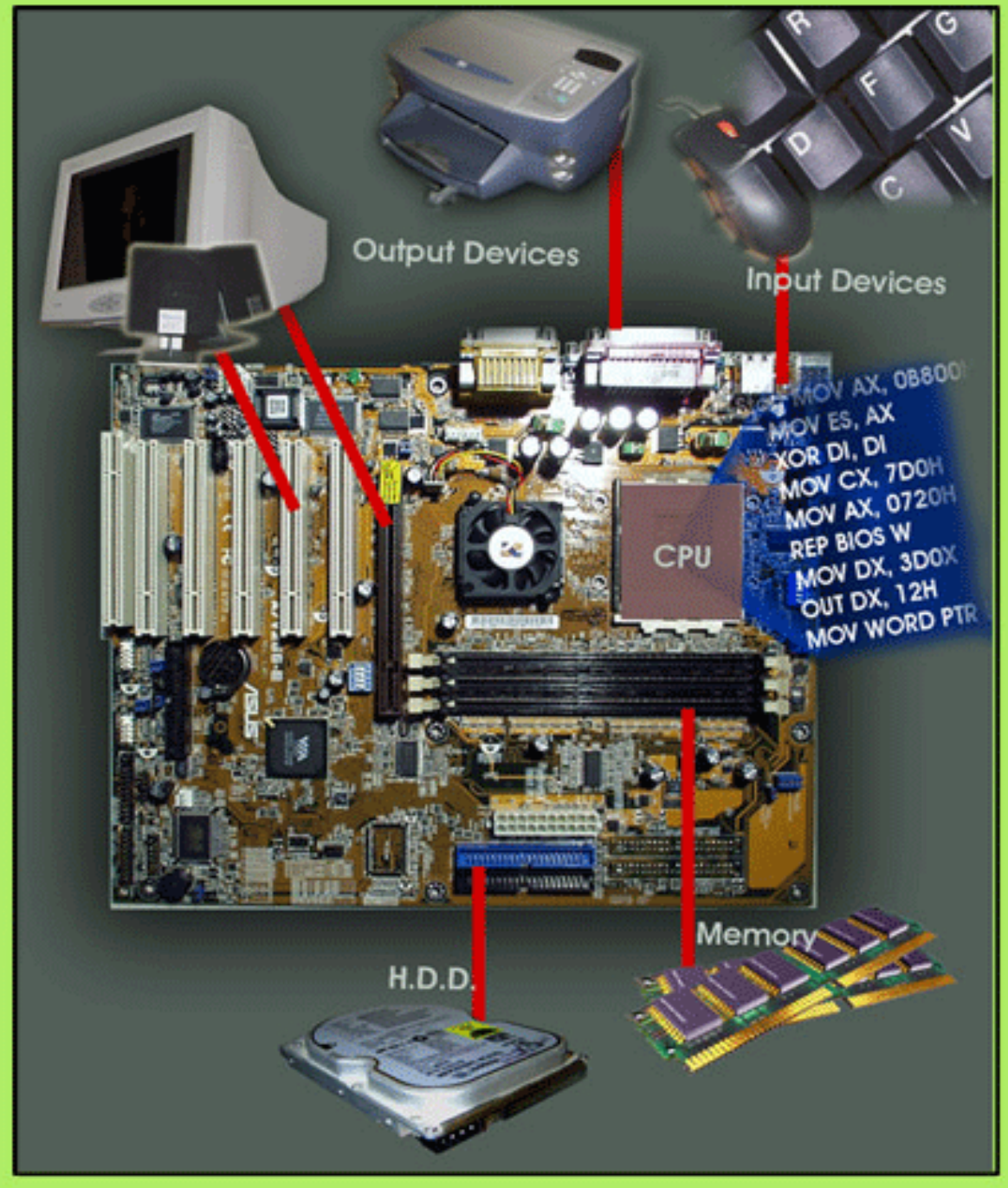
Indira Gandhi National Open University
School of Computer and Information Sciences

MCS - 012

Computer Organization and Assembly Language Programming



2

BASIC COMPUTER ORGANISATION

Indira Gandhi
National Open University
School of Computer and
Information Sciences

MCS-012
COMPUTER ORGANISATION & ASSEMBLY LANGUAGE PROGRAMMING

Block

2

BASIC COMPUTER ORGANISATION

Programme / Course Design Committee

Prof. Sanjeev K. Aggarwal, IIT, Kanpur
Prof. M. Balakrishnan, IIT , Delhi
Prof. Harish Karnick, IIT, Kanpur
Prof. C. Pandurangan, IIT, Madras
Dr. Om Vikas, Sr. Director, MIT
Prof. P. S. Grover, Sr. Consultant, SOCIS, IGNOU

Faculty of School of Computer and Information Sciences
Shri Shashi Bhushan
Shri Akshay Kumar
Prof Manohar Lal
Shri V.V. Subrahmanyam
Shri P.Venkata Suresh

Block Preparation Team

Ms. Anupama Jha
Deptt. of Computer Science
Rajdhani College
New Delhi

Mr. Amitabh Trehan
IT Consultant, Sarai/CSDS
New Delhi

Mr. Akshay Kumar
SOCIS, IGNOU

Prof. MPS Bhatia (Content Editor)
(NSIT), New Delhi

Prof.A.K.Verma (Language Editor)

Course Coordinator: Shri Akshay Kumar

Block Production Team

Shri H.K. Som, SOCIS

July, 2004

ISBN—81-266-1343-2

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, SOCIS.

BLOCK INTRODUCTION

In the first block of this course, we have discussed the basic concepts relating to Computer Organization, including von-Neumann architecture, data representation, instruction execution, digital logic circuit design and other similar concepts.

In this Block, we will discuss the Memory and Input-Output system and technologies of the Computer systems.

Unit 1 covers aspects relating to the memory system of the computer. The memory system is quite extensive in terms of technologies; therefore, this unit primarily focuses on the basic concepts and technologies.

Unit 2 revolves round the Input/Output System; it covers aspects relating to the Input-Output techniques used for transfer of information from the system to the external world. It also covers the concepts of Direct Memory Access and Input / Output processor.

Unit 3 provides details on the existing technologies available in the market with respect to the secondary storage devices. Thus, this unit provides an overall picture of market trends in this area.

Unit 4 is a detailed description of various Input/Output technologies available. This provides some of the basic input about the I/O technologies such as keyboard, mouse, printers, display etc.

Thus, this block provides information about the theoretical and practical aspects of the Memory system and about Input /Output technologies.

A course on computers can never be complete because of the existing diversities of computer systems. Therefore, you are advised to read through further readings to enhance the basic understanding you will acquire from the block. Although you may refer to the same textbooks as given in the previous block, please keep on locating and visiting various concepts from the websites in order to update your information.

Further Readings

1. Stallings W, *Computer Organization & Architecture: Designing For Performance*, 6th Edition, Prentice Hall of India Publication, 2002/ Pearson Education Asia 2003
2. Mano M Morris, *Computer System Architecture*, 3rd Edition, Prentice Hall of India Publication, 2001 / Pearson Education Asia 2003
3. Hennessy/ Patterson, *Computer Architecture: A Quantitative. Approach; 3rd Edition, Morgan Kaufmann, 2003.*

Note: You must try to obtain the latest editions of these books.

UNIT 1 THE MEMORY SYSTEM

Structure **Page Nos.**

1.0 INTRODUCTION

In the previous Block, we have touched upon the basic foundation of computers, which include concepts on von Neumann machine, instruction, execution, the digital data representation and logic circuits. In this Block we will define some of the most important component units of a computer, which are the memory unit and the input-output units. In this unit we will discuss various components of the memory system of a computer system. Computer memory is organised into a hierarchy to minimise cost. Also, it does not compromise the overall speed of access. Memory hierarchy include cache memory, main memory and other secondary storage technologies. In this Unit, we will discuss the main memory, the secondary memory and high-speed memories such as cache memory, and the memory system of microcomputer.

1.1 OBJECTIVES

After going though this Unit, you will be able to:

- describe the key characteristics of the memory system;
- distinguish among various types of random access memories;
- describe the latest secondary storage technologies;
- describe the importance of cache memory and other high-speed memories; and
- describe the different memory chips of micro computers.

1.2 THE MEMORY HIERARCHY

Memory in a computer system is required for storage and subsequent retrieval of the instructions and data. A computer system uses a variety of devices for storing these instructions and data that are required for its operation. Normally we classify the information to be stored into two basic categories: Data and Instructions. But what is a memory system?

"The storage devices along with the algorithm or information on how to control and manage these storage devices constitute the memory system of a computer."

A memory system is a very simple system, yet it exhibits a wide range of technology and types. The basic objective of a computer system is to increase the speed of computation. Likewise the basic objective of a memory system is to provide fast, uninterrupted access by the processor to the memory such that the processor can operate at the speed it is expected to work.

But does this kind of technology where there is no speed gap between processor and memory speed exist? The answer is yes, it does. Unfortunately as the access time (time taken by CPU to access a location in memory) becomes less the cost per bit of memory becomes higher. In addition, normally these memories require power supply till the information needs to be stored. Both these things are not very convenient, but on the other hand the memories with smaller cost have very high access time that will result is in slower operation of the CPU. Thus, the cost versus access time anomaly has led to a hierarchy of memories where we supplement fast memories with larger, cheaper, slower memories. These memory units may have very different physical and operational characteristics; therefore, the memory system is very diverse in type, cost, organisation, technology and performance. This memory hierarchy will work only if the frequency of access to the slower memories is significantly less than the faster memories. The memory hierarchy system consists of all storage devices employed in a computer system from the slow but high capacity auxiliary memory to a relatively faster main memory, to an even smaller and faster cache memory accessible to the high speed registers and processing logic. Figure 1 illustrates the components of a typical memory system.

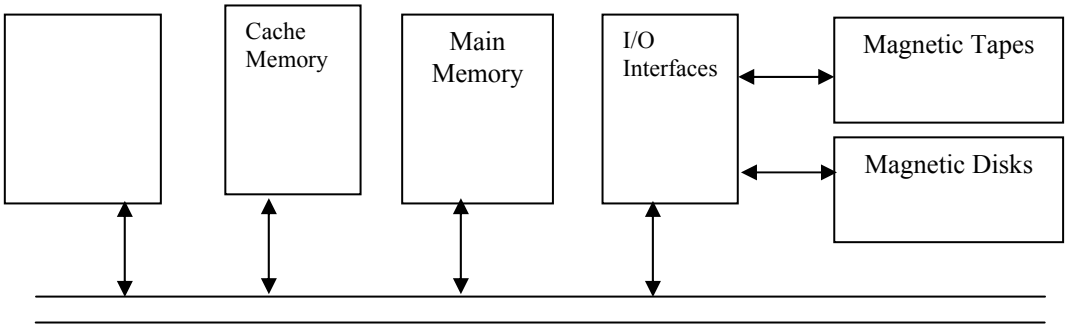


Figure 1: The Memory Hierarchy (Block Diagram)

A typical storage hierarchy is shown in Figure 1 above. Although Figure 1 shows the block diagram, it includes the storage hierarchy:

Register
Cache memory
Main memory
Secondary Storage and
Mass Storage.

As we move up the hierarchy, we encounter storage elements that have faster access time, higher cost per bit stored, and slower access time as a result of moving down the hierarchy. Thus, cache memory generally has the fastest access time, the smallest storage capacity, and the highest cost per bit stored. The primary memory (main memory) falls next in the storage hierarchy list. On-line, direct-access secondary storage devices such as magnetic hard disks make up the level of hierarchy just below the main memory. Off-line, direct-access and sequential access secondary storage devices such as magnetic tape, floppy disk, zip disk, WORM disk, etc. fall next in the storage hierarchy. Mass storage devices, often referred to as archival storage, are at

the bottom of the storage hierarchy. They are cost-effective for the storage of very large quantities of data when fast access time is not necessary.

Please note two important points here:

- The size of the memory increases as we move down the hierarchy.
- The quantum of data that is transferred between two consecutive memory layers at a time also increases as we go from a higher to lower side. For example, from main memory to Cache transfer one or few memory words are accessed at a time, whereas in a hard disk to main memory transfer, a block of about 1 Megabyte is transferred in a single access. You will learn more about this in the later sections of the unit.

Let us now discuss various forms of memories in the memory hierarchy in more details.

1.3 RAM, ROM, DRAM, FLASH MEMORY

RAM (Random Access Memory)

The main memory is a random access memory. It is normally organised as words of fixed length. The length of a word is called word length. Each of these memory words has an independent address and each has the same number of bits. Normally the total numbers of words in memory are some power of 2. Some typical memory word sizes are 8 bits, 16 bits, 32 bits etc. The main memory can be both read and written into, therefore it is called read-write memory.

The access time and cycle time in RAMs are constant and independent of the location accessed. How does this happen? To answer this, let us first discuss how a bit can be stored using a sequential circuit. The logic diagram of a binary cell is shown in Figure 2a:

Figure 2(a) Logic Diagram of RAM cell

The construction shown in Figure 2(a) is made up of one JK flip-flop and 3 AND gates. The two inputs to the system are one input bit and read/write signal. Input is fed in complemented form to AND gate ‘a’. The read/write signal has a value of 1 if it is a read operation. Therefore, during the read operation the AND gate ‘c’ has the read/write input as 1. Since AND gate ‘a’ and ’b’ have 0 read/write input, and if the

chip is selected i.e. this cell is currently being selected, then output will become equal to the state of flip-flop. In other words the data value stored in flip-flop has been read. In write operation only 'a' and 'b' gates get a read/write value of 1 and they set or clear the JK flip-flop depending on the data input value. If the data input is 0, the flip-flop will go to clear state and if data input is 1, the flip-flop will go to set state. In effect, the input data is reflected in the state of the flip-flop. Thus, we say that the input data has been stored in flip-flop or binary cell.

Figure 2(b) Internal Organisation of a 32 × 4 RAM

A 32 × 4 RAM means that this RAM has 32 words, 5 address lines ($2^5 = 32$), and 4 bit data word size. Please note that we can represent a RAM using $2^A \times D$, where A is the number of address lines and D is the number of Data lines. Figure 2 (b) is the extension of the binary cell to an integrated 32 × 4 RAM circuit where a 5 × 32 bit decoder is used. The 4 bit data inputs come through an input buffer and the 4-bit data output is stored in the output buffer.

A chip select ($\overline{CS}$) control signal is used as a memory enable input. When CS = 0 that is $\overline{CS} = 1$, it enables the entire chip for read or write operation. A R/W signal can be used for read or write operation. The word that is selected will determine the overall output. Since all the above is a logic circuit of equal length that can be accessed in equal time, thus, the word RAM.

DRAM (Dynamic Random Access Memory)

RAM technology is divided into two technologies: dynamic and static. A dynamic RAM (DRAM) is made with cells that store data as charge on capacitors. The presence or absence of charge on capacitor is interpreted as binary 1 or 0. Because capacitors have a natural tendency to discharge, dynamic RAM requires periodic charge refreshing to maintain data storage. The term dynamic refers to this tendency of the stored charge to leak away, even with power continuously applied.
Figure 3(a) is a typical DRAM structure for an individual cell that stores one bit. The address line is activated when the bit value from this cell is to be read or written. The transistor acts as

a switch that is closed (allowing current to flow) if a voltage is applied to the address line and open (no current flows) if no voltage is present on the address line.

Figure 3(a): DRAM Cell

Figure 3(b): Typical 16 Megabit DRAM

For the write operation (please refer to Figure 3 (a), a voltage signal is applied to the bit line; a high voltage represents 1, and a low voltage represents 0. A signal is then applied to the address line, allowing a charge to be transferred to the capacitor.

