

# CHAPTER 1: INTRODUCTION TO COMPUTER

The term computer is derived from the word compute. The word compute means to calculate. A computer is an electronic machine that accepts data from the user, processes the data by performing calculations and operations on it, and generates the desired output results. Computer performs both simple and complex operations, with speed and accuracy. This chapter discusses the history and evolution of computer, the concept of input-process-output and the characteristics of computer. This chapter also discusses the classification of digital computers based on their size and type, and the application of computer in different domain areas.

## 1.2 DIGITAL AND ANALOG COMPUTERS

A digital computer uses distinct values to represent the data internally. All information are represented using the digits 0s and 1s. The computers that we use at our homes and offices are digital computers. Analog computer is another kind of a computer that represents data as variable across a continuous range of values. The earliest computers were analog computers. Analog computers are used for measuring of parameters that vary continuously in real time, such as temperature, pressure and voltage. Analog computers may be more flexible but generally less precise than digital computers. Slide rule is an example of an analog computer. This book deals only with the digital computer and uses the term computer for them.

1.3 CHARACTERISTICS OF COMPUTER Speed, accuracy, diligence, storage capability and versatility are some of the key characteristics of a computer. A brief overview of these characteristics are—

- Speed

The computer can process data very fast, at the rate of millions of instructions per second. Some calculations that would have taken hours and days to complete otherwise, can be completed in a few seconds using the computer. For example, calculation and generation of salary slips of thousands of employees of an organization, weather forecasting that requires analysis of a large amount of data related to temperature, pressure and humidity of various places, etc.

- Accuracy

Computer provides a high degree of accuracy. For example, the computer can accurately give the result of division of any two numbers up to 10 decimal places.

- Diligence

When used for a longer period of time, the computer does not get tired or fatigued. It can perform long and complex calculations with the same speed and accuracy from the start till the end.

- Storage Capability

Large volumes of data and information can be stored in the computer and also retrieved whenever required. A limited amount of data can be stored, temporarily, in the primary memory. Secondary storage devices like floppy disk and compact disk can store a large amount of data permanently.

- Versatility

Computer is versatile in nature. It can perform different types of tasks with the same ease. At one moment you can use the computer to prepare a letter document and in the next moment you may play music or print a document. Computers have several limitations too. Computer can only perform tasks that it has been programmed to do. Computer cannot do any work without instructions from the user. It executes instructions as specified by the user and does not take its own decisions.

## 1.4 HISTORY OF COMPUTER

Until the development of the first generation computers based on vacuum tubes, there had been several developments in the computing technology related to the mechanical computing devices. The key developments that took place till the first computer was developed are as follows—

- Calculating Machines

ABACUS was the first mechanical calculating device for counting of large numbers. The word ABACUS means calculating board. It consists of bars in horizontal positions on which sets of beads are inserted. The horizontal bars have 10 beads each, representing units, tens, hundreds, etc. An abacus is shown in Figure 1.1

- Napier's Bones was a mechanical device built for the purpose of multiplication in 1617 ad. by an English mathematician John Napier.

- Slide Rule was developed by an English mathematician Edmund Gunter in the 16th century. Using the slide rule, one could perform operations like addition, subtraction, multiplication and division. It was used extensively till late 1970s. Figure 1.2 shows a slide rule.

- Pascal's Adding and Subtraction Machine was developed by Blaise Pascal. It could add and subtract. The machine consisted of wheels, gears and cylinders.

- Leibniz's Multiplication and Dividing Machine was a mechanical device that could both multiply and divide. The German philosopher and mathematician Gottfried Leibniz built it around 1673.

- Punch Card System was developed by Jacquard to control the power loom in 1801. He invented the punched card reader that could recognize the presence of hole in the punched card as binary one and the absence of the hole as binary zero. The 0s and 1s are the basis of the modern digital computer. A punched card is shown in Figure 1.3.

- Babbage's Analytical Engine An English man Charles Babbage built a mechanical machine to do complex mathematical calculations, in the year 1823. The machine was called as difference engine. Later, Charles Babbage and Lady Ada Lovelace developed a general-purpose calculating machine, the analytical engine. Charles Babbage is also called the father of computer.

- Hollerith's Punched Card Tabulating Machine was invented by Herman Hollerith. The machine could read the information from a punched card and process it electronically. The developments discussed above and several others not discussed here, resulted in the development of the first computer in the 1940s.

### 1.5 GENERATIONS OF COMPUTER

The computer has evolved from a large—sized simple calculating machine to a smaller but much more powerful machine. The evolution of computer to the current state

is defined in terms of the generations of computer. Each generation of computer is designed based on a new technological development, resulting in better, cheaper and smaller computers that are more powerful, faster and efficient than their predecessors. Currently, there are five generations of computer. In the following subsections, we will discuss the generations of computer in terms of— 1. the technology used by them (hardware and software), 2. computing characteristics (speed, i.e., number of instructions executed per second), 3. physical appearance, and 4. their applications.

#### **1.5.1 First Generation (1940 to 1956): Using Vacuum Tubes**

- **Hardware Technology** The first generation of computers used vacuum tubes (Figure 1.4) for circuitry and magnetic drums for memory. The input to the computer was through punched cards and paper tapes. The output was displayed as printouts.
- **Software Technology** The instructions were written in machine language. Machine language uses 0s and 1s for coding of the instructions. The first generation computers could solve one problem at a time.
- **Computing Characteristics** The computation time was in milliseconds.
- **Physical Appearance** These computers were enormous in size and required a large room for installation.
- **Application** They were used for scientific applications as they were the fastest computing device of their time.
- **Examples** UNIVersal Automatic Computer (UNIVAC), Electronic Numerical Integrator And Calculator (ENIAC), and Electronic Discrete Variable Automatic Computer (EDVAC). The first generation computers used a large number of vacuum tubes and thus generated a lot of heat. They consumed a great deal of electricity and were expensive to operate. The machines were prone to frequent malfunctioning and required constant maintenance. Since first generation computers used machine language, they were difficult to program.

#### **1.5.2 Second Generation (1956 to 1963): Using Transistors**

- **Hardware Technology** Transistors (Figure 1.5) replaced the vacuum tubes of the first generation of computers. Transistors allowed computers to become smaller, faster, cheaper, energy efficient and reliable. The second generation computers used magnetic core technology for primary memory. They used magnetic tapes and magnetic disks for secondary storage. The input was still through punched cards and the output using printouts. They used the concept of a stored program, where instructions were stored in the memory of computer. Figure 1.5 Transistors
- **Software Technology** The instructions were written using the assembly language. Assembly language uses mnemonics like ADD for addition and SUB for subtraction for coding of the instructions. It is easier to write instructions in assembly language, as compared to writing instructions in machine language. High-level programming languages, such as early versions of COBOL and FORTRAN were also developed during this period.
- **Computing Characteristics** the computation time was in microseconds.

- **Physical Appearance** Transistors are smaller in size compared to vacuum tubes, thus, the size of the computer was also reduced.
- **Application** The cost of commercial production of these computers was very high, though less than the first generation computers. The transistors had to be assembled manually in second generation computers.
- **Examples** PDP-8, IBM 1401 and CDC 1604. Second generation computers generated a lot of heat but much less than the first generation computers. They required less maintenance than the first generation computers.

### **1.5.3 Third Generation (1964 to 1971): Using Integrated Circuits**

- **Hardware Technology:** The third generation computers used the Integrated Circuit (IC) chips. Figure 1.6 shows IC chips. In an IC chip, multiple transistors are placed on a silicon chip. Silicon is a type of semiconductor. The use of IC chip increased the speed and the efficiency of computer, manifold. The keyboard and monitor were used to interact with the third generation computer, instead of the punched card and printouts. Figure 1.6 IC chips
- **Software Technology:** The keyboard and the monitor were interfaced through the operating system. Operating system allowed different applications to run at the same time. High-level languages were used extensively for programming, instead of machine language and assembly language.
- **Computing Characteristics:** The computation time was in nanoseconds.
- **Physical Appearance** The size of these computers was quite small compared to the second generation computers.
- **Application** Computers became accessible to mass audience. Computers were produced commercially, and were smaller and cheaper than their predecessors.
- **Examples** IBM 370, PDP 11. The third generation computers used less power and generated less heat than the second generation computers. The cost of the computer reduced significantly, as individual components of the computer were not required to be assembled manually. The maintenance cost of the computers was also less compared to their predecessors.

### **1.5.4 Fourth Generation (1971 to present): Using Microprocessors**

- **Hardware Technology** They use the Large Scale Integration (LSI) and the Very Large Scale Integration (VLSI) technology. Thousands of transistors are integrated on a small silicon chip using LSI technology. VLSI allows hundreds of thousands of components to be integrated in a small chip. This era is marked by the development of microprocessor. Microprocessor is a chip containing millions of transistors and components, and, designed using LSI and VLSI technology. A microprocessor chip is shown in Figure 1.7. This generation of computers gave rise to Personal Computer (PC). Semiconductor memory replaced the earlier magnetic core memory, resulting in fast random access to memory. Secondary storage device like magnetic disks became smaller in physical size and larger in capacity. The linking of computers is another key development of this era. The computers were linked to form networks that led to the emergence of the Internet. This generation also saw the development of pointing devices like mouse, and handheld devices. Figure 1.7 Microprocessors

- **Software Technology** Several new operating systems like the MS-DOS and MS- Windows developed during this time. This generation of computers supported Graphical User Interface (GUI). GUI is a user-friendly interface that allows user to interact with the computer via menus and icons. High-level programming languages are used for the writing of programs.
  - **Computing Characteristics** The computation time is in picoseconds.
  - **Physical Appearance** They are smaller than the computers of the previous generation. Some can even fit into the palm of the hand.
  - **Application** They became widely available for commercial purposes. Personal computers became available to the home user.
  - **Examples** The Intel 4004 chip was the first microprocessor. The components of the computer like Central Processing Unit (CPU) and memory were located on a single chip. In 1981, IBM introduced the first computer for home use. In 1984, Apple introduced the Macintosh. The microprocessor has resulted in the fourth generation computers being smaller and cheaper than their predecessors. The fourth generation computers are also portable and more reliable. They generate much lesser heat and require less maintenance compared to their predecessors. GUI and pointing devices facilitate easy use and learning on the computer. Networking has resulted in resource sharing and communication among different computers.
- 1.5.5 Fifth Generation (Present and Next): Using Artificial Intelligence** The goal of fifth generation computing is to develop computers that are capable of learning and self-organization. The fifth generation computers use Super Large Scale Integrated (SLSI) chips that are able to store millions of components on a single chip. These computers have large memory requirements. This generation of computers uses parallel processing that allows several instructions to be executed in parallel, instead of serial execution. Parallel processing results in faster processing speed. The Intel dualcore microprocessor uses parallel processing. The fifth generation computers are based on Artificial Intelligence (AI). They try to simulate the human way of thinking and reasoning. Artificial Intelligence includes areas like Expert System (ES), Natural Language Processing (NLP), speech recognition, voice recognition, robotics, etc.

## **1.6 CLASSIFICATION OF COMPUTER**

The digital computers that are available nowadays vary in their sizes and types. The computers are broadly classified into four categories (Figure 1.8) based on their size and type—(1) Microcomputers, (2) Minicomputers, (3) Mainframe computers, and (4) Supercomputer. Figure 1.8 Classification of computers based on size and type

### **1.6.1 Microcomputers:**

Microcomputers are small, low-cost and single-user digital computer. They consist of CPU, input unit, output unit, storage unit and the software. Although microcomputers are stand-alone machines, they can be connected together to create a network of computers that can serve more than one user. IBM PC based on Pentium microprocessor and Apple Macintosh are some examples of microcomputers. Microcomputers include desktop computers, notebook computers or laptop, tablet computer, handheld computer, smart phones and netbook, as shown in Figure 1.9. Figure 1.9 Microcomputers

- **Desktop Computer or Personal Computer (PC)** is the most common type of microcomputer. It is a stand-alone machine that can be placed on the desk. Externally, it consists of three units—keyboard, monitor,

and a system unit containing the CPU, memory, hard disk drive, etc. It is not very expensive and is suited to the needs of a single user at home, small business units, and organizations. Apple, Microsoft, HP, Dell and Lenovo are some of the PC manufacturers.

- Notebook Computers or Laptop resemble a notebook. They are portable and have all the features of a desktop computer. The advantage of the laptop is that it is small in size (can be put inside a briefcase), can be carried anywhere, has a battery backup and has all the functionality of the desktop. Laptops can be placed on the lap while working (hence the name). Laptops are costlier than the desktop machines.
- Netbook These are smaller notebooks optimized for low weight and low cost, and are designed for accessing web-based applications. Starting with the earliest netbook in late 2007, they have gained significant popularity now. Netbooks deliver the performance needed to enjoy popular activities like streaming videos or music, emailing, Web surfing or instant messaging. The word netbook was created as a blend of Internet and notebook.
- Tablet Computer has features of the notebook computer but it can accept input from a stylus or a pen instead of the keyboard or mouse. It is a portable computer. Tablet computer are the new kind of PCs.
- Handheld Computer or Personal Digital Assistant (PDA) is a small computer that can be held on the top of the palm. It is small in size. Instead of the keyboard, PDA uses a pen or a stylus for input. PDAs do not have a disk drive. They have a limited memory and are less powerful. PDAs can be connected to the Internet via a wireless connection. Casio and Apple are some of the manufacturers of PDA. Over the last few years, PDAs have merged into mobile phones to create smart phones.
- Smart Phones are cellular phones that function both as a phone and as a small PC. They may use a stylus or a pen, or may have a small keyboard. They can be connected to the Internet wirelessly. They are used to access the electronic-mail, download music, play games, etc. Blackberry, Apple, HTC, Nokia and LG are some of the manufacturers of smart phones.

### **1.6.2 Minicomputers:**

Minicomputers are digital computers, generally used in multi-user systems. They have high processing speed and high storage capacity than the microcomputers. Minicomputers can support 4–200 users simultaneously. The users can access the minicomputer through their PCs or terminal. They are used for real-time applications in industries, research centers, etc. PDP 11, IBM (8000 series) are some of the widely used minicomputers.

### **1.6.3 Mainframe Computers:**

Mainframe computers are multi-user, multi-programming and high performance computers. They operate at a very high speed, have very large storage capacity and can handle the workload of many users. Mainframe computers are large and powerful systems generally used in centralized databases. The user accesses the mainframe computer via a terminal that may be a dumb terminal, an intelligent terminal or a PC. A dumb terminal cannot store data or do processing of its own. It has the input and output device only. An intelligent terminal has the input and output device, can do processing, but, cannot store data of its own. The dumb and the intelligent terminal use the processing power and the storage facility of the mainframe computer. Mainframe computers are used in organizations like banks or companies, where

many people require frequent access to the same data. Some examples of mainframes are CDC 6600 and IBM ES000 series. Figure 1.11 Mainframe computer



•(1) Hardware, (2) Software, (3) Data, and (4) Users. The parts of computer system are shown in Figure 1.13. Hardware consists of the mechanical parts that make up the computer as a machine. The hardware consists of physical devices of the computer. The devices are required for input, output, storage and processing of the data. Keyboard, monitor, hard disk drive, floppy disk drive, printer, processor and motherboard are some of the hardware devices. Figure 1.13 Parts of computer system Software is a set of instructions that tells the computer about the tasks to be performed and how these tasks are to be performed. Program is a set of instructions, written in a language understood by the computer, to perform

a specific task. A set of programs and documents are collectively called software. The hardware of the computer system cannot perform any task on its own. The hardware needs to be instructed about the task to be performed. Software instructs the computer about the task to be performed. The hardware carries out these tasks. Different software can be loaded on the same hardware to perform different kinds of tasks. Data are isolated values or raw facts, which by themselves have no much significance. For example, the data like 29, January, and 1994 just represent values. The data is provided as input to the computer, which is processed to generate some meaningful information. For example, 29, January and 1994 are processed by the computer to give the date of birth of a person. Users are people who write computer programs or interact with the computer. They are also known as skinware, liveware, human ware or people ware. Programmers, data entry operators, system analyst and computer hardware engineers fall into this category.

#### **1.7.1 The Input-Process-Output Concept:**

A computer is an electronic device that (1) accepts data, (2) processes data, (3) generates output, and (4) stores data. The concept of generating output information from the input 4 data is also referred to as input-process-output concept. The input-process-output concept of the computer is explained as follows—

##### **• Input**

The computer accepts input data from the user via an input device like keyboard. The input data can be characters, word, text, sound, images, document, etc.

##### **• Process**

The computer processes the input data. For this, it performs some actions on the data by using the instructions or program given by the user of the data. The action could be an arithmetic or logic calculation, editing, modifying a document, etc. During processing, the data, instructions and the output are stored temporarily in the computer's main memory.

##### **• Output**

The output is the result generated after the processing of data. The output may be in the form of text, sound, image, document, etc. The computer may display the output on a monitor, send output to the printer for printing, play the output, etc.

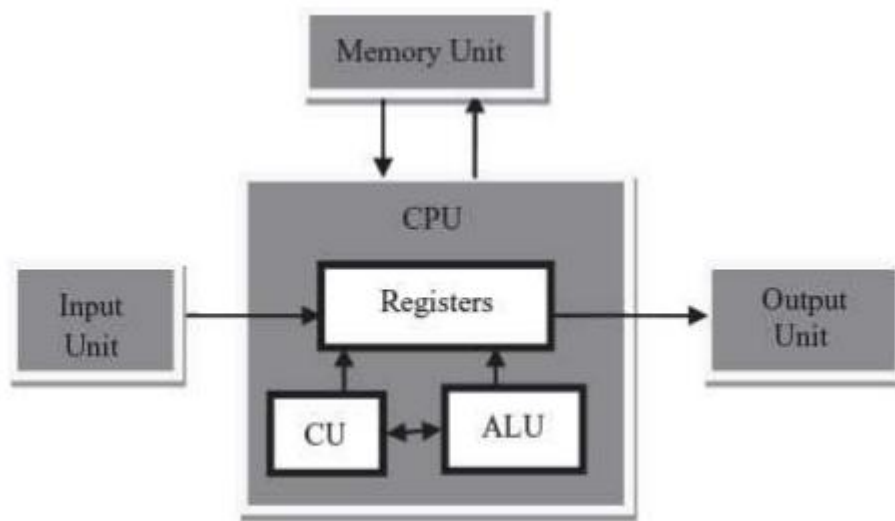
##### **• Storage**

The input data, instructions and output are stored permanently in the secondary storage devices like disk or tape. The stored data can be retrieved later, whenever needed.

#### **1.7.2 Components of Computer Hardware**

The computer system hardware comprises of three main components — 1. Input/output (I/O) Unit, 2. Central Processing Unit (CPU), and 3. Memory Unit. The I/O unit consists of the input unit and the output unit. CPU performs calculations and processing on the input data, to generate the output. The memory unit is used to store the data, the instructions and the output information. Figure 1.14 illustrates the typical interaction among the different components of the computer.

Figure 1.14 the computer system interaction



**Figure 1.14** The computer system interaction

- **Input/Output Unit**

The user interacts with the computer via the I/O unit. The Input unit accepts data from the user and the Output unit provides the processed data i.e. the information to the user. The Input unit converts the data that it accepts from the user, into a form that is understandable by the computer. Similarly, the Output unit provides the output in a form that is understandable by the user. The input is provided to the computer using input devices like keyboard, trackball and mouse. Some of the commonly used output devices are monitor and printer.

- **Central Processing Unit**

CPU controls, coordinates and supervises the operations of the computer. It is responsible for processing of the input data. CPU consists of Arithmetic Logic Unit (ALU) and Control Unit (CU).

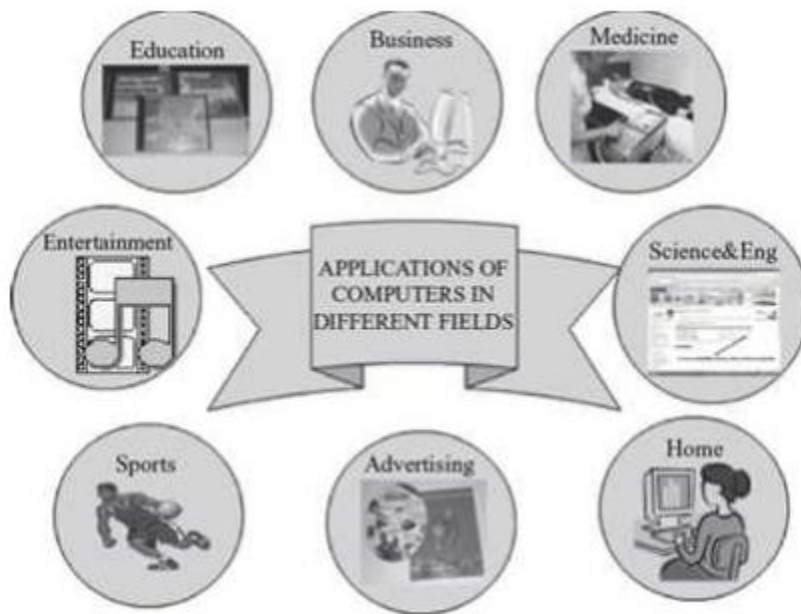
ALU performs all the arithmetic and logic operations on the input data.

CU controls the overall operations of the computer i.e. it checks the sequence of execution of instructions, and, controls and coordinates the overall functioning of the units of computer. Additionally, CPU also has a set of registers for temporary storage of data, instructions, addresses and intermediate results of calculation.

- **Memory Unit**

Memory unit stores the data, instructions, intermediate results and output, temporarily, during the processing of data. This memory is also called the main memory or primary memory of the computer. The input data that is to be processed is brought into the main memory before processing. The instructions required for processing of data and any intermediate results are also stored in the main memory. The output is stored in memory before being transferred to the output device. CPU can work with the information stored in the main memory. Another kind of storage unit is also referred to as the secondary

memory of the computer. The data, the programs and the output are stored permanently in the storage unit of the computer. Magnetic disks, optical disks and magnetic tapes are examples of secondary memory. 1.8 APPLICATION OF COMPUTERS Computers have proliferated into various areas of our lives. For a user, computer is a tool that provides the desired information, whenever needed. You may use computer to get information about the reservation of tickets (railways, airplanes and cinema halls), books in a library, medical history of a person, a place in a map, or the dictionary meaning of a word. The information may be presented to you in the form of text, images, video clips, etc.



**Figure 1.15** Applications of computer

Figure 1.15 shows some of the applications of computer. Some of the application areas of the computer are listed below—

- **Education** Computers are extensively used, as a tool and as an aid, for imparting education. Educators use computers to prepare notes and presentations of their lectures. Computers are used to develop computer-based training packages, to provide distance education using the e-learning software, and to conduct online examinations. Researchers use computers to get easy access to conference and journal details and to get global access to the research material.
- **Entertainment** Computers have had a major impact on the entertainment industry. The user can download and view movies, play games, chat, book tickets for cinema halls, use multimedia for making movies, incorporate visual and sound effects using computers, etc. The users can also listen to music, download and share music, create music using computers, etc.
- **Sports** A computer can be used to watch a game, view the scores, improve the game, play games (like chess, etc.) and create games. They are also used for the purposes of training players.
- **Advertising** Computer is a powerful advertising media. Advertisement can be displayed on different websites, electronic-mails can be sent and reviews of a product by different customers can be posted.

Computers are also used to create an advertisement using the visual and the sound effects. For the advertisers, computer is a medium via which the advertisements can be viewed globally. Web advertising has become a significant factor in the marketing plans of almost all companies. In fact, the business model of Google is mainly dependent on web advertising for generating revenues.

- **Medicine** Medical researchers and practitioners use computers to access information about the advances in medical research or to take opinion of doctors globally. The medical history of patients is stored in the computers. Computers are also an integral part of various kinds of sophisticated medical equipments like ultrasound machine, CAT scan machine, MRI scan machine, etc. Computers also provide assistance to the medical surgeons during critical surgery operations like laparoscopic operations, etc.
- **Science and Engineering** Scientists and engineers use computers for performing complex scientific calculations, for designing and making drawings (CAD/CAM applications) and also for simulating and testing the designs. Computers are used for storing the complex data, performing complex calculations and for visualizing 3– dimensional objects. Complex scientific applications like the launch of the rockets, space exploration, etc., are not possible without the computers.
- **Government** The government uses computers to manage its own operations and also for e-governance. The websites of the different government departments provide information to the users. Computers are used for the filing of income tax return, paying taxes, online submission of water and electricity bills, for the access of land record details, etc. The police department uses computers to search for criminals using fingerprint matching, etc.
- **Home** Computers have now become an integral part of home equipment. At home, people use computers to play games, to maintain the home accounts, for communicating with friends and relatives via Internet, for paying bills, for education and learning, etc. Microprocessors are embedded in house hold utilities like, washing machines, TVs, food processors, home theatres, security devices, etc. The list of applications of computers is so long that it is not possible to discuss all of them here. In addition to the applications of the computers discussed above, computers have also proliferated into areas like banks, investments, stock trading, accounting, ticket reservation, military operations, meteorological predictions, social networking, business organizations, police department, video conferencing, telepresence, book publishing, web newspapers, and information sharing.

## 2 THE COMPUTER SYSTEM HARDWARE

### 2.1 INTRODUCTION

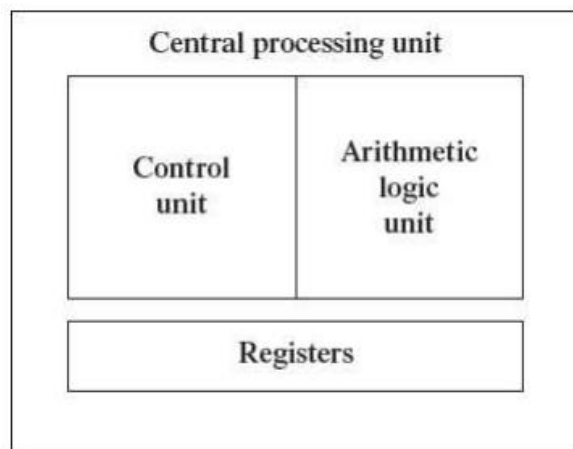
A computer consists of three main components—(1) Input/output (I/O) Unit, (2) Central Processing Unit (CPU), and (3) Memory Unit. The computer user interacts with the computer via the I/O unit. The purpose of I/O unit is to provide data and instructions as input to the computer and to present relevant information as output from the computer. CPU controls the operations of the computer and processes the received input to generate the relevant output. The memory unit stores the instructions and the data during the input activity, to make instructions readily available to CPU during processing. It also stores the processed output. This chapter discusses the hardware components of the computer and the interaction between them.

### 2.2 CENTRAL PROCESSING UNIT

Central Processing Unit (CPU) or the processor is also often called the brain of computer. CPU (Figure 2.1) consists of Arithmetic Logic Unit (ALU) and Control Unit (CU). In addition, CPU also has a set of registers which are temporary storage areas for holding data, and instructions. ALU performs the arithmetic and logic operations on the data that is made available to it. CU is responsible for organizing the processing of data and instructions. CU controls and coordinates the activity of the other units of computer. CPU uses the registers to store the data, instructions during processing.

CPU executes the stored program instructions, i.e. instructions and data are stored in memory before execution. For processing, CPU gets data and instructions from the memory. It interprets the program instructions and performs the arithmetic and logic operations required for the processing of data. Then, it sends the processed data or result to the memory. CPU also acts as an administrator and is responsible for supervising operations of other parts of the computer.

The CPU is fabricated as a single Integrated Circuit (IC) chip, and is also known as the microprocessor. The microprocessor is plugged into the motherboard of the computer (Motherboard is a circuit board that has electronic circuit etched on it and connects the microprocessor with the other hardware components).



**Figure 2.1** CPU

### 2.2.1 Arithmetic Logic Unit

- ALU consists of two units—arithmetic unit and logic unit.
- The arithmetic unit performs arithmetic operations on the data that is made available to it. Some of the arithmetic operations supported by the arithmetic unit are—addition, subtraction, multiplication and division.
- The logic unit of ALU is responsible for performing logic operations. Logic unit performs comparisons of numbers, letters and special characters. Logic operations include testing for greater than, less than or equal to condition.
- ALU performs arithmetic and logic operations, and uses registers to hold the data that is being processed.

### 2.2.2 Registers

- Registers are high-speed storage areas within the CPU, but have the least storage capacity. Registers are not referenced by their address, but are directly accessed and manipulated by the CPU during instruction execution.
- Registers store data, instructions, addresses and intermediate results of processing. Registers are often referred to as the CPU's working memory.
- The data and instructions that require processing must be brought in the registers of CPU before they can be processed. For example, if two numbers are to be added, both numbers are brought in the registers, added and the result is also placed in a register.
- Registers are used for different purposes, with each register serving a specific purpose. Some of the important registers in CPU (Figure 2.2) are as follows—
  - o Accumulator (ACC) stores the result of arithmetic and logic operations.
  - o Instruction Register (IR) contains the current instruction most recently fetched.
  - o Program Counter (PC) contains the address of next instruction to be processed.
  - o Memory Address Register (MAR) contains the address of next location in the memory to be accessed.
  - o Memory Buffer Register (MBR) temporarily stores data from memory or the data to be sent to memory.
  - o Data Register (DR) stores the operands and any other data.



**Figure 2.2** CPU registers

- The number of registers and the size of each (number of bits) register in a CPU helps to determine the power and the speed of a CPU.
- The overall number of registers can vary from about ten to many hundreds, depending on the type and complexity of the processor.
- The size of register, also called word size, indicates the amount of data with which the computer can work at any given time. The bigger the size, the more quickly it can process data. The size of a register may be 8, 16, 32 or 64 bits. For example, a 32-bit CPU is one in which each register is 32 bits wide and its CPU can manipulate 32 bits of data at a time. Nowadays, PCs have 32-bit or 64-bit registers.
- 32-bit processor and 64-bit processor are the terms used to refer to the size of the registers. Other factors remaining the same, a 64-bit processor can process the data twice as fast as one with 32-bit processor.

### **2.2.3 Control Unit**

- The control unit of a computer does not do any actual processing of data. It organizes the processing of data and instructions. It acts as a supervisor and, controls and coordinates the activity of the other units of computer.
  - CU coordinates the input and output devices of a computer. It directs the computer to carry out stored program instructions by communicating with the ALU and the registers. CU uses the instructions in the Instruction Register (IR) to decide which circuit needs to be activated. It also instructs the ALU to perform the arithmetic or logic operations. When a program is run, the Program Counter (PC) register keeps track of the program instruction to be executed next.
  - CU tells when to fetch the data and instructions, what to do, where to store the results, the sequencing of events during processing etc.
  - CU also holds the CPU's Instruction Set, which is a list of all operations that the CPU can perform. The function of a (CU) can be considered synonymous with that of a conductor of an orchestra. The conductor in an orchestra does not perform any work by itself but manages the orchestra and ensures that the members of orchestra work in proper coordination.
- ### **2.3 MEMORY UNIT**
- The memory unit consists of cache memory and primary memory. Primary memory or main memory of the computer is used to store the

data and instructions during execution of the instructions. Random Access Memory (RAM) and Read Only Memory (ROM) are the primary memory. In addition to the main memory, there is another kind of storage device known as the secondary memory. Secondary memory is non-volatile and is used for permanent storage of data and programs. A program or data that has to be executed is brought into the RAM from the secondary memory.

