CHAPTER

10

COMPUTER PERIPHERALS

CHAPTER OBJECTIVES

In this chapter you will learn about:

- How computer input and output devices work
- The operation of scanners and printers
- Graphics cards and processing of graphical images
- Synchronous and asynchronous serial data links
- High-speed Internet connections using ADSL and cable modems

In previous chapters, we discussed hardware and software features of processors, memories, disks, and CD ROMs. We also discussed the means by which a computer communicates with external devices, including the hardware and software facilities that support program-controlled I/O, interrupts, and direct memory access. This chapter presents the characteristics of some commonly used computer peripherals and how they are connected in a computer system.

The name *peripheral* refers to any external device connected to a computer. In this context, the computer consists only of the processor and its memory. Computer peripherals can be divided into two categories, according to function. The first category contains devices that perform input and output operations, such as the keyboard, mouse, trackball, printer, and video display. The second category contains devices intended primarily for secondary storage of data, with primary storage being provided by the main memory of the computer. Some mass storage devices, in particular magnetic disks, are used for *on-line storage* of data. In others, such as optical disks, floppy disks, and magnetic tapes, the storage medium can be removed from the drive unit for transferring data from one computer system to another. For example, the device most often used for distributing software is the optical disk, also know as a CD ROM. Secondary storage devices are discussed in Chapter 5.

Today, among the most important computer peripherals are devices that provide connection to the Internet. Much of the tremendous growth in the computer field in recent years is a result of the synergy between computers and communications, and the emergence of many innovative applications on the World Wide Web. These developments have touched every aspect of our lives, from business to entertainment and education.

In this chapter, we present an overview of the variety of input and output devices used in modern computer systems and give a brief description of the underlying technology. Devices that are outside the computer box often use a serial link, either wired or wireless, to communicate with the processor. We will present some of the basic ideas for serial communications.

10.1 INPUT DEVICES

Input devices include keyboards and devices used to move the cursor on the screen, such as the mouse, trackball, and joystick. Scanners and digital cameras are also extensively used to capture images and feed them into the computer in the form of digital data.

10.1.1 KEYBOARD

The most commonly encountered input device is a keyboard, usually complemented by a mouse or a trackball. Together with a video display as an output device, they are used for direct human interaction with the computer.

Keyboards are available in two types. One type consists of an array of mechanical switches mounted on a printed-circuit board. The switches are organized in rows and columns and connected to a microcontroller on the board. When a switch is pressed, the controller identifies the row and column, and thus determines which key is being pressed. After correcting for switch bounce (see Chapter 4), the controller generates a code representing that switch and sends it over a serial link to the computer.

The second type uses a flat structure consisting of three layers. The top layer is a plasticized material, with key positions marked on the top surface and conducting traces deposited on the underside. The middle layer is made of rubber, with holes at key positions. The bottom layer is metallic, with raised bumps at key positions. When pressure is applied to the top layer at a key position, the trace underneath comes in contact with the corresponding bump on the bottom layer, thus completing an electrical circuit in the same way as a mechanical switch. The current that flows in this circuit is sensed by the microcontroller. This arrangement provides a low-cost keyboard that also has the advantage of being robust and immune to problems caused by spilt food or drink. Such keyboards are commonly encountered in applications such as point-of-sale terminals.

10.1.2 MOUSE

The invention of the mouse in 1968 represented an important step in the development of new means for people to communicate with computers. Up to that point, text was the primary form of data entry. The mouse made it possible to enter graphic information directly, by drawing the desired objects, and opened the door to many new and powerful ideas, including windows and pull-down menus.

The mouse is a device shaped to fit comfortably in the operator's hand, such that it can be moved over a flat surface. An electronic circuit senses this movement and sends some measure of the distance traveled in the X and Y directions to the computer. Movement is monitored either mechanically or optically. A mechanical mouse is fitted with a ball mounted such that it can rotate freely as the mouse is moved. The rotation of the ball is sensed and used to advance two counters, one for each of the two axes of motion. The mouse is also fitted with two or three pushbuttons. The information from the counters and the buttons is collected by a microcontroller, encoded as a 3-byte packet, and sent to the computer over a serial link.

An optical mouse uses a light-emitting diode (LED) to illuminate the surface on which the mouse is placed, and a light-sensitive device senses the light reflected from the surface. In some models, the mouse must be placed on a special pad that has a pattern of vertical and horizontal lines. The reflected light changes as the mouse moves from light to dark areas on the surface underneath, and the mouse measures the distance traveled by counting these changes.

A much more sophisticated optical mouse, dubbed IntelliMouse, was introduced by Microsoft in 1999. It can be used with almost any surface. Instead of a simple light sensor, the image of a small area of the surface underneath is focused on a tiny digital camera, which converts the image into a digital representation. The camera takes 1500 such pictures every second. Unless the surface is perfectly uniform and smooth, such as a mirror, its image will contain features such as lines, changes in brightness, and so on. By comparing successive images, a processor inside the mouse is able to measure the distance traveled with considerable accuracy. The processor uses a signal-processing technique known as correlation to determine the distance traveled from one picture to the next. This is a computationally intensive task that must be repeated 1500 times each second. It is made possible only by the availability of powerful yet low-cost embedded processors. The processor used executes 18 million instructions per second.

Since the invention of the mouse, a number of devices have been introduced that perform the same function. These include the trackball, the joystick and the touchpad.

10.1.3 TRACKBALL, JOYSTICK, AND TOUCHPAD

The mouse enables an operator to move a cursor on a computer screen. A host of innovative input devices have been developed to perform a similar function, to suit various application environments and user preferences.

The operating principles of a *trackball* are very similar to those of a mechanical mouse. A ball is mounted in a shallow well on the keyboard. The user rotates the ball to indicate the desired movement of the cursor on the screen.

The *joystick* is a short, pivoted stick that can be moved by hand to point in any direction in the X-Y plane. When this information is sent to the computer, the software moves the cursor on the screen in the same direction.

The position of the stick can be sensed by a suitable linear or angular position transducer, such as the potentiometer arrangement shown in Figure 10.1. The voltage outputs of the X and Y potentiometers are fed to two analog-to-digital (A/D) converters, whose outputs determine the position of the joystick and, thus, the desired direction of motion.

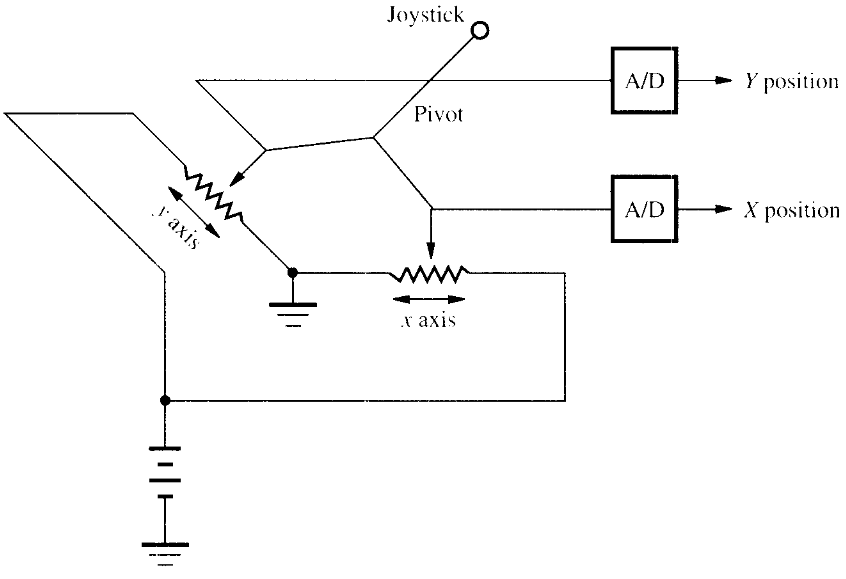

Figure 10.1 Joystick, using potentiometers as position transducers.