For the read operation, when the address line is selected, the transistor turns on and the charge stored on the capacitor is fed out onto a bit line and to the sense amplifier. The sense amplifier compares the capacitor voltage to a reference value and determines if the cell contains logic 1 or logic 0. The read out from the cell discharges the capacitor, which **must be restored** to complete the operation.

Although the DRAM cell is used to store a single bit (0 or 1), it is essentially an analog device. The capacitor can store any charge value within a range; a threshold value determines whether the charge is interpreted as 1 or 0.

Organisation of DRAM Chip

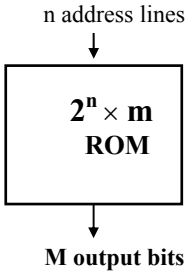
The Figure 3(b) is a typical organisation of 16 mega bit DRAM. It shows a typical organisation of 2048 × 2048 × 4 bit DRAM chip. The memory array in this organisation is a square array that is (2048 × 2048) words of 4 bits each.

Each element, which consists of 4 bits of array, is connected by horizontal row lines and vertical column lines. The horizontal lines are connected to the select input in a row, whereas the vertical line is connected to the output signal through a sense amplifier or data in signal through data bit line driver. Please note that the selection of input from this chip requires:

- Row address selection specifying the present address values A0 to A10 (11 address lines only). For the rows, it is stored in the row address buffer through decoder.
- Row decoder selects the required row.
- The column address buffer is loaded with the column address values, which are also applied to through A0 to A10 lines only. Please note that these lines should contain values for the column.
- This job will be done through a change in external signal $\overline{\text{RAS}}$ (Row address Strobe) because this signal is high at the rising edge of the clock.
- $\overline{\text{CAS}}$ (Column address Strobe) causes the column address to be loaded with these values.
- Each column is of 4 bits, that is, those require 4 bit data lines from input/output buffer. On memory write operation data in bit lines being activated while on read sense lines being activated.
- This chip requires 11 address lines (instead of 22), 4 data in and out lines and other control lines.
- As there are 11 row address lines and 11 column address lines and each column is of 4 bits, therefore, the size of the chip is $2^{11} \times 2^{11} \times 4 = 2048 \times 2048 \times 4 = 16$ mega bits. On increasing address lines from 11 to 12 we have $2^{12} \times 2^{12} \times 4 = 64$ mega bits, an increase of a factor of 4. Thus, possible sizes of such chips may be 16K, 256K, 1M, 4M, 16M, and so on.
- Refreshing of the chip is done periodically using a refresh counter. One simple technique of refreshing may be to disable read-write for some time and refresh all the rows one by one.

ROM (Read-Only Memory)

A ROM is essentially a memory or storage device in which a fixed set of binary information is stored. A block diagram of ROM is as shown in Figure 4(a). It consists of n input lines and m output lines. Each bit combination of the input variables is called an ***address***. Each bit combination that comes out of the output lines is called a ***word***. The number of bits per word is equal to the number of output lines m. The number of distinct addresses possible with n input variables is $2^{n.}$



(a) **ROM Block diagram**

Input		Output	
I_1	I_2	O_1	O_2
0	0	0	1
0	1	1	0
1	0	1	1
1	1	0	0

(b) Truth table

(c) A Sample ROM

Figure 4: ROM

A ROM is characterised by the number of words (2^n) and the number of bits (m) per word. For example, a 32 × 8 ROM which can be written as $2^5 \times 8$ consists of 32 words of 8 bit each, which means there are 8 output lines and 32 distinct words stored in the unit. There are only 5 input lines because $32 = 2^5$and with 5 binary variables, we can specify 32 addresses.

A ROM is basically a combinational circuit and can be constructed as shown in Figure 4(c). On applying an Input $I_1 = 0$, $I_2 = 0$, the 00 line of the decoder is selected and we will get $O_1 = 0$ and $O_2 = 1$; on applying $I_1 = 0$ and $I_2 = 1$ we will get $O_1 = 1$ AND $O_2 = 0$. This same logic can be used for constructing larger ROMs.

ROMs are the memories on which it is not possible to write the data when they are on-line to the computer. They can only be read. This is the reason why it is called read-only memory (ROM). Since ROM chips are non-volatile, the data stored inside a ROM are not lost when the power supply is switched off, unlike the case of a volatile RAM chip. ROMs are also known as permanent stores.

The ROMs can be used for storing micro-programs, system programs and subroutines. ROMs are non-volatile in nature and need not be loaded in a secondary storage device. ROMs are fabricated in large numbers in a way where there is no room for even a single error. But, this is an inflexible process and requires mass production. Therefore, a new kind of ROM called PROM was designed which is also non-volatile and can be written only once and hence the name Programmable ROM(PROM). The supplier or the customer can perform the writing process in PROM electrically. Special equipment is needed to perform this writing operation. Therefore, PROMs are more flexible and convenient than ROMs.

The ROMs / PROMs can be written just once, but in both the cases whatever is written once cannot be changed. But what about a case where you read mostly but write only very few times? This led to the concepts of read mostly memories and the best example of these are EPROMs (Erasable PROMs) and EEPROMs (Electrically Erasable PROMs).

The EPROMs can be read and written electrically. But, the write operation is not simple. It requires erasure of whole storage cells by exposing the chip to ultra violet light, thus bringing them to the same initial state. Once all the cells have been brought to same initial state, then the EPROM can be written electrically. EEPROMs are becoming increasingly popular, as they do not require prior erasure of previous

contents. However, in EEPROMS the writing time is considerably higher than the reading time. The biggest advantage of EEPROM is that it is non-volatile memory and can be updated easily, while the disadvantages are the high cost and at present they are not completely non-volatile and the write operation takes considerable time. But all these advantages are disappearing with growth in technology. In general, ROMs are made of cheaper and slower technology than RAMs.

Flash Memory

This memory is another form of semiconductor memory, which was first introduced in the mid-1980. These memories can be reprogrammed at high speed and hence the name flash. This is a type of non-volatile, electronic random access memory.

Basically this memory falls in between EPROM and EEPROM. In flash memory the entire memory can be erased in a few seconds by using electric erasing technology. Flash memory is used in many I/O and storage devices. *Flash memory is also used to store data and programming algorithms in cell phones, digital cameras and MP3 music players.*

Flash memory serves as a hard drive for consumer devices. Music, phone lists, applications, operating systems and other data are generally stored on stored on flash chips. *Unlike the computer memory, data are not erased when the device is turned off.*

There are two basic kinds of flash memory:

Code Storage Flash made by Intel, AMD, Atmel, etc. It stores programming algorithms and is largely found in cell phones.

Data Storage Flash made by San Disk, Toshiba, etc. It stores data and comes in digital cameras and MP3 players.

The feature of semiconductor memories are summarised in the Figure 5.

Memory	Category	Erasure	Write Mechanism	Volatility
Random-access Memory (RAM)	Read–write memory	Electrically, byte level	Electrically	Volatile
Read –only Memory (ROM)	Read–only memory	Not possible	Masks	Non-volatile
Programmable ROM (PROM)	Read–only memory	Not possible	Electrically	Non-volatile
Erasable PROM (EPROM)	Read-mostly memory	UV light chip level	Electrically	Non-volatile
Electrically Erasable (EEPROM)	Read-mostly memory	Electrically, byte level	Electrically	Non-volatile
Flash memory	Read-mostly memory	Electrically, block level	Electrically	Non-volatile

Figure 5: Features of Semiconductor Memories

1.4 SECONDARY MEMORY AND CHARACTERISTICS

It is desirable that the operating speed of the primary storage of a computer system be as fast as possible because most of the data transfer to and from the processing unit is via the main memory. For this reason, storage devices with fast access times, such as semiconductors, are generally used for the design of primary storage. These high-speed storage devices are expensive and hence the cost per bit of storage is also high for a primary storage. *But the primary memory has the following limitations*:

a) **Limited capacity:** The storage capacity of the primary storage of today 's computers is not sufficient to store the large volume of data handled by most of the data processing organisations.

b) **Volatile:** The primary storage is volatile and the data stored in it is lost when the electric power is turned off. However, the computer systems need to store data on a permanent basis for several days, months or even several years.

The result is that an additional memory called secondary storage is used with most of the computer systems. Some popular memories are described in this section.

1.4.1 Hard Disk Drive

This is one of the components of today's personal computer, having a capacity of the order of several Giga Bytes and above. A magnetic disk has to be mounted on a disk drive before it can be used for reading or writing of information. A disk drive contains all the mechanical, electrical and electronic components for holding one or more disks and for reading or writing of information on it. That is, it contains the central shaft on which the disks are mounted, the access arms, the read/write head and the motors to rotate the disks and to move the access arms assembly. Now-a-days, the disk drive assembly is packed in very small casing although having very high capacity. Now let us know about what a magnetic disk is.

Magnetic Disk

A disk is circular platter constructed of nonmagnetic material, called the substrate, coated with a magnetisable material. This is used for storing large amount of data. Traditionally, the substrate has been an aluminium or aluminium alloy material; more recently, glass substrates have been introduced. The glass substrate has a number of benefits, including the following:

- Improvement in the uniformity of the magnetic film surface to increase disk reliability.
- A significant reduction in overall surface to help reduce read-write errors.
- Ability to support lower fly heights.
- Better stiffness to reduce disk dynamics.
- Greater ability to withstand shock and damage.

Magnetic Read and Write Mechanisms

Data are recorded on and later retrieved from the disk via a conducting coil named the head; in many systems there are two heads, a read head and a write head.
During a read or write operation, the head is stationary while the platter rotates beneath it.

Figure 6: Read /write Heads

The write mechanism is based on the fact that electricity flowing through a coil produces a magnetic field. Pulses are sent to the write head, and magnetic patterns are recorded on the surface below, with different patterns for positive and negative currents. The write head itself is made of easily magnetisable material and is in the shape of a rectangular doughnut with a gap along one side and a few turns of conducting wire along the opposite side (Figure 6). An electric current in the wire induces a magnetic field across the gap, which in turn magnetizes a small area of the recording medium. Reversing the direction of the current reverses the direction of the magnetization on the recording medium.

The traditional read mechanism is based on the fact that a magnetic field moving relative to a coil produces an electrical current in the coil. When the surface of the disk passes under the head, it generates a current of the same polarity as the one already recorded. The structure of the head for reading is in this case essentially the same as for writing and therefore the same head can be used for both. Such single heads are used in floppy disk systems and in older rigid disk systems.

Data Organization and Formatting

The head is a relatively small device capable of reading from or writing to a portion of the platter rotating beneath it. This gives rise to the organization of data on the platter in a concentric set of rings, called tracks; each track is of the same width as the head. There are thousands of tracks per surface.

Figure 7 depicts this data layout. Adjacent tracks are separated by gaps. This prevents, or at least minimizes, errors due to misalignment of the head. Data are transferred to and from the disk in sectors. To identify the sector position normally there may be a starting point of a track and a starting and end point of each sector. But the question is how is a sector of a track recognised? A disk is formatted to record control data on it such that some extra data are stored on it for identical purpose. This control data is

Figure 7: Layout of Magnetic Disk

accessible only to the disk drive and not to the user. Please note that in Figure 7 as we move away from the centre of a disk the physical size of the track is increasing. Does it mean we store more data on the outside tracks? No. A disk rotates at a constant angular velocity. But, as we move away from centre the liner velocity is more than the liner velocity nearer to centre. Thus, the density of storage of information decreases as we move away from the centre of the disk. This results in larger physical sector size. Thus, all the sectors in the disk store same amount of data.

An example of disk formatting is shown in Figure 8. In this case, each track contains 30 fixed–length sectors of 600 bytes each. Each sector holds 512 bytes of data plus control information useful to the disk controller. The ID field is a unique identifier or address used to locate a particular sector. The SYNC byte is a special bit pattern that delimits the beginning of the field. The track number identifies a track on a surface. The head number identifies a head, because this disk has multiple surfaces. The ID and data fields each contain an error-detecting code.

Figure 8: A typical Track Format for Winchester Disk

Physical Characteristics

Figure 9 lists the major characteristics that differentiate among the various types of magnetic disks. First, the head may either be fixed or movable either respect to the radial direction of the platter. In a fixed-head disk, there is one read-write head per track. All of the heads are mounted on a rigid arm that extends across all tracks; such systems are rare today. In a movable-head disk, there is only one read-write head. Again, the head is mounted on an arm. Because the head must be able to be positioned above any track, the arm can be extended or retracted for this purpose.

Head Motion	**Platters**
Fixed head (one per track)	Single platter
Moveable head (one per surface)	Multiple platter
Disk Portability	**Head mechanism**
Non-removable	Disk Contact (floppy)

Removable disk	Fixed gap Aerodynamic gap (Winchester)
Sides Single sided Double sided	

Figure 9: Physical characteristics of Disk Systems

The disk itself is mounted in a disk drive, which consists of the arm, a shaft that rotates the disk, and the electronics needed for input and output binary data. A non-removable disk is permanently mounted in the disk drive; the hard disk in a personal computer is a non-removable disk. A removable disk can be removed and replaced with another disk. The advantage of the latter type is that unlimited amounts of data are available with a limited number of disk systems. Furthermore, ZIP cartridge disks are examples of removable disks. Figure 10 shows other components of the disks.

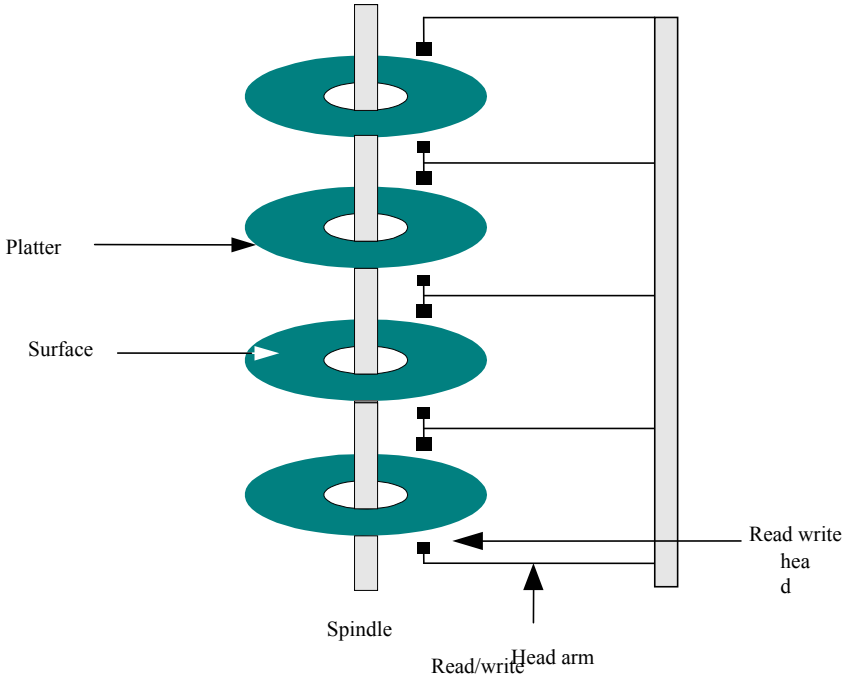


Figure 10: The Disk Components

The head mechanism provides a classification of disks into three types. Traditionally, the read-write head has been positioned at a fixed distance above the platter, allowing an air gap. At the other extreme is a head mechanism that actually comes into physical contact with the medium during a read or write operation. This mechanism is used with the floppy disk, which is a small, flexible platter and the least expensive type of disk.

To understand the third type of disk, we need to comment on the relationship between data density and the distance of head from the surface. The head generates or senses an electromagnetic field of sufficient magnitude to write and read properly. The narrower the head is, the closer it must be to the platter surface to function. A narrower head means narrower tracks and therefore greater data density, which is desirable. However, the closer the head is to the disk, the greater are the risks of errors from impurities or imperfections.