### **2.3.1 Cache Memory**

- The data and instructions that are required during the processing of data are brought from the secondary storage devices and stored in the RAM. For processing, it is required that the data and instructions are accessed from the RAM and stored in the registers. The time taken to move the data between RAM and CPU registers is large. This affects the speed of processing of computer, and results in decreasing the performance of CPU.
  - Cache memory is a very high speed memory placed in between RAM and CPU. Cache memory increases the speed of processing.
  - Cache memory is a storage buffer that stores the data that is used more often, temporarily, and makes them available to CPU at a fast rate. During processing, CPU first checks cache for the required data. If data is not found in cache, then it looks in the RAM for data.
  - To access the cache memory, CPU does not have to use the motherboard's system bus for data transfer. (The data transfer speed slows to the motherboard's capability, when data is passed through system bus. CPU can process data at a much faster rate by avoiding the system bus.) Figure 2.3 Illustration of cache memory
  - Cache memory is built into the processor, and may also be located next to it on a separate chip between the CPU and RAM. Cache built into the CPU is faster than separate cache, running at the speed of the microprocessor itself. However, separate cache is roughly twice as fast as RAM.
  - The CPU has a built-in Level 1 (L1) cache and Level2 (L2) cache, as shown in Figure 2.3. In addition to the built-in L1 and L2 cache, some CPUs have a separate cache chip on the motherboard. This cache on the motherboard is called Level 3 (L3) cache. Nowadays, high-end processor comes with built-in L3 cache, like in Intel core i7. The L1, L2 and L3 cache store the most recently run instructions, the next ones and the possible ones, respectively. Typically, CPUs have cache size varying from 256KB (L1), 6 MB (L2), to 12MB (L3) cache.
  - Cache memory is very expensive, so it is smaller in size. Generally, computers have cache memory of sizes 256 KB to 2 MB.
- ### **2.3.2 Primary Memory**
- Primary memory is the main memory of computer. It is used to store data and instructions during the processing of data. Primary memory is semiconductor memory.
  - Primary memory is of two kinds—Random Access Memory (RAM) and Read Only Memory (ROM).
  - RAM is volatile. It stores data when the computer is on. The information stored in RAM gets erased when the computer is turned off. RAM provides temporary storage for data and instructions.

- ROM is non-volatile memory, but is a read only memory. The storage in ROM is permanent in nature, and is used for storing standard processing programs that permanently reside in the computer. ROM comes programmed by the manufacturer.
- RAM stores data and instructions during the execution of instructions. The data and instructions that require processing are brought into the RAM from the storage devices like hard disk. CPU accesses the data and the instructions from RAM, as it can access it at a fast speed than the storage devices connected to the input and output unit.
- The input data that is entered using the input unit is stored in RAM, to be made available during the processing of data. Similarly, the output data generated after processing is stored in RAM before being sent to the output device. Any intermediate results generated during the processing of program are stored in RAM.
- RAM provides a limited storage capacity, due to its high cost. Figure 2.4 Interaction of CPU with memory

### **2.3.3 Secondary Memory**

- The secondary memory stores data and instructions permanently. The information can be stored in secondary memory for a long time (years), and is generally permanent in nature unless erased by the user. It is a non-volatile memory.
- It provides back-up storage for data and instructions. Hard disk drive, floppy drive and optical disk drives are some examples of storage devices.
- The data and instructions that are currently not being used by CPU, but may be required later for processing, are stored in secondary memory.
- Secondary memory has a high storage capacity than the primary memory.
- Secondary memory is also cheaper than the primary memory.
- It takes longer time to access the data and instructions stored in secondary memory than in primary memory. Magnetic tape drives, disk drives and optical disk drives are the different types of storage devices.

## **2.4 INSTRUCTION FORMAT**

A computer program is a set of instructions that describe the steps to be performed for carrying out a computational task. The program and the data, on which the program operates, are stored in main memory, waiting to be processed by the processor. This is also called the stored program concept. An instruction is designed to perform a task and is an elementary operation that the processor can accomplish. An instruction is divided into groups called fields. The common fields of an instruction are— Operation (op) code and Operand code (Figure 2.5). The remainder of the instruction fields differs from one computer type to other. The operation code represents action that the processor must execute. It tells the processor what basic operations to perform. The operand code defines the parameters of the action and depends on the operation. It specifies the locations of the data or the operand on which the operation is to be performed. It can be data or a memory address.



**Figure 2.5** Instruction format



**Figure 2.6** Instruction format for ADD command

Instruction format for ADD command The number of bits in an instruction varies according to the type of data (could be between 8 and 32 bits).

Figure 2.6 shows the instruction format for ADD command.

## 2.5 INSTRUCTION SET

A processor has a set of instructions that it understands, called as instruction set. An instruction set or an instruction set architecture is a part of the computer architecture. It relates to programming, instructions, registers, addressing modes, memory architecture, etc. An Instruction Set is the set of all the basic operations that a processor can accomplish. Examples of some instructions are shown in Figure 2.7.

LOAD R1, A

ADD R1, B

STORE R1, X

**Figure 2.7** Examples of some instructions

The instructions in the instruction set are the language that a processor understands. All programs have to communicate with the processor using these instructions. An instruction in the instruction set involves a series of logical operations (may be thousands) that are performed to complete each task. The instruction set is embedded in the processor (hardwired), which determines the machine language for the processor. All programs written in a high-level language are compiled and translated into machine code before execution, which is understood by the processor for which the program has been coded. Figure 2.7 Examples of some instructions Two processors are different if they have different instruction sets. A program run on one computer may not run on another computer having a different processor. Two processors are compatible if the same machine level program can run on both the processors. Therefore, the system software is developed within the processor's instruction set. Microarchitecture is the processor design technique used for implementing the Instruction Set. Computers having different microarchitecture can have a common Instruction Set. Pentium and Athlon CPU chips implement the x86 instruction set, but have different internal designs.

## 2.6 INSTRUCTION CYCLE

The primary responsibility of a computer processor is to execute a sequential set of instructions that constitute a program. CPU executes each instruction in a series of steps, called instruction cycle .

- A instruction cycle involves four steps

- o Fetching The processor fetches the instruction from the memory. The fetched instruction is placed in the Instruction Register. Program Counter holds the address of next instruction to be fetched and is incremented after each fetch.

- o Decoding The instruction that is fetched is broken down into parts or decoded. The instruction is translated into commands so that they correspond to those in the CPU's instruction set. The instruction set architecture of the CPU defines the way in which an instruction is decoded.

- o Executing The decoded instruction or the command is executed. CPU performs the operation implied by the program instruction. For example, if it is an ADD instruction, addition is performed.

- o Storing CPU writes back the results of execution, to the computer's memory. Figure 2.8 Instruction cycle  
Figure 2.9 Steps in instruction cycle

- Instructions are of different categories. Some categories of instructions are—

- o Memory access or transfer of data between registers.

- o Arithmetic operations like addition and subtraction.

- o Logic operations such as AND, OR and NOT.

- o Control the sequence, conditional connections, etc. A CPU performance is measured by the number of instructions it executes in a second, i.e., MIPS (million instructions per second), or BIPS (billion instructions per second).

## 2.7 MICROPROCESSOR

A processor's instruction set is a determining factor in its architecture. On the basis of the instruction set, microprocessors are classified as—Reduced Instruction Set Computer (RISC), and Complex Instruction Set Computer (CISC). The x86 instruction set of the original Intel 8086 processor is of the CISC type. The PCs are based on the x86 instruction set.

- CISC architecture hardwires the processor with complex instructions, which are difficult to create otherwise using basic instructions. CISC combines the different instructions into one single CPU.

- o CISC has a large instruction set that includes simple and fast instructions for performing basic tasks, as well as complex instructions that correspond to statements in the high level language.

- o An increased number of instructions (200 to 300) results in a much more complex processor, requiring millions of transistors. o Instructions are of variable lengths, using 8, 16 or 32 bits for storage. This results in the processor's time being spent in calculating where each instruction begins and ends.

- o With large number of application software programs being written for the processor, a new processor has to be backwards compatible to the older version of processors.

- o AMD and Cyrix are based on CISC.

- RISC has simple, single-cycle instructions, which performs only basic instructions. RISC architecture does not have hardwired advanced functions. All high-level language support is done in the software.

- o RISC has fewer instructions and requires fewer transistors, which results in the reduced manufacturing cost of processor.

- o The instruction size is fixed (32 bits). The processor need not spend time in finding out where each instruction begins and ends.

- o RISC architecture has a reduced production cost compared to CISC processors.

- o The instructions, simple in nature, are executed in just one clock cycle, which speeds up the program execution when compared to CISC processors.

- o RISC processors can handle multiple instructions simultaneously by processing them in parallel.

- o Apple Mac G3 and PowerPC are based on RISC. Processors like Athlon XP and Pentium IV use a hybrid of both technologies.

Pipelining improves instruction execution speed by putting the execution steps into parallel. A CPU can receive a single instruction, begin executing it, and receive another instruction before it has completed the first. This allows for more instructions to be performed, about, one instruction per clock cycle. Parallel Processing is the simultaneous execution of instructions from the same program on different processors. A program is divided into multiple processes that are handled in parallel in order to reduce execution time.

## **2.8 INTERCONNECTING THE UNITS OF A COMPUTER**

CPU sends data, instructions and information to the components inside the computer as well as to the peripherals and devices attached to it. Bus is a set of electronic signal pathways that allows information and signals to travel between components inside or outside of a computer. The different components of computer, i.e., CPU, I/O unit, and memory unit are connected with each other by a bus. The data, instructions and the signals are carried between the different components via a bus. The features and functionality of a bus are as follows—

- A bus is a set of wires used for interconnection, where each wire can carry one bit of data.

- A bus width is defined by the number of wires in the bus.

- A computer bus can be divided into two types—Internal Bus and External Bus.

- The Internal Bus connects components inside the motherboard like, CPU and system memory. It is also called the System Bus. Figure 2.10 shows interaction between processor and memory. Figure 2.10 Interaction between CPU and memory

- The External Bus connects the different external devices, peripherals, expansion slots, I/O ports and drive connections to the rest of computer. The external bus allows various devices to be attached to the computer. It allows for the expansion of computer's capabilities. It is generally slower than the system bus. It is also referred to as the Expansion Bus.

- A system bus or expansion bus comprise of three kinds of buses — data bus, address bus and control bus.
- The interaction of CPU with memory and I/O devices involves all the three buses.
  - o The command to access the memory or the I/O device is carried by the control bus.
  - o The address of I/O device or memory is carried by the address bus.
  - o The data to be transferred is carried by the data bus. Figure 2.11 shows interaction between processor, memory and the peripheral devices.

### **2.8.1 System Bus**

The functions of data bus, address bus and control bus, in the system bus, are as follows—

- Data Bus transfers data between the CPU and memory. The bus width of a data bus affects the speed of computer. The size of data bus defines the size of the processor. A processor can be 8, 16, 32 or 64-bit processor. An 8-bit processor has 8 wire data bus to carry 1 byte of data. In a 16-bit processor, 16-wire bus can carry 16 bits of data, i.e., transfer 2 bytes, etc. Figure 2.11 Interaction between CPU, memory and peripheral devices
- Address Bus connects CPU and RAM with set of wires similar to data bus. The width of address bus determines the maximum number of memory locations the computer can address. Currently, Pentium Pro, II, III, IV have 36-bit address bus that can address 236 bytes or 64 GB of memory.
- Control Bus specifies whether data is to be read or written to the memory, etc.

### **2.8.2 Expansion Bus**

The functions of data bus, address bus and control bus, in the expansion bus, are as follows—

- The expansion bus connects external devices to the rest of computer. The external devices like monitor, keyboard and printer connect to ports on the back of computer. These ports are actually a part of the small circuit board or expansion card that fits into an expansion slot on the motherboard. Expansion slots are easy to recognize on the motherboard.
  - Expansion slots make up a row of long plastic connectors at the back of the computer with tiny copper ‘finger slots’ in a narrow channel that grab the connectors on the expansion cards. The slots are attached to tiny copper pathways on the motherboard (the expansion bus), which allows the device to communicate with the rest of computer.
- Data Bus is used to transfer data between I/O devices and CPU. The exchange of data between CPU and I/O devices is according to the industry standard data buses. The most commonly used standard is Extended Industry Standard Architecture (EISA) which is a 32-bit bus architecture. Some of the common bus technologies are—
  - o Peripheral Component Interconnect (PCI) bus for hard disks, sound cards, network cards and graphics cards,
  - o Accelerated Graphics Port (AGP) bus for 3-D and full motion video,
  - o Universal Serial Bus (USB) to connect and disconnect different devices.
- Address Bus carries the addresses of different I/O devices to be accessed like the hard disk, CD ROM, etc.

- Control Bus is used to carry read/write commands, status of I/O devices, etc.

**2.8.3 External Ports** The peripheral devices interact with the CPU of the computer via the bus. The connections to the bus from the peripheral devices are made via the ports and sockets provided at the sides of the computer. The different ports and sockets facilitate the connection of different devices to the computer. Some of the standard port connections available on the outer sides of the computer are— port for mouse, keyboard, monitor, network, modem, and, audio port, serial port, parallel port and USB port. The different ports are physically identifiable by their different shapes, size of contact pins and number of pins. Figure 2.12 shows the interaction of serial and parallel port interfaces with the devices. Interaction of serial and parallel port interfaces

## **2.9 PERFORMANCE OF A COMPUTER**

There are a number of factors involved that are related to the CPU and have an effect on the overall speed and performance of the computer. Some of the factors that affect the performance of the computer include—

- Registers

The size of the register (word size) indicates the amount of data with which the computer can work at any given time. The bigger the size, the more quickly it can process data. A 32-bit CPU is one in which each register is 32 bits wide.

- RAM It is used to store data and instructions during execution of the instructions. Anything you do on your computer requires RAM. When the computer is switched on, the operating system, device drivers, the active files and running programs are loaded into RAM. If RAM is less, then the CPU waits each time the new information is swapped into memory from the slower devices. Larger the RAM size, the better it is. PCs nowadays usually have 1 GB to 4 GB of RAM.

- System Clock The clock speed of a CPU is defined as the frequency with which a processor executes instructions or the data is processed. Higher clock frequencies mean more clock ticks per second. The computer's operating speed is linked to the speed of the system clock. The clock frequency is measured in millions of cycles per second or megahertz (MHz) or gigahertz (GHz) which is billions of cycles per second. A CPU's performance is measured by the number of instructions it executes in a second, i.e., MIPS or BIPS. PCs nowadays come with a clock speed of more than 1 GHz. In Windows OS, you can select the System Properties dialog box to see the processor name and clock frequency.

- Bus Data bus is used for transferring data between CPU and memory. The data bus width affects the speed of computer. In a 16-bit processor, 16-bit wire bus can carry 16 bits of data.

The bus speed is measured in MHz. Higher the bus speed the better it is. Address bus connects CPU and RAM with a set of wires similar to data bus. The address bus width determines the maximum number of memory locations the computer can address.

Pentium Pro, II, III, IV have 36-bit address bus that can address 236 bytes or 64 GB of memory. PCs nowadays have a bus speed varying from 100 MHz to 400 MHz.

- Cache Memory Two of the main factors that affect a cache's performance are its size (amount of cache memory) and level L1, L2 and L3. Larger the size of cache, the better it is. PCs nowadays have a L1 cache

of 256KB and L2 cache of 1MB. Figure 2.13 shows the general information about a computer as displayed in the system properties window in Windows XP Professional.

## **2.10 INSIDE A COMPUTER CABINET**

The computer cabinet encloses the components that are required for the running of the computer. The components inside a computer cabinet include the power supply, motherboard, memory chips, expansion slots, ports and interface, processor, cables and storage devices.

### **2.10.1 Motherboard**

The computer is built up around a motherboard. The motherboard is the most important component in the PC. It is a large Printed Circuit Board (PCB), having many chips, connectors and other electronics mounted on it. The motherboard is the hub, which is used to connect all the essential components of a computer.

The RAM, hard drive, disk drives and optical drives are all plugged into interfaces on the motherboard. The motherboard contains the processor, memory chips, interfaces and sockets, etc. The motherboard may be characterized by the form factor, chipset and type of processor socket used. Form factor refers to the motherboard's geometry, dimensions, arrangement and electrical requirements. Different standards have been developed to build motherboards, which can be used in different brands of cases.

Advanced Technology Extended (ATX) is the most common design of motherboard for desktop computers. Chipset is a circuit, which controls the majority of resources (including the bus interface with the processor, cache memory and RAM, expansion cards, etc.) Chipset's job is to coordinate data transfers between the various components of the computer (including the processor and memory). As the chipset is integrated into the motherboard, it is important to choose a motherboard, which includes a recent chipset, in order to maximize the computer's upgradeability.

The processor socket may be a rectangular connector into which the processor is mounted vertically (slot), or a square-shaped connector with many small connectors into which the processor is directly inserted (socket). The Basic Input Output System (BIOS) and Complementary Metal-Oxide Semiconductor (CMOS) are present on the motherboard. ROM BIOS

- BIOS

It is the basic program used as an interface between the operating system and the motherboard. The BIOS (Figure 2.14) is stored in the ROM and cannot be rewritten. When the computer is switched on, it needs instructions to start. BIOS contain the instructions for the starting up of the computer. The BIOS runs when the computer is switched on. It performs a Power On Self Test (POST) that checks that the hardware is functioning properly and the hardware devices are present. It checks whether the operating system is present on the hard drive. BIOS invokes the bootstrap loader to load the operating system into memory. BIOS can be configured using an interface named BIOS setup, which can be accessed when the computer is booting up (by pressing the DEL key).

- CMOS Chip

BIOS ROMs are accompanied by a smaller CMOS (CMOS is a type of memory technology) memory chip. When the computer is turned off, the power supply stops providing electricity to the motherboard. When

the computer is turned on again, the system still displays the correct clock time. This is because the CMOS chip saves some system information, such as time, system date and essential system settings. CMOS is kept powered by a button battery located on the motherboard (Figure 2.15). The CMOS chip is working even when the computer power is switched off. Information of the hardware installed in the computer (such as the number of tracks or sectors on each hard drive) is stored in the CMOS chip. Figure 2.15 Battery for CMOS chip

2.10.2 Ports and Interfaces Motherboard has a certain number of I/O sockets that are connected to the ports and interfaces found on the rear side of a computer (Figure 2.16). You can connect external devices to the ports and interfaces, which get connected to the computer's motherboard.

- Serial Port— to connect old peripherals.
- Parallel Port— to connect old printers. Figure 2.16 Ports on the rear side of a PC
- USB Ports—to connect newer peripherals like cameras, scanners and printers to the computer. It uses a thin wire to connect to the devices, and many devices can share that wire simultaneously.
- Firewire is another bus, used today mostly for video cameras and external hard drives.
- RJ45 connector (called LAN or Ethernet port) is used to connect the computer to a network. It corresponds to a network card integrated into the motherboard.
- VGA connector for connecting a monitor. This connector interfaces with the built-in graphics card.
- Audio plugs (line-in, line-out and microphone), for connecting sound speakers and the microphone. This connector interfaces with the built-in sound card.
- PS/2 port to connect mouse and keyboard into PC.
- SCSI port for connecting the hard disk drives and network connectors.

### 2.10.3 Expansion Slots

The expansion slots are located on the motherboard. The expansion cards are inserted in the expansion slots. These cards give the computer new features or increased performance. There are several types of slots:

- ISA (Industry Standard Architecture) slot—To connect modem and input devices.
- PCI (Peripheral Component InterConnect) slot—To connect audio, video and graphics. They are much faster than ISA cards.
- AGP (Accelerated Graphic Port) slot—A fast port for a graphics card.
- PCI (Peripheral Component InterConnect) Express slot—Faster bus architecture than AGP and PCI buses.
- PC Card—It is used in laptop computers. It includes Wi-Fi card, network card and external modem.

### 2.10.4 Ribbon Cables

Ribbon cables are flat, insulated and consist of several tiny wires moulded together that carry data to different components on the motherboard. There is a wire for each bit of the word or byte and additional wires to coordinate the activity of moving information. They also connect the floppy drives, disk drives

and CD-ROM drives to the connectors in the motherboard. Nowadays, Serial Advanced Technology Attachment (SATA) cables have replaced the ribbon cables to connect the drives to the motherboard.

#### **2.10.5 Memory Chips**

The RAM consists of chips on a small circuit board. Two types of memory chips— Single In-line Memory Module (SIMM) and Dual In-line Memory Module (DIMM) are used in desktop computers. The CPU can retrieve information from DIMM chip at 64 bits compared to 32 bits or 16 bits transfer with SIMM chips.

DIMM chips are used in Pentium 4 onwards to increase the access speed. RAM memory chip

**2.10.6 Storage Devices** The disk drives are present inside the machine. The common disk drives in a machine are hard disk drive, floppy drive and CD drive or DVD drive. High-storage devices like hard disk, floppy disk and CDs are inserted into the hard disk drive, floppy drive and CD drive, respectively. These storage devices can store large amounts of data, permanently.

**2.10.7 Processor** The processor or the CPU is the main component of the computer. Select a processor based on factors like its speed, performance, reliability and motherboard support. Pentium Pro, Pentium 2 and Pentium 4 are some of the processors.

Figure 2.20 Storage devices (i) Hard disk drive, (ii) DVD drive, (iii) Floppy disk, (iv) CD

## 3. COMPUTER MEMORY

### 3.1 INTRODUCTION

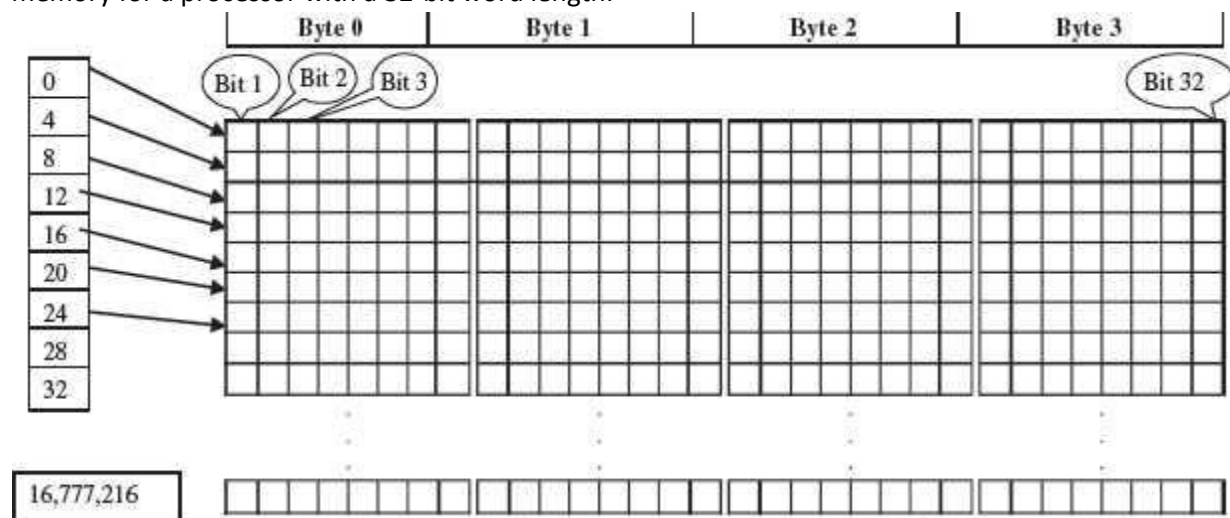
The computer's memory stores data, instructions required during the processing of data, and output results. Storage may be required for a limited period of time, instantly, or, for an extended period of time. Different types of memories, each having its own unique features, are available for use in a computer. The cache memory, registers, and RAM are fast memories and store the data and instructions temporarily during the processing of data and instructions. The secondary memory like magnetic disks and optical disks have large storage capacities and store the data and instructions permanently, but are slow memory devices. The memories are organized in the computer in a manner to achieve high levels of performance at the minimum cost.

In this chapter, we discuss different types of memories, their characteristics and their use in the computer.

### 3.2 MEMORY REPRESENTATION

The computer memory stores different kinds of data like input data, output data, intermediate results, etc., and the instructions. **Binary digit** or **bit** is the basic unit of memory. A *bit* is a single binary digit, i.e., 0 or 1. A bit is the smallest unit of representation of data in a computer. However, the data is handled by the computer as a combination of bits. A group of 8 bits form a **byte**. One byte is the smallest unit of data that is handled by the computer. One byte can store  $2^8$ , i.e., 256 different combinations of bits, and thus can be used to represent 256 different symbols. In a byte, the different combinations of bits fall in the range 00000000 to 11111111. A group of bytes can be further combined to form a **word**. A word can be a group of 2, 4 or 8 bytes. 1 bit = 0 or 1

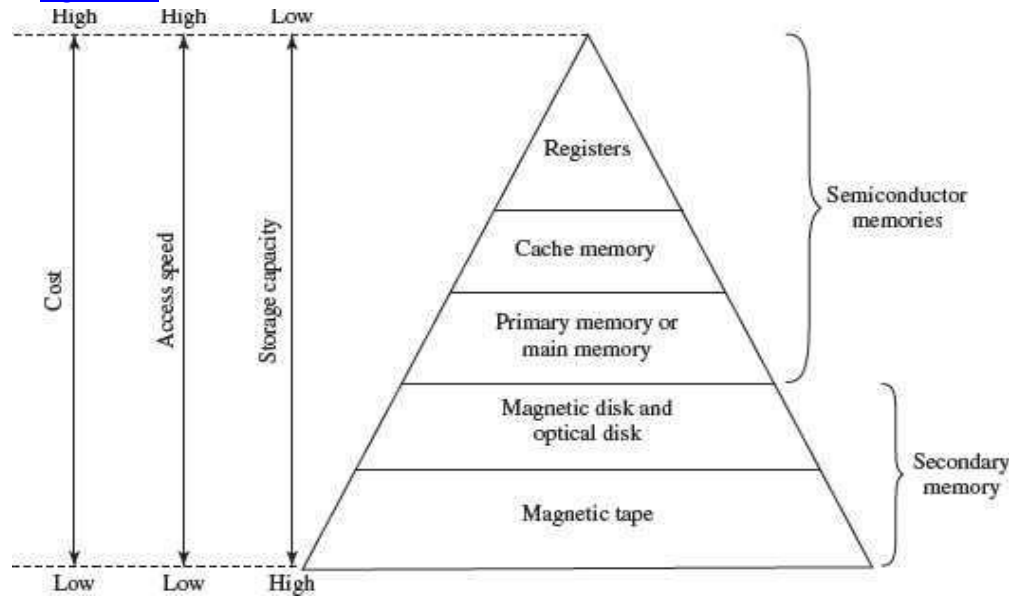
Memory is logically organized as a linear array of locations. For a processor, the range of the memory addresses is 0 to the maximum size of memory. [Figure 3.1](#) shows the organization of a 16 MB block of memory for a processor with a 32-bit word length.



**Figure 3.1** Organization of memory

### 3.3 MEMORY HIERARCHY

The memory is characterized on the basis of two key factors—capacity and access time. *Capacity* is the amount of information (in bits) that a memory can store. *Access time* is the time interval between the read/ write request and the availability of data. The lesser the access time, the faster is the *speed of memory*. Ideally, we want the memory with *fastest speed and largest capacity*. However, the cost of fast memory is very high. The computer uses a hierarchy of memory that is organized in a manner to enable the fastest speed and largest capacity of memory. The hierarchy of the different memory types is shown in [Figure 3.2](#).



**Figure 3.2** Memory hierarchy

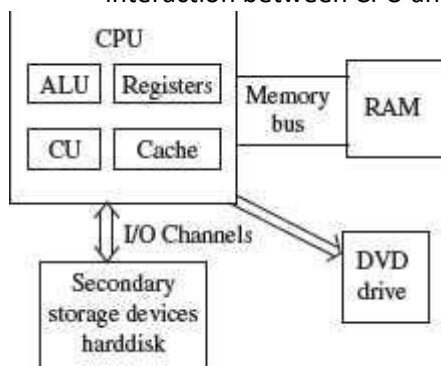
The internal memory and external memory are the two broad categories of memory used in the computer. The internal memory consists of the CPU registers, cache memory and primary memory. The internal memory is used by the CPU to perform the computing tasks. The external memory is also called the secondary memory. The secondary memory is used to store the large amount of data and the software.

In general, referring to the computer memory usually means the internal memory.

- **Internal Memory**—The key features of internal memory are—(1) limited storage capacity, (2) temporary storage, (3) fast access, and (4) high cost. Registers, cache memory, and primary memory constitute the internal memory. The primary memory is further of two kinds—RAM and ROM. Registers are the fastest and the most expensive among all the memory types. The registers are located inside the CPU, and are directly accessible by the CPU. The speed of registers is between 1—2 ns (nanosecond). The sum of the size of registers is about 200B. Cache memory is next in the hierarchy and is placed between the CPU and the main memory. The speed of cache

is between 2—10 ns. The cache size varies between 32 KB to 4MB. Any program or data that has to be executed must be brought into RAM from the secondary memory. Primary memory is relatively slower than the cache memory. The speed of RAM is around 60ns. The RAM size varies from 512KB to 3GB.

- **Secondary Memory**—The key features of secondary memory storage devices are—(1) very high storage capacity, (2) permanent storage (non-volatile), unless erased by user, (3) relatively slower access, (4) stores data and instructions that are not currently being used by CPU but may be required later for processing, and (5) cheapest among all memory. The storage devices consist of two parts—drive and device. For example, magnetic tape drive and magnetic tape, magnetic disk drive and disk, and, optical disk drive and disk. The speed of magnetic disk is around 60ms. The capacity of a hard disk ranges from 160 GB to 1,600 GB (1.6 Tera Bytes). [Figure 3.3](#) shows the interaction between CPU and memory.



**Figure 3.3** CPU and the memory

To get the fastest speed of memory with largest capacity and least cost, the fast memory is located close to the processor. The secondary memory, which is not as fast, is used to store information permanently, and is placed farthest from the processor. With respect to CPU, the memory is organized as follows—

- Registers are placed inside the CPU (small capacity, high cost, very high speed)
- Cache memory is placed next in the hierarchy (inside and outside the CPU)
- Primary memory is placed next in the hierarchy
- Secondary memory is the farthest from CPU (large capacity, low cost, low speed)

The speed of memories is dependent on the kind of technology used for the memory. The registers, cache memory and primary memory are semiconductor memories. They do not have any moving parts and are fast memories. The secondary memory is magnetic or optical memory, has moving parts and has slow speed.

### 3.4 CPU REGISTERS

- Registers are very high-speed storage areas located inside the CPU. After CPU gets the data and instructions from the cache or RAM, the data and instructions are moved to the registers for processing. Registers are manipulated directly by the control unit of CPU during instruction execution. That is why registers are often referred to as the CPU's

*working memory*. Since CPU uses registers for the processing of data, the number of registers in a CPU and the size of each register affect the power and speed of a CPU. The more the number of registers (ten to hundreds) and bigger the size of each register (8 bits to 64 bits), the better it is.

### 3.5 CACHE MEMORY

☐ Cache memory is placed in between the CPU and the RAM. Cache memory is a fast memory, faster than the RAM. When the CPU needs an instruction or data during processing, it first looks in the cache. If the information is present in the cache, it is called a *cache hit*, and the data or instruction is retrieved from the cache. If the information is not present in cache, then it is called a *cache miss* and the information is then retrieved from RAM. The content of cache is decided by the cache controller (a circuit on the motherboard). The most recently accessed information or instructions help the controller to guess the RAM locations that may be accessed next. To get good system performance, the number of hits must far outnumber the misses. The two main factors that affect the performance of cache are its size and level (L1, L2 and L3).

The CPU registers and the cache memory have been discussed in detail in the previous chapter.

### 3.6 PRIMARY MEMORY

Primary memory is the main memory of computer. It is a chip mounted on the motherboard of computer. Primary memory is categorized into two main types-

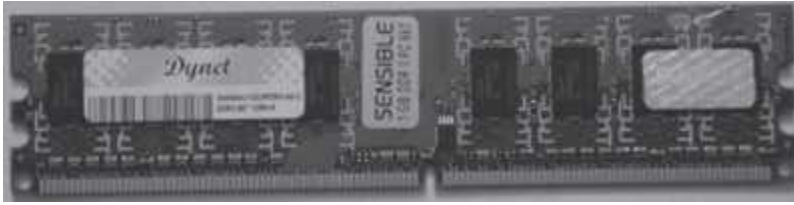
- Random Access Memory (RAM), and
- Read Only Memory (ROM)

RAM is used for the temporary storage of input data, output data and intermediate results. The input data entered into the computer using the input device, is stored in RAM for processing. After processing, the output data is stored in RAM before being sent to the output device. Any intermediate results generated during the processing of program are also stored in RAM. Unlike RAM, the data once stored in ROM either cannot be changed or can only be changed using some special operations. Therefore, ROM is used to store the data that does not require a change. *Flash memory* is another form of rewritable read-only memory that is compact, portable, and requires little energy.