Joysticks are found in notebook computers and video games. In the case of a notebook computer, the joystick is mounted among the keys of the keyboard and sticks slightly above them. By virtue of its positioning, the joystick has the advantage of not requiring the operator's hand to be moved off the keyboard. It can be pushed with one finger to position the cursor on the screen. It is also mechanically robust and requires very little space. For use in video games, the joystick is shaped into a handle that suits the nature of the game. It is usually equipped with pushbuttons to be used for such purposes as shooting a ball or firing a gun.

Another very useful input device is the touchpad and its close relative, the touch screen. The touchpad is a small pad made of pressure-sensitive material. When the user's finger or the tip of a pen touches some point on the pad, the pressure causes a change in the electrical characteristics of the material at that spot. The location of the spot is detected and communicated to the computer. By moving a finger across the pad, the user can instruct the software to move the cursor on the screen in the same direction. This makes the touchpad a low-cost replacement for the mouse or the trackball, with a high degree of robustness and reliability because it contains no moving parts. Touchpads are well suited to notebook computers.

Many new materials have been developed for use as touchpads. Perhaps the most innovative is one that has a large number of tiny optical fibers embedded in it. The material can identify the location of an object touching it as well as the amount of pressure being applied. This material was developed for robotic applications in space. It is rapidly finding other applications, for example as an input device replacing and expanding the role of the piano keyboard.

A touchpad can be combined with a liquid-crystal display to produce a touch-sensitive screen that can be used for both input and output operations. This type of screen is commonly found in personal digital assistants (PDAs), such as the Palm Pilot. Another form of touch screen uses a cathode-ray tube (CRT). The change in capacitance caused by a finger touching the screen is sensed as the electron beam scans the screen to display an image. This arrangement is commonly found in cash registers and point-of-sale terminals.

10.1.4 SCANNERS

Scanners transform printed material and photographs into digital representations. In early scanners, the page to be scanned was mounted on a glass cylinder which rotates around the sensors. Most scanners today use a flat-bed arrangement, in which the page being scanned is placed over a flat glass surface. A source of light scans the page, and the reflected light is focused on a linear array of charge-coupled devices (CCDs). When a CCD device is exposed to light, an electrical charge is stored in a tiny capacitor associated with it such that the amount of charge is proportional to the intensity of the light. This charge is collected by appropriate circuitry and converted to a digital representation using an analog-to-digital converter. For color scanners, red, green, and blue filters are used to separate the primary colors and process them separately. As the light source moves across the page, the sensor array is read repeatedly, thus sampling successive lines of pixels of the image. It should be noted that these techniques are also used in digital copiers. A digital copier is, in effect, a combination of a scanner and a laser printer.

After scanning a printed page into a computer, the image is represented in the memory as an array of pixels. In its simplest form, each pixel is represented by one bit, indicating whether the corresponding spot on the image is light or dark. For higher quality images, more information is stored for each pixel to represent the color and intensity of light at that spot. Up to three bytes of information per pixel may be used, with one byte for each of the three primary colors.

Consider the case of a page of text. The dark areas of the image correspond to the printed characters. Many character recognition techniques have been developed that make it possible to analyze the pixel map stored in the memory and recognize the characters in the image. Thus, it becomes possible to create a text file that describes the contents of the printed page, with each character translated to a suitable binary code, such as ASCII. The resulting file may then be processed by a text-processing program such as Word.

10.2 OUTPUT DEVICES

Computer output may take a variety of forms, including alphanumeric text, graphical images, or sound. We describe below some of the commonly used output devices.

10.2.1 VIDEO DISPLAYS

Video displays are used whenever visual representation of computer output is needed. The most common display device uses a cathode-ray tube (CRT).

Let us start by describing how a picture is formed on a CRT. A focused beam of electrons strikes a fluorescent screen, causing emission of light that is seen as a bright spot against a dark background. The dot thus formed disappears when the beam is turned off or moved to another spot. Thus, in general, three independent variables need to be specified at all times, representing the position and intensity of the beam. The position of the beam corresponds to the X and Y coordinates of the spot on the screen. Its intensity, which is usually referred to as the Z-axis control, provides the *gray scale* or brightness information at that spot. The smallest addressable spot on the screen is called a pixel. It consists of a number of smaller dots of different sizes, arranged in some geometrical pattern. By illuminating these dots in various combinations, different levels of brightness can be obtained. This technique is known as *half-toning*. In color displays, each pixel has three different colors of fluorescent dots, red, green, and blue. Different colors are obtained by exciting these dots in different combinations.

The size of the spot formed on the screen by the electron beam determines the total number of pixels in an image. This is usually in the range of 700 to 2500 points along each of the X and Y coordinates. The Z-axis information is described in up to 24 bits, consisting of one byte for each color. This is judged to yield the highest color resolution that can be perceived by the human eye. The most common standard for computer video displays is VGA (Video Graphics Array) and its higher-quality variants. The basic VGA display has 640 × 480 pixels. Variations of this standard specify displays with higher resolution, such as 1024 × 768 (XVGA) and 1600 × 1200 (UXGA).

Both alphanumeric text and graphical pictures can be constructed using a technique called *raster scan*. The electron beam is swept successively across each row of pixels from left to right, until all rows have been scanned from the top to the bottom of the screen. Many video displays use *interlacing* to increase the perceived rate at which the screen is refreshed. The beam scans the screen in two passes, once for all odd-numbered rows and once for even-numbered rows. The image being displayed is stored in a *display buffer* memory that provides the Z-axis information during scanning. In the simplest representation, a *bit map* consisting of one bit for each pixel can be used to describe the image to be displayed on the screen. Hence, a screen with 1024 × 1024 pixels requires a 1M-bit display buffer memory. To refresh the display at the rate of 60 times per second, the data rate is 60 megabits/s. Today's high-quality displays use 32 bits per pixel, thus requiring much larger display buffers and communication bandwidth. Usually, only 24 of the 32 bits store color information. The fourth byte provides for compatibility with the word length of the host processor. It also provides room for future enhancements. Modern systems have the capability to overlap multiple distinct screen images, as in the case of window-based operating systems, and hence require several separate display buffers.

