CHAPTER 1

INTRODUCTION

"To understand a science it is necessary to know its history"

—Auguste Comte (1798–1857)

1.1 *Historical Background*

With this quotation from Auguste Comte in mind, we begin this introductory study of communication systems with a historical account of this discipline that touches our daily lives in one way or another.[1] Each subsection in this section focuses on some important and related events in the historical evolution of communication.

Telegraph

The telegraph was perfected by Samuel Morse, a painter. With the words "What hath God wrought," transmitted by Morse's electric telegraph between Washington, D.C., and Baltimore, Maryland, in 1844, a completely revolutionary means of real-time, long-distance communications was triggered. The *telegraph*, ideally suited for manual keying, is the forerunner of digital communications. Specifically, the *Morse code* is a *variable-length* code using an alphabet of four symbols: a dot, a dash, a letter space, and a word space; short sequences represent frequent letters, whereas long sequences represent infrequent letters.

Radio

In 1864, James Clerk Maxwell formulated the *electromagnetic theory* of light and predicted the existence of radio waves; the underlying set of equations bears his name. The existence of radio waves was confirmed experimentally by Heinrich Hertz in 1887. In 1894, Oliver Lodge demonstrated wireless communication over a relatively short distance (150 yards). Then, on December 12, 1901, Guglielmo Marconi received a *radio signal* at Signal Hill in Newfoundland; the radio signal had originated in Cornwall, England, 1700 miles away across the Atlantic. The way was thereby opened toward a tremendous broadening of the scope of communications. In 1906, Reginald Fessenden, a self-educated academic, made history by conducting the first radio broadcast.

In 1918, Edwin H. Armstrong invented the *superheterodyne radio* receiver; to this day, almost all radio receivers are of this type. In 1933, Armstrong demonstrated another revolutionary concept—namely, a modulation scheme that he called *frequency modulation* (FM). Armstrong's paper making the case for FM radio was published in 1936.

[1] This historical background is adapted from Haykin's book (2001).

Telephone

In 1875, the *telephone* was invented by Alexander Graham Bell, a teacher of the deaf. The telephone made real-time transmission of speech by electrical encoding and replication of sound a practical reality. The first version of the telephone was crude and weak, enabling people to talk over short distances only. When telephone service was only a few years old, interest developed in automating it. Notably, in 1897, A. B. Strowger, an undertaker from Kansas City, Missouri, devised the automatic *step-by-step switch* that bears his name. Of all the electromechanical switches devised over the years, the Strowger switch was the most popular and widely used.

Electronics

In 1904, John Ambrose Fleming invented the *vacuum-tube diode*, which paved the way for the invention of the *vacuum-tube triode* by Lee de Forest in 1906. The discovery of the triode was instrumental in the development of transcontinental telephony in 1913 and signaled the dawn of wireless voice communications. Indeed, until the invention and perfection of the transistor, the triode was the supreme device for the design of electronic amplifiers.

The *transistor* was invented in 1948 by Walter H. Brattain, John Bardeen, and William Shockley at Bell Laboratories. The first silicon integrated circuit (IC) was produced by Robert Noyce in 1958. These landmark innovations in solid-state devices and integrated circuits led to the development of *very-large-scale integrated* (VLSI) circuits and single-chip *microprocessors*, and with them the nature of signal processing and the telecommunications industry changed forever.

Television

The first all-electronic *television* system was demonstrated by Philo T. Farnsworth in 1928, and then by Vladimir K. Zworykin in 1929. By 1939, the British Broadcasting Corporation (BBC) was broadcasting television on a commercial basis.

Digital Communications

In 1928, Harry Nyquist published a classic paper on the theory of signal transmission in telegraphy. In particular, Nyquist developed criteria for the correct reception of telegraph signals transmitted over dispersive channels in the absence of noise. Much of Nyquist's early work was applied later to the transmission of digital data over dispersive channels.

In 1937, Alex Reeves invented *pulse-code modulation* (PCM) for the digital encoding of speech signals. The technique was developed during World War II to enable the encryption of speech signals; indeed, a full-scale, 24-channel system was used in the field by the United States military at the end of the war. However, PCM had to await the discovery of the transistor and the subsequent development of large-scale integration of circuits for its commercial exploitation.

The invention of the transistor in 1948 spurred the application of electronics to switching and digital communications. The motivation was to improve reliability, increase capacity, and reduce cost. The first call through a stored-program system was placed in March 1958 at Bell Laboratories, and the first commercial telephone service with digital switching began in Morris, Illinois, in June 1960. The first *T-1 carrier system* transmission was installed in 1962 by Bell Laboratories.

In 1943, D. O. North devised the *matched filter* for the optimum detection of a known signal in additive white noise. A similar result was obtained in 1946 independently by J. H. Van Vleck and D. Middleton, who coined the term *matched filter*.

In 1948, the theoretical foundations of digital communications were laid by Claude Shannon in a paper entitled "A Mathematical Theory of Communication." Shannon's paper was received with immediate and enthusiastic acclaim. It was perhaps this response that emboldened Shannon to amend the title of his paper to "The Mathematical Theory of Communications" when it was reprinted a year later in a book co-authored with Warren Weaver. It is noteworthy that prior to the publication of Shannon's 1948 classic paper, it was believed that increasing the rate of information transmission over a channel would increase the probability of error. The communication theory community was taken by surprise when Shannon proved that this was not true, provided the transmission rate was below the channel capacity.