#### 3.6.1 Random Access Memory

- RAM is used to *store data and instructions during the operation of computer*.
  - The data and instructions that need to be operated upon by CPU are first brought to RAM from the secondary storage devices like the hard disk.
  - CPU interacts with RAM to get the data and instructions for processing.
- RAM loses information when the computer is powered off. It is a *volatile memory*. When the power is turned on, again, all files that are required by the CPU are loaded from the hard disk to RAM. Since RAM is a volatile memory, any information that needs to be saved for a longer duration of time must not be stored in RAM.

- RAM provides *random access* to the stored bytes, words, or larger data units. This means that it requires same amount of time to access information from RAM, irrespective of where it is located in it.
- RAM can be *read from and written to* with the same speed.
- The *size of RAM is limited due to its high cost*. The size of RAM is measured in MB or GB.
- The performance of RAM is affected by—
  - Access speed (how *quickly* information can be retrieved). The speed of RAM is expressed in nanoseconds.
  - Data transfer unit size (how *much* information can be retrieved in one request).
- RAM affects the speed and power of a computer. More the RAM, the better it is. Nowadays, computers generally have 512 MB to 4 GB of RAM.
- RAM is a microchip implemented using semiconductors.
- There are two categories of RAM, depending on the technology used to construct a RAM— (1) Dynamic RAM (DRAM), and (2) Static RAM (SRAM).
- **DRAM** is the most common type of memory chip. DRAM is mostly used as main memory since it is small and cheap.
  - It uses transistors and capacitors. The transistors are arranged in a matrix of rows and columns. The capacitor holds the bit of information 0 and 1. The transistor and capacitor are paired to make a *memory cell*. The transistor acts as a switch that lets the control circuitry on the memory chip read the capacitor or change its state.
  - DRAM must be refreshed continually to store information. For this, a memory controller is used. The memory controller recharges all the capacitors holding a 1 before they discharge. To do this, the memory controller reads the memory and then writes it right back.
  - DRAM gets its name from the refresh operation that it requires to store the information; otherwise it will lose what it is holding. The refresh operation occurs automatically thousands of times per second. DRAM is slow because the refreshing takes time.
  - Access speed of DRAM ranges from 50 to 150 ns.
- **SRAM** chip is usually used in *cache memory* due to its high speed.
  - SRAM uses multiple transistors (four to six), for each memory cell. It does not have a capacitor in each cell.
  - A SRAM memory cell has more parts so it takes more space on a chip than DRAM cell.
  - It does not need constant refreshing and therefore is faster than DRAM. ◦ SRAM is more expensive than DRAM, and it takes up more space.
  - It stores information as long as it is supplied with power.
  - SRAM are easier to use and very fast. The access speed of SRAM ranges from 2– 10 nanosecond.
- The memory chips ([Figure 3.4](#)) are available on a separate Printed Circuit Board (PCB) that is plugged into a special connector on the motherboard. Memory chips are generally available as part of a card called a *memory module*. There are generally two types of RAM modules—Single Inline Memory Module (SIMM) and Dual Inline Memory Module (DIMM).



**Figure 3.4** PCB containing RAM chip of 1 GB

- SIMM modules have memory chip on one side of the PCB. SIMM modules can store 8 bits to 32 bits of data simultaneously.
- DIMM modules have memory chips on both sides of the PCB. DIMM format are 64-bit memories. Smaller modules known as Small Outline DIMM (SO DIMM) are designed for portable computers. SO DIMM modules have 32-bit memory.

### 3.6.2 Read Only Memory

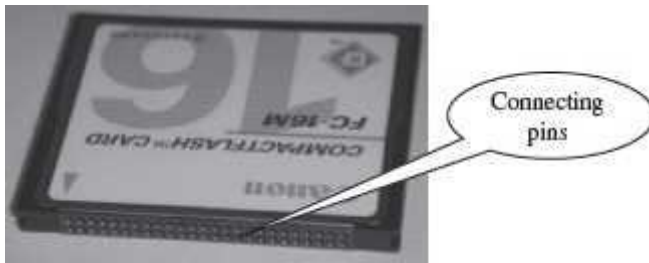
ROM is a *non-volatile* primary memory. It does not lose its content when the power is switched off. The features of ROM are described as follows—

- ROM, as the name implies, has only read capability and no write capability. After the information is stored in ROM, it is permanent and cannot be corrected.
- ROM comes programmed by the manufacturer. It stores standard processing programs that permanently reside in the computer. ROM stores the data needed for the start up of the computer. The instructions that are required for initializing the devices attached to a computer are stored in ROM.
- The ROM memory chip ([Figure 3.5](#)) stores the *Basic Input Output System (BIOS)*. BIOS provides the processor with the information required to boot the system. It provides the system with the settings and resources that are available on the system. BIOS is a permanent part of the computer. It does not load from disk but instead is stored in a ROM memory chip. The program code in the BIOS differs from ordinary software since it acts as an integral part of the computer. When the computer is turned on, the BIOS does the following things—



**Figure 3.5** ROM BIOS and CMOS battery on a motherboard

- *Power On Self Test (POST)* is a program that runs automatically when the system is booted. BIOS performs the power-on self-test. It checks that the major hardware components are working properly.
  - BIOS setup program, which is a built-in utility in BIOS, lets the user set the many functions that control how the computer works. BIOS displays the system settings and finds the bootable devices. It loads the interrupt handlers and device drivers. It also initializes the registers.
  - *Bootstrap Loader* is a program whose purpose is to start the computer software for operation when the power is turned on. It loads the operating system into RAM and launches it. It generally seeks the operating system on the hard disk. The bootstrap loader resides in the ROM. The BIOS initiates the bootstrap sequence.
- ROMs are of different kinds. They have evolved from the fixed read only memory to the ones that can be programmed and re-programmed. They vary in the number of re-writes and the method used for the re-writing. Programmable ROM (PROM), Erasable Programmable ROM (EPROM) and Electrically Erasable Programmable ROM (EEPROM) are some of the ROMs. All the different kinds of ROM retain their content when the power is turned off.
  - **PROM** can be programmed with a special tool, but after it has been programmed the contents cannot be changed. PROM memories have thousands of fuses (or diodes). High voltage (12 V) is applied to the fuses to be burnt. The burnt fuses correspond to 0 and the others to 1.
  - **EPROM** can be programmed in a similar way as PROM, but it can be erased by exposing it to ultra violet light and re-programmed. EPROM chips have to be removed from the computer for re-writing.
  - **EEPROM** memories can be erased by electric charge and re-programmed. EEPROM chips do not have to be removed from the computer for re-writing.
- **Flash Memory** is a kind of semiconductor-based non-volatile, rewritable computer memory that can be electrically erased and reprogrammed ([Figure 3.6](#)). It is a specific type of EEPROM.



**Figure 3.6** Flash memory

- It combines the features of RAM and ROM. It is a random access memory and its content can be stored in it at any time. However, like ROM, the data is not lost when the machine is turned off or the electric power is cut. Flash memory stores bits of data in memory cells.
- Flash memories are high-speed memories, durable, and have low-energy consumption. Since flash memory has no moving part, it is very shock-resistant. Due to these features, flash memory is used in devices such as digital camera, mobile phone, printer, laptop computer, and record and play back sound devices, such as MP3 players.

### 3.7 SECONDARY MEMORY

In the previous section, we saw that RAM is expensive and has a limited storage capacity. Since it is a volatile memory, it cannot retain information after the computer is powered off. Thus, in addition to primary memory, an auxiliary or secondary memory is required by a computer. The secondary memory is also called the storage device of computer. *In this chapter, the terms secondary memory and storage device are used interchangeably.* In comparison to the primary memory, the secondary memory stores much larger amounts of data and information (for example, an entire software program) for extended periods of time. The data and instructions stored in secondary memory must be fetched into RAM before processing is done by CPU.

Magnetic tape drives, magnetic disk drives, optical disk drives and magneto-optical disk drives are the different types of storage devices.

### 3.8 ACCESS TYPES OF STORAGE DEVICES

The information stored in storage devices can be accessed in two ways—

1. Sequential access
2. Direct access

#### 3.8.1 Sequential Access Devices

Sequential access means that computer must run through the data in sequence, starting from the beginning, in order to locate a particular piece of data. Magnetic tape is an example of sequential access device. Let us suppose that magnetic tape consists of 80 records. To access the 25th record, the computer

starts from first record, then reaches second, third etc. until it reaches the 25th record. Sequential access devices are generally slow devices.

### 3.8.2 Direct Access Devices

Direct access devices are the ones in which any piece of data can be retrieved in a non-sequential manner by locating it using the data's address. It accesses the data directly, from a desired location. Magnetic disks and optical disks are examples of direct access devices. There is no predefined order in which one can read and write data from a direct access device. In a magnetic disk consisting of 80 records, to access the 25th record, the computer can directly access the 25th record, without going past the first 24 records. Based on access, magnetic tapes are sequential access devices, and, magnetic disks, optical disk and magneto-optical disks are direct access devices.

### 3.9 MAGNETIC TAPE

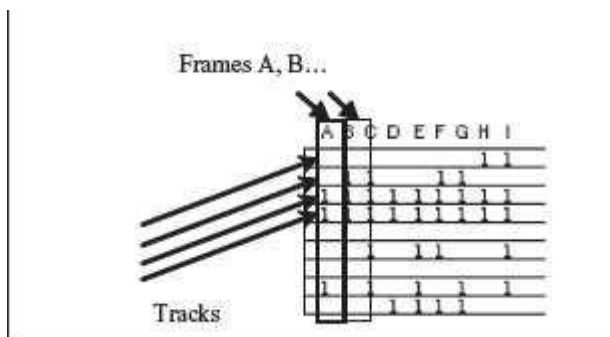
Magnetic tape is a plastic tape with magnetic coating ([Figure 3.7](#)). It is a storage medium on a large open reel or in a smaller cartridge or cassette (like a music cassette). Magnetic tapes are cheaper storage media. They are durable, can be written, erased, and re-written. Magnetic tapes are sequential access devices, which mean that the tape needs to rewind or move forward to the location where the requested data is positioned in the magnetic tape. Due to their sequential nature, magnetic tapes are not suitable for data files that need to be revised or updated often. They are generally used to store back-up data that is not frequently used or to transfer data from one system to other.



**Figure 3.7** A 10.5-inch reel of 9-track tape

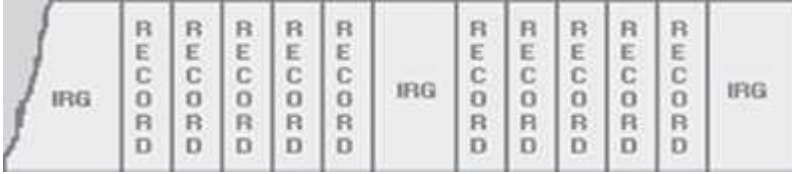
The **working of magnetic tape** is explained as follows—

- Magnetic tape is divided horizontally into tracks (7 or 9) and vertically into frames ([Figure 3.8](#)). A frame stores one byte of data, and a track in a frame stores one bit. Data is stored in successive frames as a string with one data (byte) per frame.



**Figure 3.8** A portion of magnetic tape

- Data is recorded on tape in the form of blocks, where a block consists of a group of data also called as records. Each block is read continually. There is an *Inter-Record Gap (IRG)* between two blocks that provides time for the tape to be stopped and started between records ([Figure 3.9](#)).



**Figure 3.9** Blocking of data in a magnetic tape

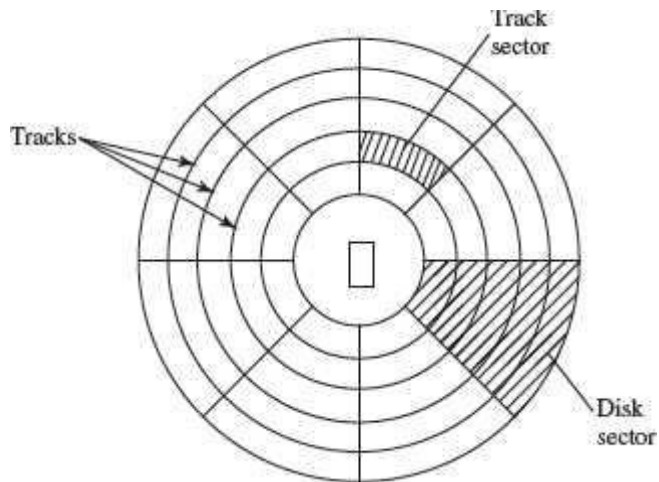
- Magnetic tape is mounted on a magnetic tape drive for access. The basic **magnetic tape drive** mechanism consists of the supply reel, take-up reel, and the read/write head assembly. The magnetic tape moves on tape drive from the supply reel to take-up reel, with its magnetic coated side passing over the read/write head.
- Tapes are categorized based on their width -  $\frac{1}{4}$  inch,  $\frac{1}{2}$  inch, etc.
- The storage capacity of the tape varies greatly. A 10-inch diameter reel of tape which is 2400 feet long can store up to 180 million characters. **The features of magnetic tape are—**
  - Inexpensive storage device
  - Can store a large amount of data
  - Easy to carry or transport
  - Not suitable for random access data
  - Slow access device
  - Needs dust prevention, as dust can harm the tape
  - Suitable for back-up storage or archiving

### 3.10 MAGNETIC DISK

Magnetic disk is a direct access secondary storage device. It is a thin plastic or metallic circular plate coated with magnetic oxide and encased in a protective cover. Data is stored on magnetic disks as magnetized spots. The presence of a magnetic spot represents the bit 1 and its absence represents the bit 0.

**The working of magnetic disk** is explained as follows—

- The surface of disk is divided into concentric circles known as **tracks**. The outermost track is numbered 0 and the innermost track is the last track. Tracks are further divided into **sectors**. A sector is a pie slice that cuts across all tracks. The data on disk is stored in sector. Sector is the smallest unit that can be read or written on a disk. A disk has eight or more sectors per track ([Figure 3.10](#)).



**Figure 3.10** Tracks and sectors of a disk

- Magnetic disk is inserted into a magnetic disk drive for access. The drive consists of a read/write head that is attached to a disk arm, which moves the head. The disk arm can move inward and outward on the disk.
- During reading or writing to disk, the motor of disk drive moves the disk at high speed (60–150 times/sec.)
- Accessing data on the disk requires the following—
  - The read/write head is positioned to the desired track where the data is to be read from or written to. The time taken to move the read/write head to the desired track is called the **seek time**.
  - Once the read/write head is at the right track, then the head waits for right sector to come under it (disk is moving at high speed). The time taken for desired sector of the track to come under read/write head is called the **latency time**.
  - Once the read/write head is positioned at the right track and sector, the data has to be written to disk or read from disk. The rate at which data is written to disk or read from disk is called **data transfer rate**.
  - The sum of seek time, latency time and time for data transfer is the **access time** of the disk.
- The storage capacity of disk drive is measured in gigabytes (GB).
- Large disk storage is created by stacking together multiple disks. A set of same tracks on all disks forms a **cylinder**. Each disk has its own read/write head which work in coordination.
- A disk can also have tracks and sectors on both sides. Such a disk is called **double-sided disk**.

The features of magnetic disk are—

- Cheap storage device
- Can store a large amount of data
- Easy to carry or transport
- Suitable for frequently read/write data

- Fast access device
- More reliable storage device
- To be prevented from dust, as the read/write head flies over the disk. Any dust particle in between can corrupt the disk.

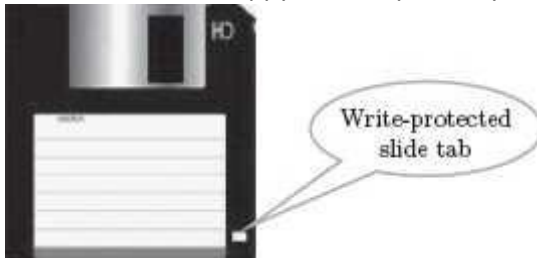
Finding data on a magnetic disk is as follows—

- In order to use a disk, it has to be formatted. Formatting includes assigning addresses to various locations on disk, assigning location of root directory and checking for defects on the surface of disk.
- During formatting, the tracks and sectors of a disk are labeled, which provides an address to each location of the disk.
- There are different methods to format a disk. File Allocation Table (FAT) is the commonly used logical format for disk formatting performed by Windows.
- Four areas are created when a disk is formatted using FAT—
  - o **Boot Sector** It contains the program that runs when the computer is started. The program checks if the disk has files required to run the operating system. It then transfers control to an operating system program which continues the startup process. Boot sector also contains information about the disk, like number of bytes per sector and number of sectors per track. This information is required by the operating system to access the data on the disk.
  - o **File Allocation Table** It records the location of each file and status of each sector. While reading or writing to disk, operating system checks the FAT to find free area or locate where data is stored on disk, respectively.
  - o **Root Directory** This is the main folder of disk. It contains other folders in it, creating a hierarchical system of folders. The root directory contains information about all folders on the disk.
  - o **Data Area** The remaining area of the disk (after boot sector, FAT, root directory) is the data area. It stores the program files and data files that are stored on the disk.
- The Windows XP and the Windows 2000 operating system use the New Technology File System (NTFS) file system. The NTFS file system offers better security and increased performance. It allows using of filenames that are more than eight characters long.

Floppy disk, hard disk and zip disk are the different types of magnetic disks.

### 3.10.1 Floppy Disk

- Floppy disk (FD) is a flat, round, single disk made of Mylar plastic and enclosed in square plastic jacket ([Figure 3.11](#)).
- Floppy Disk Drive (FDD) is the disk drive for floppy disk.
- The floppy disk is inserted into the floppy disk drive to read or write data to it.
- Floppy disk has a write-protect slide tab that prevents a user from writing to it.
- A floppy disk may be single-sided or double-sided disk, i.e., data can be read and written on one and both sides of floppy disk, respectively.

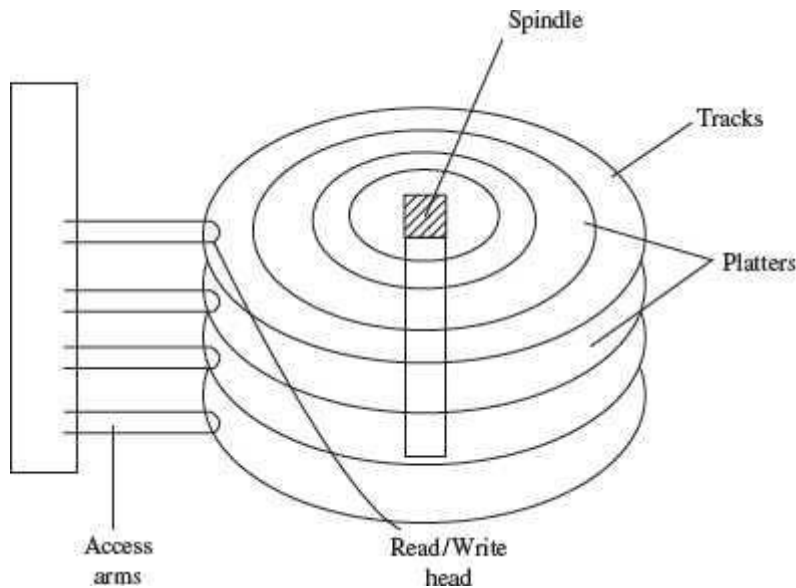


**Figure 3.11** Floppy disk

- They are portable. They can be removed from the disk drive, carried or stored separately.
- They are small and inexpensive.
- Floppy disks are slower to access than hard disk. They have less storage capacity and are less expensive than hard disk.
- They come in two basic sizes—5-¼ inch and 3-½ inch.
- The 5-¼ inch disk came around 1987. It can store 360 KB to 1.2 MB of data.
- The 3-½ inch disk has capacity of 400 KB to 1.44 MB. It usually contains 40 tracks and 18 sectors per track and can store 512 bytes per sector.

### 3.10.2 Hard Disk

- A hard disk (HD) consists of one or more platters divided into concentric tracks and sectors. It is mounted on a central spindle, like a stack. It can be read by a read/write head that pivots across the rotating disks. The data is stored on the platters covered with magnetic coating ([Figure 3.12](#)).



**Figure 3.12** Parts of hard disk

- Hard disk is a fixed disk. The disk is not removable from the drive, unlike floppy disk. □ The hard disk and Hard Disk Drive (HDD) is a single unit.
- Hard disk can store much more data than floppy disk. The data in hard disk are packed more closely (because fast spinning uses smaller magnetic charges) and they have multiple platters, with data being stored on both sides of each platter. Large capacity hard disks may have 12 or more platters.
- Unlike floppy disk, the read/write head of hard disk does not touch the disk during accessing.
- Hard disk can spin at the speed of up to 10,000 revolutions per minute and have an access time of 9–14 ms. It stores 512 bytes per sector but the number of sectors are more per track (54 or more) than floppy disk.
- Nowadays, hard disks are available that can store up to 500 GB of data. Generally, PCs come with 160 GB hard disk.
- Hard disk is the key secondary storage device of computer. The operating system is stored on the hard disk. The performance of computer like speed of computer boot up, loading of programs to primary memory, loading of large files like images, video, audio etc., is also dependent on the hard disk.
- Nowadays, *portable external hard disk drive* is available which can be attached to the USB drive of the computer. They come in the storage capacities of 80 GB to 500 GB.

### 3.10.3 Zip Disk

- They are high-capacity removable disk and drive.
- They have the speed and capacity of hard disk and portability of floppy disk.
- Zip disk are of the same size as floppy disk, i.e., 3½ inch but have a much higher capacity than the floppy disk ([Figure 3.13](#)).

□

Zip disk and drive were made by Iomega Corp. It comes as a complete unit—disk, drive, connection cable, power cord and operating system. It can be connected to the computer system externally using a parallel chord or SCSI cable.

- Their capacity ranges from 100 MB to 750 MB. They can be used to store large files, audio and video data.



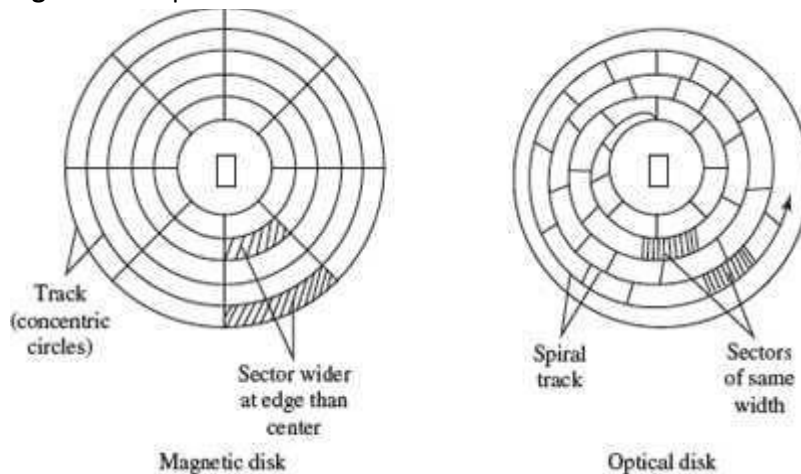
**Figure 3.13** Zip disk

### 3.11 OPTICAL DISK

Optical disk ([Figure 3.14](#)) is a flat and circular disk which is coated with reflective plastic material that can be altered by laser light. Optical disk does not use magnetism. The bits 1 and 0 are stored as spots that are relatively bright and light, respectively.

- An optical disk consists of a single spiral track that starts from the edge to the centre of disk. Due to its spiral shape, it can access large amount of data sequentially, for example music and video. The random access on optical disk is slower than that of magnetic disk, due to its spiral shape.
- The tracks on optical disk are further divided into sectors which are of same length. Thus, the sectors near the centre of disk wrap around the disk longer than the sectors on the edges of disk. Reading the disk thus requires spinning the disk faster when reading near the centre and slower when reading near the edge of disk. Optical disks are generally slower than hard disks. [Figure 3.15](#) shows the tracks and sectors in a magnetic disk and optical disk.



**Figure 3.14** Optical disk**Figure 3.15** Sectors and track in magnetic disk and optical disk

- Optical disks can store large amount of data, up to 6 GB, in a small space. Commonly used optical disks store 600–700 MB of data.
- The access time for an optical disk ranges from 100 to 200 ms.
- There are two most common categories of optical disks—read-only optical disks and recordable optical disks.

### 3.11.1 CD-ROM

- Originally, Compact Disk (CD) was a popular medium for storing music. Now, it is used in computers to store data and is called Compact Disk-Read Only Memory (CD-ROM).
- As the name suggests, CD-ROM ([Figure 3.16](#)) is an optical disk that can only be read and not written on. CD-ROM is written on by the manufacturer of the CD-ROM using the laser light.
- A CD-ROM drive reads data from the compact disk. Data is stored as pits (depressions) and lands (flat area) on CD-ROM disk. When the laser light is focused on the disk, the pits scatter the light (interpreted as 0) and the lands reflect the light to a sensor (interpreted as 1).

As CD-ROM is read only, no changes can be made into the data contained in it.

- Since there is no head touching the disk, but a laser light, CD-ROM does not get worn out easily.
- The storage density of CD-ROM is very high and cost is low as compared to floppy disk and hard disk.
- Access time of CD-ROM is less. CD-ROM drives can read data at 150Kbps. They come in multiples of this speed like—2x, 4x, 52x, 75x, etc.
- It is a commonly used medium for distributing software and large data.

### 3.11.2 DVD-ROM

□

- Digital Video Disk-Read Only Memory (DVD-ROM) is an optical storage device used to store digital video or computer data ([Figure 3.17](#)).
- DVDs look like CDs, in shape and physical size.



**Figure 3.16** CD-ROM



**Figure 3.17** DVDs

- It improves on CD technology.
- It is a high-density medium with increased track and bit density.
- DVD-ROM uses both sides of the disk and special data compression technologies. The tracks for storing data are extremely small.

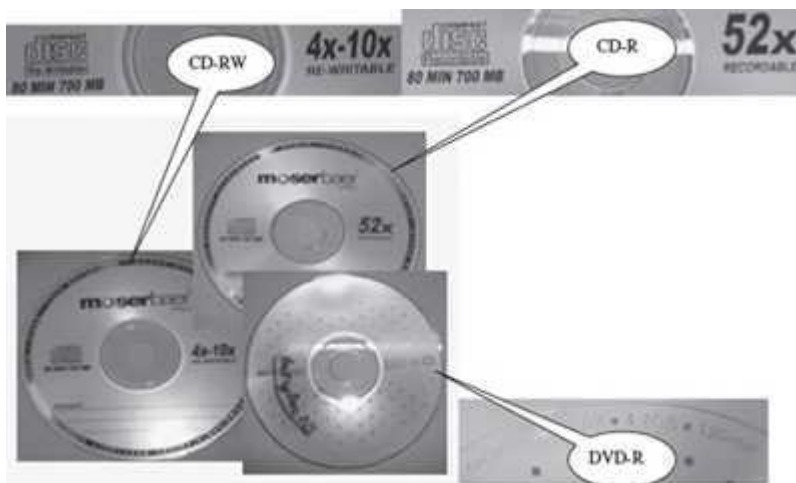
□

- A full-length movie can be stored on a single disk.
- Each side of DVD-ROM can store 4.7 GB of data, so a single DVD can store 9.4 GB of data.
- New DVD-ROMs use layers of data track, to double its capacity. Such dual layer disks can store 17 GB of data.

### 3.11.3 Recordable Optical Disk

In addition to the read only CDs and DVDs, recordable optical disks ([Figure 3.18](#)) are also available. Users can record music, video, audio and data on it. The recordable optical disks are—

- **Compact Disk-Recordable (CD-R)** is a Write Once-Read Many (WORM) disk. A CD-R disk allows the user to write data permanently on to the disk. Once the data is written, it cannot be erased. CD-R disk uses a laser that burns pits into the disk surface. It looks like a CD disk externally. To write to a CD-R disk, a device named CD-Writer or CD-burner is required. A CD-R disk can store 700 MB of data that can run for 80 minutes. CD-R is used to create music CDs in home computers, back up data from other storage devices, archives of large data, etc.
- **Compact Disk-ReWritable (CD-RW)** allows data to be written, erased and re-written on. The capacity of CD-RW is same as a CD. They generally do not play on all CD-ROM drives.
- **Digital Video Disk-Recordable (DVD-R)** allows recording of data on a DVD. A DVD writer device is required to write the data to DVD. The data once written on a DVD cannot be erased or changed



**Figure 3.18** CD-R, CD-RW and DVD-R

### 3.12 MAGNETO-OPTICAL DISK

- Magneto-optical disks use laser beam to read data and magnetic field to write data to disk. • These are optical disks where data can be written, erased and re-written.
- They are expensive and outdated. They were used during the mid 1990s. They have now been replaced by CD-RW and DVD-R.

### 3.13 USING THE COMPUTER MEMORY

The computer starts using the memory from the moment the computer is switched on, till the time it is switched off. The list of steps that the computer performs from the time it is switched

on are—

- Turn the computer on.
- The computer loads data from ROM. It makes sure that all the major components of the computer are functioning properly.
- The computer loads the BIOS from ROM. The BIOS provides the most basic information about storage devices, boot sequence, security, plug and play capability and other items. ¶ The computer loads the OS from the hard drive into the system's RAM. CPU has immediate access to the OS as the critical parts of the OS are maintained in RAM as long as the computer is on. This enhances the performance and functionality of the overall system.
- Now the system is ready for use.
- When you load or open an application it is loaded in the RAM. Since the CPU looks for information in the RAM, any data and instructions that are required for processing (read, write or update) is brought into RAM. To conserve RAM usage, many applications load only the essential parts of the program initially and then load other pieces as needed. Any files that are opened for use in that application are also loaded into RAM.
- The CPU requests the data it needs from RAM, processes it and writes new data back to RAM in a continuous cycle. The shuffling of data between the CPU and RAM happens millions of times every second.
- When you save a file and close the application, the file is written to the secondary memory as specified by you. The application and any accompanying files usually get deleted from RAM to make space for new data.
- If the files are not saved to a storage device before being closed, they are lost.

*Sometimes, when you write a program and the power goes off, your program is lost if you have not saved it. This is because your program was in the RAM and was not saved on the secondary memory; the content of the RAM gets erased when the power is switched off.*

## 5. DATA REPRESENTATION

### 5.1 INTRODUCTION

The data stored in the computer may be of different kinds, as follows—

- Numeric data (0, 1, 2, ..., 9)
- Alphabetic data (A, B, C, ..., Z)
- Alphanumeric data—Combination of any of the symbols—(A, B, C... Z), (0, 1... 9), or special characters (+, −, Blank), etc.

All kinds of data, be it alphabets, numbers, symbols, sound data or video data, is represented in terms of 0s and 1s, in the computer. Each symbol is represented as a unique combination of 0s and 1s.

This chapter discusses the number systems that are commonly used in the computer. The number systems discussed in this chapter are—(1) Decimal number system, (2) Binary number system, (3) Octal number system, and (4) Hexadecimal number system. The number conversions described in this chapter are—

- Decimal (Integer, Fraction, Integer. Fraction) to Binary, Octal, Hexadecimal □ Binary, Octal, Hexadecimal (Integer, Fraction, Integer. Fraction) to Decimal
- Binary to Octal, Hexadecimal □ Octal, Hexadecimal to Binary

The chapter also discusses the binary arithmetic operations and the representation of signed and unsigned numbers in the computer. The representation of numbers using binary coding schemes and the logic gates used for the manipulation of data are also discussed.

### 5.2 NUMBER SYSTEM

A number system in *base r or radix r* uses unique symbols for  $r$  digits. One or more digits are combined to get a number. The base of the number decides the valid digits that are used to make a number. In a number, the *position* of digit starts from the right-hand side of the number. The rightmost digit has position 0, the next digit on its left has position 1, and so on. The digits of a number have two kinds of values—

- Face value, and □ Position value.

The **face value** of a digit is the digit located at that position. For example, in decimal number 52, face value at position 0 is 2 and face value at position 1 is 5.

The **position value** of a digit is ( $\text{base}^{\text{position}}$ ). For example, in decimal number 52, the position value of digit 2 is  $10^0$  and the position value of digit 5 is  $10^1$ . Decimal numbers have a base of 10.

The **number** is calculated as the sum of, face value \*  $\text{base}^{\text{position}}$ , of each of the digits. For decimal number 52, the number is  $5 * 10^1 + 2 * 10^0 = 50 + 2 = 52$

In computers, we are concerned with four kinds of number systems, as follows—

- Decimal Number System —Base 10
- Binary Number System —Base 2
- Octal Number System —Base 8
- Hexadecimal Number System—Base 16

The numbers given as input to computer and the numbers given as output from the computer, are generally in decimal number system, and are most easily understood by humans. However, computer understands the binary number system, i.e., numbers in terms of 0s and 1s. The binary data is also represented, internally, as octal numbers and hexadecimal numbers due to their ease of use.