10.2.2 Flat-Panel Displays

Although cathode-ray tube technology has dominated display applications, flat-panel displays are becoming increasingly popular. They are thinner and lighter in weight. They provide better linearity and, in some cases, even higher resolution. Several types of flat-panel displays have been developed, including liquid-crystal panels, plasma panels, and electroluminescent panels. The availability of low-cost flat-panel displays has been instrumental in the development of notebook computers.

Liquid-crystal panels are constructed by sandwiching a thin layer of liquid crystal — a liquid that exhibits crystalline properties — between two transparent plates. The top plate has transparent electrodes deposited on it, and the back plate is a mirror. By applying appropriate electrical signals across the plates, various segments of the liquid crystal can be activated, causing changes in their light-diffusing or polarizing properties. Thus, these segments either transmit or block the light. An image is produced by passing light through selected segments of the liquid crystal and then reflecting it back from the mirror to the viewer. Liquid-crystal displays are found in watches, calculators, notebook computers, and many other devices.

Liquid-crystal displays come in two varieties. Static displays have a simple structure in which electrodes are deposited along one axis on the top plate and along an orthogonal axis on the lower plate, thus defining columns and rows. To illuminate a particular segment, a voltage is applied between one row electrode and one column electrode. This creates an electric field that causes the liquid crystal at their intersection point to turn on, and a bright spot is displayed. Displays that use this arrangement are simple to build and inexpensive, but the quality of the images they produce is low. The illuminated area is not well defined, so edges in the image are not sharp. Also, the long electrodes have a large capacitance, hence the speed with which a spot can be turned on or off is low. For example, if the cursor is moved across the screen quickly, the slow response causes a tail to be seen following the cursor.

A higher-quality display is produced by introducing a transistor at each intersection point. This provides faster response and better control over the area to be illuminated. The transistors are prepared on a thin film deposited on one of the plates. Hence this type of display is called a *thin-film transistor* (TFT) display. It is also known as an *active matrix* display. This is the type of display most commonly found in high-quality notebook computers.

Plasma panels consist of two glass plates separated by a thin gap filled with a gas such as neon. Each plate has several parallel electrodes running across it. The electrodes on the two plates run at right angles to each other. A voltage pulse applied between two electrodes, one on each plate, causes a small segment of gas at the intersection of the two electrodes to glow. The glow of gas segments is maintained by a lower voltage that is continuously applied to all electrodes. A similar pulsing arrangement is used to selectively turn points off. Plasma displays can provide high resolution but are rather expensive. They are found in applications where display quality is important and the bulky size of a cathode-ray tube is undesirable.

Electroluminescent panels use a thin layer of phosphor between two electrically conducting panels. The image is created by applying electrical signals to the plates, making the phosphor glow.

The viability of flat-panel displays for different applications is closely linked to developments in the competing cathode-ray tube display technology, which continues to provide a good combination of price and performance and permits easy implementation of color displays.

10.2.3 PRINTERS

Printers are used to produce hard copy of output data or text. They are usually classified as being either an impact or nonimpact type, depending on the nature of the printing mechanism used. Impact printers use mechanical printing mechanisms, and nonimpact printers rely on optical, ink-jet, or electrostatic techniques.

Nonimpact printers have few moving parts and can be operated at high speed. Laser printers use the same technology as photocopiers. A drum coated with positively charged photoconductive material is scanned by a laser beam. The positive charges that are illuminated by the beam are dissipated. Then a negatively charged toner powder is spread over the drum. It adheres to the positive charges, thus creating a page image that is then transferred to the paper. The drum is cleaned of any excess toner material to prepare it for printing the next page.

Other types of nonimpact printers use ink jets, in which droplets of different color inks are fired at the paper from tiny nozzles, to generate color output. A variety of techniques are used to fire the ink droplets. For example, in a *bubble ink-jet printer,* the nozzle is attached to a small chamber to which a heat pulse is applied. This causes the ink in the chamber to evaporate, forming a gas bubble that pushes a small amount of ink out of the nozzle. As the gas in the chamber cools down, it creates a vacuum that sucks in a new charge of ink. Ink-jet printers are generally more expensive than laser printers, and they produce higher-quality images.

Most printers form characters and graphic images in the same way that images are formed on a video screen, that is, by printing dots in matrices. This arrangement can easily accommodate a variety of fonts and can also be used for printing graphical images. However, because of the sensitivity of the human eye to regular patterns, a regular dot matrix is easily detected and interferes with the perceived quality of the image. High-quality printers use a technique called *dithering* to overcome this difficulty. Recall that a pixel consists of several dots, each having one of three colors. Dithering means that the geometrical arrangement of the dots in a pixel and the assignment of colors to each dot are varied. This breaks the monotony of a regular pattern and gives the appearance of having more color choices.

The highest quality printing is needed in applications such as graphic arts and photographic printing. Ink-jet printers that use a technique known as dye sublimation are suitable for these applications. They are also the most expensive. In this case, the temperature to which the ink is heated is controlled to change the amount of ink fired at the paper. Thus, the color intensity of each dot can be varied continuously. Also, special paper is used in which the ink diffuses, producing precisely controlled colors.

10.2.4 GRAPHICS ACCELERATORS

Many computer applications involve high-quality graphic images. Perhaps the most familiar use of graphics is in video games. Other applications include artistic work, medical imaging, and animated films. A high-quality image requires a large number of pixels to be displayed. Before an image is sent to a display screen, the color of each of these pixels has to be computed and stored in a memory buffer. From there, the information is sent to the screen at a rate of at least 30 times per second to keep the displayed image refreshed.

The task of computing pixel intensity and color can be done in software. The resulting image can be stored in a screen buffer in the computer's main memory, from where it can be sent to the display over the computer bus. However, the shear volume of data that need to be handled is such that this approach can easily swamp the processor and leave little computing power for other tasks. Also, using the computer bus to transfer the contents of the screen buffer to the display would consume a considerable portion of the bus bandwidth. With 32 bits per pixel, a 1024 × 1024-pixel image is represented by 4M bytes of data, which would create a minimum of 120 megabytes/s of traffic on the memory bus.

Most graphics applications require the ability to display three-dimensional (3D) objects. In computer games, for example, an artificial 3D world is created, with full-video images, entirely in software. The task of creating these images is computationally intensive. The most practical solution is to provide by a special-purpose processor, designed specifically to handle these intensive computations. Such a processor, known as a graphics-processing unit (GPU), is the basis of the popular graphics cards installed in most personal computers. The graphics card also includes a large high-speed memory, typically ranging from 8M to 64M bytes. This memory is used by the GPU while performing the computations, and it also stores the resulting image to be sent to the display screen. The display is connected directly to the graphics card, so that data transfers for refreshing the screen do not use the computer bus. A high-quality graphics card is capable of refreshing the screen between 75 and 200 times per second.

Graphics Port

The graphics card may be plugged into a computer bus such as PCI. More often, the computer motherboard includes a special connection slot known as the Accelerated Graphics Port (AGP), into which the graphics card is inserted. This is a 32-bit port that is capable of supporting higher data transfer rates than can be achieved on the PCI bus. It is usually described as AGP 1x, 2x, 4x, or 8x, where AGP 1x is the original standard which provides a data transfer rate of 264 megabytes/s. Later standards support multiples of this rate, with AGP 8x providing 2 gigabytes/s.

