importance of ATM is decreasing. ATM was also defined to be the initial network technology for UMTS, introduced in Chapter 5. It will be replaced by evolving IP technology later in the evolution of the third generation mobile networks.

6.9   Problems and Review Questions

Problem 6.1

Compare parallel and serial data transmission principles and applications.

Problem 6.2

Explain what is meant by asynchronous and synchronous data transmission.

Problem 6.3

ASCII strings "A" [1000001] and "2" [1000110] (first bit on the left) are sent and a parity bit for even parity is added in the end of each character. What are the transmitted 8-bit strings?

Problem 6.4

Synchronous frames use bit sequences or flags (0111110) to indicate the start and end of the frame. Explain why user data sequence ...01111110... is not detected as the end flag.

Problem 6.5

If the data bit string 0111101111101111110 is bit stuffed, what is the output string to be framed with flags?

Problem 6.6

Compare circuit-switched service and packet-switched networks. What are their advantages and disadvantages?

Problem 6.7

What do we mean by physical circuits and virtual circuits? How does the packet-switching principle based on datagram transmission differ from the switching based on virtual circuits? Compare their advantages and disadvantages.

Problem 6.8

Explain how frames are switched in a packet-switched node of ATM or frame-relay network.

Problem 6.9

List three examples of both circuit- and packet-switched networks.

Problem 6.10

What do we mean by *polling* in data communications?

Problem 6.11

What does *protocol* mean in data communications? Give some examples of protocols.

Problem 6.12

Why do we use a layered protocol structure in data communications?

Problem 6.13

Explain the basic principle and structure of the OSI reference model.

Problem 6.14

Compare the TCP/IP stack with the OSI reference model.

Problem 6.15

Explain the principle of data flow in the layered protocol hierarchy from the application layer through the lower layers to the physical channel. Explain also what happens in the protocol stack of the receiving computer until the data arrive at the application software at the other end of the connection.

Problem 6.16

List the most important Internet access technologies and compare their characteristics.

Problem 6.17

Explain the purpose and basic operation of a voice-band modem. What additional function does it provide in addition to actual data transmission?

Problem 6.18

Explain how the frequency band of the subscriber loop is used by ADSL. How does it transmit data over the subscriber loop? What are its main characteristics?

Problem 6.19

What are the main modifications needed in cable TV network when it is upgraded to provide Internet access to residential customers?

Problem 6.20

Explain the basic structure, operation, and characteristics of LANs.

Problem 6.21

What does the LAN protocol CDMA/CD refer to, that is, how does it operate in principle?

Problem 6.22

Explain the Ethernet frame structure.

Problem 6.23

How do DIX Ethernet and IEEE 803.2 frames differ?

Problem 6.24

Explain how collision detection in Ethernet operates.

Problem 6.25

Why is an Ethernet frame defined to be at least 64 bytes long? Why can't it be shorter?

Problem 6.26

Assume that three workstations, A, B, and C, are connected to the same hub and they start to transmit simultaneously. How is this situation resolved so that all frames will be successfully transmitted after a while?

Problem 6.27

Assume that the only delay in a LAN is cable propagation delay. Signal speed in cable is 70% of the speed of light. What would be the maximum distance between computers in an Ethernet LAN connected to the same hubs in the same collision domain for a network data rate of (a) 10 Mbps, (b) 100 Mbps, and (c) 1 Gbps? The minimum frame length is 64 bytes. (Assume here that no carrier extension is used.)

Problem 6.28

Explain how operation of a LAN hub and switch differ. How does the auto-learning switch or bridge know to which port it should send a frame?

Problem 6.29

Explain the structure of an IPv4 address.

Problem 6.30

What is the maximum number of IP hosts in networks with class A, class B, class C, and class D IP addresses, respectively?

Problem 6.31

What is the subnetwork mask and how is it used?

Problem 6.32

A class B IP address is assigned to a company. There are 12 LANs in its network. Define the subnetwork mask and host number ranges in binary, hexadecimal, and dotted decimal notation for each subnet. How many hosts can exist in each subnet?

Problem 6.33

Describe the structure of an IP packet. Explain the purpose of each field in its header.

Problem 6.34

What are the main advantages of IPv6 compared with IPv4?

Problem 6.35

Explain briefly mobile IP operation.

Problem 6.36

What is the ARP and how does it operate?

Problem 6.37

What is the RARP land how does it operate?

Problem 6.38

What is the purpose of the ICMP?

Problem 6.39

What are the host-to-host (or transport) layer protocols of TCP/IP and what are their basic differences?

Problem 6.40

Explain the TCP connection setup procedure.

Problem 6.41

Explain the sliding window principle used by TCP. Why are there different windows for transmission and reception?

Problem 6.42

Describe the TCP packet header and explain the purpose of its fields.

Problem 6.43

What is the Domain Name Server? Explain its hierarchy and basic operation.

Problem 6.44

What is a URL? Describe its structure.

Problem 6.45

What is the HTTP? Describe its basic operation.

Problem 6.46

What is HTML? For what is it used?

Problem 6.47

What are the new features that Java provides for Web page designers?