A number in a particular base is written as (number)<sub>base of number</sub> for example, (23)<sub>10</sub> means that the number 23 is a decimal number, and (345)<sub>8</sub> shows that 345 is an octal number.

### 5.2.1 Decimal Number System

- It consists of 10 digits—0, 1, 2, 3, 4, 5, 6, 7, 8 and 9.
- All numbers in this number system are represented as combination of digits 0—9. For example, 34, 5965 and 867321.
- The position value and quantity of a digit at different positions in a number are as follows—

Position:	3	2	1	0	.	-1	-2	-3
Position Value:	$10^3$	$10^2$	$10^1$	$10^0$		$10^{-1}$	$10^{-2}$	$10^{-3}$
Quantity:	1000	100	10	1		1/10	1/100	1/1000

### 5.2.2 Binary Number System

- The binary number system consists of two digits—0 and 1.
- All binary numbers are formed using combination of 0 and 1. For example, 1001, 11000011 and 10110101.
- The position value and quantity of a digit at different positions in a number are as follows—

Position:	3	2	1	0	.	-1	-2	-3
Position Value:	$2^3$	$2^2$	$2^1$	$2^0$		$2^{-1}$	$2^{-2}$	$2^{-3}$
Quantity:	8	4	2	1		1/2	1/4	1/8

### 5.2.3 Octal Number System

- The octal number system consists of eight digits—0 to 7.
- All octal numbers are represented using these eight digits. For example, 273, 103, 2375, etc.
- The position value and quantity of a digit at different positions in a number are as follows—

Position:	3	2	1	0	.	-1	-2	-3
Position Value:	$8^3$	$8^2$	$8^1$	$8^0$		$8^{-1}$	$8^{-2}$	$8^{-3}$
Quantity:	512	64	8	1		1/8	1/64	1/512

## 5.2.4 Hexadecimal Number System

- The hexadecimal number system consists of sixteen digits—0 to 9, A, B, C, D, E, F, where (A is for 10, B is for 11, C-12, D-13, E-14, F-15).
- All hexadecimal numbers are represented using these 16 digits. For example, 3FA, 87B, 113, etc.
- The position value and quantity of a digit at different positions in a number are as follows—

Position:	3	2	1	0		-1	-2	-3
Position Value:	$16^3$	$16^2$	$16^1$	$16^0$		$16^{-1}$	$16^{-2}$	$16^{-3}$
Quantity:	4096	256	16	1		1/16	1/256	1/4096

[Table 5.1](#) summarizes the base, digits and largest digit for the above discussed number systems. [Table 5.2](#) shows the binary, octal and hexadecimal equivalents of the decimal numbers 0–16.

	Base	Digits	Largest Digit
Decimal	10	0–9	9
Binary	2	0,1	1
Octal	8	0–7	7
Hexadecimal	16	0–9, A, B, C, D, E, F	F (15)

**Table 5.1** Summary of number system

Decimal	Binary	Octal	Hexadecimal
0	0000	000	0
1	0001	001	1
2	0010	002	2
3	0011	003	3
4	0100	004	4
5	0101	005	5
6	0110	006	6
7	0111	007	7
8	1000	010	8
9	1001	011	9
10	1010	012	A
11	1011	013	B
12	1100	014	C
13	1101	015	D
14	1110	016	E
15	1111	017	F
16	10000	020	10

**Table 5.2** Decimal, binary, octal and hexadecimal equivalents

### 5.3 CONVERSION FROM DECIMAL TO BINARY, OCTAL, HEXADECIMAL

A decimal number has two parts—integer part and fraction part. For example, in the decimal number 23.0786, 23 is the integer part and .0786 is the fraction part. The method used for the conversion of the integer part of a decimal number is different from the one used for the fraction part. In the following subsections, we shall discuss the conversion of decimal integer, decimal fraction and decimal integer.fraction number into binary, octal and hexadecimal number.

#### 5.3.1 Converting Decimal *Integer* to Binary, Octal, Hexadecimal

A decimal integer is converted to any other base, by using the division operation. To

convert a decimal integer to—

- binary-divide by 2,
- octal-divide by 8, and,
- hexadecimal-divide by 16.

Let us now understand this conversion with the help of some examples.

**Example 1:** Convert 25 from Base 10 to Base 2.

1. Make a table as shown below. Write the number in centre and toBase on the left side.

	to Base	Number	Remainder
		(Quotient)	
2		25	


2. Divide the number with *toBase*. After each division, write the remainder on right-side column and quotient in the next line in the middle column. Continue dividing till the quotient is 0.

	to Base	Number	Remainder
		(Quotient)	
2		25	
2		12	1
2		6	0
2		3	0
2		1	1

0                      1

3. Write the digits in *remainder column* starting from *downwards to upwards*,

to Base	Number (Quotient)	Remainder
2	25	
2	12	1
2	6	0
2	3	0
2	1	1
	0	1



The binary equivalent of number  $(25)_{10}$  is  $(11001)_2$

The steps shown above are followed to convert a decimal integer to a number in any other base.

**Example 2:** Convert 23 from Base 10 to Base 2, 8, 16.

to Base	Number (Quotient)	Remainder	to Base	Number (Quotient)	Remainder	to Base	Number (Quotient)	Remainder
2	23		8	23		16	23	
2	11	1	8	2	7	16	1	7
2	5	1		0	2		0	1
2	2	1	The octal equivalent of $(23)_{10}$ is $(27)_8$			The hexadecimal equivalent of $(23)_{10}$ is $(17)_{16}$		
2	1	0						
	0	1						
The binary equivalent of $(23)_{10}$ is $(10111)_2$								

**Example 3:** Convert 147 from Base 10 to Base 2, 8 and 16.

to Base	Number (Quotient)	Remainder
2	147	
2	73	1
2	36	1
2	18	0
2	9	0
2	4	1
2	2	0
2	1	0
	0	1

The binary equivalent of  $(147)_{10}$  is  $(10010011)_2$

to Base	Number (Quotient)	Remainder
8	147	
8	18	3
8	2	2
	0	2

The octal equivalent of  $(147)_{10}$  is  $(223)_8$

to Base	Number (Quotient)	Remainder
16	147	
16	9	3
	0	9

The hexadecimal equivalent of  $(147)_{10}$  is  $(93)_{16}$

**Example 4:** Convert 94 from Base 10 to Base 2, 8 and 16.

to Base	Number (Quotient)	Remainder
2	94	
2	47	0
2	23	1
2	11	1
2	5	1
2	2	1
2	1	0
	0	1

The binary equivalent of  $(94)_{10}$  is  $(1011110)_2$

to Base	Number (Quotient)	Remainder
8	94	
8	11	6
8	1	3
	0	1

The octal equivalent of  $(94)_{10}$  is  $(136)_8$

to Base	Number (Quotient)	Remainder
16	94	
16	5	14
	0	5

The number 14 in hexadecimal is E.

The hexadecimal equivalent of  $(94)_{10}$  is  $(5E)_{16}$

### 5.3.2 Converting Decimal Fraction to Binary, Octal, Hexadecimal

A fractional number is a number less than 1. It may be .5, .00453, .564, etc. We use the multiplication operation to convert decimal fraction to any other base. To convert a decimal fraction to—

- binary-multiply by 2,
- Octal-multiply by 8, and, □
- hexadecimal-multiply by 16.

**Steps for conversion of a decimal fraction to any other base are—**

1. Multiply the fractional number with the to *Base*, to get a resulting number.
2. The resulting number has two parts, non-fractional part and fractional part.
3. Record the non-fractional part of the resulting number.
4. Repeat the above steps at least four times.
5. Write the digits in the non-fractional part starting from upwards to downwards.

**Example 5:** Convert 0.2345 from Base 10 to Base 2.

$$\begin{array}{r}
 0.2345 \\
 \times 2 \\
 \hline
 0.4690 \\
 .4690 \\
 \times 2 \\
 \hline
 0.9380 \\
 .9380 \\
 \times 2 \\
 \hline
 1.8760 \\
 .8760 \\
 \times 2 \\
 \hline
 1.7520 \\
 .7520 \\
 \times 2 \\
 \hline
 1.5040 \\
 .5040 \\
 \times 2 \\
 \hline
 1.0080
 \end{array}$$

The binary equivalent of  $(0.2345)_{10}$  is  $(0.001111)_2$

**Example 5a:** Convert 0.865 from Base 10 to Base 2, 8 and 16.

$$\begin{array}{r}
 0.865 \\
 \times 2 \\
 \hline
 1.730 \\
 \times 2 \\
 \hline
 1.460 \\
 \times 2 \\
 \hline
 0.920 \\
 \times 2 \\
 \hline
 1.840 \\
 \times 2 \\
 \hline
 1.680 \\
 \times 2 \\
 \hline
 1.360
 \end{array}$$

The binary equivalent of  $(.865)_{10}$  is  $(.110111)_2$

$$\begin{array}{r}
 0.865 \\
 \times 8 \\
 \hline
 6.920 \\
 \times 8 \\
 \hline
 7.360 \\
 \times 8 \\
 \hline
 2.880 \\
 \times 8 \\
 \hline
 7.040
 \end{array}$$

The octal equivalent of  $(0.865)_{10}$  is  $(.6727)_8$

$$\begin{array}{r}
 0.865 \\
 \times 16 \\
 \hline
 5190 \\
 865 \times \\
 \hline
 13.840 \\
 \times 16 \\
 \hline
 5040 \\
 840 \times \\
 \hline
 13.440 \\
 \times 16 \\
 \hline
 2640 \\
 440 \times \\
 \hline
 7.040
 \end{array}$$

The number 13 in hexadecimal is D.

The hexadecimal equivalent of  $(0.865)_{10}$  is  $(.DD7)_{16}$

**5.3.3 Converting Decimal Integer. Fraction to Binary, Octal, Hexadecimal** A decimal *integer. Fraction* number has both integer part and fraction part. The steps for conversion of a decimal *integer. Fraction* to any other base are—

1. Convert decimal integer part to the desired base following the steps shown in [section 5.3.1](#).
2. Convert decimal fraction part to the desired base following the steps shown in [section 5.3.2](#).
3. The integer and fraction part in the desired base is combined to get integer. Fraction.

**Example 6:** Convert 34.4674 from Base 10 to Base 2.

to Base	Number (Quotient)	Remainder
2	34	
2	17	0
2	8	1
2	4	0
2	2	0
2	1	0
	0	1

The binary equivalent of  $(34)_{10}$  is  $(100010)_2$

0.4674  
 $\times 2$   
 0.9348  
 $\times 2$   
 1.8696  
 $\times 2$   
 1.7392  
 $\times 2$   
 1.4784  
 $\times 2$   
 0.9568  
 $\times 2$   
 1.8136

The binary equivalent of  $(0.4674)_{10}$  is  $(.011101)_2$

---

The binary equivalent of  $(34.4674)_{10}$  is  $(100010.011101)_2$

---

**Example 7:** Convert 34.4674 from Base 10 to Base 8.

to Base	Number (Quotient)	Remainder
8	34	
8	4	2
	0	4

The octal equivalent of  $(34)_{10}$  is  $(42)_8$

0.4674  
 $\times 8$   
 3.7392  
 $\times 8$   
 5.9136  
 $\times 8$   
 7.3088  
 $\times 8$   
 2.4704

The octal equivalent of  $(0.4674)_{10}$  is  $(.3572)_8$

---

The octal equivalent of  $(34.4674)_{10}$  is  $(42.3572)_8$

---

**Example 8:** Convert 34.4674 from Base 10 to Base 16.

to Base	Number (Quotient)	Remainder
16	34	
16	4	2
	0	2

The hexadecimal equivalent of  $(34)_{10}$  is  $(22)_{16}$

```

0.4674
x 16
-----
28044
4674x
-----
9.4784
x 16
-----
28704
4784x
-----
7.6544
x 16
-----
39264
6544x
-----
10.4904
x 16
-----
29424
4904x
-----
7.8464

```

The hexadecimal equivalent of  $(0.4674)_{10}$  is  $(.97A7)_{16}$

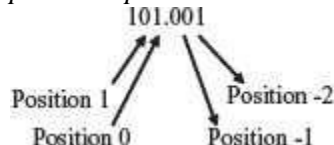
The hexadecimal equivalent of  $(34.4674)_{10}$  is  $(22.97A7)_{16}$

#### 5.4 CONVERSION OF BINARY, OCTAL, HEXADECIMAL TO DECIMAL

A binary, octal or hexadecimal number has two parts—integer part and fraction part. For example, a binary number could be 10011, 0.011001 or 10011.0111. The numbers 45, .362 or 245.362 are octal numbers. A hexadecimal number could be A2, .4C2 or A1.34.

The method used for the conversion of integer part and fraction part of binary, octal or hexadecimal number to decimal number is the same; multiplication operation is used for the conversion. The conversion mechanism uses the face value and position value of digits. The steps for conversion are as follows—

- Find the sum of the **Face Value \* (from Base)<sup>position</sup>** for each digit in the number.
  - In a non-fractional number, the rightmost digit has position 0 and the position increases as we go towards the left.*
  - In a fractional number, the first digit to the left of decimal point has position 0 and the position increases as we go towards the left. The first digit to the right of the decimal point has position -1 and it decreases as we go towards the right (-2, -3, etc.)*



**Example 9:** Convert 1011 from Base 2 to Base 10.

Convert 62 from Base 8 to Base 10.

Convert C15 from Base 16 to Base 10.

<p>1011 from Base 2 to Base 10</p> $1011 = 1*2^3 + 0*2^2 + 1*2^1 + 1*2^0$ $= 1*8 + 0*4 + 1*2 + 1*1$ $= 8 + 0 + 2 + 1$ $= 11$ <p>The decimal equivalent of <math>(1011)_2</math> is 11.</p>	<p>62 from Base 8 to Base 10</p> $62 = 6*8^1 + 2*8^0$ $= 6*8 + 2*1$ $= 48 + 2$ $= 50$ <p>The decimal equivalent of <math>(62)_8</math> is 50.</p>	<p>C15 from Base 16 to Base 10</p> $C15 = C*16^2 + 1*16^1 + 5*16^0$ $= 12*256 + 1*16 + 5*1$ $= 3072 + 16 + 5$ $= 3093$ <p>The decimal equivalent of <math>(C15)_{16}</math> is 3093</p>
--	---	---

**Example 10:** Convert .1101 from Base 2 to Base 10.

Convert .345 from Base 8 to Base 10. Convert

.15 from Base 16 to Base 10.

<p>.1101 from Base 2 to Base 10</p> $.1101 = 1*2^{-1} + 1*2^{-2} + 0*2^{-3} + 1*2^{-4}$ $= 1/2 + 1/4 + 0 + 1/16$ $= 13/16$ $= .8125$ <p>The decimal equivalent of <math>(.1101)_2</math> is .8125</p>	<p>.345 from Base 8 to Base 10</p> $.345 = 3*8^{-1} + 4*8^{-2} + 5*8^{-3}$ $= 3/8 + 4/64 + 5/512$ $= 229/512$ $= .447$ <p>The decimal equivalent of <math>(.345)_8</math> is .447</p>	<p>.15 from Base 16 to Base 10</p> $.15 = 1*16^{-1} + 5*16^{-2}$ $= 1/16 + 5/256$ $= 21/256$ $= .082$ <p>The decimal equivalent of <math>(.15)_{16}</math> is .082</p>
---	---	--

**Example 11:** Convert 1011.1001 from Base 2 to Base 10.

Convert 24.36 from Base 8 to Base 10.

Convert 4D.21 from Base 16 to Base 10.

1011.1001 from Base 2 to Base 10	24.36 from Base 8 to Base 10	4D.21 from Base 16 to Base 10
$  \begin{aligned}  1011.1001 &= 1*2^3 + 0*2^2 \\  &\quad + 1*2^1 + 1*2^0 \\  &\quad + 1*2^{-1} + 0*2^{-2} \\  &\quad + 0*2^{-3} + 1*2^{-4} \\  &= 8 + 0 + 2 + 1 + \\  &\quad 1/2 + 0 + 0 + 1/16 \\  &= 11 + 9/16 \\  &= 11.5625  \end{aligned}  $	$  \begin{aligned}  24.36 &= 2*8^1 + 4*8^0 + \\  &\quad 3*8^{-1} + 6*8^{-2} \\  &= 16 + 4 + 3/8 + 6/64 \\  &= 20 + 30/64 \\  &= 20.4687  \end{aligned}  $	$  \begin{aligned}  4D.21 &= 4*16^1 + D*16^0 + \\  &\quad 2*16^{-1} + 1*16^{-2} \\  &= 64 + 13 + 2/16 \\  &\quad + 1/256 \\  &= 77 + 33/256 \\  &= 77.1289  \end{aligned}  $
The decimal equivalent of $(1011.1001)_2$ is 11.5625	The decimal equivalent of $(24.36)_8$ is 20.4687	The decimal equivalent of $(4D.21)_{16}$ is 77.1289

## 5.5 CONVERSION OF BINARY TO OCTAL, HEXADECIMAL

A binary number can be converted into octal or hexadecimal number using a shortcut method. The shortcut method is based on the following information—

- An octal digit from 0 to 7 can be represented as a combination of 3 bits, since  $2^3 = 8$ .
- A hexadecimal digit from 0 to 15 can be represented as a combination of 4 bits, since  $2^4 = 16$ .

*The Steps for Binary to Octal Conversion are—*

1. Partition the binary number in groups of three bits, starting from the right-most side.
2. For each group of three bits, find its octal number.
3. The result is the number formed by the combination of the octal numbers.

*The Steps for Binary to Hexadecimal Conversion are—*

1. Partition the binary number in groups of four bits, starting from the right-most side.
2. For each group of four bits, find its hexadecimal number.
3. The result is the number formed by the combination of the hexadecimal numbers.

**Example 12:** Convert the binary number 1110101100110 to octal.

Given binary number

1110101100110

1. Partition binary number in groups of three bits, starting from the right-most side.

1    110    101    100    110

- For each group find its octal number.

1	1 1 0	1 0 1	1 0 0	1 1 0
1	6	5	4	6

- The octal number is 16546.

**Example 13:** Convert the binary number 1110101100110 to hexadecimal

Given binary number

1110101100110

- Partition binary number in groups of four bits, starting from the right-most side.

1	1 1 0 1	0 1 1 0	0 1 1 0
---	---------	---------	---------

- For each group find its hexadecimal number.

1	1 1 0 1	0 1 1 0	0 1 1 0
1	D	6	6

- The hexadecimal number is 1D66.

## 5.6 CONVERSION OF OCTAL, HEXADECIMAL TO BINARY

The conversion of a number from octal and hexadecimal to binary uses the inverse of the steps defined for the conversion of binary to octal and hexadecimal.

*The Steps for Octal to Binary Conversion are—*

- Convert each octal number into a three-digit binary number.
- The result is the number formed by the combination of all the bits.

*The Steps for Hexadecimal to Binary Conversion are—*

- Convert each hexadecimal number into a four-digit binary number.
- The result is the number formed by the combination of all the bits.

**Example 14:** Convert the hexadecimal number 2BA3 to binary.

- Given number is 2BA3
- Convert each hexadecimal digit into four digit binary number.

2	B	A	3
0010	1011	1010	0011

- Combine all the bits to get the result 0010101110100011.

**Example 15:** Convert the octal number 473 to binary.

1. Given number is 473
2. Convert each octal digit into three digit binary number.

4	7	3
100	111	011

3. Combine all the bits to get the result 100111011.

## 5.7 BINARY ARITHMETIC

The arithmetic operations—addition, subtraction, multiplication and division, performed on the binary numbers is called *binary arithmetic*. In computer, the basic arithmetic operations performed on the binary numbers is—

□ Binary addition, and □  
Binary subtraction.

In the following subsections, we discuss the binary addition and the binary subtraction operations.

### 5.7.1 Binary Addition

Binary addition involves addition of two or more binary numbers. The *binary addition* rules are used while performing the binary addition. [Table 5.3](#) shows the binary addition rules.

Input 1	Input 2		Sum	Carry
0	0	→	0	No carry
0	1	→	1	No carry
1	0	→	1	No carry
1	1	→	0	1

**Table 5.3** Binary addition rules

Binary addition of three inputs follows the rule shown in [Table 5.4](#).

Input 1	Input 2	Input 3		Sum	Carry
0	0	0	→	0	No Carry
0	0	1	→	1	No Carry
0	1	0	→	1	No Carry
0	1	1	→	0	1
1	0	0	→	1	No Carry
1	0	1	→	0	1
1	1	0	→	0	1
1	1	1	→	1	1

**Table 5.4** Binary addition of three inputs

Addition of the binary numbers involves the following steps—

1. Start addition by adding the bits in unit column (the right-most column). Use the rules of binary addition.
2. The result of adding bits of a column is a sum with or without a carry.
3. Write the sum in the result of that column.
4. If a carry is present, the carry is carried-over to the addition of the next left column.
5. Repeat steps 2–4 for each column, i.e., the tens column, hundreds column and so on.

Let us now understand binary addition with the help of some examples.

**Example 1:** Add 10 and 01. Verify the answer with the help of decimal addition.

When we add 0 and 1 in the unit column, sum is 1 and there is no carry. The sum 1 is written in the unit column of the result. In the tens column, we add 1 and 0 to get the sum 1. There is no carry. The sum 1 is written in the tens column of the result.

Binary Addition	Decimal Addition
$\begin{array}{r} 10 \\ + 01 \\ \hline \text{Result } 11 \end{array}$	$\begin{array}{r} 2 \\ + 1 \\ \hline 3 \end{array}$
$11_2 = 3_{10}$	

**Example 2:** Add 01 and 11. Verify the answer with the help of decimal addition.

When we add 1 and 1 in the unit column, sum is 0 and carry is 1. The sum 0 is written in the unit column of the result. The carry is carried-over to the next column, i.e., the tens column. In the tens column, we add 0, 1 and the carried-over 1, to get sum 0 and carry 1. The sum 0 is written in the tens column of the result. The carry 1 is carried-over to the hundreds column. In the hundreds column, the result is 1.

Binary Addition	Decimal Addition
$  \begin{array}{r}  11 \leftarrow \text{Carry} \\  01 \\  + 11 \\  \hline  \text{Result } 100  \end{array}  $	$  \begin{array}{r}  1 \\  + 3 \\  \hline  4  \end{array}  $
$100_2 = 4_{10}$	

**Example 3:** Add 11 and 11. Verify the answer with the help of decimal addition.

Binary Addition	Decimal Addition
$  \begin{array}{r}  11 \leftarrow \text{Carry} \\  11 \\  + 11 \\  \hline  \text{Result } 110  \end{array}  $	$  \begin{array}{r}  3 \\  + 3 \\  \hline  6  \end{array}  $
$110_2 = 6_{10}$	

**Example 4:** Add 1101 and 1111. Verify the answer with the help of decimal addition.

Binary Addition	Decimal Addition
$  \begin{array}{r}  1111 \leftarrow \text{Carry} \\  1001 \\  + 1111 \\  \hline  11000  \end{array}  $	$  \begin{array}{r}  9 \\  + 15 \\  \hline  24  \end{array}  $
$11000_2 = 24_{10}$	

**Example 5:** Add 10111, 11100 and 11. Verify the answer with the help of decimal addition.

Binary Addition	Decimal Addition
$  \begin{array}{r}  11111 \leftarrow \text{Carry} \\  10111 \\  + 11000 \\  \hline  111 \\  \hline  110110  \end{array}  $	$  \begin{array}{r}  23 \\  + 24 \\  \hline  7 \\  \hline  54  \end{array}  $
$110110_2 = 54_{10}$	

## 5.7.2 Binary Subtraction

Binary subtraction involves subtracting of two binary numbers. The *binary subtraction rules* are used while performing the binary subtraction. The binary subtraction rules are shown in [Table 5.5](#), where “Input 2” is subtracted from “Input 1.”

Input 1	Input 2		Difference	Borrow
0	0	→	0	No borrow
0	1	→	1	1
1	0	→	1	No borrow
1	1	→	0	No borrow

**Table 5.5** Binary subtraction rules

The steps for performing subtraction of the binary numbers are as follows—

1. Start subtraction by subtracting the bit in the lower row from the upper row, in the unit column.
2. Use the binary subtraction rules. If the bit in the upper row is less than lower row, *borrow 1* from the upper row of the next column (on the left side). The result of subtracting two bits is the *difference*.
3. Write the *difference* in the result of that column.
4. Repeat steps 2 and 3 for each column, i.e., the tens column, hundreds column and so on.

Let us now understand binary subtraction with the help of some examples.

**Example 1:** Subtract 01 from 11. Verify the answer with the help of decimal subtraction.

When we subtract 1 from 1 in the unit column, the difference is 0. Write the difference in the unit column of the result. In the tens column, subtract 0 from 1 to get the difference 1. Write the difference in the tens column of the result.

Binary Subtraction	Decimal Subtraction
$\begin{array}{r} 11 \\ - 01 \\ \hline \text{Result } 10 \end{array}$	$\begin{array}{r} 3 \\ - 1 \\ \hline 2 \end{array}$
$10_2 = 2_{10}$	

**Example 2:** Subtract 01 from 10. Verify the answer with the help of decimal subtraction.

When we subtract 1 from 0 in the unit column, we have to borrow 1 from the left column since 0 is less than 1. After borrowing from the left column, 0 in the unit column becomes 10, and, 1 in the left column becomes 0. We perform 10-1 to get the difference 1. We write the difference in the unit column of the result. In the tens column, subtract 0 from 0 to get the difference 0. We write the difference 0 in the tens column of the result.

Binary Subtraction	Decimal Subtraction
$  \begin{array}{r}  010 \\  + 0 \\  - 01 \\  \hline  01  \end{array}  $	$  \begin{array}{r}  2 \\  - 1 \\  \hline  1  \end{array}  $
$01_2 = 1_{10}$	

**Example 3:** Subtract 0111 from 1110. Verify the answer with the help of decimal subtraction.

When we do 0-1 in the unit column, we have to borrow 1 from the left column since 0 is less than 1. After borrowing from the left column, 0 in the unit column becomes 10, and, 1 in the left column becomes 0. We perform 10-1 to get the difference 1. We write the difference in the unit column of the result. In the tens column, when we do 0-1, we again borrow 1 from the left column. We perform 10-1 to get the difference 1. We write the difference in the tens column of the result. In the hundreds column, when we do 0-1, we again borrow 1 from the left column. We perform 10-1 to get the difference 1. We write the difference in the hundreds column of the result. In the thousands column, 0-0 is 0. We write the difference 0 in the thousands column of the result.

Binary Subtraction	Decimal Subtraction
$  \begin{array}{r}  010 \\  010 \\  010 \\  \left. \begin{array}{l} \\ \\ \end{array} \right\} \text{Borrow} \\  + + + 0 \\  - 0111 \\  \hline  0111  \end{array}  $	$  \begin{array}{r}  14 \\  - 07 \\  \hline  7  \end{array}  $
$0111_2 = 7_{10}$	

**Example 4:** Subtract 100110 from 110001. Verify the answer with the help of decimal subtraction.

Binary Subtraction	Decimal Subtraction
$  \begin{array}{r}  11 \\  + 1010 \\  10001 \\  100110 \\  \left. \begin{array}{l} \\ \\ \end{array} \right\} \text{Borrow} \\  \hline  001011  \end{array}  $	$  \begin{array}{r}  49 \\  - 38 \\  \hline  11  \end{array}  $
$001011_2 = 11_{10}$	

## 5.8 SIGNED AND UNSIGNED NUMBERS

A binary number may be positive or negative. Generally, we use the symbol “+” and “-” to represent positive and negative numbers, respectively. The sign of a binary number has to be represented using 0 and 1, in the computer. An *n-bit signed binary number* consists of two parts—sign bit and magnitude. The left most bit, also called the Most Significant Bit (MSB) is the sign bit. The remaining n-1 bits denote the *magnitude* of the number.

In signed binary numbers, the sign bit is 0 for a positive number and 1 for a negative number. For example, 01100011 is a positive number since its sign bit is 0, and, 11001011 is a negative number since its sign bit is 1. An 8-bit signed number can represent data in the range  $-128$  to  $+127$  ( $-2^7$  to  $+2^7-1$ ). The left-most bit is the sign bit.



In an  $n$ -bit *unsigned binary number*, the magnitude of the number  $n$  is stored in  $n$  bits. An 8-bit unsigned number can represent data in the range 0 to 255 ( $2^8 = 256$ ).

### 5.8.1 Complement of Binary Numbers

Complements are used in computer for the simplification of the subtraction operation. For any number in base  $r$ , there exist two complements—(1)  $r$ 's complement and (2)  $r-1$ 's complement.

Number System	Base	Complements possible
Binary	2	1's complement and 2's complement
Octal	8	7's complement and 8's complement
Decimal	10	9's complement and 10's complement
Hexadecimal	16	15's complement and 16's complement

Let us now see how to find the complement of a binary number. There are two types of complements for the binary number system—1's complement and 2's complement.

- **1's Complement of Binary Number** is computed by changing the bits 1 to 0 and the bits 0 to 1. For example,

1's complement of 101 is 010

1's complement of 1011 is 0100

1's complement of 1101100 is 0010011

- **2's Complement of Binary Number** is computed by adding 1 to the 1's complement of the binary number. For example, 2's complement of 101 is  $010 + 1 = 011$

2's complement of 1011 is  $0100 + 1 = 0101$

2's complement of 1101100 is  $0010011 + 1 = 0010100$

The rule to find the complement of any number  $N$  in base  $r$  having  $n$  digits is

$(r-1)$ 's complement— $(r^n - 1) - N$

$(r)$ 's complement— $(r^n - 1) - N + 1 = (r^n - N)$

## 5.9 BINARY DATA REPRESENTATION

A binary number may also have a binary point, in addition to the sign. The binary point is used for representing fractions, integers and integer-fraction numbers. *Registers* are high-speed storage areas within the Central Processing Unit (CPU) of the computer. All data are brought into a register before it can be processed. For example, if two numbers are to be added, both the numbers are brought in registers, added, and the result is also placed in a register. There are two ways of representing the position of the binary point in the register—fixed point number representation and floating point number representation.

The *fixed point number representation* assumes that the binary point is fixed at one position either at the extreme left to make the number a fraction, or at the extreme right to make the number an integer. In both cases, the binary point is not stored in the register, but the number is treated as a fraction or integer. For example, if the binary point is assumed to be at extreme left, the number 1100 is actually treated as 0.1100.

The *floating point number representation* uses two registers. The first register stores the number without the binary point. The second register stores a number that indicates the position of the binary point in the first register.

We shall now discuss representation of data in the fixed point number representation and floating point number representation.

### 5.9.1 Fixed Point Number Representation

The integer binary signed number is represented as follows—

- For a positive integer binary number, the sign bit is 0 and the magnitude is a positive binary number.
- For a negative integer binary number, the sign bit is 1. The magnitude is represented in any one of the three ways—
  - **Signed Magnitude Representation**—The magnitude is the positive binary number itself.
  - **Signed 1's Complement Representation**—The magnitude is the 1's complement of the positive binary number.
  - **Signed 2's Complement Representation**—The magnitude is the 2's complement of the positive binary number.

[Table 5.6](#) shows the representation of the signed number 18.

+18	0 0010010	Sign bit is 0. 0010010 is binary equivalent of +18
	Signed magnitude representation	1 0010010
		Sign bit is 1. 0010010 is binary equivalent of +18
-18	Signed 1's complement representation	1 1101101
		Sign bit is 1. 1101101 is 1's complement of +18
	Signed 2's complement representation	1 1101110
		Sign bit is 1. 1101110 is 2's complement of +18

**Table 5.6** Fixed point representation of the signed number 18

Signed magnitude and signed 1's complement representation are seldom used in computer arithmetic.

Let us now perform arithmetic operations on the signed binary numbers. We use the signed's complement representation to represent the negative numbers.