Computer Networks

During the period 1943 to 1946, the first electronic digital computer, called the ENIAC, was built at the Moore School of Electrical Engineering of the University of Pennsylvania under the technical direction of J. Presper Eckert, Jr., and John W. Mauchly. However, John von Neumann's contributions were among the earliest and most fundamental to the theory, design, and application of digital computers, which go back to the first draft of a report written in 1945. Computers and terminals started communicating with each other over long distances in the early 1950s. The links used were initially voice-grade telephone channels operating at low speeds (300 to 1200 b/s). Various factors have contributed to a dramatic increase in data transmission rates; notable among them are the idea of *adaptive equalization*, pioneered by Robert Lucky in 1965, and efficient modulation techniques, pioneered by G. Ungerboeck in 1982. Another idea widely employed in computer communications is that of *automatic repeat-request* (ARQ). The ARQ method was originally devised by H. C. A. van Duuren during World War II and published in 1946. It was used to improve radio-telephony for telex transmission over long distances.

From 1950 to 1970, various studies were made on *computer networks*. However, the most significant of them in terms of impact on computer communications was the Advanced Research Projects Agency Network (ARPANET), first put into service in 1971. The development of ARPANET was sponsored by the Advanced Research Projects Agency of the U. S. Department of Defense. The pioneering work in *packet switching* was done on ARPANET. In 1985, ARPANET was renamed the *Internet*. The turning point in the evolution of the Internet occurred in 1990 when Tim Berners-Lee proposed a hypermedia software interface to the Internet, which he named the *World Wide Web*. In the space of only about two years, the Web went from nonexistence to worldwide popularity, culminating in its commercialization in 1994. We may explain the explosive growth of the Internet by offering these reasons:

- Before the Web exploded into existence, the ingredients for its creation were already in place. In particular, thanks to VLSI, personal computers (PCs) had already become ubiquitous in homes throughout the world, and they were increasingly equipped with modems for interconnectivity to the outside world.
- For about two decades, the Internet had grown steadily (albeit within a confined community of users), reaching a critical threshold of electronic mail and file transfer.
- Standards for document description and transfer, hypertext markup language (HTML), and hypertext transfer protocol (HTTP) had been adopted.

Thus, everything needed for creating the Web was already in place except for two critical ingredients: a simple user interface and a brilliant service concept.

Satellite Communications

In 1955, John R. Pierce proposed the use of satellites for communications. This proposal was preceded, however, by an earlier paper by Arthur C. Clark that was published in 1945, also proposing the idea of using an *Earth-orbiting* satellite as a relay point for communication between two Earth stations. In 1957, the Soviet Union launched Sputnik I, which transmitted telemetry signals for 21 days. This was followed shortly by the launching of Explorer I by the United States in 1958, which transmitted telemetry signals for about five months. A major experimental step in communications satellite technology was taken with the launching of Telstar I from Cape Canaveral on July 10, 1962. The Telstar satellite was built by Bell Laboratories, which had acquired considerable knowledge from pioneering work by Pierce. The satellite was capable of relaying TV programs across the Atlantic; this was made possible only through the use of maser receivers and large antennas.

Optical Communications

The use of optical means (e.g., smoke and fire signals) for the transmission of information dates back to prehistoric times. However, no major breakthrough in optical communications was made until 1966, when K. C. Kao and G. A. Hockham of Standard Telephone Laboratories, U. K., proposed the use of a clad glass fiber as a dielectric waveguide. The *laser* (an acronym for light amplification by stimulated emission of radiation) had been invented and developed in 1959 and 1960. Kao and Hockham pointed out that (1) the attenuation in an optical fiber was due to impurities in the glass, and (2) the intrinsic loss, determined by Rayleigh scattering, is very low. Indeed, they predicted that a loss of 20 dB/km should be attainable. This remarkable prediction, made at a time when the power loss in a glass fiber was about 1000 dB/km, was to be demonstrated later. Nowadays, transmission losses as low as 0.1 dB/km are achievable.

The spectacular advances in microelectronics, digital computers, and lightwave systems that we have witnessed to date, and that will continue into the future, are all responsible for dramatic changes in the telecommunications environment. Many of these changes are already in place, and more changes will occur over time.

1.2 *Applications*

The historical background of Section 1.1 touches many of the applications of communication systems, some of which are exemplified by the telegraph that has come and gone, while others exemplified by the Internet are of recent origin. In what follows, we will focus on radio, communication networks exemplified by the telephone, and the Internet, which dominate the means by which we communicate in one of two basic ways or both, as summarized here:

- *Broadcasting*, which involves the use of a single powerful transmitter and numerous receivers that are relatively inexpensive to build. In this class of communication systems, information-bearing signals flow only in one direction, from the transmitter to each of the receivers out there in the field.
- *Point-to-point communications*, in which the communication process takes place over a link between a single transmitter and a single receiver. In this second class of communication systems, there is usually a bidirectional flow of information-bearing

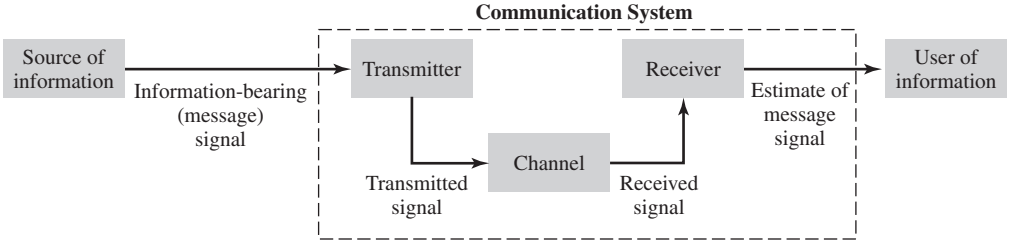

FIGURE 1.1 Elements of a communication system.

signals, which, in effect, requires the use of a transmitter and receiver (i.e., transceiver) at each end of the link.