To push the technology further, the Winchester disk was developed. Winchester heads are used in sealed drive assemblies that are almost free of contaminants. They are designed to operate closer to the disk's surface than conventional rigid disk heads, thus allowing greater data density. The head is actually an aerodynamic foil that rests

lightly on the platter's surface when the disk is motionless. The air pressure generated by a spinning disk is enough to make the foil rise above the surface. The resulting non-contact system can be engineered to use narrower heads that operate closer to the platter's surface than conventional rigid disk heads.

Accessing the Disk Data

Disks operate is semi-random mode of operation and normally are referenced block wise. The data access time on a disk consists of two main components:

- **Seek time:** Time to position the head on a specific track. On a fixed head disks it is the time taken by the electronic circuit to select the required head while in movable head disks it is the time required to move the head to a particular track.
- **Latency time:** This is the time required by a sector to reach below the read/write head. On an average it is half of the time taken for a rotation by the disk.

In addition to the seek and latency time, the time taken to transfer a (read/write) block of words can be considered but normally it is too small in comparison to latency and seek time and in general the disk access time is considered to be the sum of seek time and latency time. Since access time of disks is large, therefore it is advisable to read a sizeable portion of data in a single go and that is why the disks are referenced block wise. In fact, you will find that in most of the computer system, the input/output involving disk is given a very high priority. The basic reason for such priority is the latency time that is needed once the block which is to be read passes below the read-write head; it may take time of the order of milliseconds to do that again, in turn delaying the Input /Output and lowering the performance of the system.

1.4.2 Optical Memories

In 1983, one of the most successful consumer products of all times was introduced: the compact disk (CD) digital audio system. This CD was a non-erasable disk that could store more than 60 minutes of audio information on one side. The huge commercial success of this CD enabled the development of low-cast optical-disk storage technology that has revolutionised computer data storage. A variety of optical-disk systems has been introduced. We briefly review each of these.

Compact Disk ROM (CD-ROM)

Both the audio CD and the CD-ROM (compact disk read-only memory) share a similar technology. The main difference is that CD-ROM players are more rugged and have error correction devices to ensure that data are properly transferred from disk to computer. Both types of disk are made the same way. The disk is formed from a resin, such as polycarbonate. Digitally recorded information (either music or computer data) is imprinted as a series of microscopic pits on the surface of the polycarbonate. The pitted surface is then coated with a highly reflective surface, usually aluminium. This shiny surface is protected against dust and scratches by a topcoat of clear acrylic. Finally, a label can be silk-screened onto the acrylic.

Figure 11: The CD Surface and Operation

Information is retrieved from a CD or CD-ROM by a low-powered laser housed in an optical-disk player, or drive unit. The laser shines through the clear polycarbonate while a motor spins the disk past it (Figure 11). The intensity of the reflected light of the laser changes as it encounters a pit. Specifically, if the laser beam falls on a pit, the light scatters and a low intensity is reflected back to the source. The areas between pits are called ***lands***. A land is a smooth surface, which reflects back at a higher intensity. The change between pits and lands is detected by a **photo sensor** and converted into a digital signal. The sensor tests the surface at regular intervals. The beginning or end of a pit represents a 1; when no change in elevation occurs between intervals, a 0 is recorded.
Data on the CD-ROM are organised as a sequence of blocks. A typical block format is shown in Figure 12 (a). It consists of the following fields:

- **Sync:** The sync field identifies the beginning of a block. It consists of a byte of all 0s, 10 bytes of all 1s, and bytes of all 0s.
- **Header:** The header contains the block address and the mode byte. Mode 0 specifies a blank data field; mode 1 specifies the use of an error-correcting code and 2048 bytes of data; mode 2 specifies 2336 bytes of user data with no error correcting code.
- **Data:** User data.
- **Auxiliary:** Additional user data in mode 2. In mode 1, this is a 288-byte error correcting code.

(a) Block Format **(b) CD-ROM Layout**

Figure 12: Block Format and Disk Layout on CD-ROM

But what is the Min (Minute), Sec (Second) and Sector fields in the Header field? The sectors of CD-ROM are not organised like the sectors in hard disks (Please refer Figure 12(b)). Rather, they are all equal length segments. If we rotate the CD drive at constant speed the linear velocity of disk surface movement will be higher at the outer side than that of the centre portions. To offset this liner speed gap, either we store less data on the outer sectors or we reduce the speed of rotation while reading outer tracks. The CD follows the later approach, that is, instead of moving the CD drive at constant velocity, it is rotated at variable velocity. The speed or rotation of disk reduces as we move away from the centre such that the sector's can be read in constant time. This method of reading is called Constant Liner Velocity (CLV).

CD-ROM is appropriate for the distribution of large amounts of data to a large number of users. CD-ROMs are a common medium these days for distributing information. Compared with traditional hard disks, *the CD-ROM has three advantages*:

1. Large data/information storage capability.
2. The optical disk together with information stored on it can be mass replicated inexpensively, unlike a magnetic disk. The database on a magnetic disk has to be reproduced by copying data from one disk to second disk, using two disk drives.
3. The optical disk is removable, allowing the disk itself to be used for archival storage. Most magnetic disks are non-removable. The information on non-removable magnetic disks must first be copied on tape before the disk drive / disk can be used to store new information.

The disadvantages of CD- ROM are as follows:

1. It is read-only and cannot be updated.
2. It has an access time much longer than that of a magnetic disk drive (as it employs CLV), as much as half a second.

Compact Disk Recordable (CD-R)

To accommodate applications in which only one or a small number of copies of a set data is needed, the write-once read-many CD, known as the CD Recordable (CD-R), has been developed. For CD-R a disk is prepared in such a way that it can be subsequently written once with a laser beam of modest intensity. Thus, with a somewhat more expensive disk controller than for CD-ROM, the customer can write once as well as read the disk.

The CD-R medium is similar to but not identical to that of a CD or CD-ROM. For CDs and CD-ROMs, information is recorded by the pitting of the surface of the medium, which changes reflectivity. For a CD-R, the medium includes a dye layer. The resulting disk can be read on a CD-R drive or a CD-ROM drive.

The CD-R optical disk is attractive for archival storage of documents and files. It provides a permanent record of large volumes of user data.

Compact Disk Rewritable (CD-RW)

The CD-RW optical disk can be repeatedly written and overwritten, as with a magnetic disk. Although a number of approaches have been tried, the only pure optical approach that has proved attractive is called phase change. The phase change disk uses a material that has two significantly different reflectivities in two different phase states. There is an amorphous state, in which the molecules exhibit a random orientation and which reflects light poorly; and a crystalline state, which has a smooth surface that reflects light well. A beam of laser light can change the material from one phase to the other. The primary disadvantage of phase change optical disks is that the material eventually and permanently loses its desirable properties. Current materials can be used for between 500,000 and 1,000,000 erase cycles.

The CDRW has the obvious advantage over CD-ROM and CD-R that it can be rewritten and thus used as a true secondary storage. As such, it competes with magnetic disk. A key advantage of the optical disk is that the engineering tolerances for optical disks are much less severe than for high-capacity magnetic disks. Thus, they exhibit higher reliability and longer life.

Digital Versatile Disk (DVD)

With the capacious digital versatile disk (DVD), the electronics industry has at last found an acceptable replacement for the videotape used in videocassette recorders (VCRs) and, more important for this discussion, replace the CD-ROM in personal computers and servers. The DVD has taken video into the digital age. It delivers movies with impressive picture quality, and it can be randomly accessed like audio CDs, which DVD machines can also play. Vast volumes of data can be crammed onto the disk, several times as much as a CD-ROM. With DVD's huge storage capacity and vivid quality, PC games will become more realistic and educational software will incorporate more video.

1.4.3 CCDs, Bubble Memories

Charge-coupled Devices (CCDs)

CCDs are used for storing information. They have arrays of cells that can hold charge packets of electron. A word is represented by a set of charge packets, the presence of each charge packet represent the bit-value 1. The charge packets do not remain stationary and the cells pass the charge to the neighbouring cells with the next clock pulse. Therefore, cells are organized in tracks with a circuitry for writing the data at the beginning and a circuitry for reading the data at the end. Logically the tracks (one for each bit position) may be conceived as loops since the read circuitry passes the information back to the write circuit, which then re-creates the bit values in the track unless new data is written to the circuit.

These devices come under the category of semi-random operation since the devices must wait till the data has reached the circuit for detection of charge packets. The access time to these devices is not very high. At present this technology is used only in specific applications and commercial products are not available.

Magnetic Bubble Memories

In certain material such as garnets on applying magnetic fields certain cylindrical areas whose direction of magnetization is opposite to that of magnetic field are created. These are called magnetic bubbles. The diameter of these bubbles is found to be in the range of 1 micrometer. These bubbles can be moved at high speed by applying a parallel magnetic field to the plate surface. Thus, the rotating field can be generated by an electromagnetic field and no mechanical motion is required.

In these devices deposition of a soft magnetic material called Perm alloy is made as a predetermined path, thus making a track. Bubbles are forced to move continuously in a fixed direction on these tracks. In these memories the presence of a bubble represents a 1 while absence represents a 0 state. For writing data into a cell, a bubble generator to introduce a bubble or a bubble annihilator to remove a bubble, are required. A bubble detector performs the read operation. Magnetic bubble memories having capacities of 1M or more bits per chip have been manufactured. The cost and performance of these memories fall between semi-conductor RAMs and magnetic disks.
These memories are non-volatile in contrast to semi-conductor RAMs. In addition, since there are no moving parts, they are more reliable than a magnetic disk. But these memories are difficult to manufacture and difficult to interface with in conventional processors. These memories at present are used in specialized applications, e.g., as a secondary memory of air or space borne computers, where extremely high reliability is required.

Check Your Progress 1

1. **State True or False:**

   a) Bubble memories are non-volatile. T/F

b) The disadvantage of DRAM over static RAM is the need to refresh the capacitor charge every few milliseconds. T/F

c) Flash memory is a volatile RAM. T/F

2. **Fill in the blanks:**

a) The EPROM is _____ erasable and _____ programmable.

b) _____ memory requires a rechargeable cycle in order to retain its information.

c) Memory elements employed specifically in computer memories are generally _____ circuits.

3. Differentiate among RAM, ROM, PROM and EPROM.

……

4. What is a flash memory? Give a few of its typical uses.

……

5. A memory has a capacity of 4K × 8
(a) How many data input and data output lines does it have?
(b) How many address lines does it have?
(c) What is the capacity in bytes?

……

……

6. Describe the internal architecture of a DRAM that stores 4K bytes chip size and uses a square register array. How many address lines will be needed? Suppose the same configuration exists for an old RAM, then how many address lines will be needed?

……

……

7. How many RAM chips of size 256K × 1 bit are required to build 1M Byte memory?

……

……

1.5 RAID AND ITS LEVELS

Researchers are constantly trying to improve the secondary storage media by increasing their capacity, performance, and reliability. However, any physical media has a limit to its capacity and performance. When it gets harder to make further

improvements in the physical storage media and its drive, researchers normally look for other methods for improvement. Mass storage devices are a result of such improvements, which researchers have resorted to for larger capacity secondary devices. The idea is to use multiple units of the storage media being used as a single secondary storage device. One such attempt in the direction of improving disk performance is to have multiple components in parallel. Some basic questions for such systems are:

How are the disks organised?
May be as an array of disks.

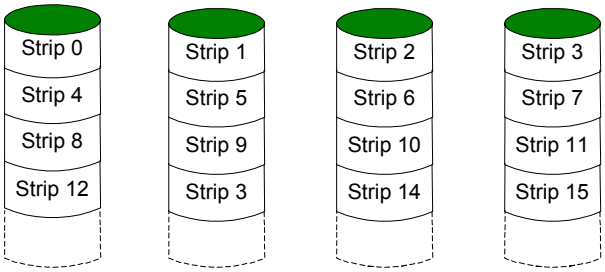
Can separate I/O requests be handled by such a system in parallel?
Yes, but only if the disk accesses are from separate disks.

Can a single I/O request be handled in parallel?
Yes, but the data block requested should be available on separate disks.

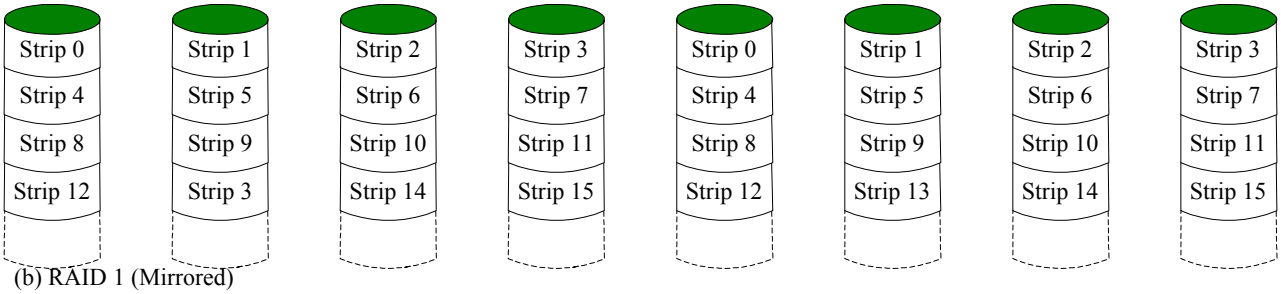
Can this array of disks be used for increasing reliability?
Yes, but for that redundancy of data is essential.

One such industrial standard, which exists for multiple-disk database schemes, is termed as RAID, i.e., Redundant Array of Independent Disks. The basic characteristics of RAID disks are:

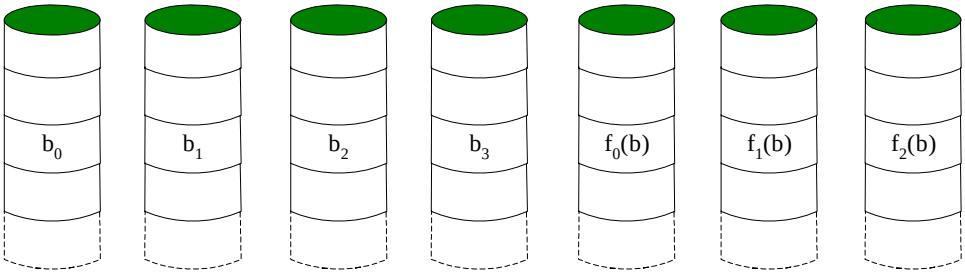
- Operating system considers the physical disks as a single logical drive.
- Data is distributed across the physical disks.
- In case of failure of a disk, the redundant information (for example, the parity bit) kept on redundant disks is used to recover the data.