- **Addition of Signed Binary Numbers**—The addition of any two signed binary numbers is performed as follows—
  - Represent the positive number in binary form. (For example, +5 is 0000 0101 and +10 is 0000 1010)
  - Represent the negative number in's complement form. (For example, —5 is 1111 1011 and —10 is 1111 0110)
  - Add the bits of the two signed binary numbers.
  - Ignore any carry out from the sign bit position.

Please note that the negative output is automatically in the's complement form.

We get the decimal equivalent of the negative output number, by finding its 2's complement, and attaching a negative sign to the obtained result.

Let us understand the addition of two signed binary numbers with the help of some examples.

**Example 1:** Add +5 and +10.

+5 in binary form, i.e., 0000 0101. +10 in binary form, i.e., 0000 1010.

Binary Addition	Decimal Addition
00000101	+ 5
00001010	+ 10
00001111	+ 15
The result is 0000 1111 <sub>2</sub> i.e., +15 <sub>10</sub>	

**Example 2:** Add –5 and +10.

–5 in 's complement form is 1111 1011. +10 in binary form is 0000 1010.

Binary Addition	Decimal Addition
11111011	– 5
00001010	+ 10
00000101	+ 5
The result is 0000 0101 <sub>2</sub> i.e., +5 <sub>10</sub>	

**Example 3:** Add +5 and –10.

+5 in binary form is 0000 0101. –10 in's complement form is 1111 0110. 1111 1011.

Binary Addition	Decimal Addition
00000101	+ 5
11110110	- 10
11111011	- 5
The result is 1111 1011 <sub>2</sub> , i.e., -5 <sub>10</sub>	

The result is in 2's complement form. To find its decimal equivalent—

Find the 2's complement of 1111 1011, i.e., 0000 0100 + 1 = 0000 0101. This is binary equivalent of + 5. Attaching a negative sign to the obtained result gives us -5.

**Example 4:** Add -5 and -10.

-5 in 's complement form is 1111 1011. -10 in 2's complement form is 1111 0110.

Binary Addition	Decimal Addition
11111011	- 5
11110110	- 10
11110001	- 15
The result is 1111 0001 <sub>2</sub> , i.e., -15 <sub>10</sub>	

The result is in 2's complement form. To find its decimal equivalent—

Find the 's complement of 1111 0001, i.e., 0000 1110 + 1 = 0000 1111. This is binary equivalent of +15. Attaching a negative sign to the obtained result gives us -15.

- **Subtraction of Signed Binary Numbers**—The subtraction of signed binary numbers is changed to the addition of two signed numbers. For this, the sign of the second number is changed before performing the addition operation.

$$(-A) - (+B) = (-A) + (-B) \quad (+B \text{ in subtraction is changed to } -B \text{ in addition})$$

$$(+A) - (+B) = (+A) + (-B) \quad (+B \text{ in subtraction is changed to } -B \text{ in addition})$$

$$(-A) - (-B) = (-A) + (+B) \quad (-B \text{ in subtraction is changed to } +B \text{ in addition})$$

$$(+A) - (-B) = (+A) + (+B) \quad (-B \text{ in subtraction is changed to } +B \text{ in addition})$$

We see that the subtraction of signed binary numbers is performed using the addition operation. The

hardware logic for the fixed point number representation is simple, when we use 's

Complement for addition and subtraction of the signed binary numbers. When two large numbers having the same sign are added, then an overflow may occur, which has to be handled.

## 5.9.2 Floating Point Number Representation

The floating point representation of a number has two parts—mantissa and exponent. The mantissa is a signed fixed point number. The exponent shows the position of the binary point in the mantissa.

For example, the binary number +11001.11 with an 8-bit mantissa and 6-bit exponent is represented as follows—

- Mantissa is 01100111. The left most 0 indicates that the number is positive.
- Exponent is 000101. This is the binary equivalent of decimal number + 5.
- □ The floating point number is Mantissa  $\times 2^{\text{exponent}}$ , i.e.,  $+ (.1100111) \times 2^{+5}$ .

The arithmetic operation with the floating point numbers is complicated, and uses complex hardware as compared to the fixed point representation. However, floating point calculations are required in scientific calculations, so, computers have a built-in hardware for performing floating point arithmetic operations.

## 5.10 BINARY CODING SCHEMES

The alphabetic data, numeric data, alphanumeric data, symbols, sound data and video data, are represented as combination of bits in the computer. The bits are grouped in a fixed size, such as 8 bits, 6 bits or 4 bits. A code is made by combining bits of definite size. *Binary Coding schemes* represent the data such as alphabets, digits 0–9, and symbols in a standard code. A combination of bits represents a unique symbol in the data. The standard code enables any programmer to use the same combination of bits to represent a symbol in the data.

The binary coding schemes that are most commonly used are—

- Extended Binary Coded Decimal Interchange Code (EBCDIC),
- American Standard Code for Information Interchange (ASCII), and
- Unicode

In the following subsections, we discuss the EBCDIC, ASCII and Unicode coding schemes.

### 5.10.1 EBCDIC

- The Extended Binary Coded Decimal Interchange Code (EBCDIC) uses 8 bits (4 bits for zone, 4 bits for digit) to represent a symbol in the data.
- EBCDIC allows  $2^8 = 256$  combinations of bits.
- 256 unique symbols are represented using EBCDIC code. It represents decimal numbers (0–9), lower case letters (a–z), uppercase letters (A–Z), Special characters, and Control characters (printable and non-printable, e.g., for cursor movement, printer vertical spacing, etc.).
- EBCDIC codes are mainly used in the mainframe computers.

### 5.10.2 ASCII

- The American Standard Code for Information Interchange (ASCII) is widely used in computers of all types.
- ASCII codes are of two types—ASCII–7 and ASCII–8.

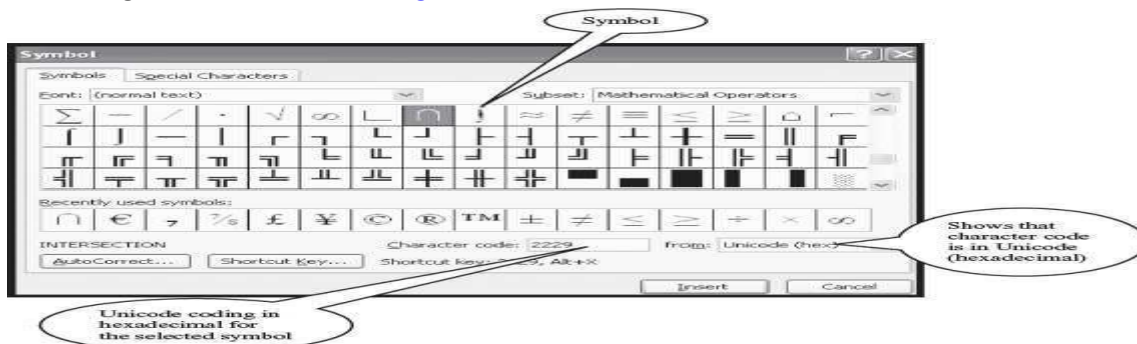
- **ASCII-7** is a 7-bit standard ASCII code. In ASCII-7, the first 3 bits are the zone bits and the next 4 bits are for the digits. ASCII-7 allows  $2^7 = 128$  combinations. 128 unique symbols are represented using ASCII-7. ASCII-7 has been modified by IBM to ASCII-8.
- **ASCII-8** is an extended version of ASCII-7. ASCII-8 is an 8-bit code having 4 bits for zone and 4 bits for the digit. ASCII-8 allows  $2^8 = 256$  combinations. ASCII-8 represents 256 unique symbols. ASCII is used widely to represent data in computers. □ The ASCII-8 code represents 256 symbols.
  - Codes 0 to 31 represent control characters (non-printable), because they are used for actions like, Carriage return (CR), Bell (BEL), etc.
  - Codes 48 to 57 stand for numeric 0–9.
  - Codes 65 to 90 stand for uppercase letters A–Z.
  - Codes 97 to 122 stand for lowercase letters a–z.
  - Codes 128 to 255 are the extended ASCII codes.

**5.10.3 Unicode** □ Unicode is a universal character encoding standard for the representation of text which includes letters, numbers and symbols in multi-lingual environments. The Unicode Consortium based in California developed the Unicode standard.

- Unicode uses 32 bits to represent a symbol in the data.
- Unicode allows  $2^{32} = 4164895296$  (~ 4 billion) combinations.
- Unicode can uniquely represent any character or symbol present in any language like Chinese, Japanese, etc. In addition to the letters; mathematical and scientific symbols are also represented in Unicode codes.
- An advantage of Unicode is that it is compatible with the ASCII-8 codes. The first 256 codes in Unicode are identical to the ASCII-8 codes.
- Unicode is implemented by different character encodings. UTF-8 is the most commonly used encoding scheme. UTF stands for Unicode Transformation Format. UTF-8 uses 8 bits to 32 bits per code.

If you wish to see the Unicode character encoding in MS-Word 2007, do as follows—

<Insert> <Symbol>. A Symbol dialog box will appear which displays the symbols, and the character codes in a coding scheme, as shown in [Figure 5.1](#).



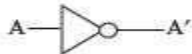


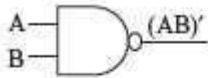
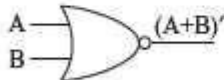


**Figure 5.1** Unicode coding

## 5.11 LOGIC GATES

The information is represented in the computer in binary form. Binary information is represented using signals in two states *off* or *on* which correspond to 0 or 1, respectively. The manipulation of the binary information is done using logic gates. Logic gates are the hardware electronic circuits which operate on the input signals to produce the output signals. Each logic gate has a unique symbol and its operation is described using algebraic expression. For each gate, the truth table shows the output that will be outputted for the different possible combinations of the input signal. The AND, OR and NOT are the basic logic gates. Some of the basic combination of gates that are widely used are—NAND, NOR, XOR and XNOR.

[Table 5.7](#) shows the different logic gates, their symbols, their algebraic function and the truth table for each logic gate. The comments list the features of each logic gate.

Operation	Symbol	Algebraic Function	Comments	Truth Table															
AND		$X = A.B$ or $X = AB$	<ul style="list-style-type: none"><li>Two or more binary inputs</li><li>The output is 1 if all the inputs are 1, otherwise the output is 0.</li><li>Represented using a multiplication symbol "."</li></ul>	<table><tr><th>A</th><th>B</th><th>A.B</th></tr><tr><td>0</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>0</td></tr><tr><td>1</td><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td><td>1</td></tr></table>	A	B	A.B	0	0	0	0	1	0	1	0	0	1	1	1
A	B	A.B																	
0	0	0																	
0	1	0																	
1	0	0																	
1	1	1																	
OR		$X = A + B$	<ul style="list-style-type: none"><li>Two or more binary inputs</li><li>The output is 1 if at least one input is 1, otherwise the output is 0.</li><li>Represented using a "+"</li></ul>	<table><tr><th>A</th><th>B</th><th>A+B</th></tr><tr><td>0</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>1</td></tr><tr><td>1</td><td>0</td><td>1</td></tr><tr><td>1</td><td>1</td><td>1</td></tr></table>	A	B	A+B	0	0	0	0	1	1	1	0	1	1	1	1
A	B	A+B																	
0	0	0																	
0	1	1																	
1	0	1																	
1	1	1																	
NOT		$A = A'$	<ul style="list-style-type: none"><li>One binary input</li><li>The output is complement (opposite) of input. If input is 1 output is 0 and if input is 0 output is 1.</li><li>Represented using a "'"</li></ul>	<table><tr><th>A</th><th>A'</th></tr><tr><td>0</td><td>1</td></tr><tr><td>1</td><td>0</td></tr></table>	A	A'	0	1	1	0									
A	A'																		
0	1																		
1	0																		

Operation	Symbol	Algebraic Function	Comments	Truth Table															
NAND		$X = (AB)'$	<ul style="list-style-type: none"><li>Two or more binary inputs</li><li>NAND is complement of AND</li></ul>	<table><tr><th>A</th><th>B</th><th><math>(A.B)'</math></th></tr><tr><td>0</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1</td><td>1</td></tr><tr><td>1</td><td>0</td><td>1</td></tr><tr><td>1</td><td>1</td><td>0</td></tr></table>	A	B	$(A.B)'$	0	0	1	0	1	1	1	0	1	1	1	0
A	B	$(A.B)'$																	
0	0	1																	
0	1	1																	
1	0	1																	
1	1	0																	
NOR		$X = (A + B)'$	<ul style="list-style-type: none"><li>Two or more binary inputs</li><li>NOR is complement of OR.</li></ul>	<table><tr><th>A</th><th>B</th><th><math>(A+B)'</math></th></tr><tr><td>0</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1</td><td>0</td></tr><tr><td>1</td><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td><td>0</td></tr></table>	A	B	$(A+B)'$	0	0	1	0	1	0	1	0	0	1	1	0
A	B	$(A+B)'$																	
0	0	1																	
0	1	0																	
1	0	0																	
1	1	0																	
XOR		$X = (A \oplus B)$	<ul style="list-style-type: none"><li>Two or more binary inputs</li><li>The output is 1 if the odd number of inputs is 1.</li><li>Represented using a " <math>\oplus</math> "</li></ul>	<table><tr><th>A</th><th>B</th><th><math>(A \oplus B)</math></th></tr><tr><td>0</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>1</td></tr><tr><td>1</td><td>0</td><td>1</td></tr><tr><td>1</td><td>1</td><td>0</td></tr></table>	A	B	$(A \oplus B)$	0	0	0	0	1	1	1	0	1	1	1	0
A	B	$(A \oplus B)$																	
0	0	0																	
0	1	1																	
1	0	1																	
1	1	0																	
XNOR		$X = (A \oplus B)'$	<ul style="list-style-type: none"><li>Two or more binary inputs</li><li>XNOR is complement of XOR.</li></ul>	<table><tr><th>A</th><th>B</th><th><math>(A \oplus B)'</math></th></tr><tr><td>0</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1</td><td>0</td></tr><tr><td>1</td><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td><td>1</td></tr></table>	A	B	$(A \oplus B)'$	0	0	1	0	1	0	1	0	0	1	1	1
A	B	$(A \oplus B)'$																	
0	0	1																	
0	1	0																	
1	0	0																	
1	1	1																	

**Table 5.7** Logic gates

## 9. DATA COMMUNICATION AND COMPUTER NETWORK

### 9.1 INTRODUCTION

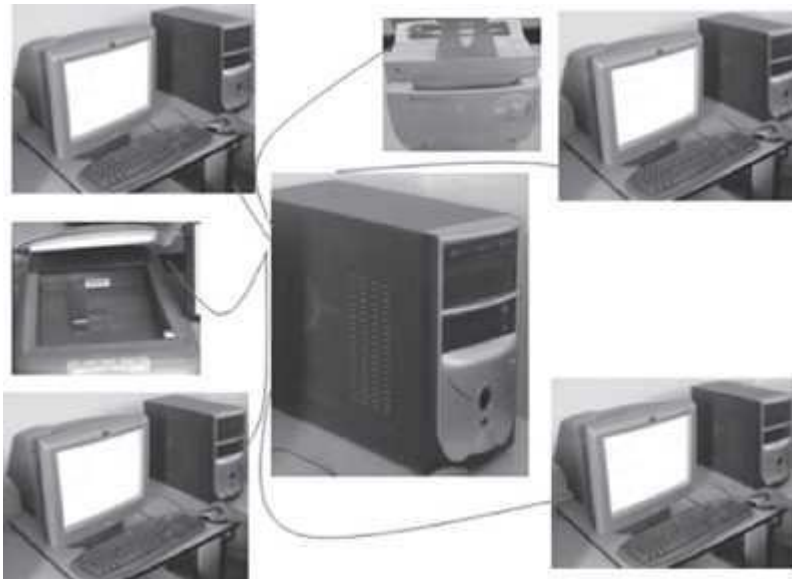
The communication process involves—sender of information, receiver of information, language used for communication, and medium used to establish the communication. Communication between computers also follows a similar process.

This chapter discusses the data communication and the computer networks. The section on data communication discusses the media used for transmission of data, how data can be transferred across the communication media and the relationship between data transmission and data networking. The section on computer network discusses different network types, network topologies, communication protocol and network communicating devices. A brief explanation of wireless networks is also provided.

### 9.2 IMPORTANCE OF NETWORKING

Networking of computers provides a communication link between the users, and provides access to information. Networking of computers has several uses, described as follows:

- **Resource Sharing**—In an organization, resources such as printers, fax machines and scanners are generally not required by each person at all times. Moreover, for small organizations it may not be feasible to provide such resources to each individual. Such resources can be made available to different users of the organization on the network. It results in availability of the resource to different users regardless of the physical location of the resource or the user, enhances optimal use of the resource, leads to easy maintenance, and saves cost too ([Figure 9.1](#)).



**Figure 9.1** A network of computers, printer and scanner

- **Sharing of Information**—In addition to the sharing of physical resources, networking facilitates sharing of information. Information stored on networked computers located at same or different physical locations, becomes accessible to the computers connected to the network.
- **As a Communication Medium**—Networking helps in sending and receiving of electronic-mail (email) messages from anywhere in the world. Data in the form of text, audio, video and pictures can be sent via e-mail. This allows the users to communicate online in a faster and cost effective manner. Video conferencing is another form of communication made possible via networking. People in distant locations can hold a meeting, and they can hear and see each other simultaneously.
- **For Back-up and Support**—Networked computers can be used to take back-up of critical data. In situations where there is a requirement of always-on computer, another computer on the network can take over in case of failure of one computer.

### 9.3 DATA TRANSMISSION MEDIA

The data is sent from one computer to another over a transmission medium. The transmission media can be grouped into guided media, and unguided media.

In the *guided media*, the data signals are sent along a specific path, through a wire or a cable. Copper wire and optical fibers are the most commonly used guided media. Copper wire transmits data as electric signals. Copper wires offer low resistance to current signal, facilitating signals to travel longer distances. To minimize the effect of external disturbance on the copper wire, two types of wiring is used—(1) Twisted Pair, and (2) Coaxial Pair. Optical fibers transmit data as light signals.

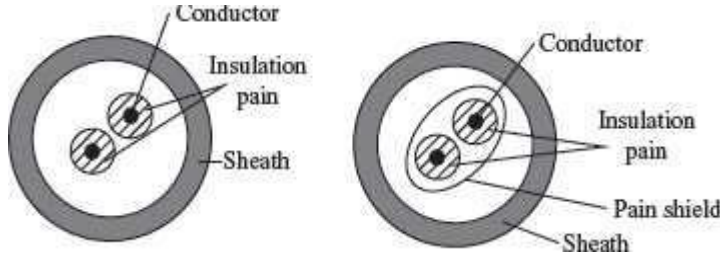
In the *unguided media*, the data signals are not bounded by a fixed channel to follow. The data signals are transmitted by air. Radio, microwave, and satellite transmissions fall into this category.

Now let's discuss both the guided and the unguided data transmission media.

#### 9.3.1 Twisted Pair

- A twisted pair cable consists of four pairs of copper wires coated with an insulating material like plastic or Teflon, twisted together. The twisting of wires reduces electromagnetic interference from external sources.
- Twisted pair cabling is often used in data networks for short and medium length connections because of its relatively lower costs compared to optical fiber and coaxial cable.
- Twisted pair is of two kinds—Shielded Twisted Pair (STP), and Unshielded Twisted Pair (UTP).
- *STP* cable has an extra layer of metal foil between the twisted pair of copper wires and the outer covering. The metal foil covering provides additional protection from external disturbances. However, the covering increases the resistance to the signal and thus decreases the length of the cable. STP is costly and is generally used in networks where cables pass closer to devices that cause external disturbances.
- *UTP* is the most commonly used medium for transmission over short distances up to 100m. Out of the four pairs of wires in a UTP cable, only two pairs are used for communication. [Figure 9.2](#) shows the cross-section of STP and UTP cables.

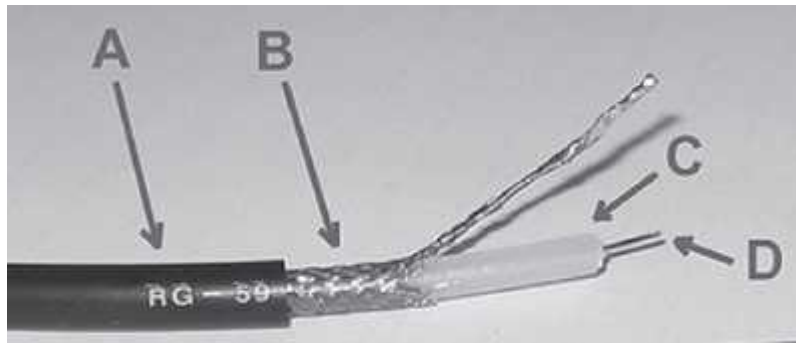
- UTP cables are defined in different categories. The commonly used UTP cable is the Cat-5 *cable* which is used with fast Ethernet.



**Figure 9.2** Cross section of (a) UTP (b) STP

### 9.3.2 Coaxial Cable

- A coaxial cable ([Figure 9.3](#)) has a single inner conductor that transmits electric signals; the outer conductor acts as a ground. The two conductors are separated by insulation. The inner conductor, insulator, and the outer conductor are wrapped in a sheath of Teflon or PVC.



**Figure 9.3** Coaxial cable (A: outer plastic sheath, B: woven copper shield, C: inner dielectric insulator, D: copper core)

- The copper wire is used for both inner and outer conductor. The signal is transmitted over the surface of the inner conductor.
- In an ideal coaxial cable the electromagnetic field carrying the signal exists only in the space between the inner and outer conductors. This allows coaxial cable runs to be installed next to metal objects such as gutters without the power losses that occur in other transmission lines, and provides protection of the signal from external electromagnetic interference.
- A thicker coaxial cable can transmit more data than a thinner one.
- The commonly used coaxial cable is *10 base 2* that transmits over a distance of 185 m, and *10 base 5* that transmits over a distance of 500 m.

### 9.3.3 Optical Fiber

- Optical fibers are being used for transmission of information over large distances more costeffectively than the copper wire connection. Communication systems are now unthinkable without fiber optics.
- Optical fiber transmits data as light signals instead of electric signals.

- An optical fiber cable ([Figure 9.4](#)) consists of (1) core—optical fiber conductor (glass) that transmits light, (2) cladding—an optical material that surrounds the core to prevent any light from escaping the core, and (3) jacket—outer covering made of plastic to protect the fiber from damage.



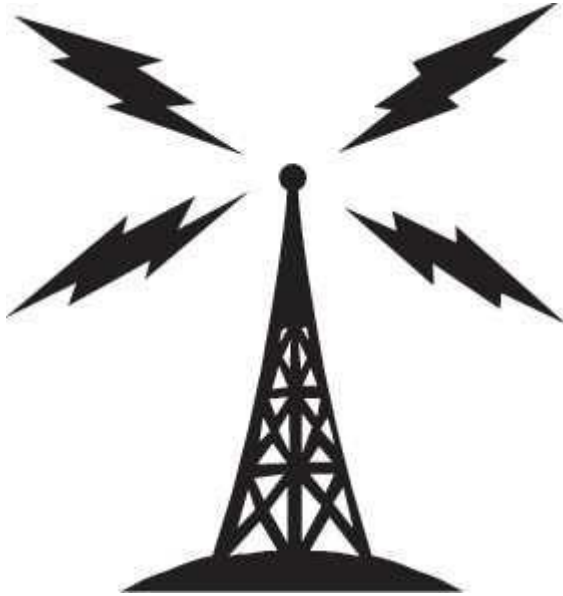
**Figure 9.4** (a) Optical fiber (b) Cross section of optical fiber

- Modern optical fiber cables can contain up to a thousand fibers in a single cable, so the performance of optical networks easily accommodate large demands for bandwidth on a point-to-point basis.
- Optical fibers come in two types: (a) Single-mode fibers, and (b) Multi-mode fibers
  - Single-mode fibers have small cores (about  $3.5 \times 10^{-4}$  inches or 9 microns in diameter) and transmit infrared laser light (wavelength = 1,300 to 1,550 nanometers).
  - Multi-mode fibers have larger cores (about  $2.5 \times 10^{-3}$  inches or 62.5 microns in diameter) and transmit infrared light (wavelength 850 to 1,300 nm) from Light Emitting Diodes (LEDs).
- The *Advantages of Optical Fibers* over wires are:
  - Optical fibers do not cause electrical interference in other cables, since they use light signals.
  - Due to much lower attenuation and interference, optical fiber has large advantages over existing copper wire in long-distance and high-demand applications.
  - A fiber can carry a pulse of light much farther than a copper wire carrying a signal.
  - Optical fiber can carry more information than a wire (light can encode more information than electrical signal).
  - A single optical fiber is required for light to travel from one computer to another (two wires are required for electric connection).
  - Because signals in optical fibers degrade less, lower-power transmitters can be used instead of the high-voltage electrical transmitters needed for copper wires. Again, this saves your provider and you, money.
  - No amplification of the optical signal is needed over distances of hundreds of kilometers. This has greatly reduced the cost of optical networking, particularly over undersea spans where the cost reliability of amplifiers is one of the key factors determining the performance of the whole cable system.
  - Optical fibers are ideally suited for carrying digital information, which is especially useful in computer networks.
  - They are highly secure as they cannot be tapped and for lack of signal radiation.
- The *Disadvantages of Optical Fiber* are:
  - Installing an optical fiber requires special equipment.
  - If a fiber breaks, finding the broken location is difficult.
  - Repairing a broken optical fiber is difficult and requires

special equipment. ○ Due to its high installation costs, they are economical when the bandwidth utilization is high.

### 9.3.4 Radio Transmission

The electromagnetic radio waves that operate at the radio frequency are also used to transmit computer data. This transmission is also known as Radio Frequency (RF) transmission ([Figure 9.5](#)). The computers using RF transmission do not require a direct physical connection like wires or cable. Each computer attaches to an antenna that can both send and receive radio transmission.



**Figure 9.5** Radio transmission

### 9.3.5 Microwave Transmission

Microwave transmission ([Figure 9.6](#)) refers to the technique of transmitting information over a microwave link. Microwaves have a higher frequency than radio waves. Microwave transmission can be aimed at a single direction, instead of broadcasting in all directions (like in radio waves).

Microwaves can carry more information than radio waves but cannot penetrate metals.

Microwaves are used where there is a clear path between the transmitter and the receiver.



**Figure 9.6** Microwave transmission

Microwave transmission has the advantage of not requiring access to all contiguous land along the path of the system, since it does not need cables. They suffer from the disadvantages: a) needing expensive towers and repeaters, and b) are subject to interference from passing airplanes and rain. Because microwave systems are line-of-sight media, radio towers must be spaced approximately every 42 km along the route.

### 9.3.6 Satellite Transmission

The communication across longer distances can be provided by combining radio frequency transmission with satellites. Geosynchronous satellites are placed in an orbit synchronized with the rotation of the earth at a distance of 36,000 km above the surface of the earth.

Geosynchronous satellites appear to be stationary when viewed from the earth. The satellite consists of transponder that can receive RF signals and transmit them back to the ground at a different angle. A ground station on one side of the ocean transmits signal to the satellite which in turn sends the signal to the ground station on the other side of the ocean ([Figure 9.7](#)).

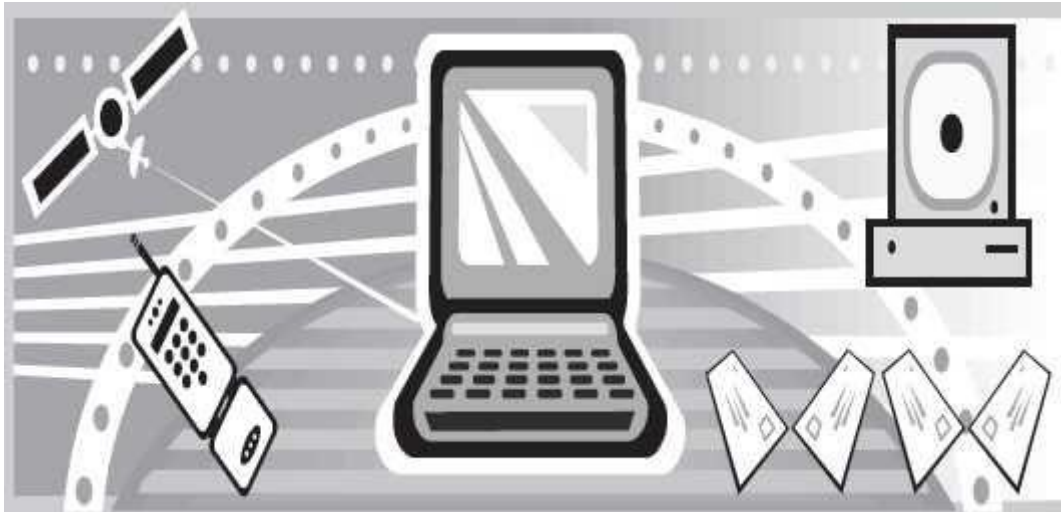
## 9.4 DATA TRANSMISSION ACROSS MEDIA

Transmitting data across media implies sending bits through the transmission medium. Physically, the data is sent as electric signals, radio waves or as light signals. Let's now discuss the use of electric current to transfer digital information. For this, the bits are encoded and sent as characters.

### 9.4.1 Transmission Modes

The direction in which data can be transmitted between any two linked devices is of three types—(1) Simplex, (2) Half-duplex, and (3) Full-duplex, or duplex. *Simplex transmission* is unidirectional data transmission. Of the two linked devices, only one of them can send data and the other one can only receive data. *Half-duplex transmission* is bi-directional data transmission, but the linked devices cannot

send and receive at the same time. When one device is sending data the other can only receive. *Full-duplex transmission* is bi-directional and the linked devices can send and receive data simultaneously. The linked devices can send data and at the same time receive data. [Figure 9.8](#) shows the different kinds of transmission modes used for interaction.



**Figure 9.7** Satellite transmission



**Figure 9.8** Transmission modes

#### 9.4.2 Transmission Speed

- When the signals are transmitted between two computers, two factors need to be considered — (1) Bandwidth, and (2) Distance.
- *Bandwidth* is the amount of data that can be transferred through the underlying hardware i.e. the communication medium, in a fixed amount of time. Bandwidth is measured in *cycles per second (cps)* or *Hertz (Hz)*. The bandwidth of the transmission medium determines the data transfer rate.
- *Throughput* is the amount of data that is actually transmitted between the two computers. Throughput is specified in *bits per second (bps)*. The throughput capability of the communication medium is also called *bandwidth*. The bandwidth of the communication medium is the upper bound on the throughput, because data cannot be sent at a rate more than the throughput of the communication medium.
- Higher throughput is achieved by using a large part of the electromagnetic spectrum (large bandwidth). Technology that uses large part of the electromagnetic spectrum to achieve higher throughput is known as *broadband technology*. The technology that uses small part of the electromagnetic spectrum is known as *baseband technology*.
- Throughput is affected by the *distance* between the connected computers or devices. Even if a transmission medium is designed for a specific bandwidth, the throughput is affected by the distance of communication.
- The bandwidth of transmission medium is limited by the distance over which the medium needs to transmit the signal. The bandwidth decreases with the increase in the distance between the connected devices. When a signal has to travel long distance, the signal strength decreases; the signal strength is utilized to overcome the resistance offered by the connecting medium (cable or wire). The gradual deterioration of signal strength across long distances is called *attenuation*.
- Moreover, with increasing distance the external disturbance increases, which causes the signal to deteriorate and results in less amount of data to be transferred. The degradation of signal due to internal or external disturbances is called *distortion*.
- The bandwidth and distance of the transmission medium is selected so that it offers minimum attenuation and minimum distortion.
- The *cat-5* UTP cable has a throughput of 100 Mbps over a distance of 100m. The 10 base2 coaxial cable has a throughput up to 10Mbps over a distance of 185 m. The *10 base 5* coaxial cable has a throughput up to 10Mbps over a distance of 500 m.

### 9.4.3 Fundamentals of Transmission

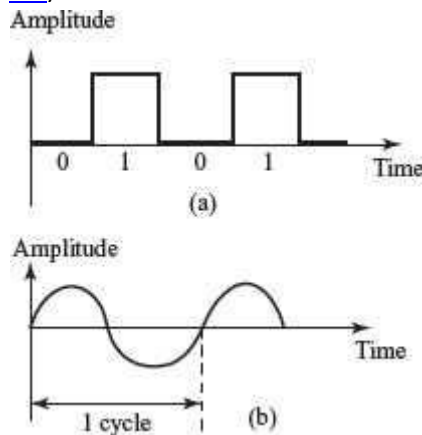
Telecommunication systems use *electromagnetic waves* to transfer information. Electromagnetic waves can travel through transmission media like copper wires, fiber optics or as radio waves. They can also travel in vacuum. Wireless communication uses electromagnetic waves for transmission of information. The transmission media through which the waves propagate are not perfect. As a result, the waves propagated via the transmission media get *attenuated and distorted*.