Graphics Processing

In computer graphics, a three-dimensional object is represented by dividing its surface into a large number of small polygons, usually triangles. The first task is to convert the 3D scene into a 2D representation that matches as closely as possible the image that would be seen by the human eye. *Projection* and *perspective* calculations determine the locations in the two-dimensional image of the vertices of the triangles representing various objects in the scene. Then, complex algorithms are used to determine appropriate color and shading for each of the triangles to create a realistic image. These computations take into account the lighting sources on the scene, reflections from various surfaces, shadows, and so on. An important step in this process is to give some texture to the surface, such as the appearance of wood grain or a brick wall. The texture is usually created using elements called *texels*. An array of texels is applied to individual image triangles to create the impression of a textured surface on the original three-dimensional object. Hidden parts of the scene are eliminated in a process known as *clipping* to save unnecessary computations. The final step is *sampling*, in which the image is sampled to determine the color and intensity of each image pixel. The entire computational process that reduces a 3D scene to a description of the pixels to be sent to the display is known as *rendering*.

For moving images, these computations must be repeated many times each second. To create the appearance of smooth motion on the screen, the image pixels must be recomputed at least 20 times per second, usually 30 to 40 times per second, to produce a high-quality video picture. This is called the *frame rate*. The ability of a video card to perform the required computations is often measured by its T&L (Transformations and Lighting) rating, which is the number of triangles per second for which the card can complete all the computations needed for projection, clipping, lighting, and sampling. Typical ratings are in the range 10 to 30 million triangles per second.

As an example, the salient characteristics of the RADEON VE graphics card manufactured by ATI Corp. are given in Table 10.1. The GeForce 2 MX graphics processor manufactured by *n*Vidia Corp. offers very similar capabilities. Both are popular for use in personal computers. Professional versions with enhanced capabilities are also available. Much more powerful processors can be expected in the near future in this rapidly expanding segment of the computer industry.

Table 10.1 RADEON VE graphics card

Feature	Description
GPU chip	RADEON VE
Bus	AGP 4x
Memory	Up to 64M bytes, DDR SDRAM
Color	32 bits, including 8 bits reserved for future use
Pixels	2048 × 1536
T&L rating	30M triangles per second
Screen refresh rate	75 to 200, where higher rates are for lower-resolution images
Additional capabilities	Provisions for use with TV, VCR, DVD, HDTV, and MPEG 2 compression

Graphics Software

Graphics cards offer a variety of sophisticated features. Making use of these features requires software designed specifically for the card. There are very few standards in this area, and the market is wide open for competition. Simply installing a better graphics card in a computer will not automatically improve the quality of the images produced. Specialized software for use with this card is needed. Some application programming interface (API) standards for graphics software are beginning to emerge. The objective of these standards is to enable hardware-independent software to be developed. Thus, the software for a computer game, for example, would work well with graphics cards manufactured by different companies and would be able to make use of the features that each provides. OpenGL (Open Graphics Language) is an example of such a standard. Increasingly, graphics cards are being designed for compatibility with this and a number of similar standards that relate to various aspects of graphics processing.

10.3 Serial Communication Links

Devices such as the keyboard and mouse are connected directly to the computer with which they are used, typically through a serial communication link. Other devices, such as printers and scanners, may be connected to a computer either directly or via a communication network, so that they may be shared among several users. As the Internet plays an important role in many computer applications, a computer is often connected to the Internet either permanently or over dialed telephone links.

In the remainder of this chapter, we discuss some of the schemes that are commonly used in serial communication links. We start by presenting some basic ideas.

Modulation and Demodulation

In a digital circuit, we represent one bit by an electrical signal that has one of two voltage values, as we have seen elsewhere in this book. When the same representation is used over a communication link, the link is said to use *baseband*. An alternative scheme in which 0s and 1s are represented by modulating a sinusoidal *carrier signal* is also widely used. This is called *broadband* transmission. For example, the signal frequency may be changed between two values, f_1 to represent a 0 and f_2 to represent a 1. In this case, the link is said to use *frequency modulation* or *frequency shift keying* (FSK). Many other modulation schemes are in use. The phase of the carrier signal may be changed to provide *phase shift keying* (PSK) or its amplitude may be changed to provide *amplitude modulation* (AM). A scheme known as *quadrature amplitude modulation* (QAM) combines amplitude and phase modulation of the carrier signal. Since two parameters are being changed, there are four possible combinations. Hence, the transmitted signal can represent two bits of information.

The signal configuration transmitted in any clock period time is called a *symbol*. Thus, in the FSK scheme, there are two possible symbols, each consisting of a sinusoidal signal with a frequency f_1 or f_2. In QAM, four possible symbols are available, defined by their amplitude and phase. The term *baud* rate refers to the number of symbols transmitted per second. Equivalently, it is the number of times the state of a signal changes per second. This is the same as the rate of data transmission in bits per second only in the case of a binary modulation scheme, such as FSK. For QAM, the bit rate is twice the baud rate, because each symbol represents two bits of information. There exist modulation schemes that use 8, 16, or more symbols. In a system with 16 symbols, each symbol represents 4 bits. Hence, the bit rate is four times the baud rate.

A device called a *modem* (MOdulator-DEModulator) is installed at each end of a communications link to perform the desired signal transformations, as shown in Figure 10.2. The figure shows a computer connected to a network server. This could be a permanent connection or a dialed connection over a telephone line.

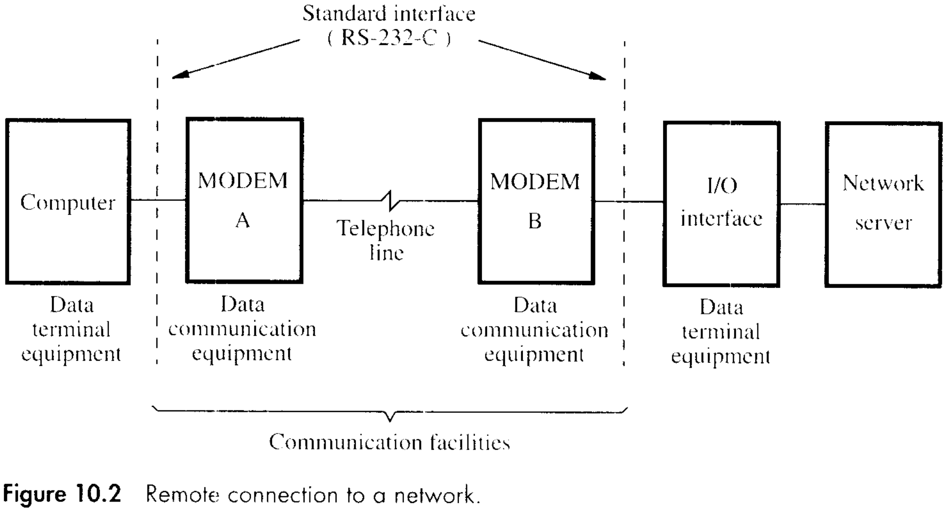

Figure 10.2 Remote connection to a network.