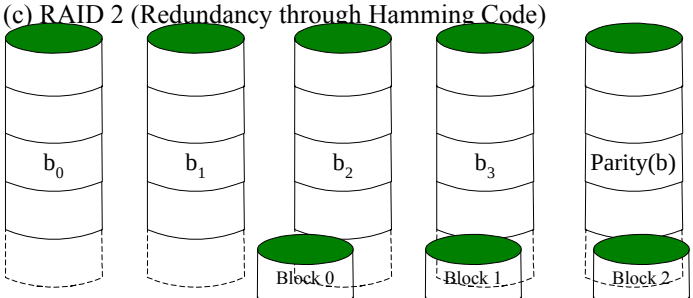
(a) RAID 0 (Non-redundant)



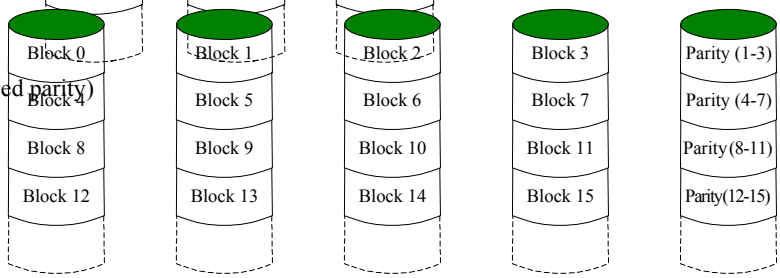
(b) RAID 1 (Mirrored)



(c) RAID 2 (Redundancy through Hamming Code)



(d) RAID 3 (Bit-interleaved parity)



(e) RAID 4 (Block level Parity)

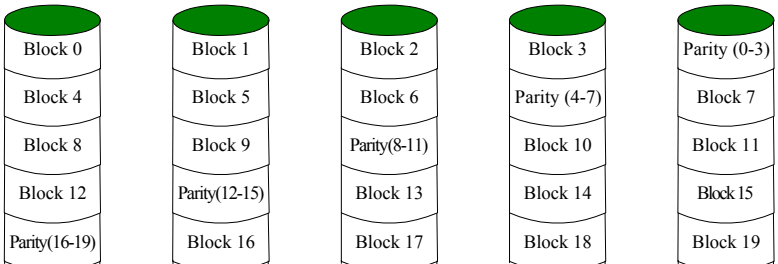


Figure 13: The RAID levels

The term RAID was coined by researchers at University of Berkley. In their paper the meaning of RAID was Redundant Array of Inexpensive Disks. However, later the term Independent was adopted instead of Inexpensive to signify performance and reliability gains.

RAID has been proposed at various levels, which are basically aimed to cater for the widening gap between the processor and on-line secondary storage technology.

The basic strategy used in RAID is to replace the large capacity disk drive with multiple smaller capacity disks. The data on these disks is distributed to allow simultaneous access, thus improving the overall input/output performance. It also allows an easy way of incrementing the capacity of the disk. Please note that one of the main features of the design is to compensate for the increase in probability of failure of multiple disks through the use of parity information. The seven levels of RAID are given in Figure 13 shown above. Please note that levels 2 and 4 are not commercially offered.

In RAID technologies have two important performance considerations:

- The Data Transfer Rate
- Input/Output Request rate

High data transfer rate is dependent on:

- Path between individual disks and Memory.
- Fulfilment of an I/O request by multiple disks of disk array. Thus, increasing the transfer rate.

Input/Output request rate is defined as the time taken to fulfil an I/O request. This is an important consideration while dealing with transaction processing systems like Railway reservation system. A high I/O request rate can be obtained through distributing data requests on multiple disks.

1.6 THE CONCEPTS OF HIGH SPEED MEMORIES

RAID Level	Category	Features	I/O Request Rate (Read /write)	Data Transfer Rate (Read /write)	Typical Application
0	Striping	a) The disk is divided into strips, maybe a block, a sector or other unit. b) Non-redundant.	Large strips: Excellent	Small strip: Excellent	Applications requiring high performance for non-critical data
1	Mirroring	a) Every disk in the array has a mirror disk that contains the same data. b) Recovery from a failure is simple. When a drive fails, the data may still be recovered from the second drive.	Good / fair	Fair /fair	System drives; critical files
2	Parallel Access	a) All member disks participate in the execution of every I/O request by synchronising the spindles of all the disks to the same position at a time. b) The strips are very small, often a single byte or word. c) Redundancy via hamming code which is able to correct single-bit errors and detect double-bit errors.	Poor	Excellent	Commercially not useful.
3	Parallel Access	a) Employs parallel access as that of level 2, with small data strips. b) A simple parity bit is computed for the set of data instead of an error-correcting code in case a disk fails.	Poor	Excellent	Large I/O request size application, such as imaging CAD
4	Independe nt access	a) Each member disk operates independently, thus enabling fulfilment of separate input/output requests in parallel. b) Data strip is large and bit by bit parity strip is created for bits of strips of each disk. c) Parity strip is stored on a separate disk.	Excellent/ fair	Fair / poor	Commercially not useful.
5	Independe nt access	a) Employs independent access as that of level 4 and distributes the parity strips across all disks. b) The distribution of parity strips across all drives avoids the potential input/output bottleneck found in level 4.	Excellent / fair	Fair / poor	High request rate read intensive, data lookup
6	Independe nt access	a) Also called the P+Q redundancy scheme, is much like level 5, but stores extra redundant information to guard against multiple disk failures. b) P and Q are two different data check algorithms. One of the two is the exclusive-or calculation used in level 4 and 5. The other one is an independent data check algorithm.	Excellent/ poor	Fair / poor	Application requiring extremely high availability

Why are high-speed memories needed? Is the main memory not a high-speed memory? The answer to the second question is definitely “No”, but why so? For this, we have to go to the fundamentals of semiconductor technology, which is beyond the scope of the Unit. Then if the memories are slower, then how slow are they? On an average it has been found that the operating speed of main memories lack by a factor

of 5 to 10 than that of the speed of processors (such as CPU or Input / Output Processors).

In addition, each instruction requires several memory accesses (it may range from 2 to 7 or even more sometimes). If an instruction requires even 2 memory accesses, even then almost 80% of the time of executing an expression, processors waits for memory access.

The question is what can be done to increase this processor-memory interface bandwidth? There are four possible answers to the question. These are:

a) Decrease the memory access time; use a faster but expensive technology for main memory.

b) Access more words in a single memory access cycle. That is, instead of accessing one word from the memory in a memory access cycle, access more words.

c) Insert a high-speed memory termed as Cache between the main memory and processor.

d) Use associative addressing in place of random access.

Hardware researchers are taking care of the first point. Let us discuss some high speed memories that are in existence at present.

1.6.1 Cache Memory

Cache memory is an extremely fast, small memory between CPU and main memory whose access time is closer to the processing speed of the CPU. It acts as a high-speed buffer between CPU and main memory and is used to temporarily store currently active data and instructions during processing. Since the cache memory is faster than main memory, the processing speed is increased by making data and instructions needed in present processing available in the cache

The obvious question that arises is how the system can know in advance which data and instruction are needed in present processing so as to make it available beforehand in the cache. The answer to this question comes from a principle known as ***locality of reference***. According to this principle, during the course of execution of most programs, memory references by the processor, for both instructions and data, tend to cluster. That is, if an instruction is executed, there is a likelihood of the nearby instruction being executed soon. Locality of reference is true not only for reference to program instruction but also for references to data. As shown in Figure 14, the cache memory acts as a small, fast-speed buffer between the processor and main memory.

Figure 14: Cache Memory Operation

It contains a copy of a portion of main memory contents. When a program is running and the CPU attempts to read a word of memory (instruction or data), a check is made to determine if the word is in the cache. If so, the word is delivered to the CPU from the cache. If not, a block of main memory, consisting of some fixed number of words including the requested word, is read into the cache and then the requested word is delivered to the CPU. Because of the feature of locality of reference, when a block of memory word is fetched into the cache to satisfy a single memory reference, it is likely that there will soon be references to other words in that block. That is, the next time the CPU attempts to read a word, it is very likely that it finds it in the cache and saves the time needed to read the word from main memory.

Many computer systems are designed to have two separate cache memories called ***instruction cache*** and ***data cache***. The instruction cache is used for storing program instruction and the data cache is used for storing data. This allows faster identification of availability of accessed word in the cache memory and helps in further improving the processor speed. Many computer systems are also designed to have multiple levels of caches (such as level one and level two caches, often referred to as **L1 and L2 caches**). L1 cache is smaller than L2 cache and is used to store more frequently accessed instruction/data as compared to those in the L2 cache.

The use of cache memory requires several design issues to be addressed. Some key design issues are briefly summarised below:

1. **Cache Size:** Cache memory is very expensive as compared to the main memory and hence its size is normally kept very small. It has been found through statistical studies that reasonably small caches can have a significant impact on processor performance. As a typical example of cache size, a system having 1 GB of main memory may have about 1 MB of cache memory. Many of today's personal computers have 64KB, 128KB, 256KB, 512KB, or 1 MB of cache memory.

2. **Block Size:** Block size refers to the unit of data (few memory words) exchanged between cache and main memory. As the block size increases from very small to larger size, the hit ratio (fraction of times that referenced instruction/data is found in cache) will at first increase because of the principle of locality since more and more useful words are brought into the cache. However, the hit ratio will begin to decrease as the block size further increases because the probability of using the newly fetched words becomes less than the probability of reusing the words that must be moved out of the cache to make room for the new block. Based on this fact, the block size is suitably chosen to maximise the hit ratio.

3. **Replacement Policy:** When a new block is to be fetched into the cache, another may have to be replaced to make room for the new block. The replacement policy decides which block to replace in such a situation. Obviously, it will be best to replace a block that is least likely to be needed again in the near future.

4. **Write Policy:** If the contents of a block in the cache are altered, then it is necessary to write it back to main memory before replacing it. The write policy decides when the altered words of a block are written back to main memory. At

one extreme, an updated word of a block is written to the main memory as soon as such updates occur in the block. At the other extreme, all updated words of the block are written to the main memory only when the block is replaced from the cache. The latter policy minimises overheads of memory write operations but temporarily leaves main memory in an inconsistent (obsolete) state.

1.6.2 Cache Organisation

Cache memories are found in almost all latest computers. They are very useful for increasing the speed of access of information from memory. Let us look into their organisation in more detail in this section.

The fundamental idea of cache organisation is that by keeping the most frequently accessed instructions and data in the fast cache memory; hence the average memory access time will approach the access time of the cache.

The basic operation of the cache is as follows. When the CPU needs to access memory, the cache is examined. If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word. A block of words is then transferred from main memory to cache memory.

The performance of cache memory is frequently measured in terms of a quantity called ***hit ratio***. When the CPU refers to the main memory and finds the word in cache, it is said to produce a ***hit***. If the word is not found in cache, it is in the main memory and it counts as a ***miss***. The ratio of the number of hits divided by the total CPU references to memory is the hit ratio.

The average memory access time of a computer system can be improved considerably by use of a cache. For example, if memory read cycle takes 100 ns and a cache read cycle takes 20 ns, then for four continuous references, the first one brings the main memory contents to cache and the next three from cache.

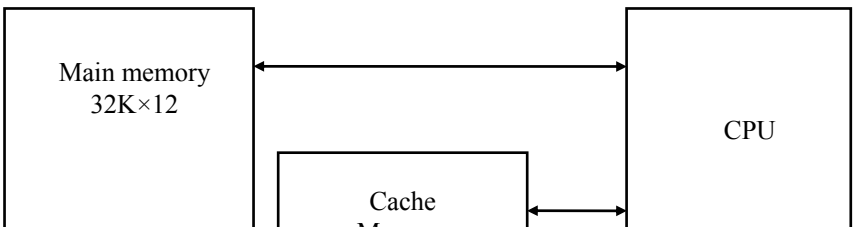
The time taken with cache = (100+20) + (20×3)
(For the first read operation) (For the last three read operations)
= 120 + 60
= 180 ns
Time taken without cache = 100×4 = 400 ns

Thus, the closer are the reference, the better is the performance of cache.

The basic characteristic of cache memory is its fast access time. Therefore, very little or no time must be wasted when searching for words in the cache. The transformation of data from main memory to cache memory is referred to as a mapping process. The mapping procedure for the cache organization is of three types:

1. Associative mapping
2. Direct mapping
3. Set-associative mapping

Let us consider an example of a memory organization as shown in Figure 15 in which the main memory can store 32K words of 12 bits each and the cache is capable of storing 512 blocks (each block in the present example is equal to 24 bits, which is equal to two main memory words) at any time.



Size of main memory address (Given word size of 12 bits) = 32 K words = 2^{15} words
$\Rightarrow$ 15 bits are needed for address
Block Size of Cache = 2 Main Memory Words

Figure 15: Cache Memory

For every word stored in cache, there is a duplicate copy in the main memory. The CPU communicates with both memories. It first sends a 15 bits ($32K = 2^5 \times 2^{10} = 2^{15}$) address to cache. If there is a hit, the CPU uses the relevant 12 bits data from 24 bit cache data. If there is a miss, the CPU reads the block containing the relevant word from the main memory. So the key here is that a cache must store the address and data portions of the main memory to ascertain whether the given information is available in the cache or not. However, let us assume the block size as 1 memory word for the following discussions.

Associative Mapping

The most flexible and fastest cache organization uses an associative memory which is shown in Figure 16. The associative memory stores both the address and data of the memory word. This permits any location in cache to store any word from the main memory. The address value of 15 bits is shown as a five-digit octal number and its corresponding 12 bits word is shown as a five digit octal number. A CPU address of 15 bits is placed in the argument register and the associative memory is searched for a matching address. If the address is found, the corresponding 12 bits data is read and sent to the CPU. If no matches are found, the main memory is accessed for the word. The address-data pair is then transferred to the associative cache memory. This address checking is done simultaneously for the complete cache in an associative way.

CPU ADDRESS 15 BITS

↓

Argument register

Address	Data
01001	3450
03767	7613
23245	1234
24250	2205

(All numbers are in octal)

Figure 16: Associative Mapping Cache

Direct Mapping

In the general case, there are 2^k words in cache memory and 2^n words in the main memory. The n-bits memory address is divided into two fields: k bits for the index field and (n - k) bits for the tag field.

The direct mapping cache organization uses the n-bit address to access the main memory and k-bit index to access the cache. The internal organization of the words in the cache memory is as shown in Figure 17. Each word in cache consists of the data word and its associated tag. When a new word is first brought into the cache, the tag bits are stored alongside the data bits. When the CPU generates a memory request, the index field is used for the address to access the cache.

Figure 17: Addressing Relationship for Main Memory and Cache

The tag field of the CPU address is compared with the tag in the word read from the cache. If the two tags match, there is a hit and the desired data word is in cache. If there is no match, there is a miss and the required word is read from the main memory.
Let us consider a numerical example shown in Figure 18. The word at address zero is at present stored in the cache (index = 000, tag = 00, data = 1456). Suppose that the CPU wants to access the word at address 02000. The index address is 000, so it is used to access the cache. The two tags are then compared. The cache tag is 00 but the address tag is 02, which does not produce a match. Therefore, the main memory is accessed and the data word 4254 is transferred to the CPU. The cache word at index address 000 is then replaced with a tag of 02 and data of 4254.

Figure 18: Direct Mapping Cache Organisation

Set-Associative Mapping

A third type of cache organization called set-associative mapping is an improvement on the direct mapping organization in that each word of cache can store two or more words of memory under the same index address. Each data word is stored together with its tag and the number of tag data items in one word of cache is said to form a set.

Let us consider an example of a set-associative cache organization for a set size of two as shown in the Figure 19. Each index address refers to two data words and their associated tags. Each tag requires six bits and each data word has 12 bits, so the word length of cache is 2(6+12) = 36 bits. An index address of nine bits can accommodate 512 words. Thus, the size of cache memory is 512 × 36. In general, a Set-Associative cache of set size K will accommodate K-words of main memory in each word of cache.

Index	Tag	Data	Tag	Data
000	01	3450	02	5670
777	02	6710	00	2340

Figure 19: Two-Way Set-Associative Mapping Cache

Write Policy: The data in cache and main memory can be written by processors or input/output devices. The main problems associated in writing with cache memories are:

a) The contents of cache and main memory can be altered by more than one device. For example, CPU can write to caches and input/output module can directly write to the main memory. This can result in inconsistencies in the values of the cache and main memory.

b) In the case of multiple CPUs with different cache, a word altered in one cache automatically invalidate the word in the other cache.