The information to be transmitted does not always exist in a form that is compatible with the transmission medium. Waves that are compatible with the transmission medium must be generated to carry information. A *signal* is a wave that is suitable for carrying information over a transmission medium. Signals can be electric signals, light signals, electromagnetic signals or radio signals. Electric signals are used to carry information through copper wires, light signals for fiber optic cables, and radio signals for carrying information in free space. Electrical signals have limited bandwidth and cannot be

used in long distance communication. They need to be amplified or regenerated. Light signals have a high bandwidth and are suited for long distance communication.

#### 9.4.3.1 Analog and Digital Signals

- Information carrying signals are of two types—(a) analog signal, and (b) digital signal ([Figure 9.9](#)).



**Figure 9.9** (a) Digital signal (b) Analog signal

- Analog Signal:** An analog signal is a wave that continuously changes its information carrying properties over time. The wave may vary in amplitude or frequency in response to changes in sound, light, heat, position, or pressure etc. For example a telephone voice signal is analog. The intensity of the voice causes electric current variations. At the receiving end, the signal is reproduced in the same proportion.
- Digital Signal:** A digital signal is a wave that takes limited number of values at discrete intervals of time. Digital signals are non-continuous, they change in individual steps. They consist of pulses or digits with discrete levels or values. The value of each pulse is constant, but there is an abrupt change from one digit to the next. Digital signals have two amplitude levels called nodes. The value of which are specified as one of two possibilities such as 1 or 0, HIGH or LOW, TRUE or FALSE, and so on.
- Analog and digital signals are compared on the basis of—(1) impact of noise, (2) loss of information, and (3) introduction of error.
- Analog signal has the potential for an infinite amount of signal resolution. Another advantage with analog signals is that they can be processed more easily than their digital equivalent. The primary disadvantage of the analog signals is the noise. The effects of noise create signal loss and distortion, which is impossible to recover, since amplifying the signal to recover attenuated parts of the signal, also amplifies the noise. Even if the resolution of an analog signal is higher than a comparable digital signal, the difference can be overshadowed by the noise in the signal. In digital systems, degradation can not only be detected, but corrected as well.
- Amplifier** is any device or a circuit that changes, usually increases, the amplitude of an analog signal.

- *Repeater* is an electronic device that receives a signal and retransmits it at a higher level and/or higher power, so that the signal can cover longer distances. With physical media like Ethernet or Wi-Fi, data transmissions can only span a limited distance before the quality of the signal degrades. Repeaters attempt to preserve signal integrity and extend the distance over which data can safely travel. Actual network devices that serve as repeaters usually have some other name. Active hubs, for example, are repeaters. Active hubs are sometimes also called “multiport repeaters,” but more commonly they are just “hubs.”

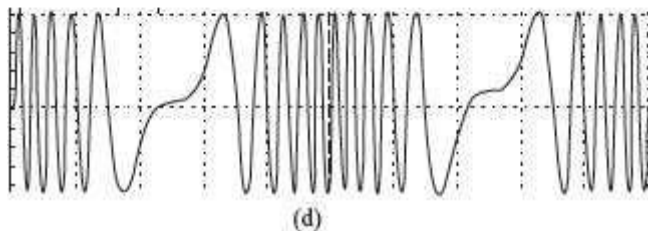
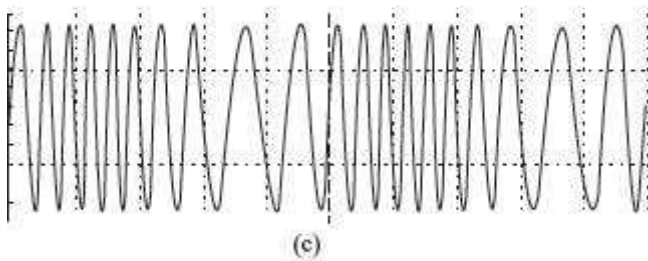
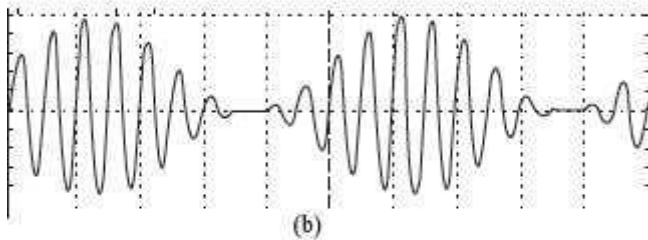
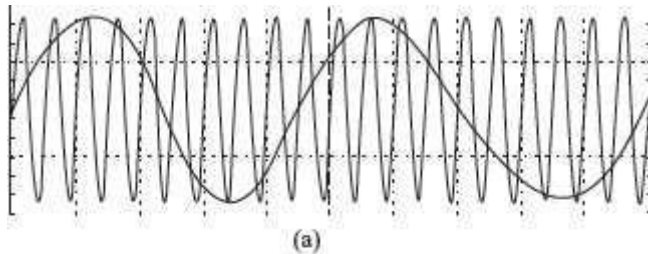
#### 9.4.3.2 Modulation and Demodulation

- *Modulation*: Signals consist of two components—the information signal and the carrier signal. The transmission of any signal over some communication medium usually involves *modulation* of a carrier. Prior to their transmission the information signal and the carrier signal are combined and the process of combining these two signals is called *modulation*. Characteristics of the carrier signal are varied in proportion to the amplitude of the information-carrying signal. Modulation results in the transfer of the signal information to higher frequency carrier signal. In simple English terms, the information signal sits on top of the carrier signal and rides on it from the receiver to the transmitter.
- *Need for Modulation*: Let’s understand the need for modulation by using a simple example. Stereophonic radio signal consist of frequency ranges from 30 Hz (Hertz) to 15 KHz (Kilo Hertz). Hence they need a bandwidth of 15 KHz. If ten different radio stations start transmitting their voice signals between 30 Hz and 15 KHz frequencies, then a combination of these signals would only create noise and the receiver would not be able to discriminate between the signals of each radio station. To overcome this, usually the FM broadcast band, used for broadcasting FM radio stations, goes from 87.5 to 108.0 MHz (Mega Hertz). For example a radio channel-1 may be broadcast using a carrier signal of 102 MHz and would typically use band of frequencies between 101.9 to 102.1 MHz. Radio channel2 using a carrier signal of 102.2 MHz would use band of frequencies between 102.1 and 102.3 MHz. Similarly for other channels, the same method of allocation would be followed. This eliminates the problem of discrimination and decoding signals of each of the radio stations at the receiving end.

There are three primary reasons which necessitate modulation:

1. To make efficient use of the lines or media used for communication
  2. To make radio communications feasible: The lower the frequency of signal, the larger is the size of the antenna needed for transmission and reception. A signal of 10 KHz would require an antenna whose dimensions are in the range of a few kilometers.
  3. To simplify signal processing: It is simpler to design electronic systems for narrow frequency bands.
- At the sending side (transmitter), the signal is superimposed on the carrier wave, which results in a modulated carrier wave. The modulated carrier wave is transmitted. At the receiving end, the receiver is configured to recognize the carrier that the sender is using. The receiver detects the modulation of the carrier wave and reconstructs the data signal.

- The process of segregating the data signal and the carrier signal from the modulated carrier wave is called *demodulation*. At the receiving end, the carrier wave is discarded after the data signal has been reconstructed.
- Modulation technique did not originate for data communication for computers, but has long been used for radio, television, and telephone communication. For long distance transmission, computer networks use modulation, whether the signals are transmitted over wires, optical fibers, microwave or radio frequency.
- Modulation is of three kinds, which are defined as follows:
  - *Amplitude Modulation*—The amplitude of the carrier wave is modified in proportion to the data signal. The frequency and phase of the carrier signal remains unchanged.
  - *Frequency Modulation*—The frequency of the carrier signal is modified in proportion to the data signal. The amplitude and phase of the carrier signal remains unchanged.
  - *Phase Shift Modulation*—The phase of the carrier signal is modified in proportion to the data signal. The amplitude and frequency of the carrier signal remains unchanged.
- For computer networks, generally phase shift modulation is used. [Figure 9.10](#) shows the different kinds of modulation of a carrier wave with signal.



**Figure 9.10** (a) Carrier wave with signal (b) Amplitude modulation (c) Frequency modulation (d) Wavelength modulation

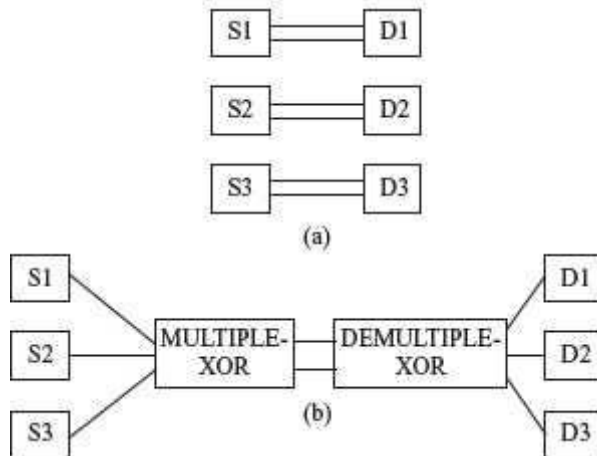
- *Modem* ([Figure 9.11](#)) is a device that has both a **modulator** and a **demodulator**. Modulator accepts data signals from the computer and modulates the carrier wave accordingly. Demodulator accepts modulated carrier wave and regenerates the original data signal from it. During data communication, modem is attached to the computer, both at the sender and the receiver side. Modems are used with all transmission media like RF modem for RF transmission and optical modem for transmission through fiber optics.



**Figure 9.11** Modems

### 9.4.3.3 Multiplexing

- Transmission medium have varying data carrying capacities. To utilize the full capacity of the transmission medium, computer networks use separate channels that allow sharing of a single physical connection for multiple communication. Multiple carrier signals are transmitted over the same medium at the same time and without interference from each other.
- The combining of multiple signals into a form that can be transmitted over a single link of a communication medium is called *multiplexing*. [Figure 9.12 \(a\)](#) shows computers connected without multiplexing and [Figure 9.12 \(b\)](#) shows computers connected via a multiplexer.
- *Demultiplexing* is a technique of separating the merged signals and sending them to the corresponding receivers. The two basic multiplexing techniques are— Frequency Division Multiplexing (FDM) and Wavelength Division Multiplexing (WDM).

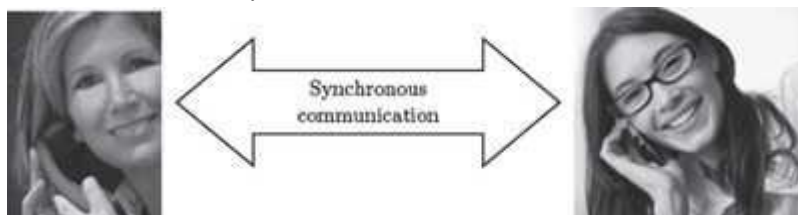


**Figure 9.12** (a) No multiplexing (b) Multiplexing

- *FDM* combines different carrier frequencies signals into a single signal of higher bandwidth. The bandwidth of the communication medium link carrying the combined signal is greater than the sum of the bandwidth of the individual signals that are combined. *FDM* is used for high band—width analog transmission systems like broadband technology.
- *WDM* is similar to *FDM* except that *FDM* involves electromagnetic spectrum below light and *WDM* involves light signals. *WDM* uses very high frequencies. *WDM* combines different light signals coming from different sources into a larger band light signal across a single optical fiber. It also enables bi-directional communications over one strand of fiber.

#### 9.4.3.4 Asynchronous and Synchronous Transmission

- One major difficulty in data transmission is that of synchronising the receiver with the sender. Whenever an electronic device transmits digital (and sometimes analog) data to another electronic device, there must be a certain rhythm established between the two devices, i.e., the receiving device must have some way of knowing, within the context of the fluctuating signal that it is receiving, where each unit of data begins, and ends. The signal must be synchronized in a way that the receiver can distinguish the bits and bytes as the transmitter intends them to be distinguished. Two approaches exist to address the problem of synchronisation—synchronous transmission, and asynchronous transmission.

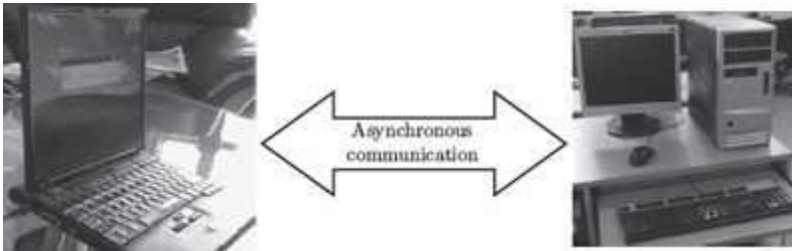


**Figure 9.13** Synchronous communication

- *Synchronous communication* is the characteristic of a communication system in which the sender must coordinate (i.e. synchronize) with the receiver before sending data ([Figure 9.13](#)).

The network is designed to move the data at the precise rate, which is not affected by the increase or decrease in network traffic. Voice system network use the synchronous transmission.

- *Asynchronous communication* is a characteristic of a communication system in which the sender and receiver do not coordinate before the transmission of data ([Figure 9.14](#)). The receiver must be prepared to accept data at any time. The sender can wait when no data is available and send when data is available for sending. Most of the data networks use asynchronous transmission. E.g.: RS-232 based serial devices use asynchronous communication.

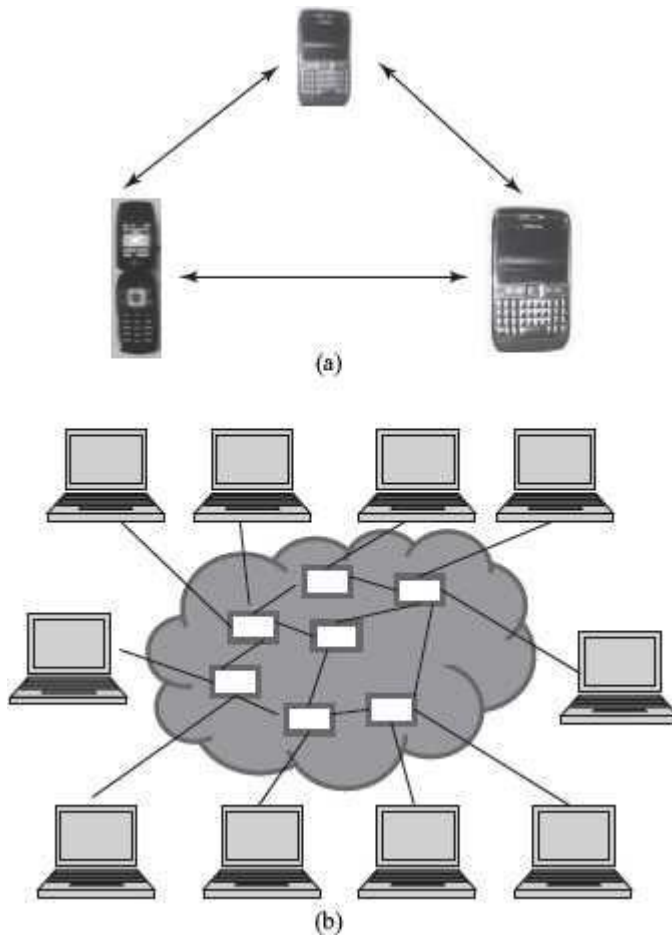


**Figure 9.14** Asynchronous communication

## 9.5 DATA TRANSMISSION AND DATA NETWORKING

Data transmission at physical level involves the hardware required for handling individual bits and encoding bits in signals. The details of the underlying hardware are generally handled by the engineers who design the hardware.

Any two devices directly linked via a communication medium (point to point communication) can send and receive data, to and from each other respectively ([Figure 9.15 a](#)). If a large number of computers need to interact with each other, point to point communication will require direct link between all the computers. This is not a practical solution. The communication circuits and the associated hardware required for communication (like modem) are expensive. Moreover, there may not be a need to transmit data all the time, which will result in the communication medium lying idle for most of the time. For long distance communication, instead of point to point connection, a network of nodes is used as a communication medium. The different computers attached to the network share the communication facility.



**Figure 9.15** (a) Point-to-point communication (b) Switching

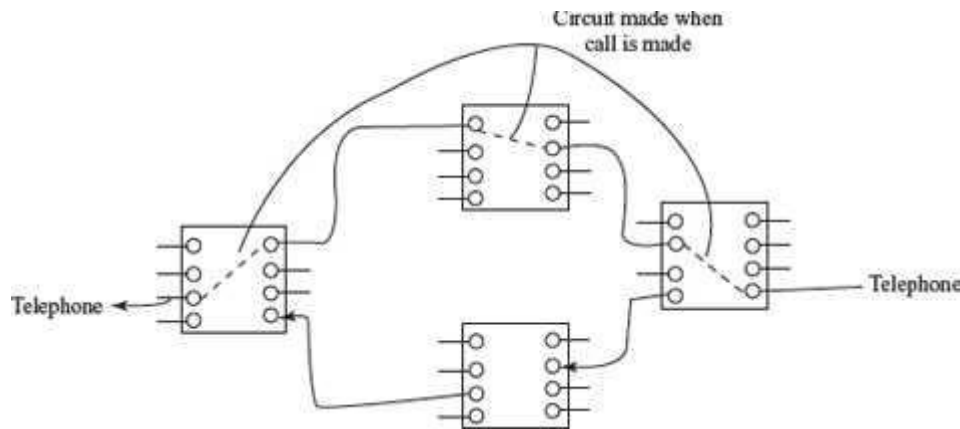
The computer network provides a convenient interface that handles sending of multiple bytes of data across the network instead of handling data transmission at physical level.

### 9.5.1 Switching

A network cannot allow or deny access to a shared communication facility. All computers attached to the network can use it to send and receive data. Networks allow sharing of communication medium using *switching* ([Figure 9.15 b](#)). Switching routes the traffic (data traffic) on the network. It sets up temporary connections between the network nodes to facilitate sending of data. Switching allows different users, fair access to the shared communication medium. There are three kinds of switching techniques—(1) Packet switching, (2) Circuit switching, and (3) Message switching. Computer networks generally use packet switching, occasionally use circuit switching but do not use message switching.

#### 9.5.1.1 Circuit Switching

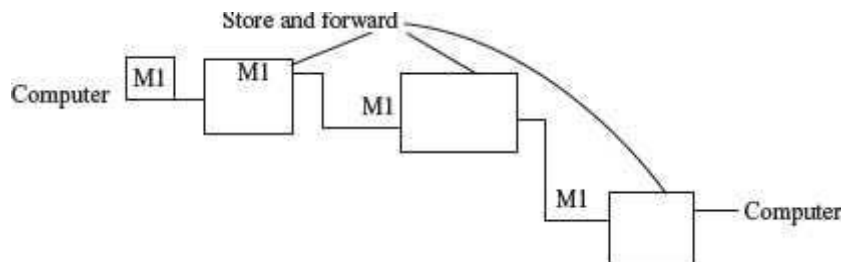
Circuit switching ([Figure 9.16](#)) sets up end-to-end communication path between the source and the destination, before the data can be sent. The path gets reserved during the duration of the connection. Circuit switching is commonly used in the telephone communication network.



**Figure 9.16** Circuit switching

### 9.5.1.2 Message Switching

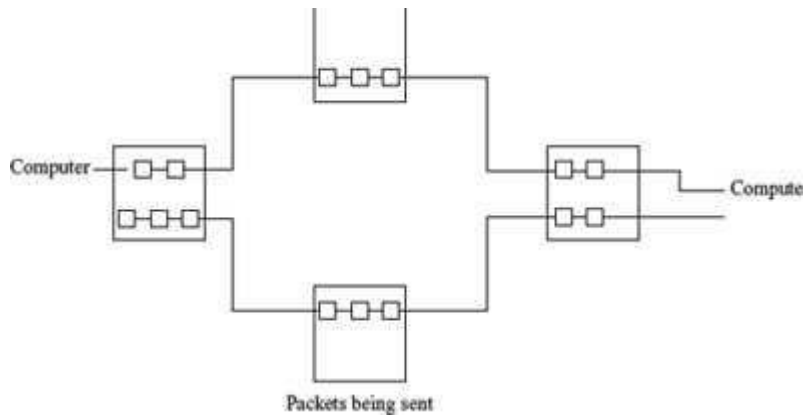
Message switching ([Figure 9.17](#)) does not establish a physical path in advance, between the sender and the receiver. It uses the 'store and forward' mechanism. In this mechanism, the network nodes have large memory storage. The message is received from the sender and stored in the network node, and when it finds a free route, it forwards the message to the next node till it reaches the destination. Message switching requires large data storage capacity and incurs delay in storing and forwarding of message. Message switching may block the network nodes for a long time. They are thus not suitable for interactive communication. Message switching is no more used in computer networks.



**Figure 9.17** Message switching

### 9.5.1.3 Packet Switching

- Like message switching, packet switching does not establish a physical path between the sender and the receiver, in advance. Packet switching ([Figure 9.18](#)) also uses the 'store and forward' mechanism. However, instead of a complete message, packets are sent over the network. Packet switching splits a message into small "packets" of defined size to be sent over the network. Each packet is numbered.



**Figure 9.18** Packet switching

- A packet is a self-contained part of data that can be sent over the network. A packet contains the data to be transmitted and a header that contains information about the packet, like the source and destination addresses, size of packet, error checking bytes etc.
- Since the path through which the packets travel is not reserved, the packets may travel through different paths in the network and may not reach the destination in order. At the destination, the received packets are reassembled (according to the packet number), and the complete message is constructed.
- Packet switching is suited for interactive traffic. Packet switching limits the size of the packet and does not block a network node for a long time. Moreover, a node can transmit a packet before the arrival of another full packet, thus reducing the delay.
- Packet switching does not require dedicated communication link, and shares the underlying resources. Packet switching is commonly used for computer networks, including the Internet.

## 9.6 COMPUTER NETWORK

A *computer network* is an interconnection of two or more computers that are able to exchange information. The computers may be connected via any data communication link, like copper wires, optical fibers, communication satellites, or radio links. The computers connected to the network may be personal computers or large main frames. The computers in a network may be located in a room, building, city, country, or anywhere in the world.

### 9.6.1 Network Types

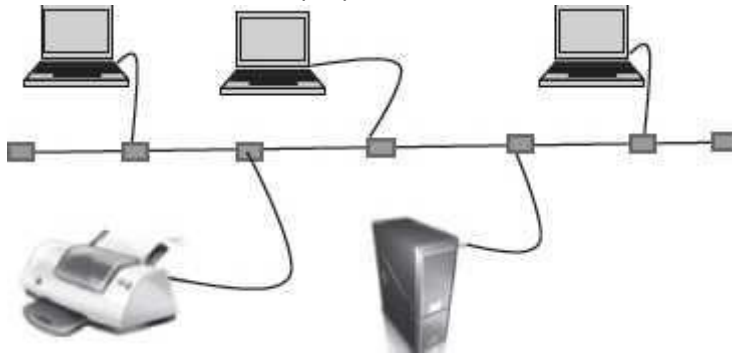
Computer network is broadly classified into three types—(1) Local Area Network (LAN), (2) Metropolitan Area Network (MAN), and (3) Wide Area Network (WAN). The different network types are distinguished from each other based on the following characteristics:

- Size of the network
- Transmission Technology
- Networking Topology

The *size of the network* refers to the area over which the network is spread. *Transmission technology* refers to the transmission media used to connect computers on the network and the transmission protocols used for connecting. *Network topology* refers to the arrangement of computers on the network or the shape of the network. The following subsections discuss the three types of networks and their characteristics.

#### 9.6.1.1 Local Area Network

LAN ([Figure 9.19](#)) is a computer network widely used for local communication. LAN connects computers in a small area like a room, building, office or a campus spread up to a few kilometers. They are privately owned networks, with a purpose to share resources and to exchange information.



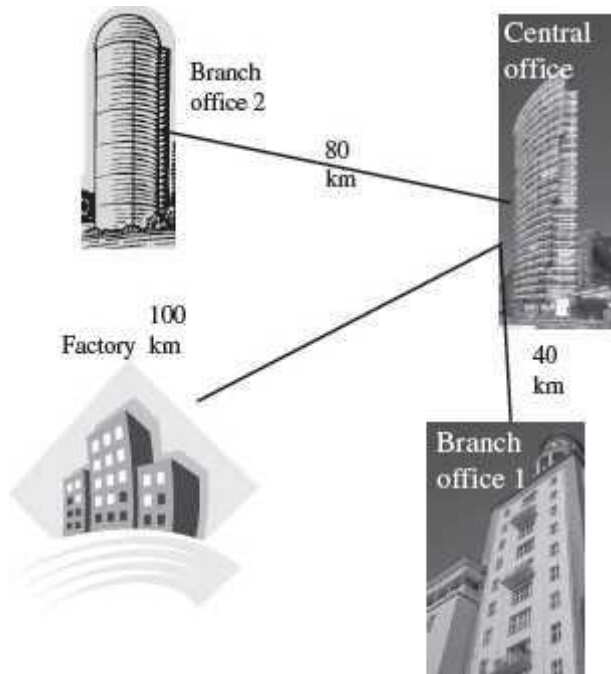
**Figure 9.19** LAN

The computers in a LAN are generally connected using cables. LAN is different from other types of network since they share the network. The different computers connected to a LAN take turns to send data packets over the cables connecting them. This requires coordination of the use of the network. Some of the transmission protocols used in LAN are Ethernet, Token bus, and FDDI ring.

Star, Bus, and Ring are some of the common LAN networking topologies. LAN runs at a speed of 10 Mbps to 100 Mbps and has low delays. A LAN based on WiFi wireless network technology is called Wireless Local Area Network (WLAN).

#### 9.6.1.2 Metropolitan Area Network

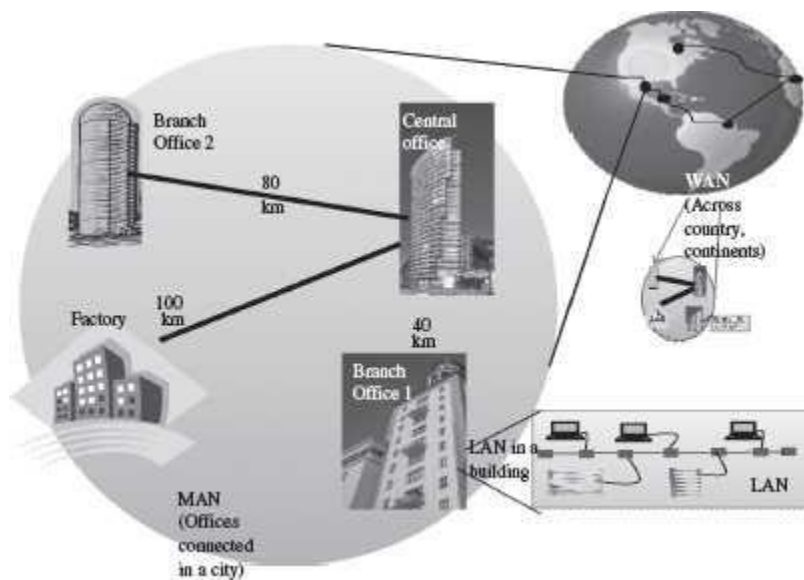
MAN ([Figure 9.20](#)) is a computer network spread over a city. Cable television network is an example of MAN. The computers in a MAN are connected using coaxial cables or fiber optic cables. MAN also connects several LAN spread over a city.



**Figure 9.20** MAN

### 9.6.1.3 Wide Area Network

WAN is a network that connects computers over long distances like cities, countries, continents, or worldwide ([Figure 9.21](#)). WAN uses public, leased, or private communication links to spread over long distances. WAN uses telephone lines, satellite link, and radio link to connect. The need to be able to connect any number of computers at any number of sites, results in WAN technologies to be different from the LAN technologies. WAN network must be able to grow itself. Internet is a common example of WAN.



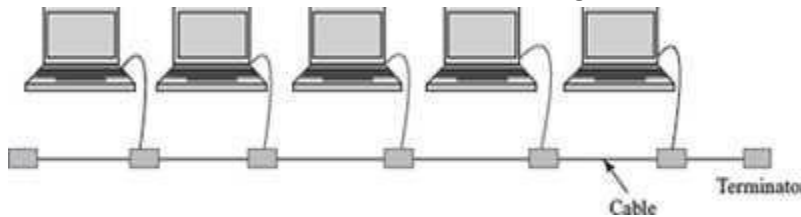
**Figure 9.21** LAN, MAN and WAN

### 9.6.2 LAN Topologies

There are different types of network topologies that are used in a network. The network topologies in the structure or the layout of the different devices and computers connected to the network. The topologies commonly used in LAN are—Bus topology, Star topology, and Ring topology.

#### 9.6.2.1 Bus Topology

- All devices on the network are connected through a central cable called a Bus ([Figure 9.22](#)).

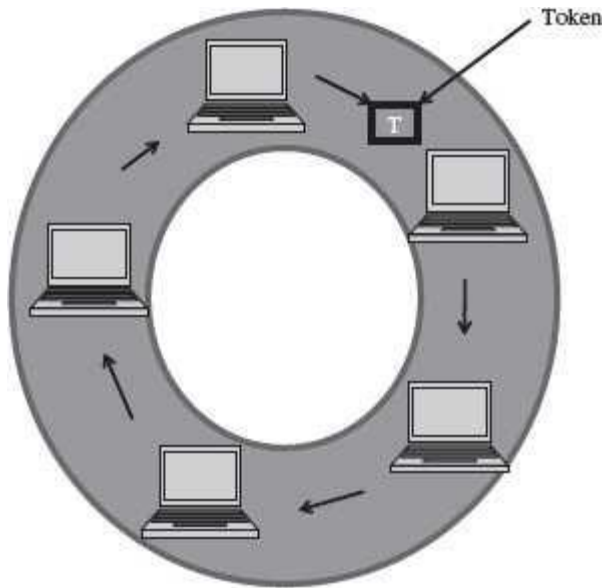


**Figure 9.22** Bus topology

- The data signal is available to all computers connected to the bus .
- The data signal carries the address of the destination computer.
- Each computer on the network checks the destination address as the data signal travels through the bus. The computer whose address matches makes a copy of the signal and converts it into data. The data signal on the bus does not get destroyed and still transmits along the bus, and is finally absorbed by the terminator attached to the end of the network.
- It is good for connecting 15–20 computers.
- A single coaxial cable is generally used in bus topology, to which the computers or devices are connected.
- Ethernet is a commonly used protocol in networks connected by bus topology.

#### 9.6.2.2 Ring Topology

- All devices in the network are connected in the form of a ring.
- Each device has a receiver and transmitter to receive the data signals and to send them to the next computer, respectively.
- Ring network does not have terminated ends, thus data signals travel in a circle.
- Ring topology ([Figure 9.23](#)) uses token passing method to provide access to the devices in the network.
- The computers or devices are connected to the ring using twisted pair cables, coaxial cables or optic fibers.
- The protocols used to implement ring topology are Token Ring and Fiber Distributed Data Interface (FDDI).



**Figure 9.23** Ring topology

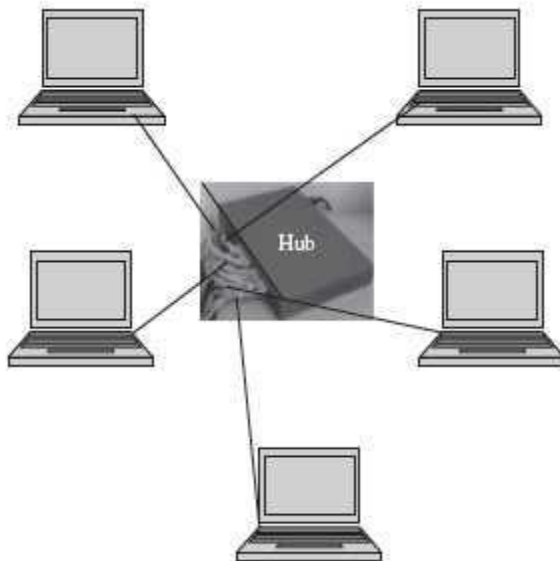
### 9.6.2.3 Star Topology

- All devices are connected through a central link forming a star-like structure.
- The central link is a hub or switch. The computers are connected to the hub or switch using twisted pair cables, coaxial cables or optic fibers.
- Star topology ([Figure 9.24](#)) is the most popular topology to connect computer and devices in network.
- The data signal is transmitted from the source computer to the destination computer via the hub or switch.
- The common protocols used in star topology are Ethernet, Token Ring, and LocalTalk.