Synchronization

Serial communications means that data are sent one bit at a time. This requires that both the transmitting and the receiving devices use the same timing information for interpretation of individual bits. When the communicating devices are physically close to each other and multiple signal paths are available, a clock signal can be transmitted along with the data. However, this is not feasible over longer links, where only one signal path is available. More importantly, even if a second path is provided, the delays encountered by the data and the clock signals could be different. For these reasons, timing information and data are encoded on one transmission channel. A variety of encoding schemes have been developed that enable the receiver to decode the received signal and recover the timing and transmitted data correctly.

There are two basic ways of realizing serial transmission. For data transmission at speeds not exceeding a few tens of kilobits per second, a simple scheme can be used in which the sender and receiver use independent clock signals having the same nominal frequency. No attempt is made to guarantee that the two clocks have exactly the same phase or frequency. Hence, this scheme is called asynchronous transmission. The start-stop scheme to be described shortly is a common example of this approach.

Synchronous transmission is needed to transmit data at higher speed. In this case, the receiver recovers the clock timing used by the transmitter by continuously observing the positions of the transitions in the received signal and adjusting the phase of its local clock accordingly. As a result, the receiver's clock is synchronized with the transmitter's clock and can be used to recover the transmitted data correctly. There is a wide range of techniques used to encode timing information over synchronous links. They vary in their ability to make use of the bandwidth of the link and hence in the data rate they can achieve.

Full-Duplex and Half-Duplex Links

A communication link may be operated according to one of the following three schemes:

Simplex allows transmission in one direction only.

Half duplex (HDX) allows transmission in either direction, but not at the same time.

Full duplex (FDX) allows simultaneous transmission in both directions.

The simplex configuration is useful only if the remote location contains an input or an output device, but not both. Hence, it is seldom used. The choice between half and full duplex is a trade-off between economy and speed of operation.

Using the most straightforward electrical circuit arrangement, a pair of wires enables transmission in one direction only, that is, simplex operation. To obtain a half-duplex link, switches at both ends must be used to connect either the transmitter or the receiver, but not both, to the line. When transmission in one direction is completed, the switches are reversed to enable transmission in the reverse direction. Control of the position of the switches is a part of the function of the devices at each end of the line.

Full-duplex operation is possible on a four-wire link, with two wires dedicated to each direction of transmission. It is also possible on a two-wire link by using two nonoverlapping frequency bands. The two bands create two independent transmission channels, one for each direction of transmission. Alternatively, full-duplex operation can be achieved in a common frequency band using a device called a *hybrid* at each end of the line. The hybrid separates the signals traveling in opposite directions so that they do not interfere with each other. Dialed telephone connections use lines of this type.

In the case of synchronous half-duplex operation, a time delay occurs whenever the direction of transmission is reversed because the transmitting modem may have to transmit an initializing sequence of signals to allow the receiving end to adapt to the conditions of the channel. The amount of delay encountered depends on the modem and the transmission facilities, and may be anywhere from a few milliseconds to over a hundred milliseconds.

The discussion above relates directly to the characteristics of the transmission link and the modems. Other important factors that influence the choice between half- and full-duplex operation are the nature of the data traffic and the way the system reacts to errors during transmission. We discuss only the first of these factors here.

Many computer applications require the computer to receive input data, perform some processing, and then return output data. A half-duplex link satisfies the requirements for such an application. However, if the messages exchanged between the two ends are short and frequent, the delay encountered in reversing the direction of transmission becomes significant. For this reason, many applications use full-duplex transmission facilities, although actual data transmission never takes place in both directions at the same time.

In some situations, simultaneous transmission in both directions can be used to considerable advantage. For example, in the system in Figure 10.2, the user of the computer may wish to communicate directly with the network server, using the computer as a video terminal. Each character entered at the keyboard should be echoed back to be displayed on the computer screen. This may be done locally by the computer or remotely by the network server. The use of remote echo provides an automatic checking capability to ensure that no errors have been introduced during transmission. If a half-duplex link is used in such a case, transmission of the next character must be delayed until the first character has been echoed back. No such restriction is necessary with full-duplex operation. Links between nodes in a high-speed computer communication network is another example where full-duplex transmission is useful. Messages traveling in opposite directions on any given link often bear no relation to each other; hence, they can be transmitted simultaneously.

10.3.1 ASYNCHRONOUS TRANSMISSION

The simplest scheme for serial communications is asynchronous transmission using a technique called *start-stop*. To facilitate timing recovery, data are organized in small groups of 6 to 8 bits, with a well defined beginning and end. In a typical arrangement, alphanumeric characters encoded in 8 bits are transmitted as shown in Figure 10.3. The line connecting the transmitter and the receiver is in the 1 state when idle. Transmission of a character is preceded by a 0 bit, referred to as the Start bit, followed by eight data bits and one or two Stop bits. The Stop bits have a logic value of 1. The Start bit alerts the receiver that data transmission is about to begin. Its leading edge is used

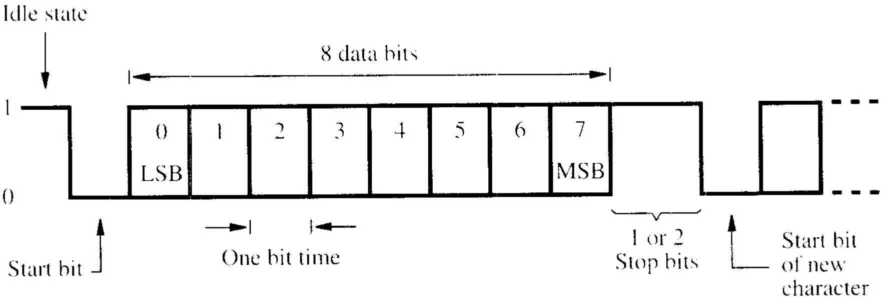

Figure 10.3 Asynchronous serial character transmission.

to synchronize the receiver clock with that of the transmitter. The Stop bits at the end delineate consecutive characters in the case of continuous transmission. When transmission stops, the line remains in the 1 state after the end of the Stop bits. It is the responsibility of the sender and receiver circuitry to insert and remove the Start and Stop bits.

To ensure proper synchronization at the receiving end, the receiver clock is derived from a local clock whose frequency is substantially higher than the transmission rate, typically 16 times higher. This means that 16 clock pulses occur during each data bit interval. This clock is used to increment a modulo-16 counter, which is reset to 0 when the leading edge of a Start bit is detected. When the count reaches 8, it indicates that the middle of the Start bit has been reached. The value of the Start bit is sampled to confirm that it is a valid Start bit, and the counter is again reset to 0. From this point onward, the incoming data signal is sampled whenever the count reaches 16, which should be close to the middle of each bit transmitted. Therefore, as long as the relative positioning of bits within a transmitted character is not in error by more than one half of a clock cycle, the receiver correctly interprets the bits of the encoded character.

A number of standard transmission rates are found in commercially available equipment, ranging from 300 to 56,000 bits per second. Start-stop transmission is used on short connections, such as the connection between the computer and the modem in Figure 10.2. For longer distances, such as for the connection between the two modems in the figure, start-stop can be used only at very low speeds. High-speed modems use the synchronous transmission schemes discussed in the next section.