The suggested techniques for writing in system with caches are:

(a) **Write through:** Write the data in cache as well as main memory. The other CPUs - Cache combination has to watch with traffic to the main memory and make suitable amendment in the contents of cache. The disadvantage of this technique is that a bottleneck is created due to large number of accesses to the main memory by various CPUs.

(b) **Write block:** In this method updates are made only in the cache, setting a bit called Update bit. Only those blocks whose update bit is set is replaced in the main memory. But here all the accesses to the main memory, whether from other CPUs or input/output modules, need to be from the cache resulting in complex circuitry.

(c) **Instruction Cache:** An instruction cache is one which is employed for accessing only the instructions and nothing else. The advantage of such a cache is that as the instructions do not change we need not write the instruction cache back to memory, unlike data storage cache.

1.6.3 Memory Interleaving

In this method, the main memory is divided into 'n' equal-size modules and the CPU has separate Memory Address Register and Memory Base register for each memory module. In addition, the CPU has 'n' instruction register and a memory access system. When a program is loaded into the main memory, its successive instructions are stored in successive memory modules. For example if n=4 and the four memory modules are M_1, M_2, M_3, and M_4 then 1st instruction will be stored in M_1, 2nd in M_2, 3rd in M_3, 4th in M_4, 5th in M1, 6th in M_2 and so on. Now during the execution of the program, when the processor issues a memory fetch command, the memory access system creates n consecutive memory addresses and places them in the Memory Address Register in the right order. A memory read command reads all the 'n' memory modules simultaneously, retrieves the 'n' consecutive instructions, and loads them into the 'n' instruction registers. Thus each fetch for a new instruction results in the loading of 'n' consecutive instructions in the 'n' instruction registers of the CPU.
Since the instructions are normally executed in the sequence in which they are written, the availability of N successive instructions in the CPU avoids memory access after each instruction execution, and the total execution time speeds up. Obviously, the fetch successive instructions are not useful when a branch instruction is encountered during the course of execution. This is because they require the new set of 'n' successive instructions, overwriting the previously stored instructions, which were loaded, but some of which were not executed. The method is quite effective in minimising the memory-processor speed mismatch because branch instructions do not occur frequently in a program.

Figure 20 illustrates the memory interleaving architecture. The Figure shows a 4- way (n=4) interleaved memory system.

Figure 20: A 4-way Interleaved Memory

1.6.4 Associative Memory

The time required to find an item stored in memory can be reduced considerably if stored data can be identified for access by the contents of the data itself rather than by an address. A memory unit accessed by content of the data is called an **associative memory** or **content addressable memory (CAM)**. This type of memory is accessed simultaneously and in parallel on the basis of data content rather than by specific address or location. When a word is written in an associative memory, no address is given. The memory is capable of finding an empty unused location to store the word. When a word is to be read from an associative memory, the content of the word, or part of the word, is specified. The memory locates all words, which match the specified content, and marks them for reading.
Because of its organization, the associative memory is uniquely suited to do parallel searches by data association. Moreover, searches can be done on an entire word or on a specific field within a word. An associative memory is more expensive than a random access memory because each cell must have storage capability as well as logic circuits for matching its content with an external argument. For this reason associative memories are used in applications where the search time is very critical and must be very short.

Hardware Organization

The block diagram of an associative memory is shown in Figure 21. It consists of a memory array and logic for m words with n bits per word. The argument register A and key register K each have n bits, one for each bit of a word. The match register M has m bits, one for each memory word. Each word in memory is compared in parallel with the content of the argument register; the words that match the bits of the argument register set a corresponding bit in the match register. After the matching process, those bits in the match register that have been set indicate the fact that their corresponding words have been matched. Reading is accomplished by a sequential access to memory for those words whose corresponding bits in the match register have been set.
The key register provides a mask for choosing a particular field or key in the argument word. The entire argument is compared with each memory word if the key register contains all 1s. Otherwise, only those bits in the argument that have 1s in their corresponding positions of the key register are compared. Thus the key provides a mask or identifying information, which specifies how reference to memory is made.

Figure 21: Associative Memory – Block Diagram

To illustrate with a numerical example, suppose that the argument register A and the key register K have the bit configuration shown below. Only the three leftmost bits of a compared with memory words because K has 1's on these positions

A	101 111100	
K	111 000000	
Word 1	100 111100	no match
Word 2	101 000001	match

Word 2 matches the unmasked argument field because the three leftmost bits of the argument and the word are equal.

Check Your Progress 2

1. What is a RAID? What are the techniques used by RAID for enhancing reliability?

   ……

   ……

   ……

2. **State True or False:**

   a) Interleaved memories are best suited for small loops and large sequential code. T/F

   b) The principle of locality of reference justifies the use of cache memory. T/F

   c) High-speed memories are needed to bridge the gap of speed between I/O device and memory. T/F

   d) Write policy is not needed for instruction cache. T/F

   e) A replacement algorithm is needed only for associative and set associative mapping of cache. T/F

3. How can the Cache memory and interleaved memory mechanisms be used to improve the overall processing speed of a Computer system?

……

……

……

4. Assume a Computer having 64 word RAM (assume 1 word = 16 bits) and cache memory of 8 blocks (block size = 32 bits). Where can we find Main Memory Location 25 in cache if (a) Associative Mapping (b) Direct mapping and (c) 2 way set associative (2 blocks per set) mapping is used.

……

……

……

5. How is a given memory word address (memory location 25 as above) located to Cache for the example above for (a) Associative (b) Direct and (c) 2 way set associative mapping.

……

……

……

6. A computer system has a 4K-word cache organised in block set associative manner with 4 blocks per set, 64 words per block. What is the number of bits in the Index and Block Offset fields of the main memory address formula?

……

……

……

1.7 VIRTUAL MEMORY

In a memory hierarchy system, programs and data are first stored in auxiliary or secondary memory. The program and its related data are brought into the main memory for execution. What if the size of Memory required for the Program is more than the size of memory? Virtual memory is a concept used in some large computer systems that permit the user to construct programs as though a large memory space were available, equal to the totality of secondary memory. Each address generated by the CPU goes through an address mapping from the so-called virtual address to a physical address in the main memory. Virtual memory is used to give programmers the illusion that they have a very large memory at their disposal, even though the computer actually has a relatively small main memory. A Virtual memory system provides a mechanism for translating program-generated addresses into correct main memory locations. This is done dynamically, while programs are being executed in the CPU. The translation or mapping is handled automatically by the hardware by means of a mapping table.

Address Space and Memory Space

An address used by a programmer will be called a virtual address, and the set of such addresses the address space. An address in the main memory is called a physical address. The set of such locations is called the memory space. Thus, the address space is the set of addresses generated by programs as they reference instructions and data; the memory space consists of the actual main memory locations directly addressable for processing.

Consider a computer with a main-memory capacity of 64K words (K=1024). 16-bits are needed to specify a physical address in memory since 64K = 2^{16}. Suppose that the computer has auxiliary memory for storing information equivalent to the capacity of 16 main memories. Let us denote the address space by N and the memory space by M, we then have for this example N = 16 × 64 K = 1024K and M = 64K.

In a multiprogramming computer system, programs and data are transferred to and from auxiliary memory and main memory based on demands imposed by the CPU. Suppose that program 1 is currently being executed in the CPU. Program 1 and a portion of its associated data are moved from secondary memory into the main memory as shown in Figure 22. Portions of programs and data need not be in contiguous locations in memory since information is being moved in and out, and empty spaces may be available in scattered locations in memory.

Figure 22: Address and Memory Space in Virtual Memory

In our example, the address field of an instruction code will consist of 20 bits but physical memory addresses must be specified with only 16-bits. Thus CPU will reference instructions and data with a 20 bits address, but the information at this address must be taken from physical memory because access to auxiliary storage for individual words will be prohibitively long. A mapping table is then needed, as shown in Figure 23, to map a virtual address of 20 bits to a physical address of 16 bits. The mapping is a dynamic operation, which means that every address is translated immediately as a word is referenced by CPU.

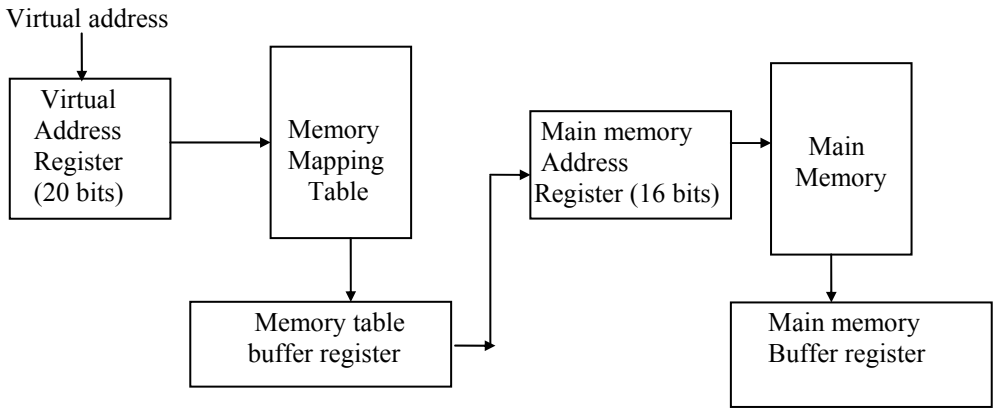


Figure 23: Memory table for mapping a virtual table

1.8 THE MEMORY SYSTEM OF MICROCOMPUTER

Till now we have discussed various memory components. But, how is the memory organised in the physical computer? Let us discuss various kinds of memory technologies used in personal computer.

1.8.1 SIMM (Single In-line Memory Modules), DIMM (Dual In line Memory Modules), etc., Memory Chips

From the early days of semiconductor memory until the early 1990s, memory was manufactured, brought and installed as a single chip. Chip density went from 1K bits to 1M bits and beyond, but each chip was a separate unit. Early PCs often had empty sockets into which additional memory chips could be plugged, if and when the purchaser needed them. At present, a different arrangement is often used called SIMM or DIMM.

A group of chips, typically 8 to 16, is mounted on a tiny printed circuit board and sold as a unit. This unit is called a SIMM or DIMM depending on whether it has a row of connectors on one side or both sides of the board.

A typical SIMM configuration might have 8 chips with 32 megabits (4MB) each on the SIMM. The entire module then holds 32MB. Many computers have room for four modules, giving a total capacity of 128MB when using 32MB SIMMs. The first SIMMs had 30 connectors and delivered 8 bits at a time. The other connectors were addressing and control. A later SIMM had 72 connectors and delivered 32 bits at a time. For a machine like Pentium, which expected 64-bits at once, 72-connectors SIMMs were paired, each one delivering half the bits needed.

A DIMM is capable of delivering 64 data bits at once. Typical DIMM capacities are 64MB and up. Each DIMM has 84 gold patted connectors on each side for a total of 168 connectors. SIMM and DIMM are shown in Figure 24 (a) and (b) respectively. How they are put on a motherboard is shown in Figure 24 (c).

SIMM

DIMM

(C) DIMM on Motherboard

Figure 24: SIMM & DIMM

1.8.2 SDRAM, RDRAM, Cache RAM Types of Memory

The basic building block of the main memory remains the DRAM chip, as it has for decades. Until recently, there had been no significant changes in DRAM architecture since the early 1970s. The traditional DRAM chip is constrained both by its internal architecture and by its interface to the processor's memory bus. The two schemes that currently dominate the market are SDRAM and RDRAM. A third one, that is Cache RAM, is also very popular.

SDRAM (Synchronous DRAM)

One of the most widely used forms of DRAM is the synchronous DRAM (SDRAM). Unlike the traditional DRAM, which is asynchronous, the SDRAM exchanges data with the processor synchronized to an external clock signal and running at the full speed of the processor /memory bus without imposing wait states.

In a typical DRAM, the processor presents addresses and control levels to the memory, indicating that a set of data at a particular location in memory should be either read from or written into the DRAM. After a delay, the access time, the DRAM either writes or reads the data during the access-time delay. The DRAM performs various internal functions, such as activating the high capacitance of the row and column lines, sensing the data and routing the data out through the output buffers. The processor must simply wait through this delay, slowing system performance.

With synchronous access, the DRAM moves data in and out under control of the system clock. The processor or other master issues the instruction and address information, which is latched on to by the DRAM. The DRAM then responds after a set number of clock cycles. Meanwhile, the master can safely do other tasks while the SDRAM is processing the request.

The SDRAM employs a burst mode to eliminate the address setup time. In burst mode, a series of data bits can be clocked out rapidly after the first bit has been accessed. The mode is useful when all the bits to be accessed are in sequence and in the same row of the array as the initial access. In addition, the SDRAM has a multiple-bank internal architecture that improves opportunities for on-chip parallelism.

The mode register and associated control logic is another key feature differentiating SDRAMs from conventional DRAMs. It provides a mechanism to customize the SDRAM to suit specific system needs. The mode register specifies the burst length, which is the number of separate units of data synchronously fed onto the bus. The register also allows the programmer to adjust the latency between receipt of a read request and the beginning of data transfer.

The SDRAM performs best when it is transferring large blocks of data serially, such as for applications like word processing, spreadsheets, and multimedia.

RDRAM (Rambus DRAM)

RDRAM, developed by Rambus, has been adopted by Intel for its Pentium and Itanium processors. It has become the main competitor to SDRAM. RDRAM chips are vertical packages, with all pins on one side. The chip exchanges data with the processor over 28 wires no more than 12 centimeters long. The bus address up to 320 RDRAM chips and is rated at 1.6 GBps.

The special RDRAM bus delivers address and control information using an asynchronous block-oriented protocol. After an initial 480 ns access time, this produces the 1.6 GBps data rate. The speed of RDRAM is due to its high speed Bus. Rather than being controlled by the explicit RAS CAS R/W, and CE signals used in conventional DRAMs an RDAR gets a memory request over the high-speed bus. This request contains the desired address, the type of operation and the number of bytes in the operation.

CDRAM (Cache DRAM)

Cache DRAM (CDRAM), developed by Mitsubishi, integrates a small SRAM cache (16Kb) onto a generic DRAM chip. The SRAM on the CDRAM can be used in two ways. First, it can be used as true cache consisting of a number of 64-bit line. The cache mode of the CDRAM is effective for ordinary random access to memory.

The SRAM on the CDRAM can also be used as a buffer to support the serial access of a block of data. For example, to refresh a bit-mapped screen, the CDRAM can prefetch the data from the DRAM into the SRAM buffer. Subsequent accesses to chip result in accesses solely to the SRAM.

Check Your Progress 3

1. Difference between

   a) SDRAM and RDRAM b) SIMM and DIMM

   ……

   ……

   ……

   ……

2. A Computer supports a virtual memory address space of 1Giga Words, but a physical Memory size of 64 Mega Words. How many bits are needed to specify an instruction address for this machine?

   ……

   ……

   ……

   ……

1.9 SUMMARY

In this unit, we have discussed the details of the memory system of the computer. First we discussed the concept and the need of the memory hierarchy. Memory hierarchy is essential in computers as it provides an optimised low-cost memory system. The unit also covers details on the basic characteristics of RAMs and different kinds of ROMs. These details include the logic diagrams of RAMs and ROMs giving basic functioning through various control signals. We have also discussed the latest secondary storage technologies such as CD-ROM, DVD-ROM, CD-R, CD-RW etc. giving details about their data formats and access mechanisms.

The importance of high-speed memories such as cache memory, interleaved memory and associative memories are also described in detail. The high-speed memory, although small, provides a very good overall speed of the system due to locality of reference. There are several other concepts such as the memory system of the microcomputer which consists of different types of chips such as SIMM, DIMM and different types of memory such as SDRAM, RDRAM also defined in easy way. The unit also contains details on Virtual Memory. For more details on the memory system you can go through further units.