The block diagram of Fig. 1.1 highlights the basic composition of a communication system. The *transmitter*, at some location in space, converts the *message signal* produced by a *source of information* into a form suitable for transmission over the channel. The *channel*, in turn, transports the message signal and delivers it to the receiver at some other location in space. However, in the course of transmission over the channel, the signal is *distorted* due to channel imperfections. Moreover, noise and interfering signals (originating from other sources) are added to the channel output, with the result that the received signal is a corrupted version of the transmitted signal. The *receiver* has the task of operating on the received signal so as to produce an *estimate* of the original message signal for the *user of information*. We say an "estimate" here because of the unavoidable deviation, however small, of the receiver output compared to the transmitter input, the deviation being attributed to channel imperfections, noise, and interference.

■ RADIO

Speaking in a generic sense, the radio embodies the means for broadcasting as well as point-to-point communications, depending on how it is used.

The *AM radio* and *FM radio* are both so familiar to all of us. (AM stands for amplitude modulation, and FM stands for frequency modulation.) The two of them are built in an integrated form inside a single unit, and we find them in every household and installed in every car. Via radio we listen to news about local, national, and international events, commentaries, music, and weather forecasts, which are transmitted from broadcasting stations that operate in our neighborhood. Traditionally, AM radio and FM radio have been built using analog electronics. However, thanks to the ever-increasing improvements and cost-effectiveness of digital electronics, *digital radio* (in both AM and FM forms) is already in current use.

Radio transmits voice by electrical signals. *Television*, which operates on similar electromagnetic and communication-theoretic principles, also transmits visual images by electrical signals. A voice signal is naturally defined as a *one-dimensional function of time*, which therefore lends itself readily to signal-processing operations. In contrast, an image with motion is a *two-dimensional function of time*, and therefore requires more detailed attention. Specifically, each image at a particular instant of time is viewed as a frame subdivided into a number of small squares called *picture elements* or *pixels*; the larger the number of pixels used to represent an image, the better the resolution of that image will be. By *scanning* the pixels in an orderly sequence, the information contained in the image is converted into an electrical signal whose magnitude is proportional to the brightness level of the individual pixels. The electrical signal generated at the output of the scanner is

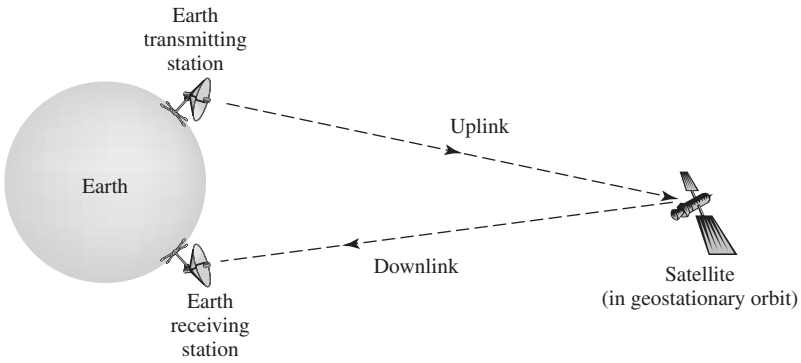

FIGURE 1.2 Satellite communication system.

the *video signal* that is transmitted. Generation of the video signal is the result of a well-defined *mapping process* known to the receiver. Hence, given the video signal, the receiver is able to reconstruct the original image. As with digital radio, television is also the beneficiary of spectacular advances in digital electronics. These advances, coupled with the application of advanced digital signal processing techniques and the demands of consumers, have motivated the development of *high-definition television* (HDTV), which provides a significant improvement in the quality of reconstructed images at the receiver output.

We turn next to the point-to-point communication scene. The radio has also touched our daily lives in highly significant ways through two avenues: satellite communications and wireless communications. *Satellite communications*, built around a satellite in *geostationary orbit*, relies on *line-of-sight radio propagation* for the operation of an uplink and a downlink. The uplink connects an Earth terminal to a transponder (i.e., electronic circuitry) on board the satellite, while the downlink connects the transponder to another Earth terminal. Thus, an information-bearing signal is transmitted from the Earth terminal to the satellite via the uplink, amplified in the transponder, and then retransmitted from the satellite via the downlink to the other Earth terminal, as illustrated in Fig. 1.2. In so doing, a satellite communication system offers a unique capability: *global coverage*.

In a loose sense, *wireless communications* operates in a manner similar to satellite communications in that it also involves a downlink and an uplink. The *downlink* is responsible for forward-link radio transmission from a *base station* to its mobile users. The uplink is responsible for reverse-link radio transmission from the mobile users to their base stations. Unlike satellite communications, the operation of wireless communications is dominated by the *multipath phenomenon* due to reflections of the transmitted signal from objects (e.g., buildings, trees, etc.) that lie in the propagation path. This phenomenon tends to degrade the receiver performance, which makes the design of the receiver a challenging task. In any event, wireless communications offers a unique capability of its own: *mobility*. Moreover, through the use of the cellular concept, the wireless communication system is enabled to *reuse* the radio spectrum over a large area as many times as possible. Within a cell, the available communication resources can be shared by the mobile users operating within that cell.