Problem 6.48

What are the main advantages and disadvantages of IP networks compared with PSTN for ordinary telephone service?

Problem 6.49

Explain the main goals behind the development of ATM.

References

[1]   Freeman, R. L., *Telecommunications System Engineering,* 3rd ed., New York: John Wiley & Sons, 1996.

[2]   Halsall, F., *Data Communications: Computer Networks and Open Systems,* 3rd ed., Reading, MA: Addison-Wesley, 1996.

[3]   Tanenbaum, A. S., *Computer Networks,* 3rd ed., Upper Saddle River, NJ: Prentice Hall, 1996.

[4]   Comer, D. E., *Internetworking with TCP/IP: Principles, Protocols, and Architecture,* 4th ed., Upper Saddle River, NJ: Prentice Hall, 2000.

[5]   Ericsson Telecom, *Understanding Telecommunications,* Vol. 2, Lund, Sweden: Ericsson Telecom, Telia, and Studentlitteratur, 1998.

[6]   Dutta-Roy, A., "An Overview of Cable Modem Technology and Market Perspective," *IEEE Communications Magazine,* June 2001, pp. 81–88..

[7]   Grilo, A., P. Estrela, and M. Nunes, "Terminal Independent Mobility," *IEEE Communications Magazine,* December 2001, pp. 34–71.

7

Future Developments in Telecommunications

The 1990s were marked by rapid development of telecommunications services, technologies, and business. The two major areas of huge expansion were the Internet and cellular telephones. Most professionals in the telecommunications business did not expect this kind of growth at the beginning of 1990s.

It is difficult to estimate which new services will gain market acceptance and which will not be successful. A technology must be available but, in addition, success depends on many other things such as how attractive the service is, how the new services are launched and charged, and what alternative services are available. In the following sections we look at some future development areas.

7.1   Information Networks

The success of the Internet was based on the graphical user interface developed at CERN that was introduced in Chapter 6. As we saw it was not a technical revolution but it made Internet service more user friendly and attractive to the public. Internet network technology itself was more than 10 years old and it had proved to be capable of providing global service for academic experts all around the world. After introduction of the Web, it became

available to anyone and the telecommunications business structures were forced to change.

The demand for Internet service has also supported development of new broadband access technologies, such as cable TV modems, DSL, and fixed radio access technologies introduced in Chapter 6. Internet access has become a new business sector for PSTN and cable TV network operators. Even new enterprises, such as ISPs, which provide Internet service only, were born.

The expansion of the Internet will continue. Broadband access technologies will improve its performance for residential users. Business usage of the Internet will continue to grow although not all early experiences were encouraging. Services available on the Internet will become richer and it will provide, for example, integrated voice services. Home shopping, for which ordinary mail has been used, can use electronic catalogs. This has an advantage over mail if the products can be seen in action and if questions can be asked about the product. Integrated telephone would allow a customer to clarify the details of an order at the same time. Even a virtual visit to, for example, a holiday site will become practical. A customer may take virtual walks in various hotels and select the one she likes the best.

7.2   Telephone Services

Another area of major growth in the 1990s was mobile or cellular communications, mainly telephone. The success of the second generation cellular systems was based on standards with wide acceptance and deregulation of telecommunications business. Mass markets made low-cost handy mobile stations available and competition reduced service costs.

Telephone communication will change to more and more personal communication as voice moves from the wired network to cellular networks. Lines for fixed telephone network will be released for broadband Internet access. When users have broadband access to the Internet, an increasing share of fixed telephone communications will be transferred to the Internet. Also in the core network of PSTN, use of Internet technology will increase.

7.3   Wireless Communications

As just mentioned, telephone service is moving from wired networks to wireless cellular networks. The second generation cellular systems provide text

messaging service, which became much more popular than initially expected. Cellular networks have been further developed to provide better and better data and information services. The success of information services is very difficult to estimate. Businesses may explore them only if attractive applications are published. Implemented packet-switched technology will allow low-cost use of those services. Location-based services, in which the information provided depends on the user's current location, can provide many business opportunities.

Cellular networks will also be used for wide-area Internet access. The packet-switched air interface makes it possible for network operators to provide high-data-rate communications at attractive fees. For better performance WLAN technologies will be integrated with cellular systems. For high-data-rate short-haul data communications, WLAN technologies will be available for travelers.

Personal area network technologies, as a less complex option to WLAN, will make our living easier by connecting electronic devices to each other in homes and offices. There will be new applications that we have not yet thought of today.

7.4   Optical Technology

Development of optical technology took a major step in the 1990s when DWDM technology and optical amplifiers began to be used. Chapter 4 provided an introduction to these technologies. DWDM offers a very economical way to increase the transmission capacity of the core network where optical cables are available. In the future, the optical core network will become more cost effective and flexible with the help of optical network elements, which allow flexible routing of optical signals without the need to convert them into electrical form. This kind of fully optical network is called an *optical transport network* (OTN).

These developments will decrease the cost of digital transmission and increase the transmission capacity of the network to meet the growing demands of data, especially Internet, communications.