1.10 SOLUTIONS / ANSWERS

Check Your Progress 1

1. a) True
   b) True
   c) False
2. a) Ultraviolet light, electrically
   b) Dynamic
   c) Sequential
3.

	RAM	**ROM and all types of ROM's**
a)	Volatile Memory	Non – volatile
b)	Faster access time	Slower than RAM
c)	Higher cost per bit storage	Lower than RAM
d)	Random access	Sequential access
e)	Less storage capacity	Higher storage capacity

4. A type of EPROM called EEPROM is known as flash memory used in many I/O and storage devices. It is commonly used memory in embedded systems.

5. (a) Eight, since the word size is 8.

   (b) 4K = 4 × 1024 = 4096 words. Hence, there are 4096 memory addresses. Since $4096 = 2^{12}$ it requires 12 bits address code to specify one of 4096 addresses.

   (c) The memory has a capacity of 4096 bytes.

6. 4K bytes is actually 4 × 1024 = 4096 bytes and the DRAM holds 4096 eight bit words. Each word can be thought of as being stored in an 8 bit register and there are 4096 registers connected to a common data bus internal to the chip. Since $4096 = (64)^2$, the registers are arranged in a 64 × 64 array, that is there are $64=2^6$ rows and $64=2^6$ columns. This requires a 6 × 64 decoder to decode six- address inputs for the row select and a second 6 × 64 decoder to decode six other address

inputs for the column select. Using the structure as shown in Figure 3 (b), it requires only 6 bit address input.

While in the case of an old RAM, the chip requires 12 address lines (Please refer to Figure 2(b)), since 4096 = 2^{12} and there are 4096 different addresses.

7. 1 M Bytes = 2^{20}. 2^3 bits = 2^{23}
   256K × 1 bit = 2^8. 2^{10} bits = 2^{18}
   Hence, total number of RAM chips of size (256K × 1) = $\frac{2^{23}}{2^{18}} = 2^5 = 32$

Check Your Progress 2

1) A disk array known as RAID systems is a mass storage device that uses a set of hard disks and hard disk drives with a controller mounted in a single box. All the disks of a disk array form a single large storage unit. The RAID systems were developed to provide large secondary storage capacity with enhanced performance and enhanced reliability. The performance is based upon the data transfer rate, which is very high rather than taking an individual disk. The reliability can be achieved by two techniques that is mirroring (the system makes exact copies of files on two hard disks) and stripping (a file is partitioned into smaller parts and different parts of the files are stored on different disks).

2. a) True
   b) True
   c) False
   d) True
   e) False

3. The Cache memory is a very fast, small memory placed between CPU and main memory whose access time is closer to the processing speed of the CPU. It acts as a high-speed buffer between CPU and main memory and is used to temporarily store data and instruction needed during current processing. In memory interleaving, the main memory is divided into n number of equal size modules. When a program is loaded in to the main memory, its successive instruction in also available for the CPU, thus, it avoids memory access after each instruction execution and the total time speeds up.

4. Main memory Size = 64 Words
   Main Memory word size = 16 bits
   Cache Memory Size = 8 Blocks
   Cache Memory Block size = 32 words
   ⇒1 Block of Cache = 2 Words of RAM
   ⇒ Memory location address 25 is equivalent to Block address 12.
   ⇒ Total number of possible Blocks in Main Memory = 64/2 = 32 blocks

   (a) Associative Mapping: The block can be anywhere in the cache.

   (b) Direct Mapping:
   Size of Cache = 8 blocks
   Location of Block 12 in Cache = 12 modulo 8 = 4

   (c) 2 Way set associative mapping:
   Number of blocks in a set = 2
   Number of sets = Size of Cache in blocks / Number of blocks in a set
   = 8 / 2 = 4
   Block 12 will be located anywhere in (12 modulo 4) set, that is set 0.

Following figure gives the details of above schemes.

Cache mapping scheme

5. The maximum size of physical memory address in the present machine
   = 6 bits (as Memory have 64 words, that is, 2^6)
   The format of address mapping diagram for Direct and Set-Associative Mapping:

Physical Memory Word Address		
Block Address		Block Offset
Tag	Index	Block Offset

The address mapping in Direct Mapping:

Memory Address	0	1	1	0	0	1	⇒ Memory Address = 25
Block Address	0	1	1	0	0	1	⇒ Block Address = 12 and . Block offset = 1
Cache Address	0	1	1	0	0	1	⇒ Tag = 1; Index = 4 and . Block offset = 1

Please note that a main memory address 24 would have a block offset as 0.

Address Mapping for 2 way set associative mapping:

Memory Address	0	1	1	0	0	1	⇒ Memory Address = 25
Block Address	0	1	1	0	0	1	⇒ Block Address = 12 and . Block offset = 1
Cache Address	0	1	1	0	0	1	⇒ Tag = 3; Index (Set . Number) = 0 and . Block offset = 1

The Tag is used here to check whether a given address is in a specified set. This cache has 2 blocks per set, thus, the name two way set associative cache. The total number of sets here is 8 / 2 = 4.

For Associative mapping the Block address is checked directly in all location of cache memory.

6. There are 64 words in a block, therefore 4K cache has (4 × 1024)/ 64 = 64 blocks. Since 1 set has 4 blocks, there are 16 sets. 16 sets need 4 bit as 2^4 = 16

representation. In a set there are 4 blocks. So, the block field needs 2 bits. Each block has 64 words. So the block offset field has 6 bits.
Index Filed is of 4 bits.
Block offset is of 6 bits.

Check Your Progress 3

1. a) SDRAM exchanges data with the processor synchronized to an external clock signal and running at the full speed of the processor/memory bus without imposing wait states.

   A RDRAM module sends data to the controller synchronously to the clock to master, and the controller sends data to an RDRAM synchronously with the clock signal in the opposite direction.

   b) In SIMM, a group of chips, typically 8 to 16, is mounted on a tiny printed circuit board with one side only. While in DIMM, this can be mounted on both sides of the printed circuit board.

   The latest SIMMs and DIMMs are capable of delivering 64 data bits at once. Each DIMM has 84 gold patted connectors on each side for a total of 168 connectors while each SIMM 72 connectors.

2. The virtual address is 1 GB = 2^{30}, thus, 30 bit Virtual address, that will be translated to physical memory address of 26 bits (64 Mega words = 2^{26}).

UNIT 2 THE INPUT/OUTPUT SYSTEM

Structure **Page No.**

2.0 INTRODUCTION

In the previous Unit, we have discussed the memory system for a computer system that contains primary memory, secondary memory, high speed memory and their technologies; the memory system of micro-computers i.e., their chips and types of memory. Another important component in addition to discussing the memory system will be the input/output system. In this unit we will discuss Input /Output controllers, device drivers, the structure of I/O interface, the I/O techniques. We will also discuss about the Input / Output processors which were quite common in mainframe computers.

2.1 OBJECTIVES

At the end of this unit you should be able to:

- identify the structure of input/output interface;
- identify the concepts of device drivers and device controllers;
- describe the input/output techniques, i.e., programmed I/O, interrupt-driven I/O and direct memory access;
- define an input/output processor;
- describe external communication interfaces such as serial and parallel interfaces; and
- define interrupt processing.

2.2 INPUT/ OUTPUT DEVICES OR EXTERNAL OR PERIPHERAL DEVICES

Before going on to discuss the input/ output sub/systems of a computer, let us discuss how a digital computer system can be implemented by a microcomputer system. A typical microcomputer system consists of a microprocessor plus memory and I/O interface. The various components that form the system are linked through buses that transfer instructions, data, addresses and control information among the components. The block diagram of a microcomputer system is shown in Figure. 1.

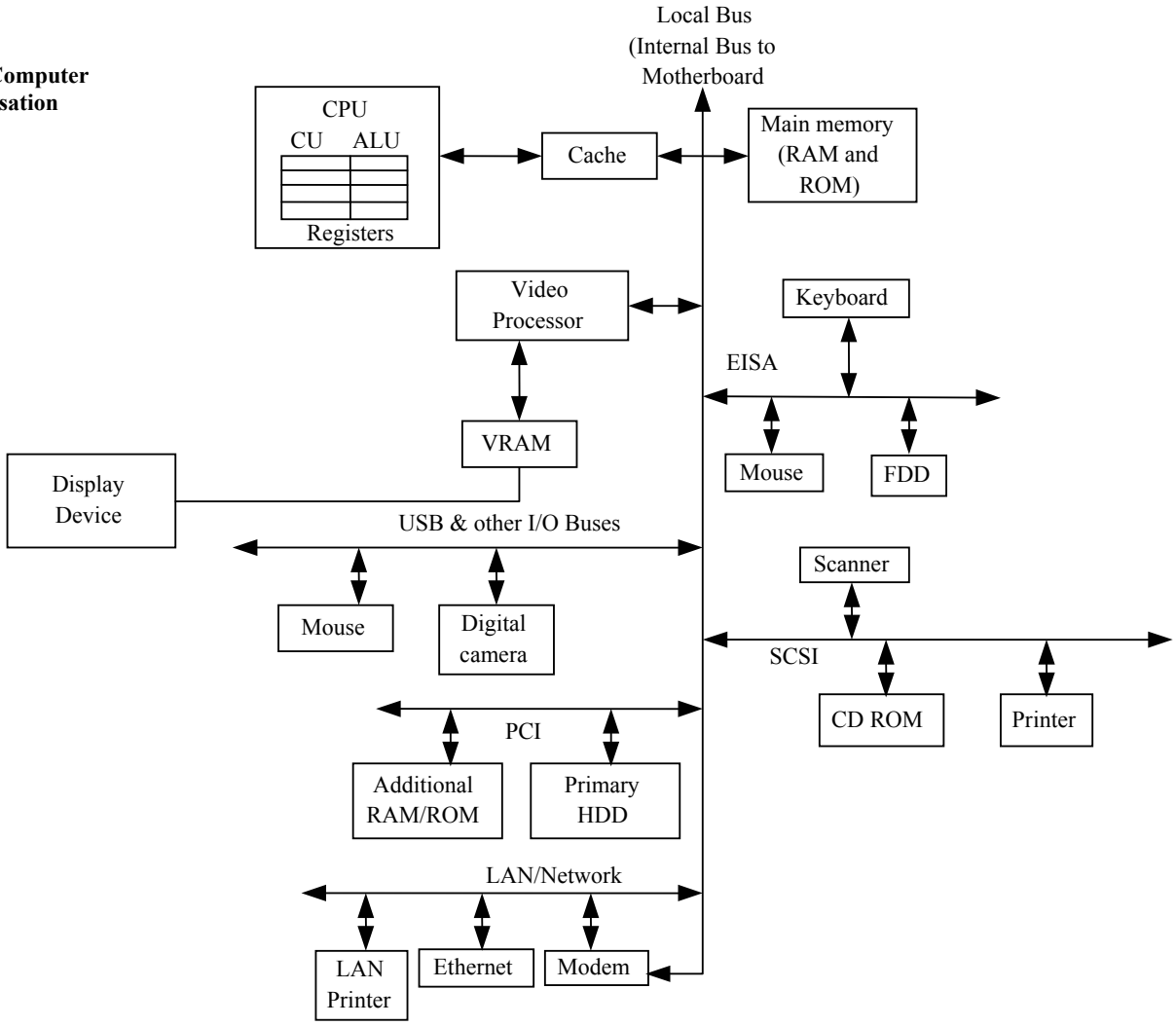


Figure 1: Block Diagram of a Microcomputer System

The microcomputer has a single microprocessor, a number of RAM and ROM chips and an interface units communicates with various external devices through the I/O Bus.

The Input / Output subsystem of a computer, referred to as I/O, provides an efficient mode of communication between the central system and the output environment. External devices that are under the direct control of the computers are said to be connected **on-line**. These devices are designed to read information into or out of the memory unit upon command from the CPU and are considered to be part of the computer system. Input / Output devices attached to the computer are also called peripherals. We can broadly classify peripherals or external devices into 3 categories:

- Human readable: suitable for communicating with the computer user, e.g., video display terminals (VDTs) & printers.
- Machine-readable: suitable for communicating with equipment, e.g., magnetic disks and tape system.
- Communication: suitable for communicating with remote devices, e.g., terminal, a machine-readable device.

2.3 THE INPUT /OUTPUT INTERFACE

The Input /Output interface provides a method for transferring information between internal storage and external I/O devices. Peripherals connected to a computer need special communication links for interfacing them with the CPU. The purpose of the

communication link is to resolve the differences that exist between the central computer and each peripheral. The major differences are:

- Peripherals are electromagnetic and electromechanical devices and their operations are different from the operation of the CPU and the memory, which are electronic devices.
- The data transfer rate of peripherals is usually slower than the transfer rate of the CPU, and consequently a synchronization mechanism may be needed.
- Data codes and formats in peripherals differ from the word format in the CPU and memory.
- The operating modes of peripherals are different from each other and each must be controlled so as not to disturb the operation of other peripherals connected to the CPU.

To resolve these differences, computer systems include special hardware component between the CPU and peripherals to supervise and synchronize all input and output transfers. These components are called *interface* units because they interface between the processor bus and the peripheral device.

Functions of I/O Interface

An I/O interface is bridge between the processor and I/O devices. It controls the data exchange between the external devices and the main memory; or external devices and processor registers. Therefore, an I/O interface provides an interface internal to the computer which connects it to the processor and main memory and an interface external to the computer connecting it to external device or peripheral. The I/O interface should not only communicate the information from processor to main I/O device, but it should also coordinate these two. In addition, since there are speed differences between processor and I/O devices, the I/O interface should have facilities like buffer and error detection mechanism. Therefore, the major functions or requirements of an I/O interface are:

It should be able to provide control and timing signals

The need of I/O from various I/O devices by the processor is quite unpredictable. In fact it depends on I/O needs of particular programs and normally does not follow any pattern. Since, the I/O interface also shares system bus and memory for data input/output, control and timing are needed to coordinate the flow of data from/to external devices to/from processor or memory. For example, the control of the transfer of data from an external device to the processor might involve the following steps:

1. The processor enquires from the I/O interface to check the status of the attached device. The status can be busy, ready or out of order.
2. The I/O interface returns the device status.
3. If the device is operational and ready to transmit, the processor requests the transfer of data by means of a command, which is a binary signal, to the I/O interface.
4. The I/O interface obtains a unit of data (e.g., 8 or 16 bits) from the external device.
5. The data is transferred from the I/O interface to the processor.

It should communicate with the processor

The above example clearly specifies the need of communication between the processor and I/O interface. This communication involves the following steps:

1. Commands such as READ SECTOR, WRITE SECTOR, SEEK track number and SCAN record-id sent over the control bus.
2. Data that are exchanged between the processor and I/O interface sent over the data bus.
3. Status: As peripherals are so slow, it is important to know the status of the I/O interface. The status signals are BUSY or READY or in an error condition from I/O interface.
4. Address recognition as each word of memory has an address, so does each I/O device. Thus an I/O interface must recognize one unique address for each peripheral it controls.

It should communicate with the I/O device

Communication between I/O interface and I/O device is needed to complete the I/O operation. This communication involves commands, status or data.

It should have a provision for data buffering

Data buffering is quite useful for the purpose of smoothing out the gaps in speed of processor and the I/O devices. The data buffers are registers, which hold the I/O information temporarily. The I/O is performed in short bursts in which data are stored in buffer area while the device can take its own time to accept them. In I/O device to processor transfer, data are first transferred to the buffer and then passed on to the processor from these buffer registers. Thus, the I/O operation does not tie up the bus for slower I/O devices.