■ COMMUNICATION NETWORKS

The *computer* was originally conceived as a machine working by itself to perform numerical calculations. However, given the natural ability of a computer to perform logical functions, it was soon recognized that the computer is ideally suited to the design of

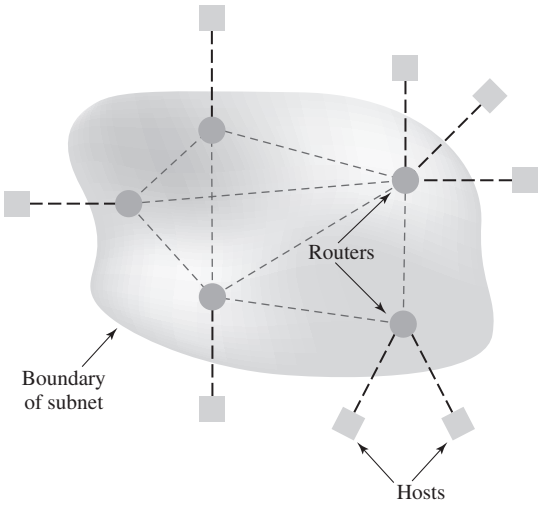

FIGURE 1.3 Communication network.

communication networks. As illustrated in Fig. 1.3, a *communication network* consists of the interconnection of a number of *routers* that are made up of intelligent processors (e.g., microprocessors). The primary purpose of these processors is to route voice or data through the network, hence the name "routers." Each router has one or more *hosts* attached to it; hosts refer to devices that communicate with one another. The purpose of a network is to provide for the delivery or exchange of voice, video, or data among its hosts, which is made possible through the use of *digital switching*. There are two principal forms of switching: circuit switching and packet switching.

In *circuit switching*, dedicated communication paths are established for the transmission of messages between two or more terminals, called *stations*. The communication path or *circuit* consists of a connected sequence of links from source to destination. For example, the links may consist of time slots (as in time-division multiplexed systems), for which a common channel is available for multiple users. The important point to note is that once it is in place, the circuit remains uninterrupted for the entire duration of transmission. Circuit switching is usually controlled by a centralized hierarchical control mechanism with knowledge of the network's entire organization. To establish a circuit-switched connection, an available path through the telephone network is seized and then dedicated to the exclusive use of the two users wishing to communicate. In particular, a call-request signal propagates all the way to the destination, whereupon it is acknowledged before communication can begin. Then, the network is effectively transparent to the users, which means that during the entire connection time the resources allocated to the circuit are essentially "owned" by the two users. This state of affairs continues until the circuit is disconnected.

Circuit switching is well suited for telephone networks, where the transmission of voice constitutes the bulk of the network's traffic. We say so because voice gives rise to a stream traffic, and voice conversations tend to be of long duration (about 2 minutes on the average) compared to the time required for setting up the circuit (about 0.1 to 0.5 seconds).

In *packet switching*,[2] on the other hand, the sharing of network resources is done on a *demand* basis. Hence, packet switching has an advantage over circuit switching in that

[2] Packet switching was invented by P. Baran in 1964 to satisfy a national defense need of the United States. The original need was to build a distributed network with different levels of redundant connections, which is *robust* in the sense that the network can withstand the destruction of many nodes due to a concerted attack, yet the surviving nodes are able to maintain intercommunication for carrying common and control information; see Baran (1990).

when a link has traffic to send, the link tends to be more fully utilized. Unlike voice signals, data tend to occur in the form of *bursts* on an occasional basis.

The network principle of packet switching is *store and forward*. Specifically, in a *packet-switched network*, any message longer than a specified size is subdivided prior to transmission into segments not exceeding the specified size. The segments so formed are called *packets*. After transporting the packets across different parts of the network, the original message is reassembled at the destination on a packet-by-packet basis. The network may thus be viewed as a pool of network resources (i.e., channel bandwidth, buffers, and switching processors), with the resources being *dynamically shared* by a community of competing hosts that wish to communicate. This dynamic sharing of network resources is in direct contrast to the circuit-switched network, where the resources are dedicated to a pair of hosts for the entire period they are in communication.

■ DATA NETWORKS

A communication network in which the hosts are all made up of computers and terminals is commonly referred to as a *data network*. The design of such a network proceeds in an orderly way by looking at the network in terms of a *layered architecture*, which is regarded as a hierarchy of nested layers. A *layer* refers to a process or device inside a computer system that is designed to perform a specific function. Naturally, the designers of a layer will be familiar with its internal details and operation. At the system level, however, a user views the layer in question merely as a "black box," which is described in terms of inputs, outputs, and the functional relation between the outputs and inputs. In the layered architecture, each layer regards the next lower layer as one or more black boxes with some given functional specification to be used by the given higher layer. In this way, the highly complex communication problem in data networks is resolved as a manageable set of well-defined interlocking functions. It is this line of reasoning that has led to the development of the *open systems interconnection* (OSI) *reference model*.[3] The term "open" refers to the ability of any two systems to interconnect, provided they conform to the reference model and its associated standards.

In the OSI reference model, the communications and related-connection functions are organized as a series of *layers* with well-defined interfaces. Each layer is built on its predecessor. In particular, each layer performs a related subset of primitive functions, and it relies on the next lower layer to perform additional primitive functions. Moreover, each layer offers certain services to the next higher layer and shields that layer from the implementation details of those services. Between each pair of layers there is an *interface*, which defines the services offered by the lower layer to the upper layer.

As illustrated in Fig. 1.4, the OSI model is composed of seven layers. The figure also includes a description of the functions of the individual layers of the model. Layer k on system A, say, communicates with a layer R on some other system B in accordance with a set of rules and conventions, which collectively constitute layer k *protocol*, where $k = 1, 2, \ldots, 7$. (The term "protocol" has been borrowed from common usage that describes conventional social behavior between human beings.) The entities that comprise the corresponding layers on different systems are referred to as *peer processes*. In other words, communication between system A and system B is achieved by having the peer processes in the two systems communicate via protocol. Physical connection between peer processes

[3] The OSI reference model was developed by a subcommittee of the International Organization for Standardization (ISO) in 1977. For a discussion of the principles involved in arriving at the original seven layers of the OSI model and a description of the layers themselves, see Tannenbaum (1996).