When transmitting characters, they are represented by the 7-bit ASCII code (see Appendix E) occupying bits 0 through 6 in Figure 10.3. The MSB, bit 7 of the transmitted byte, is usually set to 0. Alternatively, it may be used as a parity bit, to aid in detecting transmission errors. Parity is the sum modulo 2 of a group of bits. Hence, it is equal to 1 if the transmitted data contains an odd number of 1s, and it is equal to 0 otherwise. When a parity bit is used, it is set by the transmitter such that the parity of the 8 bits transmitted is always the same, either odd or even. If a transmission error causes the value of one bit to change, the receiver will detect an incorrect parity and hence will be able to determine that an error has occurred.

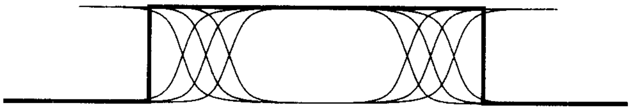

Figure 10.4 Overlapped transitions from successive bits showing the eye pattern.

The ASCII character set consists of letters, numbers, and special symbols such as $, +, and >. A number of nonprinting characters are also provided, for example, EOT (end of transmission) and CR (carriage return). These characters may be used to request specific actions, particularly when transmitting or receiving messages to or from a remote computer.

10.3.2 SYNCHRONOUS TRANSMISSION

In the start-stop scheme described above, the position of the 1-to-0 transition at the beginning of the start bit in Figure 10.3 is the key to obtaining correct timing. Hence, this scheme is useful only where the speed of transmission is sufficiently low and the conditions on the transmission link are such that the square waveforms shown in the figure maintain their shape. For higher speed and longer lines, much signal degradation takes place. Figure 10.4 shows a number of bits overlapped on top of each other to illustrate how the waveform may change from one bit position to another. Signal degradation is a result of such factors as signal distortion introduced by the line and the transmission equipment, by nearby sources of interference, jitter (random variations in the position of signal transitions), and so on. Because of the shape of the open area in the middle of a bit position, this figure is called the *eye pattern* of the transmission link. Sophisticated encoding and decoding schemes are used to help the receiver in determining the center point of the eye pattern, where 1s and 0s are farthest apart. This is the best point to sample the received signal.

In synchronous transmission, data are transmitted in blocks consisting of several hundreds or thousands of bits each. The beginning and end of each block are marked by appropriate codes, and data within a block are organized according to an agreed upon set of rules. Modems require a significant start-up time to complete such operations as transmitting and detecting carrier frequencies and establishing synchronization. In some modems, the start-up time is also used to adapt the modem circuits to the transmission properties of the link.

Network Connections — ADSL

In recent years, and particularly with the wide-spread use of the world-wide web, there has been an increasing demand for connecting computers at home and in the office via high-speed links to the Internet. Until recently, existing modems could not achieve the desired performance. A conventional modem converts digital signals to analog form using frequencies within the 4-kHz voice band of a telephone line. When a computer is using such a modem to communicate with another computer, the line is not available to make an ordinary telephone call. More importantly, the speed of transmission is limited to a few tens of kilobits per second. This is much less than the speeds needed to connect to a remote server or to the Internet.

Traditional telephone technology makes use of a small portion of the information carrying capacity of a telephone line. Depending on distance and the condition of the line, modern communication methods make it possible to transmit upwards of 50 megabits/s over the twisted-pair wire used in the telephone system. Many schemes have been developed to tap into this unused capacity by transmitting information in digital form directly between a subscriber's location at home or business and the central office of the telephone company. Telephone companies refer to the connection between a central office and a subscriber as the subscriber loop. Hence, when digital transmission is used, the scheme is called *digital subscriber loop* (DSL).

Computer communications is critically dependent on compatibility among equipment and services provided by several parties, including computer companies, modem manufacturers, and network service providers. Hence, agreement on a few standards that are accepted by all parties is essential. In the DSL domain, a few such standards have been developed. These include SDSL (Symmetric DSL), HDSL (High-speed DSL), and ADSL (Asymmetric DSL). Of these, ADSL is the most widely used for connecting home PCs to the Internet. We will discuss briefly the main features of this scheme.

The asymmetry in ADSL refers to the difference between transmission speeds in the upstream and downstream directions. Most of the time the information sent from a computer to a server on the Internet (the upstream direction) consists of input from the user. A low-speed connection is sufficient in this case. On the other hand, the information flowing to the user, such as an image to be displayed on the computer screen, requires transmission at high-speed to provide good response. For this reason, the speed of transmission in the downstream direction in ADSL is considerably higher than that used for upstream transmission.

The ADSL scheme uses different frequency bands and a technique called time-division multiplexing to create several channels of communication. One of these is allocated to regular telephone service. The others are allocated to the transmission of data in the upstream and downstream directions. A typical arrangement is shown in Figure 10.5. A single twisted pair wire carries information between the central office and the subscriber. At each end, a splitter separates data traffic from voice signals. At the subscriber end, data are directed to the computer over an appropriate data link, such as an Ethernet or a USB connection. Voice signals are sent to the telephone. At the central office, data are sent to a router device connected to the Internet, and voice signals are directed to the telephone switch. (A router is a switching device used to direct traffic in a data network.) This arrangement makes it possible to have the computer connected to the Internet all the time, without the need for dialing. At the same time, regular dialed telephone service continues to be available.

Cable Modems

Cable modems provide an alternative means for connecting a home computer to the Internet. They use the cable TV connection instead of the telephone connection.