Error detection mechanism should be in-built

The error detection mechanism may involve checking the mechanical as well as data communication errors. These errors should be reported to the processor. The examples of the kind of mechanical errors that can occur in devices are paper jam in printer, mechanical failure, electrical failure etc. The data communication errors may be checked by using parity bit.

2.4 THE DEVICE CONTROLLERS AND ITS STRUCTURE

All the components of the computer communicate with the processor through the system bus. That means the I/O devices need to be attached to the system bus. However, I/O devices are not connected directly to the computer's system bus. Instead they are connected to an intermediate electronic device interface called a device controller, which in turn is connected to the system bus. Hence a device controller is an interface between an I/O device and the system bus. On one side, it knows how to communicate with the I/O device connected to it, and on the other it knows how to communicate with the computer's CPU or processor and memory through the system bus.

2.4.1 Device Controller

A device controller need not necessarily control a single device. It can usually control multiple I/O devices. It comes in the form of an electronic circuit board that plugs directly into the system bus, and there is a cable from the controller to each device it controls. The cables coming out of the controller are usually terminated at the back panel of the main computer box in the form of connectors known as ports.

The Figure 2 below illustrates how I/O devices are connected to a computer system through device controllers. Please note the following points in the diagram:

- Each I/O device is linked through a hardware interface called I/O Port.
- Single and Multi-port device controls single or multi-devices.
- The communication between I/O controller and Memory is through bus only in case of Direct Memory Access (DMA), whereas the path passes through the CPU for such communication in case of non-DMA.

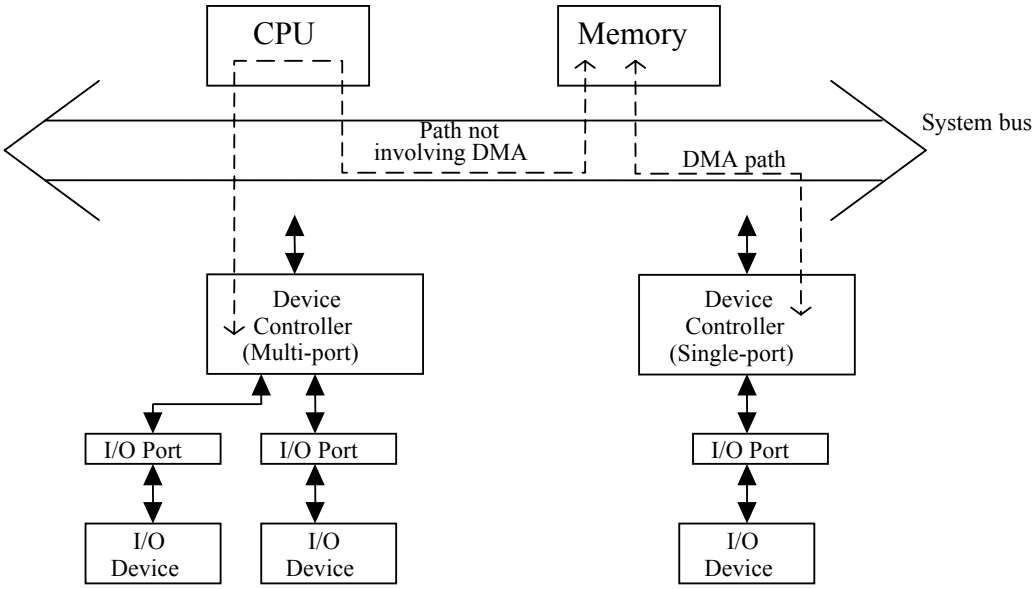


Figure 2: Connecting I/O Devices using Device Controller

Using device controllers for connecting I/O devices to a computer system instead of connecting them directly to the system bus has the following advantages:

- A device controller can be shared among multiple I/O devices allowing many I/O devices to be connected to the system.
- I/O devices can be easily upgraded or changed without any change in the computer system.
- I/O devices of manufacturers other than the computer manufacturer can be easily plugged in to the computer system. This provides more flexibility to the users in buying I/O devices of their choice.

2.4.2 Structure of an I/O Interface

Due to the complexity and the number of external devices that the I/O interface control, there is no standard structure of I/O interface. Let us give a general structure to an I/O interfaces:

- There is a need of I/O logic, which should interpret and execute dialogue between the processor and I/O interface. Therefore, there need to be control lines between processors and I/O interface.
- The data line connecting I/O interface to the system bus must exist. These lines serve the purpose of data transfer.
- Data registers may act as buffer between processor and I/O interface.
- The I/O interface contains logic specific to the interface with each device that it controls.

Figure 3: The General Structure of an I/O

Figure 3 above is a typical diagram of an I/O interface which in addition to all the registers as defined above has status/control registers which are used to pass on the status information or the control information.

2.5 DEVICE DRIVERS

A device driver is software interface which manages the communication with, and the control of, a specific I/O device, or type of device. It is the task of the device driver to convert the logical requests from the user into specific commands directed to the device itself. For example, a user request to write a record to a floppy disk would be realised within the device driver as a series of actions, such as checking for the presence of a disk in the drive, locating the file via the disk directory, positioning the heads, etc.

Device Drivers in UNIX, MS-DOS and Windows System

Although device drivers are in effect add-on modules, they are nevertheless considered to be part of the system since they are closely integrated with the Input/ Output Control System, which deals with I/O related system calls.

In UNIX the device drivers are usually linked onto the object code of the kernel (the core of the operating system). This means that when a new device is to be used, which was not included in the original construction of the operating system, the UNIX kernel has to be re-linked with the new device driver object code. This technique has the advantages of run-time efficiency and simplicity, but the disadvantage is that the addition of a new device requires regeneration of the kernel. In UNIX, each entry in the /dev directory is associated with a device driver which manages the communication with the related device. A list of some device names is as shown below:

Device name	Description
/dev/console	system console
/dev/tty01	user terminal 1
/dev/tty02	user terminal 2
/dev/lp	line printer
/dev/dsk/f03h	1.44 MB floppy drive

In MS-DOS, device drivers are installed and loaded dynamically, i.e., they are loaded into memory when the computer is started or re-booted and accessed by the operating system as required. The technique has the advantage that it makes addition of a new driver much simpler, so that it could be done by relatively unskilled users. The additional merit is that only those drivers which are actually required need to be loaded into the main memory. The device drivers to be loaded are defined in a special file called CONFIG.SYS, which must reside in the root directory. This file is automatically read by MS-DOS at start-up of the system, and its contents acted upon. A list of some device name is as shown below:

Device name	Description
con:	keyboard/screen
com1:	serial port1
com2:	serial port2
lpt1:	printer port1
A:	first disk drive
C:	hard disk drive

In the Windows system, device drivers are implemented as dynamic link libraries (DLLs). This technique has the advantages that DLLs contains shareable code which means that only one copy of the code needs to be loaded into memory. Secondly, a driver for a new device can be implemented by a software or hardware vendor without the need to modify or affect the Windows code, and lastly a range of optional drivers can be made available and configured for particular devices.

In the Windows system, the idea of Plug and Play device installation is required to add a new device such as a CD drive, etc. The objective is to make this process largely automatic; the device would be attached and the driver software loaded. Thereafter, the installation would be automatic; the settings would be chosen to suit the host computer configuration.

Check Your Progress 1

1. What are the functions of an I/O interface?

   ……

   ……

   ……

2. **State True or False:**

   (a) Com1 is a UNIX port. T/F
   (b) The buffering is done by data register. T/F
   (c) Device controller is shareable among devices. T/F
   (d) I/O system is basically needed for better system efficiency T/F
   (e) Device drives can be provided using software libraries. T/F
   (f) The devices are normally connected directly to the system bus. T/F
   (g) Data buffering is helpful for smoothing out the speed differences between CPU and input/output devices. T/F
   (h) Input/ output module is needed only for slower I/O devices T/F

3. What is a device driver? Differentiate between device controller and device drivers.

……

……

……

……

……

2.6 INPUT-OUTPUT TECHNIQUES

After going through the details of the device interfaces, the next point to be discussed is how the interface may be used to support input/output from devices. Binary information received from an external device is usually stored in memory for later processing. Information transferred from the central computer into an external device originates in the memory unit. Data transfer between the central computer and I/O devices may be handled in a variety of modes. Three techniques are possible for I/O operation. These are:

- Programmed input/output
- Interrupt driven input/output
- Direct memory access

Figure 4 gives an overview of these three techniques

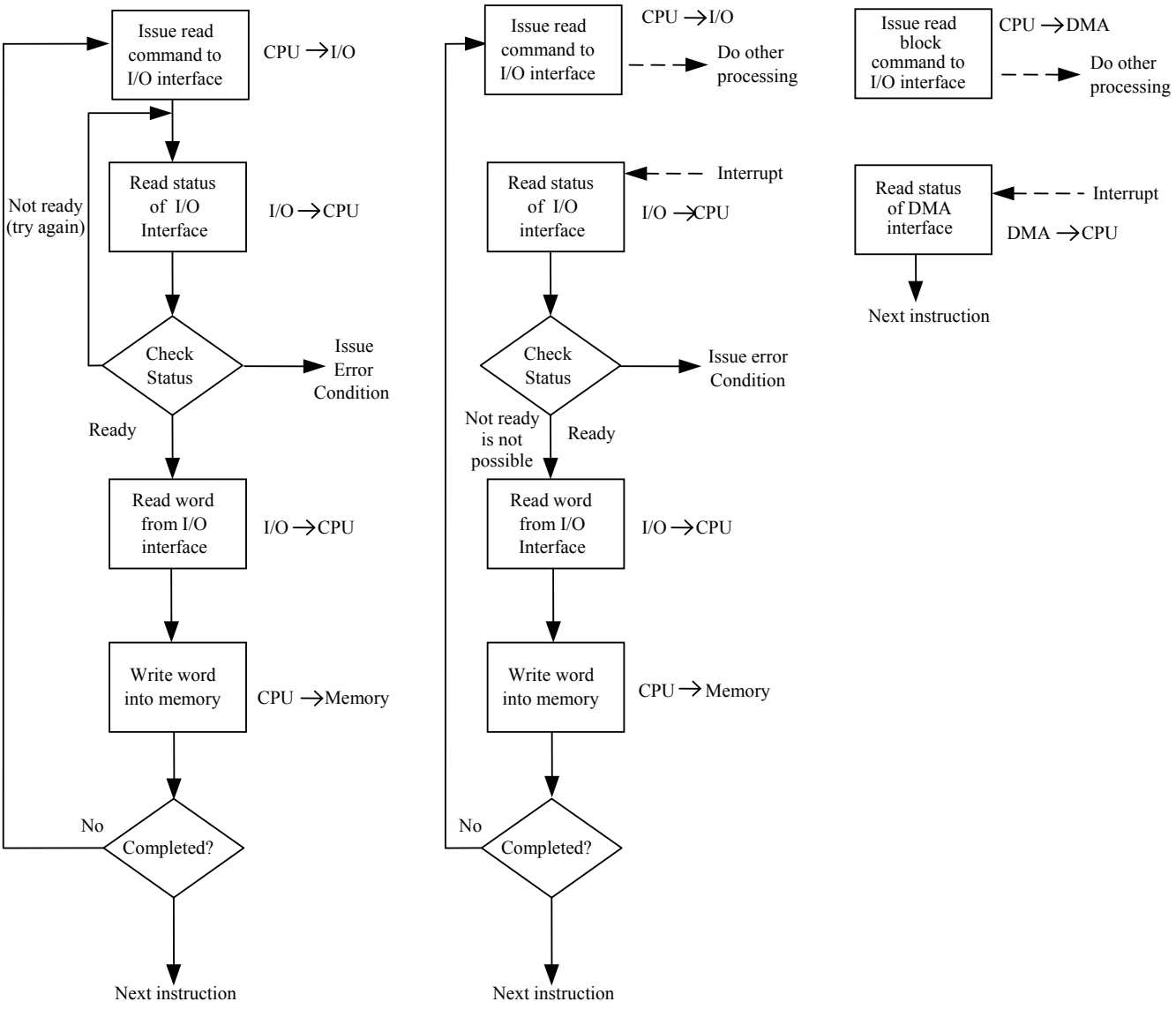
	Interrupt Required	**I/O interface to/from memory transfer (refer Figure 2)**
Programmed I/O	No	Through CPU
Interrupt-driven I/O	Yes	Through CPU
DMA	Yes	Direct to Memory

Figure 4: Overview of the three Input/ Output

In programmed I/O, the I/O operations are completely controlled by the processor. The processor executes a program that initiates, directs and terminate an I/O operation. It requires a little special I/O hardware, but is quite time consuming for the processor since the processor has to wait for slower I/O operations to complete.

With interrupt driven I/O, when the interface determines that the device is ready for data transfer, it generates an interrupt request to the computer. Upon detecting the external interrupt signal, the processor stops the task it is processing, branches to a service program to process the I/O transfer, and then returns to the task it was originally performing which results in the waiting time by the processor being reduced.

With both programmed and interrupt-driven I/O, the processor is responsible for extracting data from the main memory for output and storing data in the main memory during input. What about having an alternative where I/O device may directly store data or retrieve data from memory? This alternative is known as direct memory access (DMA). In this mode, the I/O interface and main memory exchange data directly, without the involvement of processor.



(a) Programmed I/O **(b) Interrupt Driven I/O** **(c) DMA**

Figure 5: Three techniques of I/O

2.6.1 Programmed Input /Output

Programmed input/output is a useful I/O method for computers where hardware costs need to be minimised. The input or output operation in such cases may involve:

a) Transfer of data from I/O device to the processor registers.

b) Transfer of data from processor registers to memory.

With the programmed I/O method, the responsibility of the processor is to constantly check the status of the I/O device to check whether it is free or it has finished inputting the data. Thus, this method is very time consuming where the processor wastes a lot of time in checking and verifying the status of an I/O device. Figure 5(a) gives an example of the use of programmed I/O to read in a block of data from a peripheral device into memory.

I/O Commands

There are four types of I/O commands that an I/O interface may receive when it is addressed by a processor:

- **Control:** These commands are device specific and are used to provide specific instructions to the device, e.g. a magnetic tape requiring rewinding and moving forward by a block.
- **Test:** This command checks the status such as if a device is ready or not or is in error condition.
- **Read:** This command is useful for input of data from input device.
- **Write:** this command is used for output of data to output device.

I/O Instructions:

An I/O instruction is stored in the memory of the computer and is fetched and executed by the processor producing an I/O-related command for the I/O interface. With programmed I/O, there is a close correspondence between the I/O-related instructions and the I/O commands that the processor issues to an I/O interface to execute the instructions.

In systems with programmed I/O, the I/O interface, the main memory and the processors normally share the system bus. Thus, each I/O interface should interpret the address lines to determine if the command is for itself. There are two methods for doing so. These are called memory-mapped I/O and isolated I/O.

With memory-mapped I/O, there is a single address space for memory locations and I/O devices. The processor treats the status and data registers of I/O interface as memory locations and uses the same machine instructions to access both memory and I/O devices. For a memory-mapped I/O only a single read and a single write line are needed for memory or I/O interface read or write operations. These lines are activated by the processor for either memory access or I/O device access. Figure 6 shows the memory-mapped I/O system structure.

Figure 6: Structure of Memory Mapped I/O

With isolated I/O, there are separate control lines for both memory and I/O device read or write operations. Thus a memory reference instruction does not affect an I/O device. In isolated I/O, the I/O devices and memory are addressed separately; hence separate input/output instructions are needed which cause data transfer between addressed I/O interface and processor. Figure 7 shows the structure of isolated I/O.