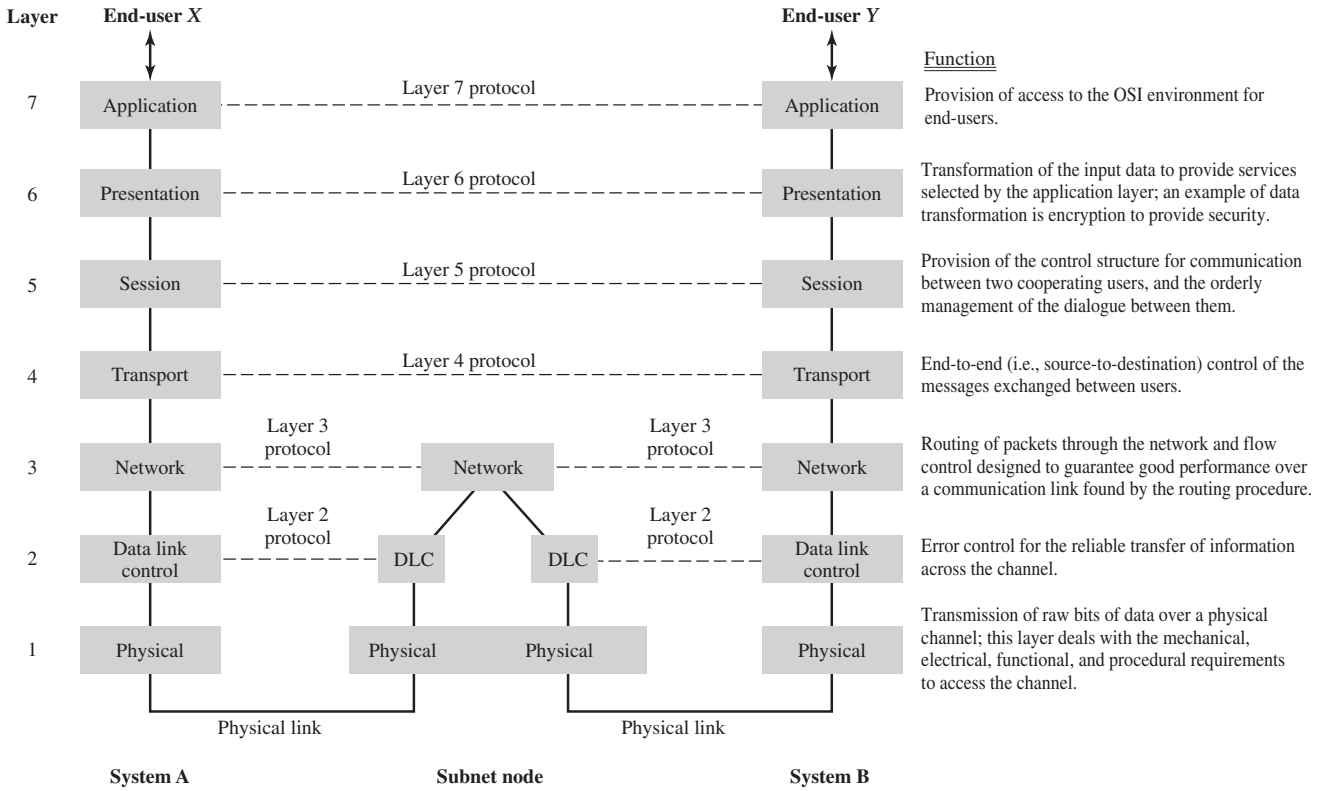

FIGURE 1.4 OSI model; the acronym DLC in the middle of the figure stands for *data link control*.

exists only at layer 1—namely, the *physical layer*. The remaining layers, 2 through 7, are in *virtual communication* with their distant peers. Each of these latter six layers exchanges data and control information with its neighboring layers (lower and above) through layer-to-layer interfaces. In Fig. 1.4, physical communication is shown by solid lines, and virtual communications are shown by dashed lines.

■ INTERNET[4]

The discussion of data networks just presented leads to the *Internet*. In the Internet paradigm, the underlying network technology is decoupled from the applications at hand by adopting an abstract definition of network service. In more specific terms, we may say the following:

- The applications are carried out independently of the technology employed to construct the network.
- By the same token, the network technology is capable of evolving without affecting the applications.

[4] For a fascinating account of the Internet, its historical evolution from the ARPANET, and international standards, see Abbate (2000). For easy-to-read essays on the Internet, see Special Issue, IEEE Communications Magazine (2002); the articles presented therein are written by pioneering contributors to the development of the Internet.

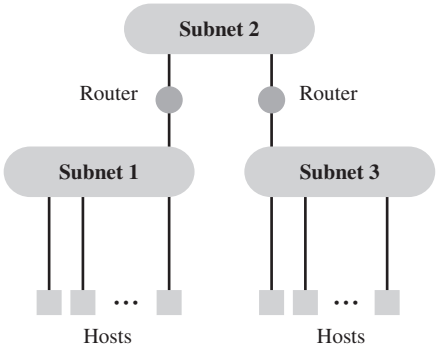

FIGURE 1.5 An interconnected network of subnets.