7.5   Digital Broadcast Systems

Current broadcasting systems such as radio and TV use technologies that were originally developed in the 1940s. Even though some updates have been

made such as color TV, stereo sound, and *radio data system* (RDS), current systems do not meet the quality requirements of the future. Another problem with these systems is that they do not utilize radio frequencies as efficiently as more modern technologies do.

Digital broadcast radio standards were approved many years ago, but digital radio has not become popular. Analog FM broadcast radio is still the main broadcast radio technology and that is because digital radio does not provide much that is new to the listener. The quality is better but FM radio quality is acceptable for most of us. Another reason is that there has not been political pressure to change from analog to digital and analog transmission is allowed to continue.

Digital TV standards are also available today. Digital TV will improve quality and provide some additional services and one of its major advantages is more efficient use of the limited resource of broadcast TV frequencies. An additional converter or a new TV set is required for digital broadcast reception. In many countries decisions have been made to phase out analog TV transmissions between 2005 and 2010. The exact time for transition depends on, among other things, the availability of digital TV sets.

7.6 Summary

We introduced in earlier chapters the physical basics of electrical communications and current data communications and telecommunications technologies. In the future we will see their integration such that high-performance voice, data, and information services will be available anywhere. Although different network technologies will still be in use in different parts of the world, multimode terminals will allow us to access similar services anywhere.

Development of telecommunications as a business area depends most of all on new applications that will be provided to customers. We cannot even guess what they will be and if they will create the same kind of boom to this business that we had in 1990s because of rapidly increased Internet and cellular telephone penetration.

About the Author

Tarmo Anttalainen graduated with a B.Sc. from the Helsinki Institute of Technology in 1975 and an M.Sc. in telecommunications in 1983 from the Helsinki University of Technology. He was a development engineer for Nokia Telecommunications/Transmission Systems in the area of digital multiplex and line equipment from 1973 to 1983. From 1983 to 1986, he was a development manager for Nokia in the Multiplex and Line Equipment division and his areas of interest included copper cable and optical systems. His activities at Nokia also included product development and technical support for marketing and customer training in Europe, the Middle East, and the Far East.

Dr. Anttalainen was the development department manager for the PDH Multiplex and Line Equipment division at Nokia from 1986 to 1989, and the department manager of Nokia's PDH and SDH Transmission Systems division, including multiplex and line systems, from 1989 to 1992. His duties there also included technical marketing all over the world, especially Europe and the Far East, including Australia and Japan. He was in charge of project management of international SDH development and holds several patents in the area of transmission systems, including SDH. Since 1992, he has been a principal lecturer in telecommunications at Espoo-Vantaa Institute of Technology in Espoo, Finland, specializing in the areas of data communications, public telecommunications networks, and cellular networks.

Index

Trunked networks, 45–46, 199–201
   analog, 200
   defined, 199
   operating principle, 199
   TETRA, 201–2
Twisted pair, 170–71
Twisted-pair Ethernet, 292–94
   10BaseT, 292, 293
   defined, 292
   illustrated, 292
   See also Ethernet

UMTS, 208–9
   core network, 209
   defined, 208
Uniform Resource Locators (URLs),
      331–32
   defined, 331
   structure, 332
   See also World Wide Web (WWW)
Universal Mobile Telecommunications
      System (UMTS) Forum, 14–15
Universal synchronous bus (USB), 248
Unshielded twisted pair (UTP), 171
User Datagram Protocol (UDP), 326–27
   datagrams, 326
   defined, 326
   headers, 326, 331
   payload, 331
   use, 327
User network interface (UNI), 344

Variable-bit-rate (VBR) service, 343
Variable delay tolerance, 80
Very-high-bit-rate DSL (VDSL), 276
Vestigal-sideband (VSB) modulation, 131,
      132–33
   defined, 132
   derivation, 133
   illustrated, 131
Video-on-demand (VoD), 273
Virtual circuit identifiers (VCIs), 245–46
Virtual LANs (VLANs), 300–301
Virtual private networks (VPNs), 52–53
   defined, 52

   extranets, 53
   firewalls, 52
   intranet, 52
   principle, 52
   See also Telecommunications network(s)
Visitors location register (VLR), 192–93
   functions, 214–15
   GSM, 214–15
   See also Home location register (HLR)
Vocoders, 107, 113
Voice-band modems. *See* Modems
Voice communication networks, 51
Voice over DSL (VoDSL), 276
Voice over IP (VoIP), 337–41
   applications, 339–40
   protocols, 340–41

Waveform coding, 111
Wavelength-division multiplexing (WDM),
      179–80
   defined, 179
   DWDM, 180
   optical fiber system and, 179
Wide-area networks (WANs), 60
Wideband CDMA (WCDMA), 209
Wireless access, 279–80
Wireless application protocol (WAP), 17
Wireless communications, 358–59
Wireless LANs (WLANs), 49, 210–11,
      280–81
   access points (APs), 211
   networks, 211
   technology, 210
Wireless telegraphy, 4
World Wide Web (WWW), 6, 302,
      331–37
   defined, 331
   HTML, 334–37
   HTTP, 332–34
   Java, 337
   URLs, 331–32
   See also Internet

X.25 network, 48–49