In addition to the bus, ring, and star topologies, there are complex topologies like the tree topology, and the mesh topology used for networking in LAN. [Table 9.1](#) lists the advantages and disadvantages of the different LAN network topologies.

### 9.6.3 Communication Protocol

Data networks are a combination of software and hardware components. The hardware includes transmission media, devices, and transmission equipments. The software allows the hardware to interact with one another and provide access to the network. The application programs that use the network do not interact with the hardware directly. The application programs interact with the protocol software, which follows the rules of the protocol while communicating. *Protocol* is a network term used to indicate the set of rules used by a network for communication.



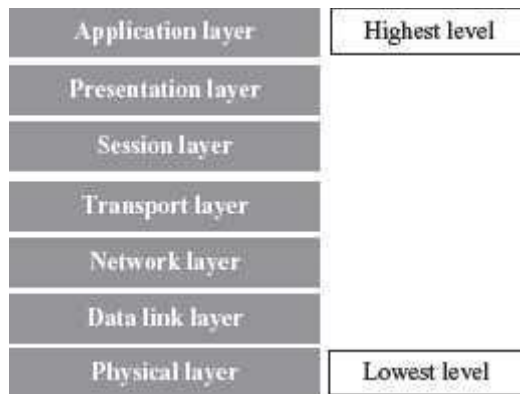
**Figure 9.24** Star topology

	Bus Topology	Ring Topology	Star Topology
Advantages	<p>Easy to implement (computers connected linearly through cable)</p> <p>Easily extendable (new devices can be easily added)</p> <p>Not very expensive</p>	<p>All computers in the ring have equal access to the ring</p> <p>Each computer in the ring gets an opportunity to transmit data.</p>	<p>Failure of a device attached to the network does not halt the complete network; only that device is down.</p> <p>Easily extendable—by attaching a new device to the hub or switch.</p> <p>No disturbance when a new device is added or removed.</p> <p>Easy to troubleshoot the network.</p>
Disadvantages	<p>If the cable gets damaged, the whole network collapses</p> <p>A computer can transmit data only if network is not being utilized</p> <p>Network slows down if additional computers are connected to the network.</p>	<p>Adding or removing devices is difficult and affects the complete network</p> <p>Failure in a node or the cable breaks down the ring and thus the network.</p> <p>The length of the ring and the number of nodes are limited</p>	<p>It is costly, since each device on the network is attached by a single cable to the central link.</p> <p>Failure of the hub or switch breaks the complete network.</p>

**Table 9.1** Advantages and disadvantages of network topologies

All the computers connected to the network use the protocol software. The network communication protocol is organized as a stack of layers with one layer built upon the other. Each layer has a specific function and interacts with the layers above and below it. The outgoing data from a computer connected to the network passes down through each layer and the incoming data passes up through each layer. The corresponding layers on the different machines are called *peers*. The peers interact with each other using the protocol.

The International Standards Organization (ISO) has developed a seven-layer reference model for data networks, known as Open System Interconnection (OSI) model. The OSI model specifies the functions of each layer. It does not specify how the protocol needs to be implemented. It is independent of the underlying architecture of the system and is thus an open system. The seven layers of the OSI model are—(1) Physical layer, (2) Data link layer, (3) Network layer, (4) Transport layer, (5) Session layer, (6) Presentation layer, and (7) Application layer. The functions of the different layers ([Figure 9.25](#)) are as follows:

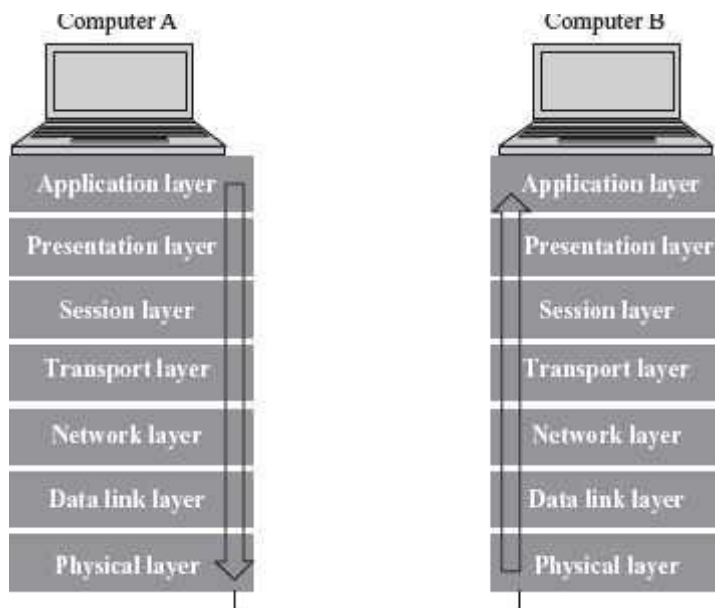


**Figure 9.25** OSI model

- **Physical Layer**—This layer specifies the basic network hardware. Some of the characteristics defined in the specification are—interface between transmission media and device, encoding of bits, bit rate, error detection parameters, network topology, and the mode of transmission (duplex, half-duplex or simplex).
- **Data Link Layer**—This layer specifies the functions required for node-to-node transmission without errors. It specifies the organization of data into frames, error detection in frames during transmission, and how to transmit frames over a network.
- **Network Layer**—The network layer specifies the assignment of addresses (address structure, length of address etc.) to the packets and forwarding of packets to the destination i.e. routing.
- **Transport Layer**—It specifies the details to handle reliable transfer of data. It handles end-to-end error control and flow control, breaking up data into frames and reassembling the frames.
- **Session Layer**—The session layer maintains a session between the communicating devices. It includes specifications for password and authentication, and maintaining synchronization between the sender and the receiver.

- **Presentation Layer**—This layer specifies the presentation and representation of data. Its functions include translation of the representation of the data into an identifiable format at the receiver end, encryption, and decryption of data etc.
- **Application Layer**—This layer specifies how an application uses a network. It deals with the services attached to the data. It contains the protocols used by users like HTTP, protocol for file transfer and electronic mail.

Each layer at the sender's side transforms the data according to the function it handles. For this it attaches headers to the data. At the receiver's side, the corresponding layer applies the inverse of the transformation that has been applied at the source ([Figure 9.26](#)). As an example, if the Data link layer at the sender's side adds an error detection code to the frame, then at the receiver's side, the Data link layer verifies the error detection code and removes it from the frame before passing it to the next higher level, i.e. the Network layer.



**Figure 9.26** Data transfer in OSI model

The 7-layer ISO reference model forms a framework for communication between the devices attached to the network. For different networks, the number of layers and their functions may vary. For example, the TCP/IP Internet protocol is organized into five layers. The X.25 Wide Area Network protocol (the first public data network) provides connectivity to Public Switched Telephone Network (PSTN) network and has three layers.

#### 9.6.4 Network Devices

The cables are used to transmit data in the form of signals from one computer to another. But cables cannot transmit signals beyond a particular distance. Moreover there is a need to connect multiple computers and devices. A *concentrator* is a device having two or more ports to which the computers and other devices can be connected. A concentrator has two main functions—(1) it amplifies the signal to restore the original strength of the signal, and (2) it provides an interface to connect multiple computers

and devices in a network. Repeater, hub, switch, bridge, and gateway are examples of network connecting devices.

Two or more LANs using different protocols may not be able to communicate with the computers attached to their network. For example, a LAN connected using Ethernet may not be able to communicate with a LAN connected using Token Ring. Bridge, Router, and Gateway are devices used to interconnect LANs.

#### 9.6.4.1 Network Interface Card

- A Network Interface Card (NIC) is a hardware device through which the computer connects to a network.
- NIC is an expansion card ([Figure 9.27](#)), it can be either ISA or PCI, or can be on-board integrated on a chipset. NIC has an appropriate connector to connect the cable to it. NIC for different LAN are different (NIC for token ring is different from NIC for Ethernet).

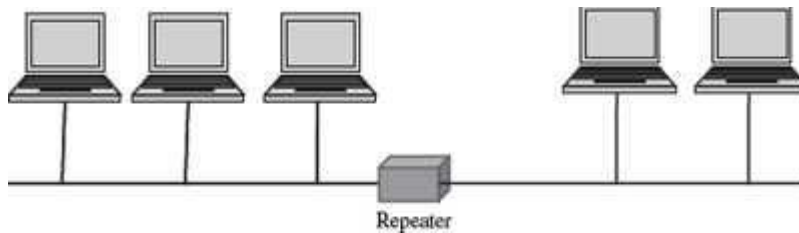


**Figure 9.27** NIC card

- NIC work at both the data link layer and physical layer of the OSI reference model.
- At the data link layer, NIC converts the data packets into data frames, adds the Media ACcess address (MAC address) to data frames. At the physical layer, it converts the data into signals and transmits it across the communication medium. The MAC address is a globally unique hardware number present on the NIC and is specified by the NIC manufacturer.
- NIC depends upon the configuration of the computer, unlike hub or switches that perform independently.

#### 9.6.4.2 Repeater

- Repeaters ([Figure 9.28](#)) are used to extend LAN. It has only two ports and can connect only two segments of a network. Multiple repeaters can be used to connect more segments. (Segment is a logical section of the same network).

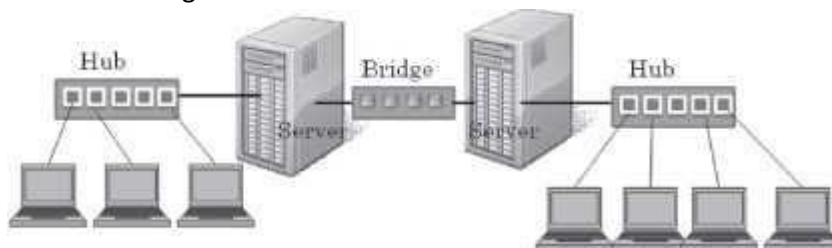


**Figure 9.28** Repeater

- Repeaters operate at the Physical layer of OSI reference model.
- They are useful when computers in a network are located far away from each other.
- Repeaters amplify the signal so that the signal is as strong as the original signal. They can thus extend the reach of a network.
- Repeaters cannot be used if multiple computers need to be interconnected or multiple segments need to be interconnected.
- Repeaters cannot identify complete frames. Thus, in addition to the valid transmissions from one segment to another, repeater also propagates any electrical interference occurring on a segment to other segment.

#### 9.6.4.3 Bridge

- Bridge ([Figure 9.29](#)) is used to connect two LAN segments like a repeater; it forwards complete and correct frames to the other segment. It does not forward any electrical interference signals to the other segment.

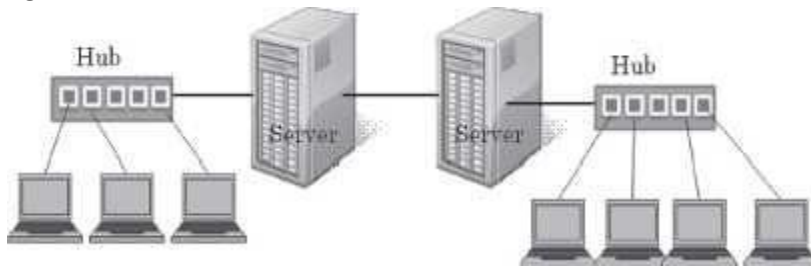


**Figure 9.29** Bridge

- Bridge forwards a copy of the frame to the other segment, only if necessary. If a frame is meant for a computer on the same segment, then bridge does not forward a copy of the frame to other segment.
- Bridge connects networks that use different protocol at the Data Link Layer. The frame format of data in the two networks is different. The bridge converts the frame format before transmitting data from one network to another, with translation software included in the bridge.
- A bridge is also used to divide a network into separate broadcast domains to reduce network traffic while maintaining connectivity between the computers.

#### 9.6.4.4 Hub

- It is like a repeater with multiple ports. But, hub does not amplify the incoming signal.
- Hub ([Figure 9.30](#)) operates at the Physical layer of OSI reference model, hence treats data as a signal.

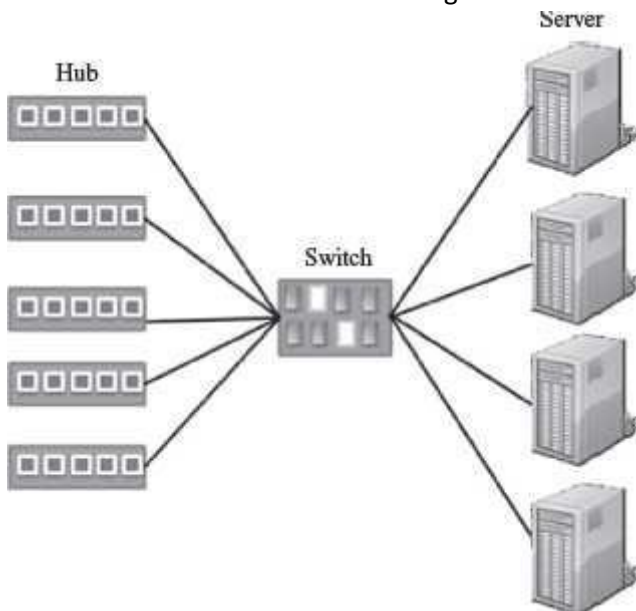


**Figure 9.30** Hub

- Hubs are used to connect multiple segments of the same network.
- Hubs are also used to connect computers to network that use Star topology.
- The port on the hubs can also be used to connect another hub, switch, bridge or router.
- Hubs increase the network traffic because they broadcast data to all the device connected all the ports of the hub.
- It is preferable to use a hub in a small LAN having about 8–10 computers connected to it.

#### 9.6.4.5 Switch

- Like hub, switch also connects multiple computers in a network or different segments of the same network. A hub simulates a single segment that is shared by all computers attached to it (hub transmits the data to all computers attached to it). In a hub, at most two computers can interact with each other at a given point of time. However, in a switch each computer attached to a switch has a simulated LAN segment.
- Switches ([Figure 9.31](#)) work at the Data Link Layer of the OSI reference model. Hence, switches consider data as frames and not as signals.

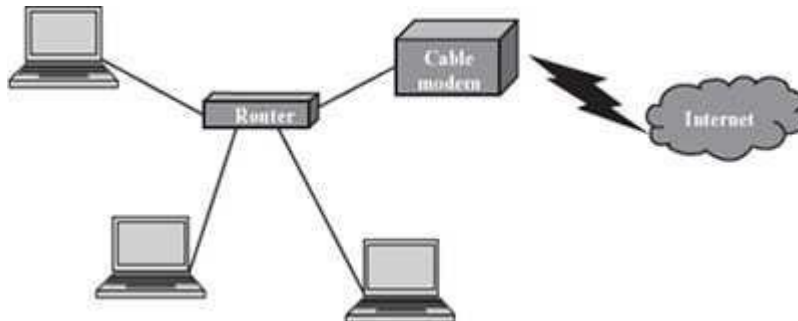


**Figure 9.31 Switch**

- A data frame contains the MAC address of the destination computer. A switch receives a signal as a data frame from a source computer on a port, checks the MAC address of the frame, forwards the frame to the port connected to the destination computer having the same MAC addresses, reconverts the frame back into signal and sends to the destination computer. (Switching is a technique that reads the MAC address of the data frame and forwards the data to the appropriate port). Switches, thus, regenerate the signals.
- Since a switch does not broadcast data, but sends the data from the source computer to the destination computer, a half of the computers attached to the switch can send data at the same time.
- Switch is also referred to as a multi-port bridge. In general, bridges are used to extend the distance of the network, and switches are primarily used for their filtering capabilities to create a multiple and smaller virtual LAN (a LAN segment can be connected to each port of the switch) from a single large LAN.

#### 9.6.4.6 Router

- Router ([Figure 9.32](#)) is used to connect heterogeneous networks.

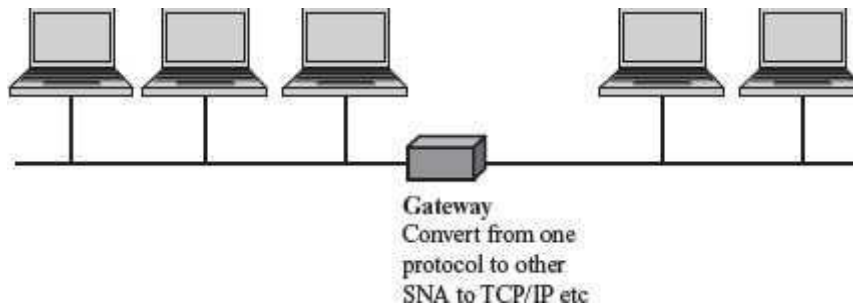


**Figure 9.32 Router**

- A router has a processor, memory, and I/O interface for each network to which it connects.
- A router connects networks that use different technologies, different media, and physical addressing schemes or frame formats.
- A router can connect two LANs, a LAN and a WAN, or two WANs.
- A router is used to interconnect the networks in the Internet.
- Router operates at the Network layer of the OSI model (layer 3).
- Physically, a router resembles a bridge, but is different from a bridge. A router determines which way is the shortest or fastest in a network, and routes packets accordingly. Since it works at the Network layer, it moves packets based on the IP addresses etc. In contrast, a bridge connects two LANs almost permanently.

#### 9.6.4.7 Gateway

- *Gateway* ([Figure 9.33](#)) is a generic term used to represent devices that connect two dissimilar networks.



**Figure 9.33** Gateway

- A gateway at the transport layer converts protocols among communications networks. It can accept a packet formatted for one protocol and convert it to a packet formatted for another protocol, before forwarding it. An application gateway can translate messages from one format to the other.
- A gateway can be implemented in hardware, software, or in both hardware and software. Generally, gateway is implemented by software installed within a router.

The network connecting devices—repeater and hub operate at the physical layer, bridge and switch operate at the data link layer, and the router operates at the network layer of the OSI model.

## 9.7 WIRELESS NETWORKING

Wireless technology, as the name suggests, is used to establish a wire-free connection or communication between two or more devices. In contrast to the wired technology where data is encoded as electric current and signals travel through wires, in wireless technology data is encoded on electromagnetic waves that travel through air. The wireless technology is used for broadcasting in radio and television communication, for communication using mobile phones and pagers, for connecting components of computers using Bluetooth technology, for Internet connection using Wi-Fi, Wireless LAN, PDA, and in remote controls for television, doors etc.

- Wireless network is a computer network connected wirelessly. The communication is done through a wireless media like radio waves, infrared or Bluetooth.
- The wireless networks have two main components—the wireless access points that include the transmitter along with the area it can cover, and the wireless clients like mobile handsets, laptops with Ethernet cards etc.
- The access point receives data frames from the computers attached to it wirelessly, checks the frames, and transmits them to their destination. The coverage area of a transmitter depends on the output power of the transmitter, its location, and the frequency used to transmit the data. Higher frequencies require a clear line of sight as compared to lower frequencies.
- The speed of wireless connection is determined by the distance of the wireless client device from the access point, the obstruction-free path (walls, trees etc.), interference, and the number of users using the network at a given time.
- Wireless networks can be divided into three categories based on their use:

- *Bluetooth technology* to connect the different components of the computer in a room, a small office or home.
- *Wireless LAN* is used to connect computers and devices wirelessly in a LAN, for example, different computers or devices in an office or campus.
- *Wireless WAN* is used to connect wide area systems, for example access to Internet via mobile devices like cell phone, PDAs and laptops.

### 9.7.1 Bluetooth Technology

The different components of the computer like the keyboard, printer, monitor etc., are connected to the computer case via wires. Bluetooth technology ([Figure 9.34](#)) is used to connect the different components wirelessly. A printer placed in a room may be connected to a computer placed in a different room using Bluetooth technology. Using Bluetooth does away with the wires required to connect the components to the computer and allows portability of components within a small area lying within the Bluetooth range.



**Figure 9.34** Bluetooth

### 9.7.2 Wireless LAN

Wireless LAN ([Figure 9.35](#)) has some benefits over the wired LANs. In wireless LAN, there is flexibility to move the computers and devices within the network. It can connect computers where cabling is not possible. It is easy to expand by using an access point. Since no physical medium is required, wireless LANs are easy to install. Since data is transmitted using radio or infrared waves, there is no attenuation or distortion of the signal due to electromagnetic interference. Wireless LANs are used at home to connect devices on different floors or to set up a home network ([Figure 9.36](#)), to provide connectivity in public places like airports, railway stations, college campus, and hotels etc., where traveling users can access the network. Wireless LANs can also be connected to a WAN thus providing access to Internet to the user. IEEE 802.11 is a standard for wireless LAN.



**Figure 9.35** (a) Wireless ethernet bridge (b) Wireless antenna in a desktop



**Figure 9.36** Computers at different floors in a house connected by wireless LAN

### 9.7.3 Wireless WAN

The radio network used for cellular telephone is an example of wireless WAN ([Figure 9.37](#)). Wireless WANs allow the users to access the Internet via their mobile devices. This provides flexibility to the user to access the Internet from any location where wireless connectivity exists.



**Figure 9.37** Wireless WAN

Almost all wireless networks are connected to the wired network at the back-end to provide access to Internet. Wireless networks also offer many challenges, like, the compatibility among different standards promoted by different companies, congested networks in case of low bandwidth, the high infrastructure and service cost, data security, battery storage capability of wireless device, and health risk.

## 14.1 INTRODUCTION

We all like to be secure in our home, office, locality, city, country, and in this world. We use different mechanisms to ensure our security. Inside our homes, we keep our valuables safely locked in a cupboard that is accessible by the elders of the house; we keep the gates of our house bolted and even have an intrusion-detection system installed. We have high walls and gates surrounding our locality and also a watchman who guards the open gates. We have police for our security within a city and armed forces for the country. We take all these measures to make ourselves and our valuables, resources, possessions secure.

The widespread use of computers has resulted in the emergence of a new area for security— security of computer. Computer security is needed to protect the computing system and to protect the data that they store and access. Transmission of data using network (Internet) and communication links has necessitated the need to protect the data during transmission over the network. Here, we use the term computer security to refer to both the computer security and the network security.

*Computer security* focuses on the security attacks, security mechanisms and security services.

- *Security attacks* are the reasons for breach of security. Security attacks comprise of all actions that breaches the computer security.
- *Security mechanisms* are the tools that include the algorithms, protocols or devices, that are designed to detect, prevent, or recover from a security attack.
- *Security services* are the services that are provided by a system for a specific kind of protection to the system resources.

The purpose of computer security is to provide reliable security services in the environments suffering security attacks, by using security mechanisms. The security services use one or more security mechanism(s).

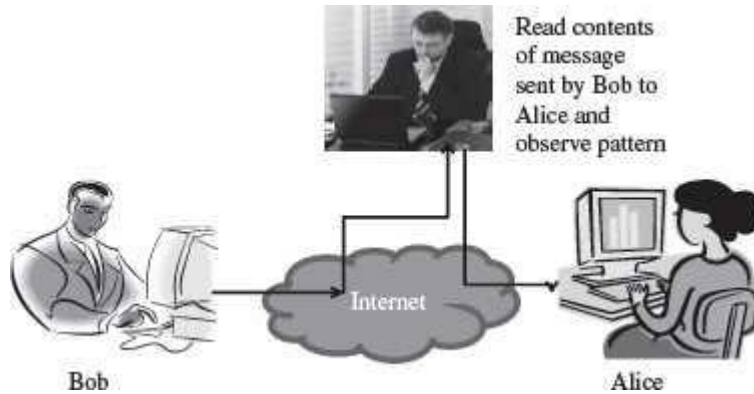
This chapter discusses the different security threats and security attacks from malicious software and hackers. The chapter highlights the security services. The security mechanisms like cryptography, digital signatures, and firewalls are discussed in detail. The need for security awareness and the security policy in an organization is also emphasized.

## 14.2 SECURITY THREAT AND SECURITY ATTACK

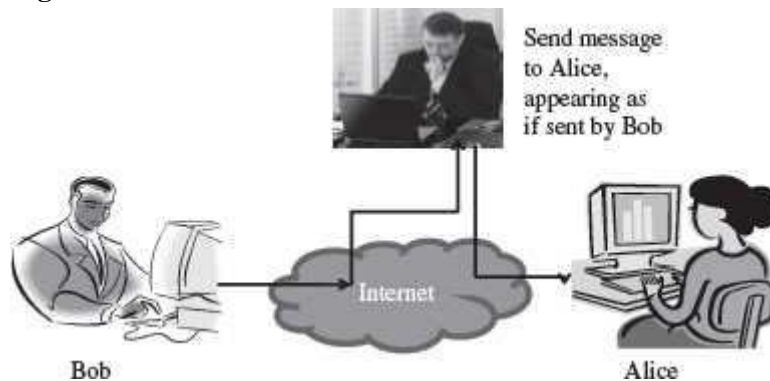
A *threat* is a potential violation of security and causes harm. A threat can be a malicious program, a natural disaster or a thief. *Vulnerability* is a weakness of system that is left unprotected. Systems that are vulnerable are exposed to threats. Threat is a possible danger that might exploit vulnerability; the actions that cause it to occur are the security attacks. For example, if we leave the house lock open—it is vulnerable to theft; an intruder in our locality (might exploit the open lock) is a security threat; the intruder comes to know of the open lock and gets inside the house—This is a security attack.

A security attack may be a passive attack or an active attack.

- The aim of a *passive attack* is to get information from the system but it does not affect the system resources. Passive attacks are similar to eavesdropping ([Figure 14.1](#)). Passive attacks may analyze the traffic to find the nature of communication that is taking place, or, release the contents of the message to a person other than the intended receiver of the message. Passive attacks are difficult to detect because they do not involve any alteration of the data. Thus, the emphasis in dealing with passive attacks is on prevention rather than detection.



**Figure 14.1** Passive attack



**Figure 14.2** Active attack (masquerade)

- An *active attack* tries to alter the system resources or affect its operations. Active attack may modify the data or create a false data ([Figure 14.2](#)). An active attack may be a masquerade (an entity pretends to be someone else), replay (capture events and replay them), modification of messages, and denial of service. Active attacks are difficult to prevent. However, an attempt is made to detect an active attack and recover from them.

Security attacks can be on users, computer hardware and computer software ([Figure 14.3](#)).

- Attacks on users* could be to the identity of user and to the privacy of user. Identity attacks result in someone else acting on your behalf by using personal information like password, PIN number in an ATM, credit card number, social security number etc. Attacks on the privacy of user involve tracking of users habits and actions—the website user visits, the buying habit of the user etc. Cookies and spam mails are used for attacking the privacy of users.

- *Attacks on computer hardware* could be due to a natural calamity like floods or earthquakes; due to power related problems like power fluctuations etc.; or by destructive actions of a burglar.
- *Software attacks* harm the data stored in the computer. Software attacks may be due to malicious software, or, due to hacking. *Malicious software or malware* is a software code included into the system with a purpose to harm the system. Hacking is intruding into another computer or network to perform an illegal act.

This chapter will discuss the malicious software and hacking in detail.



**Figure 14.3** Security attacks

### 14.3 MALICIOUS SOFTWARE

Malicious users use different methods to break into the systems. The software that is intentionally included into a system with the intention to harm the system is called *malicious software*. Viruses, Trojan horse, and Worms are examples of malicious programs. Javascripts and Java applets written with the purpose of attacking, are also malicious programs.

#### 14.3.1 Virus

Virus is a software program that is destructive in nature. Virus programs have the following properties:

- It can attach itself to other healthy programs.
- It can replicate itself and thus can spread across a network.
- It is difficult to trace a virus after it has spread across a network.
- Viruses harm the computer in many ways—
  - corrupt or delete data or files on the computer,
  - change the functionality of software applications,
  - use e-mail program to spread itself to other computers,
  - erase everything on the hard disk, or,
  - degrade performance of the system by utilizing resources such as memory or disk space.
- Virus infects an executable file or program. The virus executes when a program infected with virus is executed or you start a computer from a disk that has infected system files.

- Once a virus is active, it loads into the computer's memory and may save itself to the hard drive or copies itself to applications or system files on the disk.
- However, viruses cannot infect write protected disks or infect written documents. Viruses do not infect an already compressed file. Viruses also do not infect computer hardware; they only infect software.
- Viruses are most easily spread by attachments in e-mail messages. Viruses also spread through download on the Internet.

Some examples of viruses are—"Melissa" and "I Love You".

### 14.3.2 Worms

*Worm* is self-replicating software that uses network and security holes to replicate itself. A copy of the worm scans the network for another machine that has a specific security hole. It copies itself to the new machine using the security hole, and then starts replicating from there, as well. A worm is however different from a virus. A worm does not modify a program like a virus, however, it replicates so much that it consumes the resources of the computer and makes it slow. Some examples of worms are—"Code Red" and "Nimda".

### 14.3.3 Trojan Horse

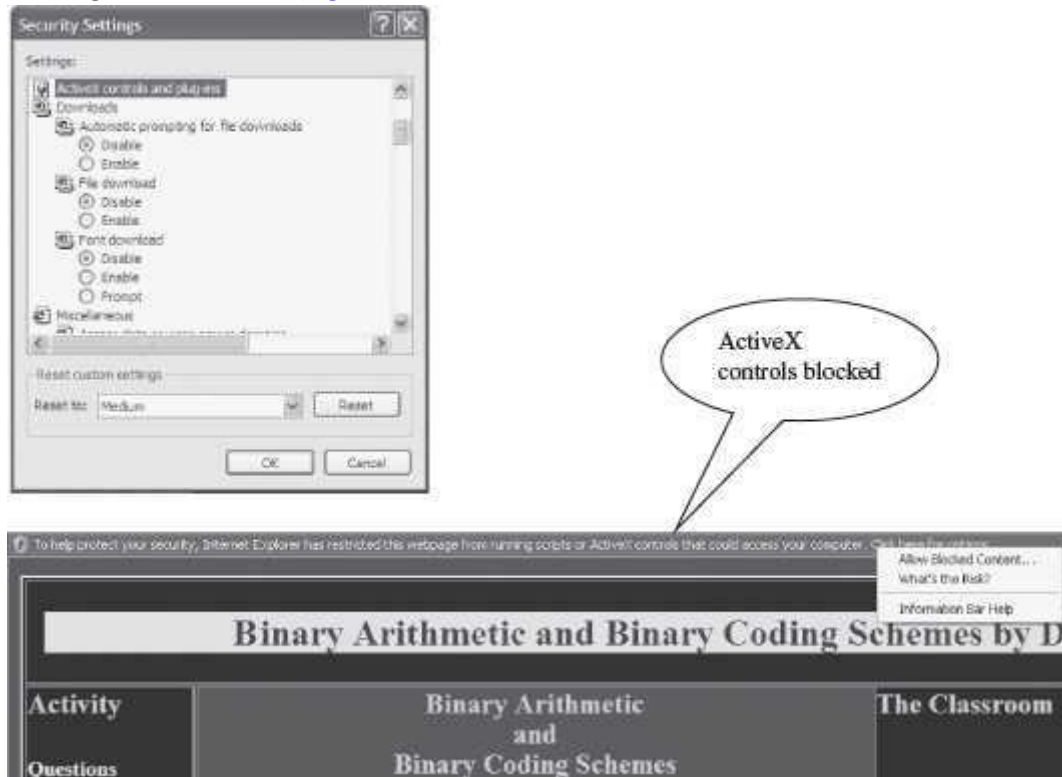
Trojan horse is destructive programs that masquerade as useful programs. The name "Trojan horse" is given because of the Greek soldiers who reached the city of Troy by hiding themselves inside a large wooden horse ([Figure 14.4](#)). The people of the city of Troy themselves pulled the horse inside their city, unaware of the fact that the Greek soldiers were hiding inside the horse. Similarly, users install Trojan horses thinking that it will serve a useful purpose such as a game or provide entertainment. However, Trojan horses contain programs that corrupt the data or damage the files. Trojan horses can corrupt software applications. They can also damage files and can contain viruses that destroy and corrupt data and programs. Trojan horse does not replicate themselves like viruses.