The Internet application depicted in Fig. 1.5 has three functional blocks: hosts, subnets, and routers. The hosts constitute nodes of the network, where data originate or where they are delivered. The routers constitute intermediate nodes that are used to cross subnet boundaries. Within a subnet, all the hosts belonging to that subnet exchange data directly; see, for example, subnets 1 and 3 in Fig. 1.5. In basic terms, the internal operation of a subnet is organized in two different ways (Tanenbaum, 1996):

1. *Connected* manner, where the connections are called *virtual circuits*, in analogy with physical circuits set up in a telephone system.
2. *Connectionless* manner, where the independent packets are called *datagrams*, in analogy with telegrams.

Like other data networks, the Internet has a layered set of protocols. In particular, the exchange of data between the hosts and routers is accomplished by means of the *Internet protocol* (IP), as illustrated in Fig. 1.6. The IP is a universal protocol that resides in the network layer (i.e., layer 3 of the OSI reference model). It is simple, defining an addressing plan with a built-in capability to transport data in the form of packets from node to node. In crossing a subnetwork boundary, the routers make the decisions as to how the packets addressed for a specified destination should be routed. This is done on the basis of routing tables that are developed through the use of custom protocols for exchanging pertinent information with other routers. The net result of using the layered set of protocols is the provision of *best effort service*. That is, the Internet offers to deliver each packet of data,

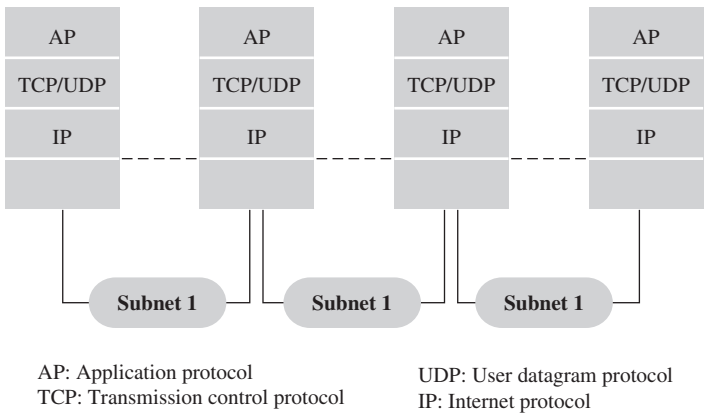

FIGURE 1.6 Illustrating the network architecture of the Internet.

but there are no guarantees on the transit time experienced in delivery or even whether the packets will be delivered to the intended recipient.

The Internet has evolved into a worldwide system, placing computers at the heart of a communication medium that is changing our daily lives in the home and workplace in profound ways. We can send an *e-mail message* from a host in North America to another host in Australia at the other end of the globe, with the message arriving at its destination in a matter of seconds. This is all the more remarkable because the packets constituting the message are quite likely to have taken entirely different paths as they are transported across the network.

Another application that demonstrates the remarkable power of the Internet is our use of it to *surf the Web*. For example, we may use a *search engine* to identify the references pertaining to a particular subject of interest. A task that used to take hours and sometimes days searching through books and journals in the library now occupies a matter of seconds!

To fully utilize the computing power of the Internet from a host located at a remote site, we need a *wideband modem* (i.e., modulator-demodulator) to provide a fast communication link between that host and its subnet. When we say "fast," we mean operating speeds on the order of megabits per second and higher. A device that satisfies this requirement is the so-called *digital subscriber line* (DSL). What makes the DSL all the more remarkable is the fact that it can operate over a linear wideband channel with an arbitrary frequency response. Such a channel is exemplified by an ordinary telephone channel built using twisted pairs for signal transmission. A *twisted pair* consists of two solid copper conductors, each of which is encased in a polyvinyl chloride (PVC) sheath. Twisted pairs are usually made up into cables, with each cable consisting of many twisted pairs in close proximity to each other. From a signal-transmission viewpoint, the DSL satisfies the challenging requirement described herein by following the well-known engineering principle of *divide and conquer*. Specifically, the given wideband channel is *approximated* by a set of narrowband channels, each of which can then be accommodated in a relatively straightforward manner.

One last comment is in order. Typically, access to the Internet is established via hosts in the form of computer terminals (i.e., servers). The access is expanded by using *hand-held devices* that act as hosts, which communicate with subnets of the Internet via wireless links. Thus, by adding mobility through the use of wireless communications to the computing power of the Internet to communicate, we have a new communication medium with enormous practical possibilities.

Integration of Telephone and Internet

One of the important challenges facing the telecommunications industry is the transmission of *Voice over Internet Protocol* (VoIP), which would make it possible to integrate telephony services with the rapidly growing Internet-based applications. The challenge is all the more profound because the IP is designed to accommodate the exchange of data between the hosts and the routers, which makes it difficult to support quality of service for VoIP. *Quality of service* (QoS) is measured in terms of two parameters:

- *Packet loss ratio*, defined as the number of packets lost in transport across the network to the total number of packets pumped into the network.
- *Connection delay*, defined as the time taken for a packet of a particular host-to-host connection to transmit across the network.

Subjective tests performed on VoIP show that in order to provide voice-grade telephone service, the packet loss ratio must be held below 1 percent, and one-way connection delay

can accumulate up to 160 ms without significant degradation of quality. Well-designed and managed VoIP networks, satisfying these provisions, are being deployed. However, the issue of *initial-echo control* remains a challenge.[5] Initial echo refers to the echo experienced at the beginning of a call on the first word or couple of words out of a user's mouth. The echo arises due to an impedance mismatch somewhere in the network, whereupon the incident signal is reflected back to the source.

Looking into the future, we may make the following remarks on internet telephony:

1. VoIP will replace *private branch exchanges* (PBXs) and other office switches; PBXs are remote switching units that have their own independent controls.[6]
2. VoIP is also currently having success with longer distance calls, but this is mainly due to the excess capacity that is now available on long-haul networks. If the loading on these long-haul networks increases, the delays will increase and a real-time service such as VoIP will be degraded. Accordingly, if long-service providers keep adding capacity so that loading is always low and response time is fast, thereby ensuring quality of service, then VoIP telephony may become mainstream and widespread.