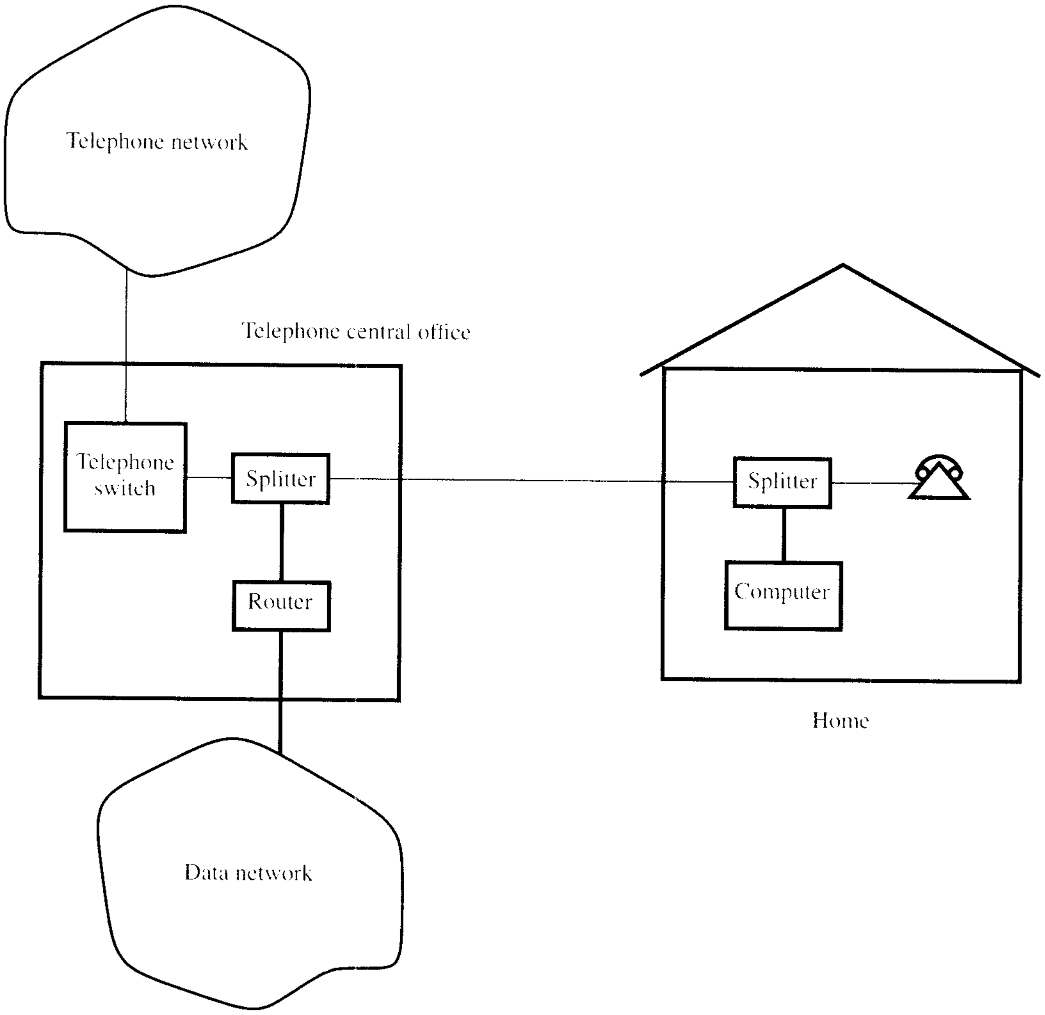

Figure 10.5 An ADSL connection.

The coaxial cable used in cable TV has a much higher bandwidth than a twisted pair wire. Hence the maximum speed possible with a cable modem is higher than what can be achieved with DSL schemes. However, cable TV uses a bus-like connection to all subscribers in a given neighborhood. Hence, the information carrying capacity of the cable is shared among all those connected to it. The full capacity is available to one user only when no other users on the same cable are active. A typical cable modem arrangement is shown in Figure 10.6.

The top speed available to any single user of a cable modem system is determined by the network service provider. It may vary from 600 kilobits/s to 10 megabits/s.

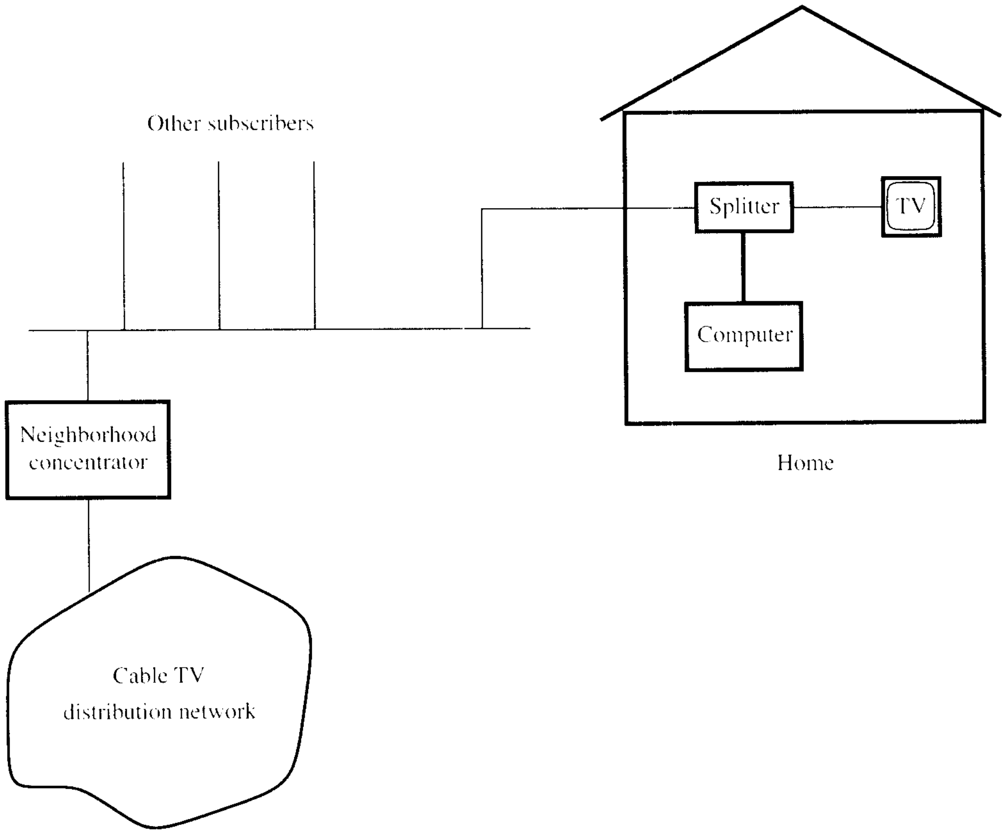

Figure 10.6 A cable-modem connection.

10.3.3 Standard Communications Interfaces

A standard interface refers to the collection of points at which two devices are connected together. One such standard that has gained wide acceptance is the EIA (Electronics Industry Association) Standard RS-232-C. Outside North America, this is known as CCITT (Comité Consultatif International Télégraphique et Téléphonique) Recommendation V24. This standard completely specifies the interface between data communication devices, such as modems, and data terminal equipment, such as computers. The RS-232-C interface consists of 25 connection points, which are described in Table 10.2.

Let us discuss a simple but common example. Consider the link in Figure 10.2, and assume that the connection is made over the dialed telephone network. The modems used are capable of going on- and off-hook, of sending dial tones, and of detecting an incoming ringing signal. They use the FSK transmission scheme described earlier and are capable of full-duplex operation. There are two transmission channels, one for each direction. One channel uses the frequencies 1275 and 1075 Hz, and the other uses frequencies 2225 and 2025 Hz to represent logic levels 1 and 0, respectively.

Table 10.2 Summary of the EIA Standard RS-232-C Signals (CCITT Recommendation V24).

Name			
EIA	CCITT	Pin* no.	Function
AA	101	1	Protective ground
AB	102	7	Signal ground-common return
BA	103	2	Transmitted data
BB	104	3	Received data
CA	105	4	Request to send
CB	106	5	Clear to send
CC	107	6	Data set ready
CD	108.2	20	Data terminal ready
CE	125	22	Ring indicator
CF	109	8	Received line signal detector
CG	110	21	Signal quality detector
CH	111	23§	Data signal rate selector (from DTE† to DCE‡)
CI	112	23§	Data signal rate selector (from DCE‡ to DTE†)
DA	113	24	Transmitter signal element timing (DTE†)
DB	114	15	Transmitter signal element timing (DCE‡)
DD	115	17	Receiver signal element timing (DCE‡)
SBA	118	14	Secondary transmitted data
SBB	119	16	Secondary received data
SCA	120	19	Secondary request to send
SCB	121	13	Secondary clear to send
SCF	122	12	Secondary received line signal detector

Pins 9 and 10 are used for testing purposes, and pins 11, 18, and 25 are spare.
†Data terminal equipment.
‡Data communication equipment.
§The name of the signal on this pin depends on the signal's direction.