**Figure 14.4** Trojan horse

### 14.3.4 Javascripts, Java Applets and ActiveX Controls

Applets (Java programs), and ActiveX controls are used with Microsoft technology, which can be inserted in a Web page and are downloaded on the client browser for execution. Applets and ActiveX controls are generally used to provide added functionality such as sound and animation. However, these programs when designed with a malicious intention can be disastrous for the client machine. Java Applets have strong security checks that define what an applet can do and what it cannot. ActiveX controls do not have such security checks. Normally, ActiveX controls must be kept disabled while working on the Internet (Figure 14.5).



**Figure 14.5** (a) Making security settings in Windows XP (b) ActiveX control popup in Internet

Javascript is a scripting language generally nested within HTML code. The client-side scripts on a HTML page execute inside the Web browser on the client computer. Javascript codes can be used to transfer files, send e-mails and write to local files. If used with a malignant intention, the scripts can be dangerous for the client machine.

## 14.4 HACKING

Hacking is the act of intruding into someone else's computer or network. A hacker is someone who does hacking. Hacking may result in a *Denial of Service (DoS) attack*. The DoS attack prevents authorized users from accessing the resources of the computer. It aims at making the computer resource unusable or unavailable to its intended users. It targets the computer and its network connections, to prevent the user from accessing email, web sites, online accounts (banking, etc.), or other services that rely on the affected computer. In a DoS attack, the services of the entire network, an Internet site or service, may be suppressed or disabled. The affected machine is flooded with spurious requests and messages so as to overload the network. As a result, the affected machine cannot process

the valid requests. This is a denial of service to the valid users. Generally, the targets of such attacks are the sites hosted on high-profile web servers such as banks and credit card payment gateways.

Packet sniffing, E-mail hacking and Password cracking are used to get the username and password of the system to gain unauthorized access to the system. These methods gather the information when the data is being transmitted over the network.

#### **14.4.1 Packet Sniffing**

The data and the address information are sent as packets over the Internet. The packets may contain data like a user name and password, e-mail messages, files etc. Packet sniffing programs are used to intercept the packets while they are being transmitted from source to destination. Once intercepted, the data in the packets is captured and recorded. Generally, packet sniffers are interested in packets carrying the username and password. Packet sniffing attacks normally go undetected. Ethereal and Zx Sniffer are some freeware packet sniffers. Telnet, FTP, SMTP are some services that are commonly sniffed.

#### **14.4.2 Password Cracking**

Cracking of password is used by hackers to gain access to systems. The password is generally stored in the system in an encrypted form. Utilities like Password cracker is used to crack the encrypted passwords. Password cracker is an application that tries to obtain a password by repeatedly generating and comparing encrypted passwords or by authenticating multiple times to an authentication source.

#### **14.4.3 E-mail Hacking**

The e-mail transmitted over the network contains the e-mail header and the content. If this header and the content are sent without encryption, the hackers may read or alter the messages in transit. Hackers may also change the header to modify the sender's name or redirect the messages to some other user. Hackers use *packet replay* to retransmit message packets over a network. Packet replay may cause serious security threats to programs that require authentication sequences. A hacker may replay the packets containing authentication data to gain access to the resources of a computer.

### **14.5 SECURITY SERVICES**

The security services provide specific kind of protection to system resources. Security services ensure Confidentiality, Integrity, Authentication, and Non-Repudiation of data or message stored on the computer, or when transmitted over the network. Additionally, it provides assurance for access control and availability of resources to its authorized users.

- **Confidentiality**—The confidentiality aspect specifies availability of information to only authorized users. In other words, it is the protection of data from unauthorized disclosure. It requires ensuring the privacy of data stored on a server or transmitted via a network, from being intercepted or stolen by unauthorized users. Data encryption stores or transmits data, in a form that unauthorized users cannot understand. Data encryption is used for ensuring confidentiality.

- **Integrity**—It assures that the received data is exactly as sent by the sender, i.e. the data has not been modified, duplicated, reordered, inserted or deleted before reaching the intended recipient. The data received is the one actually sent and is not modified in transit.
- **Authentication**—Authentication is the process of ensuring and confirming the identity of the user before revealing any information to the user. Authentication provides confidence in the identity of the user or the entity connected. It also assures that the source of the received data is as claimed. Authentication is facilitated by the use of username and password, smart cards, biometric methods like retina scanning and fingerprints.
- **Non-Repudiation** prevents either sender or receiver from denying a transmitted message. For a message that is transmitted, proofs are available that the message was sent by the alleged sender and the message was received by the intended recipient. For example, if a sender places an order for a certain product to be purchased in a particular quantity, the receiver knows that it came from a specified sender. Non-repudiation deals with signatures.
- **Access Control**—It is the prevention of unauthorized use of a resource. This specifies the users who can have access to the resource, and what are the users permitted to do once access is allowed.
- **Availability**—It assures that the data and resources requested by authorized users are available to them when requested.

#### 14.6 SECURITY MECHANISMS

Security mechanisms deal with prevention, detection, and recovery from a security attack. Prevention involves mechanisms to prevent the computer from being damaged. Detection requires mechanisms that allow detection of when, how, and by whom an attack occurred. Recovery involves mechanism to stop the attack, assess the damage done, and then repair the damage.

Security mechanisms are built using personnel and technology.

- Personnel are used to frame security policy and procedures, and for training and awareness.
- Security mechanisms use technologies like cryptography, digital signature, firewall, user identification and authentication, and other measures like intrusion detection, virus protection, and, data and information backup, as countermeasures for security attack.

#### 14.7 CRYPTOGRAPHY

Cryptography is the science of writing information in a “hidden” or “secret” form and is an ancient art. Cryptography is necessary when communicating data over any network, particularly the Internet. It protects the data in transit and also the data stored on the disk. Some terms commonly used in cryptography are:

- Plaintext is the original message that is an input, i.e. unencrypted data.
- *Cipher and Code*—Cipher is a bit-by-bit or character-by-character transformation without regard to the meaning of the message. Code replaces one word with another word or symbol. Codes are not used any more.
- *Cipher text*—It is the coded message or the encrypted data.

- **Encryption**—It is the process of converting plaintext to cipher text, using an encryption algorithm.
- **Decryption**—It is the reverse of encryption, i.e. converting cipher text to plaintext, using a decryption algorithm.

Cryptography uses different schemes for the encryption of data. These schemes constitute a pair of algorithms which creates the encryption and decryption, and a key.

**Key** is a secret parameter (string of bits) for a specific message exchange context. Keys are important, as algorithms without keys are not useful. The encrypted data cannot be accessed without the appropriate key. The size of key is also important. The larger the key, the harder it is to crack a block of encrypted data. The algorithms differ based on the number of keys that are used for encryption and decryption. The three cryptographic schemes are as follows:

- **Secret Key Cryptography (SKC)**: Uses a single key for both encryption and decryption,
- **Public Key Cryptography (PKC)**: Uses one key for encryption and another for decryption, □ **Hash Functions**: Uses a mathematical transformation to irreversibly encrypt information.

*In all these schemes, algorithms encrypt the plaintext into cipher text, which in turn is decrypted into plaintext.*

#### 14.7.1 Secret Key Cryptography

- Secret key cryptography uses a single key for both encryption and decryption. The sender uses the key to encrypt the plaintext and sends the cipher text to the receiver. The receiver applies the same key to decrypt the message and recover the plaintext ([Figure 14.6](#)). Since a single key is used for encryption and decryption, secret key cryptography is also called *symmetric encryption*.



**Figure 14.6** Secret key cryptography (uses a single key for both encryption and decryption)

- Secret key cryptography schemes are generally categorized as *stream ciphers* or *block ciphers*.
- *Stream ciphers* operate on a single bit (byte or computer word) at a time and implement some form of feedback mechanism so that the key is constantly changing.
- *Block cipher* encrypts one block of data at a time using the same key on each block. In general, the same plaintext block will always encrypt to the same cipher text when using a same key in a block cipher.
- Secret key cryptography requires that the key must be known to both the sender and the receiver. The drawback of using this approach is the distribution of the key. Any person who has the key can use it to decrypt a message. So, the key must be sent securely to the receiver, which is a problem if the receiver and the sender are at different physical locations.

- Data Encryption Standard (DES) and Advanced Encryption Standard (AES) are some of the secret key cryptography algorithms that are in use nowadays.

### 14.7.2 Public-Key Cryptography

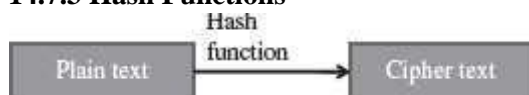
- Public-key cryptography facilitates secure communication over a non-secure communication channel without having to share a secret key.
- Public-key cryptography uses two keys—one public key and one private key.
- The public key can be shared freely and may be known publicly.
- The private key is never revealed to anyone and is kept secret.
- The two keys are mathematically related although knowledge of one key does not allow someone to easily determine the other key.



**Figure 14.7** Public key cryptography (uses two keys—one for encryption and other for decryption)

- The plaintext can be encrypted using the public key and decrypted with the private key and conversely the plaintext can be encrypted with the private key and decrypted with the public key. Both keys are required for the process to work ([Figure 14.7](#)). Because a pair of keys is required for encryption and decryption; public-key cryptography is also called *asymmetric encryption*.
- Rivest, Shamir, Adleman (RSA) is the first and the most common public-key cryptography algorithm in use today. It is used in several software products for key exchange, digital signatures, or encryption of small blocks of data. The Digital Signature Algorithm (DSA) is used to provide digital signature capability for the authentication of messages.

### 14.7.3 Hash Functions



**Figure 14.8** Hash function (have no key since plain text is not recoverable from cipher text)

- Hash functions are one-way encryption algorithms that, in some sense, use no key. This scheme computes a fixed-length hash value based upon the plaintext. Once a hash function is used, it is difficult to recover the contents or length of the plaintext ([Figure 14.8](#)).
- Hash functions are generally used to ensure that the file has not been altered by an intruder or virus. Any change made to the contents of a message will result in the receiver calculating a different hash value than the one placed in the transmission by the sender.
- Hash functions are commonly employed by many operating systems to encrypt passwords. Message Digest (MD) algorithm and Secure Hash Algorithm (SHA) are some of the common used hash algorithms.

The different cryptographic schemes are often used in combination for a secure transmission. Cryptography is used in applications like, security of ATM cards, computer passwords, and electronic commerce. Cryptography is used to protect data from theft or alteration, and also for user authentication.

*Certification Authorities (CA)* are necessary for widespread use of cryptography for e-commerce applications. CAs are trusted third parties that issue digital certificates for use by other parties. A CA issues digital certificates which contains a public key, a name, an expiration date, the name of authority that issued the certificate, a serial number, any policies describing how the certificate was issued, how the certificate may be used, the digital signature of the certificate issuer, and any other information.

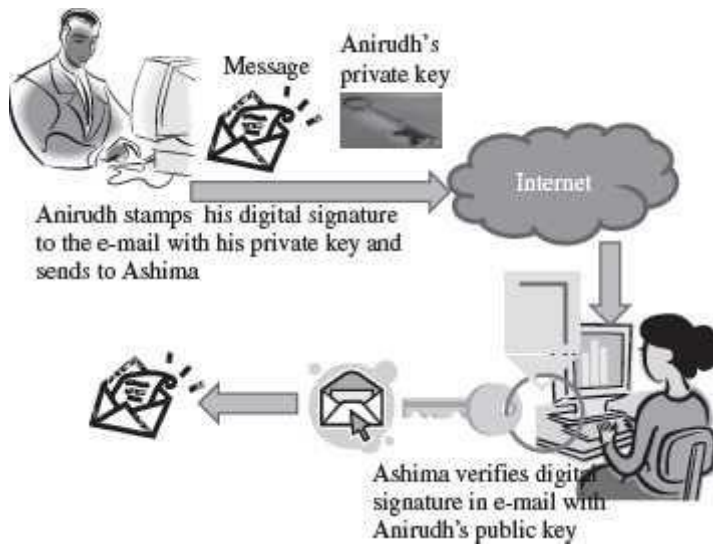
## 14.8 DIGITAL SIGNATURE

A signature on a legal, financial or any other document authenticates the document. A photocopy of that document does not count. For computerized documents, the conditions that a signed document must hold are—(1) The receiver is able to verify the sender (as claimed), (2) The sender cannot later repudiate the contents of the message, (3) The receiver cannot concoct the message himself. A digital signature is used to sign a computerized document. The properties of a digital signature are same as that of ordinary signature on a paper. Digital signatures are easy for a user to produce, but difficult for anyone else to forge. Digital signatures can be permanently tied to the content of the message being signed and then cannot be moved from one document to another, as such an attempt will be detectable.

Digital signature scheme is a type of asymmetric cryptography. Digital signatures use the publickey cryptography, which employs two keys—private key and public key. The digital signature scheme typically consists of three algorithms:

- *Key generation algorithm*—The algorithm outputs private key and a corresponding public key. □
- *Signing algorithm*—It takes, message + private key, as input, and, outputs a digital signature.
- *Signature verifying algorithm*—It takes, message + public key + digital signature, as input, and, accepts or rejects digital signature.

The use of digital signatures typically consists of two processes—Digital signature creation and Digital signature verification ([Figure 14.9](#)). Two methods are commonly used for creation and verification of the digital signatures.



**Figure 14.9** Digital signature

- In the First Method, the signer has a private key and a public key. For a message to be sent, the signer generates the digital signature by using the private key to encrypt the message. The digital signature along with the message is sent to the receiver. The receiver uses the public key (known to the receiver) to verify the digital signature. This method is used to verify the digital signature. Even if many people may know the public key of a given signer and use it to verify that signer's signature, they cannot generate the signer's private key and use it to forge digital signatures.
- In the Second Method, a hash function is used for digital signature. It works as follows:
  - Digital signature creation
    - The signer has a private key and a public key.
    - For a message to be sent, a hash function in the signer's software computes an "original hash result" unique to the "original message".
    - The signer uses signing algorithm to generate a unique digital signature.  
"original hash result" + signer's private key = digital signature.
  - The generated digital signature is attached to its "original message" and transmitted with it. ○ *Digital signature verification* uses digital signature, "received message" and signer's public key.
    - A "new hash result" of the "received message" is computed using the same hash function used for the creation of the digital signature.
    - The verification software verifies two things—whether the digital signature was created using the signer's private key and, whether the "received message" is unaltered. For this, the signer's public key verifies the digital signature (signer's public key can only verify a digital signature created with the signer's private key). Once the key is verified, the "original hash result" of the digital signature is available. It compares "original hash result" with the "new hash result". When the verification software verifies both the steps as "true"; it verifies the received message.

The digital signature accomplish the effects desired of a signature for many legal purposes:

- **Signer Authentication:** The digital signature cannot be forged, unless the signer loses control of the private key.
- **Message Authentication:** The digital signature verification reveals any tampering, since the comparison of the hash results shows whether the message is the same as when signed.
- **Efficiency.** The digital signatures yield a high degree of assurance (as compared to paper methods like checking specimen signatures) without adding much to the resources required for processing.

The likelihood of malfunction or a security problem in a digital signature cryptosystem, designed and implemented as prescribed in the industry standards, is extremely remote. Digital signatures have been accepted in several national and international standards developed in cooperation with and accepted by many corporations, banks, and government agencies. In India “Information Technology Act 2000” provides legal recognition for transactions carried out by means of electronic data interchange and other means of electronic communication, commonly referred to as “electronic commerce”, which involves the use of alternatives to paper based methods of communication and storage of information, to facilitate electronic filing of documents with the government agencies.

#### 14.9 FIREWALL

A firewall is a security mechanism to protect a local network from the threats it may face while interacting with other networks (Internet). A firewall can be a hardware component, a software component, or a combination of both. It prevents computers in one network domain from communicating directly with other network domains. All communication takes place through the firewall, which examines all incoming data before allowing it to enter the local network ([Figure 14.10](#)).

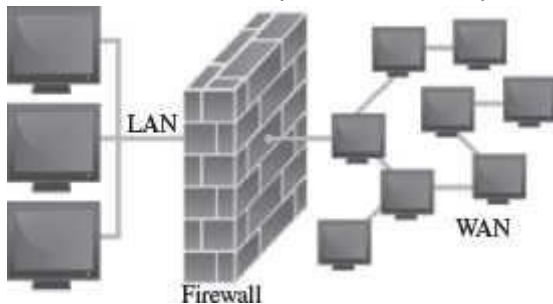
**Functions of Firewall**—The main purpose of firewall is to protect computers of an organization (local network) from unauthorized access. Some of the basic functions of firewall are:

- Firewalls provide security by examining the incoming data packets and allowing them to enter the local network only if the conditions are met ([Figure 14.11](#)).
- Firewalls provide user authentication by verifying the username and password. This ensures that only authorized users have access to the local network.
- Firewalls can be used for hiding the structure and contents of a local network from external users. Network Address Translation (NAT) conceals the internal network addresses and replaces all the IP addresses of the local network with one or more public IP addresses.



**Figure 14.10** (a) Windows firewall icon in control panel (b) Windows firewall setting (c) Security center

The local network uses a single network interface to interact with the server. Local network clients use IP addresses that are not attached to any computer. When a client sends a packet to the Internet, the masquerading server replaces the IP address of the packet with its own IP address. When a packet is received by local network, the server replaces the IP address of the packet with the masqueraded address and sends the packet to the respective client.



**Figure 14.11** Firewall

**Working of Firewall**—The working of firewall is based on a filtering mechanism. The filtering mechanism keeps track of source address of data, destination address of data and contents of data. The filtering mechanism allows information to be passed to the Internet from a local network without any authentication. It makes sure that the downloading of information from the Internet to a local network happens based only on a request by an authorized user.

#### **Firewall Related Terminology:**

- **Gateway**—The computer that helps to establish a connection between two networks is called gateway. A firewall gateway is used for exchanging information between a local network and the Internet.
- **Proxy Server**—A proxy server masks the local network's IP address with the proxy server IP address, thus concealing the identity of local network from the external network. Web proxy and application-level gateway are some examples of proxy servers. A firewall can be deployed with the proxy for protecting the local network from external network.

- *Screening Routers*—They are special types of router with filters, which are used along with the various firewalls. Screening routers check the incoming and outgoing traffic based on the IP address, and ports.

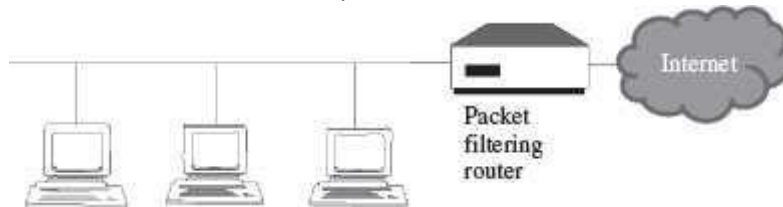
### 14.9.1 Types of Firewall

All the data that enter a local network must come through a firewall. The type of firewall used varies from network to network. The following are the various types of firewalls generally used:

- Packet filter Firewall
- Circuit Filter Firewall
- Proxy server or Application-level Gateway

#### 14.9.1.1 Packet Filter Firewall

Packet Filter Firewall is usually deployed on the routers ([Figure 14.12](#)). It is the simplest kind of mechanism used in firewall protection.



**Figure 14.12** Packet filtering

- It is implemented at the network level to check incoming and outgoing packets.
- The IP packet header is checked for the source and the destination IP addresses and the port combinations.
- After checking, the filtering rules are applied to the data packets for filtering. The filtering rules are set by an organization based on its security policies.
- If the packet is found valid, then it is allowed to enter or exit the local network.
- Packet filtering is fast, easy to use, simple and cost effective.
- A majority of routers in the market provide packet filtering capability. It is used in small and medium businesses.
- Packet filter firewall does not provide a complete solution.

#### 14.9.1.2 Circuit Filter Firewall

Circuit filter firewalls provide more protection than packet filter firewalls. Circuit filter firewall is also known as a “stateful inspection” firewall.

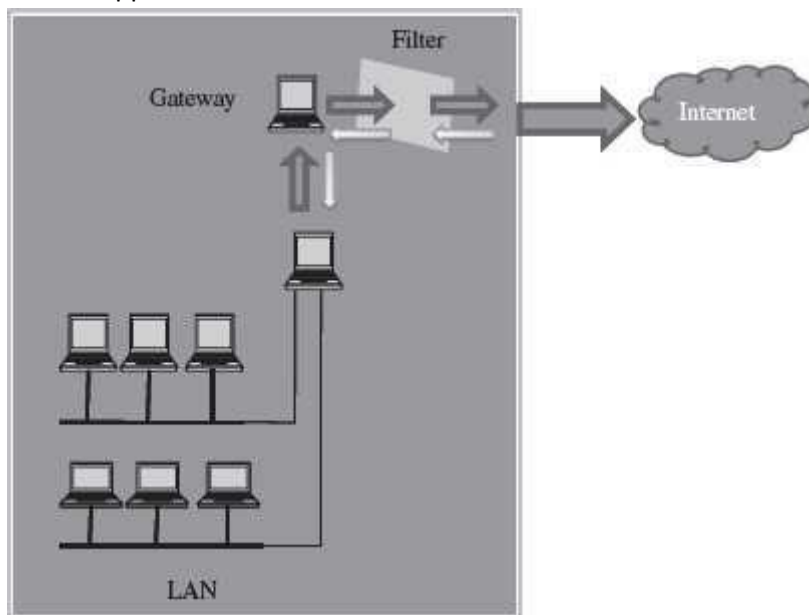
- It prevents transfer of suspected packets by checking them at the network layer.
- It checks for all the connections made to the local network, in contrast, to the packet filter firewall which makes a filtering decision based on individual packets.

- It takes its decision by checking all the packets that are passed through the network layer and using this information to generate a decision table. The circuit level filter uses these decisions tables to keep track of the connections that go through the firewall.
- For example, when an application that uses TCP creates a session with the remote host, the TCP port number for the remote application is less than 1024 and the TCP port number for the local client is between 1024 and 65535. A packet filter firewall will allow any packet which has a port number within the range 1024 and 65535. However, the circuit filter firewall creates a directory of all outbound TCP connections. An incoming packet is allowed if its profile matches with an entry in the directory for the TCP port numbers.

#### 14.9.1.3 Application-Level Gateway

An application-level gateway or a proxy server protects all the client applications running on a local network from the Internet by using the firewall itself as the gateway ([Figure 14.13](#)).

- A proxy server creates a virtual connection between the source and the destination hosts.
- A proxy firewall operates on the application layer. The proxy ensures that a direct connection from an external computer to local network never takes place.
- The proxy automatically segregates all the packets depending upon the protocols used for them. A proxy server must support various protocols. It checks each application or service, like Telnet or e-mail, when they are passed through it.
- A proxy server is easy to implement on a local network.
- Application level gateways or proxy server tend to be more secure than packet filters. Instead of checking the TCP and IP combinations that are to be allowed, it checks the allowable applications.



**Figure 14.13** Application-level gateway

## 14.10 USERS IDENTIFICATION AND AUTHENTICATION

*Identification* is the process whereby a system recognizes a valid user's identity. Authentication is the process of verifying the claimed identity of a user. For example, a system uses userpassword for identification. The user enters his password for identification. Authentication is the system which verifies that the password is correct, and thus the user is a valid user. Before granting access to a system, the user's identity needs to be authenticated. If users are not properly authenticated then the system is potentially vulnerable to access by unauthorized users. If strong identification and authentication mechanisms are used, then the risk that unauthorized users will gain access to a system is significantly decreased. Authentication is done using one or more combinations of—what you have (like smartcards), what you know (Password), and what you are (Biometrics like Fingerprints, retina scans).

We will now briefly discuss the following authentication mechanisms:

- User name and password
- Smart Card
- Biometrics—Fingerprints, Iris/retina scan

Once the user is authenticated, the access controls for the user are also defined. Access controls is what the user can access once he is authenticated.

### 14.10.1 User Name and Password

The combination of username and password is the most common method of user identification and authentication. The systems that use password authentication first require the user to have a username and a password. Next time, when the user uses the system, user enters their username and password. The system checks the username and password by comparing it to the stored password for that username. If it matches, the user is authenticated and is granted access to the system ([Figure 14.14](#)).

However, there are several security issues with the use of password, like, any invalid user if gets to know of a valid password can get access to the system, a simple password can be easily cracked etc. According to CERT, approximately 80% of all network security issues are caused by bad passwords. Some actions that can be taken to make the passwords safer are as follows:

- Nearly all modern multiuser computer and network operating systems, at the very least, employ passwords to protect and authenticate users accessing computer and network resources. The passwords are not kept in plaintext, but are generally encrypted using some sort of hash scheme. For example, In Unix/ Linux, all passwords are hashed and stored as a 13-byte string. In Windows NT, all passwords are hashed resulting in a 16-byte hash value.

A smart card is a pocket-sized card with embedded integrated circuits which can process data. With an embedded microcontroller, smart cards have the unique ability to store large amounts of data, carry out

their own on-card functions (e.g. encryption and mutual authentication) and interact intelligently with a smart card reader. A smart card inserted into a smart card reader makes a direct connection to a conductive contact plate on the surface of the card (typically gold plated). Transmission of commands, data, and card status takes place over these physical contact points.

The smart card is made of plastic, generally PVC. The card may embed a hologram. Using smart cards is a strong security authentication for single sign-on within large companies and organizations. Smart cards are used in secure identity applications like employee-ID badges, citizen-ID documents, electronic passports, driver license and online authentication devices.

### 14.10.3 Biometric Techniques

Biometrics is the science and technology of measuring and statistically analyzing biological data. In information technology, biometrics refers to technologies that measures and analyzes human traits for authentication. This can include fingerprints, eye retinas and irises, voice patterns, facial patterns and hand measurements, for authentication purposes. [Figure 14.15](#) shows a fingerprint biometric device.

Biometrics is still not widely used, though it may play a critical role in future computers. For example, many PCs nowadays include a fingerprint scanner where you could place your index finger. The computer analyzes the fingerprint to determine your identity and authenticate you. Biometric systems are relatively costly and are used in environments requiring high-level security.

In the Hindi movie *Krishh*, the computer identified and authenticated the heartbeat (Biometric) of *Hrithik Roshan* to start working.



**Figure 14.15** Biometric device (fingerprint)

### 14.11 OTHER SECURITY MEASURES

In addition to the above discussed security techniques, several other security techniques are used for security purposes. Some of these are listed below:

- **Intrusion Detection Systems**—They complement firewalls to detect if internal assets are being hacked or exploited. A Network-based Intrusion Detection monitors real-time network traffic for malicious activity and sends alarms for network traffic that meets certain attack patterns or signatures. A Host-based Intrusion Detection monitors computer or server files for anomalies and sends alarms for network traffic that meets a predetermined attack signature.
- **Virus Protection Software**—They should be installed on all network servers, as well as computers. They screen all software coming into your computer or network system (files, attachments, programs, etc.) preventing a virus from entering into the system.
- **Data and Information Backups**—It is required for disaster recovery and business continuity. Back-ups should be taken daily and periodically (weekly) and should be kept for at least 30 days while rotating stockpile.
- **Secure Socket Layer (SSL)** is an algorithm developed by Netscape Communications to provide application-independent security and privacy over the Internet. SSL is designed so that protocols such as HTTP, FTP, and Telnet can operate over it transparently. SSL allows both server authentication (mandatory) and client authentication (optional). It uses public-key cryptography (RSA algorithm). *HTTP Secure* (HTTPS) is an extension to HTTP to provide secure exchange of documents over the WWW
- **IP Security (IPsec) Protocol**—The IPsec protocol suite is used to provide privacy and authentication services at the Internet layer. IPv4 is currently the dominant Internet Protocol version. IPv6 is the next-generation Internet Layer protocol for the Internet. IPv6 protocol stacks include IPsec, which allows authentication, encryption, and compression of IP traffic. IPsec can be used to protect any application traffic across the Internet. Applications need not be specifically designed to use IPsec, unlike SSL where the use of SSL must be incorporated into the design of application.

#### 14.12 SECURITY AWARENESS

The aim of the security awareness is to enhance the security of the organization's resources by improving the awareness of the need to secure the system resources. Staff members play a critical role in protecting the integrity, confidentiality, and availability of IT systems and networks. It is necessary for an organization to train their staff for security awareness and accepted computer practices. Security of resources can be ensured when the people using it are aware of the need to secure their resources. Security awareness of staff includes the knowledge of practices that must be adhered to, for ensuring the security and the possible consequences of not using those security practices. For example, not disclosing your password to unauthorized users is a security practice, but if the users are not aware of the possible consequences of disclosing the password, they may disclose their password to other users, unintentionally, thus making their systems prone to security attack. In order to make the users and people in an organization aware of the security practices to be followed, regular training programs are conducted in organizations. Awareness is also promoted by regular security awareness sessions, videotapes, newsletters, posters, and flyers. [Figure 14.16](#) shows a poster for security awareness.



**Figure 14.16** Security awareness (A poster)

### 14.13 SECURITY POLICY

- A *security policy* is a formal statement that embodies the organization's overall security expectations, goals, and objectives with regard to the organization's technology, system and information.
- To be practical and implementable, policies must be defined by standards, guidelines, and procedures. Standards, guidelines, and procedures provide specific interpretation of policies and instruct users, customers, technicians, management, and others on how to implement the policies.
- The security policy states what is, and what is not allowed. A security policy must be comprehensive, up-to-date, complete, delivered effectively, and available to all staff. A security policy must also be enforceable. To accomplish this, the security policy can mention that strict action will be taken against employees who violate it, like disclosing a password.

- Generally, security policies are included within a *security plan*. A security plan details how the rules put forward by the security policy will be implemented. The statements within a security plan can ensure that each employee knows the boundaries and the penalties of overstepping those boundaries. For example, some rules could be included in the security policy of an organization, such as, to log off the system before leaving the workstation, or not to share the password with other users.
- The security policy also includes physical security of the computers. Some of the measures taken to ensure the physical security of a computer are—taking regular backups to prevent data loss from natural calamity, virus attack or theft, securing the backup media, keeping valuable hardware resources in locked room (like servers), to avoid theft of systems and storage media.

#### 14.13.1 Formulating a Security Policy

Security policies are defined based on an organization's needs. A security policy includes approaches and techniques that an organization is going to apply or include in order to secure its resources. The steps followed while formulating the security policy are:

- *Analyzing Current Security Policies*—The vulnerabilities and the current security policies must be analyzed by the security administrators before defining an effective security policy. The security administrator is required to study the existing documents containing details of the physical security policies, network security policies, data security policies, disaster recovery plans, and contingency plans.
- *Identifying IT Assets that Need to be Secure*—The security administrator must identify the IT resources of an organization that need to be secure. It may include the following:
  - Physical resources like computers, servers like database servers and web servers, local networks that are used to share the local computer with the remote computer, private networks shared by two or more organizations, corporate network permanently connected to the Internet, laptop, manuals, backup media, communication equipment, network cables, and CDs.
  - Information resources like password, data, or applications. The data of an organization can be classified for security purposes based upon the sensitivity and the integrity of data. For example, public information, internal information, confidential information, and secret information
- *Identifying Security Threats and Likely Security Attacks*—After identifying the IT assets and classifying them, a security administrator must identify the various security threats to the assets. For example, in a bank the security threat to the database storing the account details of the customers may be—unauthorized access to information, attacks of viruses, worms and Trojan horses, natural disasters like earthquake, fire etc.
- *Defining the Proactive and Reactive Security Strategies*—A *proactive strategy* is a pre-attack strategy. It involves identifying possible damage from each type of attack, determining the vulnerabilities that each type of attack can exploit, minimizing those vulnerabilities and making a contingency plan. A contingency plan specifies the actions to be taken in case an attack penetrates into a system and damages the IT assets of the organization. A contingency plan aims at keeping the computer functional and ensuring the availability, integrity, and confidentiality of data. However, it is not possible for the security administrator to prepare a computer against all attacks. A *reactive strategy* is implemented on the failure of the proactive strategy. It defines

the steps to be taken after the attack. It aims at identifying the cause of attack, vulnerabilities used to attack the system, damage caused by the attack, and repairing of the damage caused by the attack.