■ DATA STORAGE

When considering important applications of digital communication principles, it is natural to think in terms of broadcasting and point-to-point communication systems. Nevertheless, the very same principles are also applied to the digital storage of audio and video signals, exemplified by *compact disc (CD)* and *digital versatile disc (DVD) players*. DVDs are refinements of CDs in that their storage capacity (in the order of tens of gigabytes) are orders of magnitude higher than that of CDs, and they can also deliver data at a much higher rate.

The digital domain is preferred over the analog domain for the storage of audio and video signals for the following compelling reasons:

(i) The quality of a digitized audio/video signal, measured in terms of frequency response, linearity, and noise, is determined by the digital-to-analog conversion (DAC) process, the parameterization of which is under the designer's control.

(ii) Once the audio/video signal is digitized, we can make use of well-developed and powerful encoding techniques for data compression to reduce bandwidth, and error-control coding to provide protection against the possibility of making errors in the course of storage.

(iii) For most practical applications, the digital storage of audio and video signals does not degrade with time.

(iv) Continued improvements in the fabrication of integrated circuits used to build CDs and DVDs ensure the ever-increasing cost-effectiveness of these digital storage devices.

With the help of the powerful encoding techniques built into their design, DVDs can hold hours of high-quality audio-visual contents, which, in turn, makes them ideally suited for interactive multimedia applications.

---

[5] The limits on QoS measures mentioned herein are taken from the overview article by James, Chen, and Garrison (2004), which appears in a Special Issue of the *IEEE Communications Magazine* devoted to voice VoIP and quality of service.

[6] PBXs are discussed in McDonald (1990).

1.3 *Primary Resources and Operational Requirements*

The communication systems described in Section 1.2 cover many diverse fields. Nevertheless, in their own individual ways, the systems are designed to provide for the *efficient* utilization of two *primary communication resources:*

- *Transmitted power*, which is defined as the average power of the transmitted signal.
- *Channel bandwidth*, which is defined by the width of the passband of the channel.

Depending on which of these two resources is considered to be the limiting factor, we may classify communication channels as follows:

(i) *Power-limited channels*, where transmitted power is at a premium. Examples of such channels include the following:
- *Wireless channels*, where it is desirable to keep the transmitted power low so as to prolong battery life.
- *Satellite channels*, where the available power on board the satellite transponder is limited, which, in turn, necessitates keeping the transmitted power on the downlink at a low level.
- *Deep-space links*, where the available power on board a probe exploring outer space is extremely limited, which again requires that the average power of information-bearing signals sent by the probe to an Earth station be maintained as low as possible.

(ii) *Band-limited channels*, where channel bandwidth is at a premium. Examples of this second category of communication channels include the following:
- *Telephone channels*, where, in a multi-user environment, the requirement is to minimize the frequency band allocated to the transmission of each voice signal while making sure that the quality of service for each user is maintained.
- *Television channels*, where the available channel bandwidth is limited by regulatory agencies and the quality of reception is assured by using a high enough transmitted power.

Another important point to keep in mind is the unavoidable presence of noise at the receiver input of a communication system. In a generic sense, *noise* refers to unwanted signals that tend to disturb the quality of the received signal in a communication system. The sources of noise may be internal or external to the system. An example of internal noise is the ubiquitous *channel noise* produced by thermal agitation of electrons in the front-end amplifier of the receiver. Examples of external noise include atmospheric noise and interference due to transmitted signals pertaining to other users.

A quantitative way to account for the beneficial effect of the transmitted power in relation to the degrading effect of noise (i.e., assess the quality of the received signal) is to think in terms of the *signal-to-noise ratio* (SNR), which is a dimensionless parameter. In particular, the SNR at the receiver input is formally defined as the ratio of the average power of the received signal (i.e., channel output) to the average power of noise measured at the receiver input. The customary practice is to express the SNR in *decibels* (dBs), which is defined as 10 times the logarithm (to base 10) of the power ratio.[7] For example, signal-to-noise ratios of 10, 100, and 1000 are 10, 20, and 30 dBs, respectively.

[7] For a discussion of the decibel, see Appendix 1.

In light of this discussion, it is now apparent that as far as performance evaluation is concerned, there are only two *system-design parameters*: signal-to-noise ratio and channel bandwidth. Stated in more concrete terms:

> The design of a communication system boils down to a tradeoff between signal-to-noise ratio and channel bandwidth.

Thus, we may improve *system performance* by following one of two alternative design strategies, depending on system constraints:

1. Signal-to-noise ratio is increased to accommodate a limitation imposed on channel bandwidth.
2. Channel bandwidth is increased to accommodate a limitation imposed on signal-to-noise ratio.

Of these two possible design approaches, we ordinarily find that strategy 1 is simpler to implement than strategy 2, because increasing signal-to-noise ratio can be accomplished simply by raising the transmitted power. On the other hand, in order to exploit increased channel bandwidth, we need to increase the bandwidth of the transmitted signal, which, in turn, requires increasing the *complexity* of both the transmitter and receiver.

1.4 *Underpinning Theories of Communication Systems*

The study of communication systems is challenging not only in technical terms but also in theoretical terms. In this section, we highlight four theories, each of which is essential for understanding a specific aspect of communication systems.[8]

■ MODULATION THEORY

Modulation is a signal-processing operation that is basic to the transmission of an information-bearing signal over a communication channel, whether in the context of digital or analog communications. This operation is accomplished by changing some parameter of a *carrier wave* in accordance with the information-bearing (message) signal. The carrier wave may take one of two basic forms, depending on the application of interest:

- *Sinusoidal carrier wave*, whose amplitude, phase, or frequency is the parameter chosen for modification by the information-bearing signal.
- *Periodic sequence of pulses*, whose amplitude, width, or position is the parameter chosen for modification by the information-bearing signal.