Figure 10.7 gives the sequence of logic signals needed to establish a connection, transmit data, and terminate the connection. The steps involved in this process are described briefly as follows:

1. When the network server is ready to accept a call, it sets the data-terminal-ready signal (CD) to 1.

2. Modem B monitors the telephone line, and when it detects the ringing current that indicates an incoming call, it signals the server by setting the ringing indicator (CE) to 1. If CD = 1 at the time the ringing current is detected, the modem

Step no.	Computer	Interface signals	Modem A	Modem B	Interface signals	Server
1				Enable automatic answering	CD	← 1
2	Dialed digits →			1 → Goes off hook 1 →	CE CC	
3		CF	← 1	← 2225 Hz 1 →	CA CB	← 1
4	1 →	CA CB CC	1275 Hz → ← 1 ← 1	1 →	CF	
5	Output data ← Input data →	BB BA	← Data 1275/1075 Hz →	← 2225/2025 Hz Data →	BA BB	← Output data → Input data
6		CF	← 0	Drop 2225 Hz and disconnect 0 → 0 → 0 →	CA CD CF CC CB	← 0 ← 0
7	(0 →) Terminate connection	CA CB CC	Drop 1275 ← 0 ← 0			
8					CD	← 1

Figure 10.7 RS-232-C standard signalling sequence.

automatically answers the call by going off-hook. It then sets the modem-ready signal (CC) to 1.

3. The server instructs modem B to start transmitting the frequency representing a 1 (2225 Hz) by setting request-to-send (CA) to 1. When this is accomplished, modem B responds by setting clear-to-send (CB) to 1. The detection of this frequency at modem A causes it to set the received-line-signal detector (CF) to 1.

4. The computer sets CA to 1. Modem A transmits the 1275-Hz signal and sets CB and CC to 1. When modem B detects the 1275-Hz frequency, it sets CF to 1.

5. A full-duplex link is now established between the server and the computer, and it can be used for transmitting data in either direction. Interface pins BA (transmitted data) and BB (received data) are used for this purpose; all other signals in the interface remain unchanged.

6. When the user signs off, the server sets the request-to-send and data-terminal-ready signals, CA and CD, to 0, causing modem B to drop the 2225-Hz signal and disconnect from the line. Signals CB, CF, and CC are set to 0 by modem B. When modem A senses the disappearance of the signal on the line, it sets the received-line-signal detector (CF) to 0.

7. Modem A removes its 1275-Hz signal from the line, sets CB and CC to 0, and goes off-hook.

8. The server sets data-terminal-ready (CD) to 1 in preparation for a new call.

The initial connection procedure used with modems involves an exchange of messages in which the two sides agree on such parameters as the encoding scheme to be used, the speed of transmission, the size of data blocks, and so on. The RS-232-C interface can provide a serial connection between any two digital devices. The interpretation of individual signals such as CA and CD depends on the functional capabilities of the devices involved. When these signals are not needed, they are simply ignored by both devices. In most applications, no more than nine of the signals in Table 10.2 are used.

10.4 CONCLUDING REMARKS

This chapter presented an overview of input and output devices and their principles of operation. I/O devices are a fundamental part of a computer system because they constitute the link for feeding information into a computer and for receiving the results. Many new and innovative devices have been introduced in recent years. High-quality output devices are now available at affordable prices for personal computers. The range of input devices available also continues to expand, including digital cameras and hand-held devices of various kinds.

This chapter also presented some aspects of computer communications, in particular the basic technologies used over serial links. Such links are commonly used to connect computers to I/O devices and to each other. Examples of high-speed links to the Internet over common-carrier facilities (the telephone network) and the cable TV network have been briefly described. The connectivity that has resulted from the widespread use of such facilities has transformed the way we use computers and opened the door to a myriad of home and business applications. It is this synergy between the fields of computers and communications that has ushered in the modern era of information technology.

PROBLEMS

10.1 The display on a video screen must be refreshed at least 30 times per second to remain flicker-free. During each full scan of the screen, the total time required to illuminate each point is 1 μs. The beam is then turned off and moved to the next point to be illuminated. On average, moving the beam from one spot to the next takes 3 μs. Because of power-dissipation limitations, the beam cannot be turned on more than 10 percent of the time. Determine the maximum number of points that can be illuminated on the screen.

10.2 Consider a communication channel that uses eight-valued signals instead of the two-valued signals used in a binary channel. If the channel is rated at 9600 baud, what is its capacity in bits per second?

10.3 The following components are provided:

- A 6-bit binary counter, with Clock and Clear inputs and six outputs
- A 3-bit serial-input–parallel-output shift register
- A clock running at eight times the input data rate
- Logic gates and D flip-flops with Preset and Clear controls

Design a circuit using these components to load 3 bits of serial data from an input data line into the shift register. Assume the data to have the format of Figure 10.3, but with only 3 bits of data instead of 8. The circuit you design should have two outputs, A and B, both initially cleared to 0. Output A should be set to 1 if a Stop bit is detected following the data bits. Otherwise, output B should be set to 1. Give an explanation of the operation of your design.

10.4 An asynchronous link between two computers uses the start-stop scheme, with one start bit and one stop bit, and a transmission rate or 38.8 kilobits/s. What is the effective transmission rate as seen by the two computers?

10.5 A communication link uses odd parity for each character transmitted. Refer to Appendix E, and give the 8-bit pattern transmitted for the characters A, P, =, and 5.

10.6 Consider a communication line modem connected to a computer through an RS-232-C interface. The control signals associated with this interface are accessed by the computer through a 16-bit register, as shown in Figure P10.1. The status change bit, b_{15}, is set to 1 whenever there is a change in the state of bits b_{12} or b_{13}, or when b_{14} is set to 1. Bit b_{15} is cleared whenever this register is accessed by the processor. Write a program for one of the processors in Chapter 3 to implement the control sequence required to establish a telephone connection according to steps 1 through 4 of Figure 10.7.

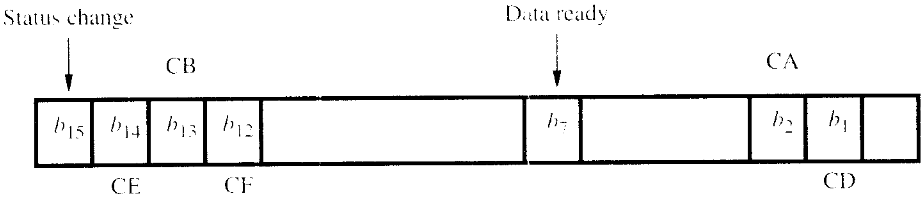

Figure P10.1 Organization of an I/O register for the modem interface in Problem 10.6.