Regardless of which particular approach is used to perform the modulation process, the issues in modulation theory that need to be addressed are:

- Time-domain description of the modulated signal.
- Frequency-domain description of the modulated signal.
- Detection of the original information-bearing signal and evaluation of the effect of noise on the receiver.

[8] One other theory—namely, Information Theory—is basic to the study of communication systems. We have not included this theory here because of its highly mathematical and therefore advanced nature, which makes it inappropriate for an introductory book.

Fourier Analysis

The *Fourier transform* is a *linear* mathematical operation that transforms the time-domain description of a signal into a frequency-domain description without loss of information, which means that the original signal can be recovered exactly from the frequency-domain description. However, for the signal to be Fourier transformable, certain conditions have to be satisfied. Fortunately, these conditions are satisfied by the kind of signals encountered in the study of communication systems.

Fourier analysis provides the mathematical basis for evaluating the following issues:

- Frequency-domain description of a modulated signal, including its transmission bandwidth.
- Transmission of a signal through a linear system exemplified by a communication channel or (frequency-selective) filter.
- Correlation (i.e., similarity) between a pair of signals.

These evaluations take on even greater importance by virtue of an algorithm known as the *fast Fourier transform*, which provides an efficient method for computing the Fourier transform.

Detection Theory

Given a received signal, which is perturbed by additive channel noise, one of the tasks that the receiver has to tackle is how to *detect* the original information-bearing signal in a reliable manner. The *signal-detection problem* is complicated by two issues:

- The presence of noise.
- Factors such as the unknown phase-shift introduced into the carrier wave due to transmission of the sinusoidally modulated signal over the channel.

Dealing with these issues in analog communications is radically different from dealing with them in digital communications. In analog communications, the usual approach focuses on *output signal-to-noise ratio* and related calculations. In digital communications, on the other hand, the signal-detection problem is viewed as one of *hypothesis testing*. For example, in the specific case of binary data transmission, given that binary symbol 1 is transmitted, what is the probability that the symbol is correctly detected, and how is that probability affected by a change in the received signal-to-noise ratio at the receiver input?

Thus, in dealing with detection theory, we address the following issues in analog communications:

- The figure of merit for assessing the noise performance of a specific modulation strategy.
- The threshold phenomenon that arises when the transmitted signal-to-noise ratio drops below a critical value.
- Performance comparison of one modulation strategy against another.

In digital communications, on the other hand, we look at:

- The average probability of symbol error at the receiver output.
- The issue of dealing with uncontrollable factors.
- Comparison of one digital modulation scheme against another.

■ PROBABILITY THEORY AND RANDOM PROCESSES

From the brief discussion just presented on the role of detection theory in the study of communication systems, it is apparent that we need to develop a good understanding of the following:

- Probability theory for describing the behavior of randomly occurring events in mathematical terms.
- Statistical characterization of random signals and noise.

Unlike a deterministic signal, a *random signal* is a signal about which there is *uncertainty* before it occurs. Because of the uncertainty, a random signal may be viewed as belonging to an *ensemble*, or a *group*, of signals, with each signal in the ensemble having a different waveform from that of the others in the ensemble. Moreover, each signal within the ensemble has a certain probability of occurrence. The ensemble of signals is referred to as a *random process* or *stochastic process*. Examples of a random process include:

- Electrical noise generated in the front-end amplifier of a radio or television receiver.
- Speech signal produced by a male or female speaker.
- Video signal transmitted by the antenna of a TV broadcasting station.

In dealing with probability theory, random signals, and noise, we address the following issues:

- Basic concepts of probability theory and probabilistic models.
- Statistical description of a random process in terms of ensemble as well as temporal averages.
- Mathematical analysis and processing of random signals.

1.5 *Concluding Remarks*

In this chapter, we have given a historical account and applications of communications and a brief survey of underlying theories of communication systems. In addition, we presented the following points to support our view that the study of this discipline is both highly challenging and truly exciting:

(i) Communication systems encompass many and highly diverse applications: radio, television, wireless communications, satellite communications, deep-space communications, telephony, data networks, Internet, and quite a few others.

(ii) Digital communication has established itself as the dominant form of communication. Much of the progress that we have witnessed in the advancement of digital communication systems can be traced to certain enabling theories and technologies, as summarized here:

- Abstract mathematical ideas that are highly relevant to a deep understanding of the processing of information-bearing signals and their transmission over physical media.
- Digital signal-processing algorithms for the efficient computation of spectra, correlation, and filtering of signals.
- Software development and novel architectures for designing microprocessors.
- Spectacular advances in the physics of solid-state devices and the fabrication of very-large-scale integrated (VLSI) chips.

(iii) The study of communication systems is a *dynamic discipline*, continually evolving by exploiting new technological innovations in other disciplines and responding to new societal needs.

(iv) Last but by no means least, communication systems touch our daily lives both at home and in the workplace, and our lives would be much poorer without the wide availability of communication devices that we take for granted.

The remainder of the book, encompassing ten chapters, provides an introductory treatment of both analog and digital kinds of communication systems. The book should prepare the reader for going on to deepen his or her knowledge of a discipline that is best described as *almost limitless* in scope. This is especially the case given the trend toward the unification of wireline and wireless networks to accommodate the integrated transmission of voice, video, and data.