Figure 7: Structure of Isolated I/O

2.6.2 Interrupt-Driven Input/Output

The problem with programmed I/O is that the processor has to wait a long time for the I/O interface to see whether a device is free or wait till the completion of I/O. The result is that the performance of the processor goes down tremendously. What is the solution? What about the processor going back to do other useful work without waiting for the I/O device to complete or get freed up? But how will the processor be intimated about the completion of I/O or a device is ready for I/O? A well-designed mechanism was conceived for this, which is referred to as interrupt-driven I/O. In this mechanism, provision of interruption of processor work, once the device has finished the I/O or when it is ready for the I/O, has been provided.

The interrupt-driven I/O mechanism for transferring a block of data is shown in Figure 5(b). Please note that after issuing a read command (for input) the CPU goes off to do other useful work while I/O interface proceeds to read data from the associated device. On the completion of an instruction cycle, the CPU checks for interrupts (which will occur when data is in data register of I/O interface and it now needs CPU's attention). Now CPU saves the important register and processor status of the executing program in a stack and requests the I/O device to provide its data, which is placed on the data bus by the I/O device. After taking the required action with the data, the CPU can go back to the program it was executing before the interrupt.

2.6.3 Interrupt-Processing

The occurrence of an interrupt fires a numbers of events, both in the processor hardware and software. Figure 8 shows a sequence.

Figure 8: Interrupt-Processing Sequence

When an I/O device completes an I/O operation, the following sequence of hardware events occurs:

1. The device issues an interrupt signal to the processor.
2. The processor finishes execution of the current instruction *before responding* to the interrupt.
3. The processor tests for the interrupts and sends an acknowledgement signal to the device that issued the interrupt.
4. The minimum information required to be stored for the task being currently executed, before the CPU starts executing the interrupt routine (using its registers) are:
   (a) The status of the processor, which is contained in the register called program status word (PSW), and
   (b) The location of the next instruction to be executed, of the currently executing program, which is contained in the program counter (PC).

5. Th e processor now loads the PC with the entry location of the interrupt-handling program that will respond to this interrupting condition. Once the PC has been loaded, the processor proceeds to execute the next instruction, that is the next instruction cycle, which begins with an instruction fetch. Because the instruction fetch is determined by the contents of the PC, the result is that control is transferred to the interrupt-handler program. The execution results in the following operations:
6. The PC & PSW relating to the interrupted program have already been saved on the system stack. In addition, the contents of the processor registers are also needed to be saved on the stack that are used by the called Interrupt Servicing Routine because these registers may be modified by the interrupt-handler. Figure 9(a) shows a simple example. Here a user program is interrupted after the instruction at location N. The contents of all of the registers plus the address of the next instruction (N+1) are pushed on to the stack.
7. The interrupt handler next processes the interrupt. This includes determining of the event that caused the interrupt and also the status information relating to the I/O operation.
8. When interrupt processing is complete, the saved register values are retrieved from the stack and restored to the registers, which are shown in Figure 9(b).
9. The final step is to restore the values of PSW and PC from the stack. As a result, the instruction to be executed will be from the previously interrupted program.

Figure 9: Interrupt Handling

Thus, interrupt handling involves interruption of the currently executing program, execution of interrupt servicing program and restart of interrupted program from the point of interruption.

Design issues: Two design issues arise in implementing interrupt-driven I/O:

1) How does the processor determine which device issued the interrupt?
2) If multiple interrupts have occurred, how does the processor decide which one to be processed first?

To solve these problems, four general categories of techniques are in common use:

- **Multiple Interrupt Lines:** The simplest solution to the problems above is to provide multiple interrupt lines, which will result in immediate recognition of the interrupting device. Priorities can be assigned to various interrupts and the interrupt with the highest priority should be selected for service in case a multiple interrupt occurs. But providing multiple interrupt lines is an impractical approach because only a few lines of the system bus can be devoted for the interrupt.
- **Software Poll:** In this scheme, on the occurrence of an interrupt, the processor jumps to an interrupt service program or routine whose job it is to poll (roll call) each I/O interface to determine which I/O interface has caused the interrupt. This may be achieved by reading the status register of the I/O interface. Once the correct interface is identified, the processor branches to a device-service routine specific to that device. The disadvantage of the software poll is that it is time consuming.
- **Daisy chain:** This scheme provides a hardware poll. With this technique, an interrupt acknowledge line is chained through various interrupt devices. All I/O interfaces share a common interrupt request line. When the processor senses an interrupt, it sends out an interrupt acknowledgement. This signal passes through all the I/O devices until it gets to the requesting device. The first device which has made the interrupt request thus senses the signal and responds by putting in a word which is normally an address of interrupt servicing program or a unique identifier on the data lines. This word is also referred to as interrupt vector. This address or identifier in turn is used for selecting an appropriate interrupt-servicing program. The daisy chaining has an in-built priority scheme, which is determined by the sequence of devices on interrupt acknowledge line.
- **Bus arbitration:** In this scheme, the I/O interface first needs to control the bus and only after that it can request for an interrupt. In this scheme, since only one of the interfaces can control the bus, therefore only one request can be made at a time. The interrupt request is acknowledged by the CPU on response of which I/O interface places the interrupt vector on the data lines. An interrupt vector normally contains the address of the interrupt serving program.

An example of an interrupt vector can be a personal computer, where there are several IRQs (Interrupt request) for a specific type of interrupt.

2.6.4 DMA (Direct Memory Access)

In both interrupt-driven and programmed I/O, the processor is busy with executing input/output instructions and the I/O transfer rate is limited by the speed with which the processor can test and service a device. What about a technique that requires minimal intervention of the CPU for input/output? These two types of drawbacks can be overcome with a more efficient technique known as DMA, which acts as if it has taken over control from the processor. Hence, the question is: why do we use DMA interface? It is used primarily when a large amount of data is to be transferred from the I/O device to the Memory.

DMA Function

Although the CPU intervention in DMA is minimised, yet it must use the path between interfaces that is the system bus. Thus, DMA involves an additional interface on the system bus. A technique called cycle stealing allows the DMA interface to transfer one data word at a time, after which it must return control of the bus to the processor. The processor merely delays its operation for one memory cycle to allow the directly memory I/O transfer to "steal" one memory cycle. When an I/O is requested, the processor issues a command to the DMA interface by sending to the DMA interface the following information (Figure 10):

- Which operations (read or write) to be performed, using the read or write control lines.
- The address of I/O devices, which is to be used, communicated on the data lines.
- The starting location on the memory where the information will be read or written to be communicated on the data lines and is stored by the DMA interface in its address register.
- The number of words to be read or written is communicated on the data lines and is stored in the data count register.

Figure 10: DMA block diagram

The DMA interface transfers the entire block of data, one word at a time, directly to or from memory, without going through the processor. When the transfer is complete, the DMA interface sends an interrupt signal to the processor. Thus, in DMA the processor involvement can be restricted at the beginning and end of the transfer, which can be shown as in the figure above. But the question is when should the DMA take control of the bus?

For this we will recall the phenomenon of execution of an instruction by the processor. Figure 11 below shows the five cycles for an instruction execution. The Figure also shows the five points where a DMA request can be responded to and a point where the interrupt request can be responded to. Please note that an interrupt request is acknowledged only at one point of an instruction cycle, and that is at the interrupt cycle.

Figure 11: DMA and Interrupt Breakpoints

The DMA mechanism can be configured into a variety of ways. Some possibilities are shown below in Figure 12(a), in which all interfaces share the same system bus. The DMA acts as the supportive processor and can use programmed I/O for exchanging data between memory and I/O interface through DMA interface. But once again this spoils the basic advantage of DMA not using extra cycles for transferring information from memory to/from DMA and DMA from/to I/O interface.

Figure 12: DMA Configuration

The Figure 12(b) configuration suggests advantages over the one shown above. In these systems a path is provided between I/O interface and DMA interface, which does not include the system bus. The DMA logic may become part of an I/O interface and can control one or more I/O interfaces. In an extended concept an I/O bus can be connected to this DMA interface. Such a configuration (shown in Figure 12 (c)) is quite flexible and can be extended very easily. In both these configurations, the added advantage is that the data between I/O interface and DMA interface is transferred off the system bus, thus eliminating the disadvantage we have witnessed for the first configuration.

Check Your Progress 2

1. Which of the I/O techniques does not require an Interrupt Signal? Is this technique useful in Multiprogramming Operating Systems? Give reason.

   ……

   ……

   ……

2. What are the techniques of identifying the device that has caused the Interrupt?

   ……

   ……

   ……

3. What are the functions of I/O interface? What is DMA?

   ……

   ……

   ……

4 State True or False:

   a) Daisy chain provides software poll. T/F

   b) I/O mapped I/O scheme requires no additional lines from CPU to I/O device except for the system bus. T/F

   c) Most of the I/O processors have their own memory while a DMA module does not have its own memory except for a register or a simple buffer area. T/F

   d) The advantage of interrupt driven I/O over programmed I/O is that in the first the interrupt mechanisms free I/O devices quickly. T/F

2.7 INPUT-OUTPUT PROCESSORS

Before discussing I/O processors, let us briefly recapitulate the development in the area of input/output functions. These can be summarised as:

1. The CPU directly controls a peripheral device.
2. Addition of I/O controller or I/O interface: The CPU uses programmed I/O without interrupts. CPU was separated from the details of external I/O interfaces.
3. Contained use of I/O controllers but with interrupts: The CPU need not spend time waiting for an I/O operation to be performed, increasing efficiency.
4. Direct access of I/O interface to the memory via DMA: CPU involvement reduced to at the beginning and at the end of DMA operation.

5. The CPU directs the I/O processors to execute an I/O program in memory. The I/O processor fetches and executes these instructions without CPU intervention. This allows the CPU to specify a sequence of I/O activities and to be interrupted only when the entire sequence has been performed. With this architecture, a large set of I/O devices can be controlled, with minimum CPU involvement.

With the last two steps (4 and 5), a major change occurs with the introduction of the concept of an I/O interface capable of executing a program. For steps 5, the I/O interface is often referred to as an I/O channel and I/O processor.

Characteristics of I/O Channels

The I/O channel represents an extension of the DMA concept. An I/O channel has the ability to execute I/O instructions, which gives complete control over the I/O operation. With such devices, the CPU does not execute I/O instructions. Such instructions are stored in the main memory to be executed by a special-purpose processor in the I/O channel itself. Thus, the CPU initiates an I/O transfer by instructing the I/O channel to execute a program in memory. Two types of I/O channels are commonly used which can be seen in Figure 13 (a and b).

(a) Selector Channel **(b) Multiplexer Channel**

Figure 13: I/O Channel Structures

A **selector channel** controls multiple high-speed devices and, at any one time, is dedicated to the transfer of data with one of those devices. Each device is handled by a controller or I/O interface. Thus the I/O channel serves in place of the CPU in controlling these I/O controllers.

A **multiplexer channel** can handle I/O with multiple devices at the same time. If the devices are slow then byte multiplexer is used. Let us explain this with an example. If we have three slow devices which need to send individual bytes as:

X1 X2 X3 X4 X5 ……
Y1 Y2 Y3 Y4 Y5……
Z1 Z2 Z3 Z4 Z5……

Then on a byte multiplexer channel they may send the bytes as X1 Y1 Z1 X2 Y2 Z2 X3 Y3 Z3…… For high-speed devices, blocks of data from several devices are interleaved. These devices are called **block multiplexer**.

2.8 EXTERNAL COMMUNICATION INTERFACES

The external interface is the interface between the I/O interface and the peripheral devices. This interface can be characterised into two main categories: (a) parallel interface and (b) serial interface.

In parallel interface multiple bits can be transferred simultaneously. The parallel interface is normally used for high-speed peripherals such as tapes and disks. The dialogues that take place across the interface include the exchange of control information and data.

In serial interface only one line is used to transmit data, therefore only one bit is transferred at a time. Serial printers are used for serial printers and terminals. With a new generation of high-speed serial interfaces, parallel interfaces are becoming less common.

In both cases, the I/O interface must engage in a dialogue with the peripheral. The dialogue for a read or write operation is as follows:

- A control signal is sent by I/O interface to the peripheral requesting the permission to send (for write) or receive (for read) data.
- The peripheral acknowledges the request.
- The data are transferred from I/O interface to peripheral (for write) or from peripheral to I/O interface (for read).
- The peripheral acknowledges receipt of the data.

The connection between an I/O interface in a computer system and external devices can be either point-to-point or multipoint. A point-to-point interface provides a dedicated line between the I/O interface and the external device. For example keyboard, printer and external modems are point-to-point links. The most common serial interfaces are RS-232C and EIA-232.

A multipoint external interface used to support external mass storage devices (such as disk and tape drives) and multimedia devices (such as CD-ROM, video, audio).

Two important examples of external interfaces are FireWire and InfiniBand.

Check Your Progress 3

1. What is the need of I/O channels?

   ……

   ……

   ……

   ……

2. What is the need of external Communication Interfaces?

   ……

   ……

   ……

   ……

2.9 SUMMARY

This unit is totally devoted to the I/O of computer system. In this unit we have discussed the identification of I/O interface, description of I/O techniques such as programmed I/O, interrupt-driven I/O and direct memory access. These techniques are useful for increasing the efficiency of the input-output transfer process. The concepts of device drivers for all types of operating systems and device controllers are also discussed with this unit. We have also defined an input/output processor, the external communication interfaces such as serial and parallel interfaces and interrupt processing. The I/O processors are the most powerful I/O interfaces that can execute the complete I/O instructions. You can always refer to further reading for detail design.

2.10 SOLUTIONS /ANSWERS

Check Your Progress 1

1. The functions of I/O interfaces are to provide:
   - Timing and control signal.
   - Communication with processor and I/O devices.
   - Support for smoothing the speed gap between CPU and Memory using buffering.
   - Error detection.

2. (a) False (b) True (c) True (d) True (e) True (f) False (g) True (h) False

3. A device driver is a software module which manages the communication with, and the control of, a specific I/O device, or type of device. The difference between device driver and controller are:
   - One device controller can control many devices, whereas drivers are device specific.
   - Device controllers are a more intelligent hardware-software combination than device drivers.
   - I/O controllers allow different types and upgradeability of devices whereas device driver is device specific.

Check Your Progress 2

1. The technique Programmed I/O does not require an Interrupt. It is very inefficient for Multiprogramming environment as the processor is busy waiting for the I/O to complete, while this time would have been used for instruction execution of other programs.

2. The techniques for recognition of interrupting device/conditions can be:

   - Multiple Interrupt Lines: Having separate line for a device, thus direct recognition.
   - Software Poll: A software driven roll call to find from devices whether it has made an interrupt request.
   - Daisy Chain: A hardware driven passing the buck type signal that moves through the devices connected serially. The device on receipt of signal on his turn, if has interrupt informs its address.
   - Bus Arbitration: In this scheme, the I/O interface requests for control of the Bus. This is a common process when I/O processors are used.

3. The functions of I/O interface are:

   - Control and timing signals
   - CPU communications
   - I/O device communication
   - Data buffering
   - In-built error-detection mechanism.

   DMA is an I/O technique that minimises the CPU intervention at the beginning and end of a time consuming I/O. One, commonplace where DMA is used is when I/O is required from a Hard Disk, since one single I/O request requires a block of data transfer which on the average may take a few milliseconds. Thus, DMA will free CPU to do other useful tasks while I/O is going on.

4. a) False
   b) False
   c) True
   d) False

Check Your Progress 3

1. The I/O channels were popular in older mainframes, which included many I/O devices and I/O requests from many users. The I/O channel takes control of all I/O instructions from the main processor and controls the I/O requests. It is mainly needed in situations having many I/O devices, which may be shared among multiple users.

2. The external interfaces are the standard interfaces that are used to connect third party or other external devices. The standardization in this area is a must.

UNIT 3 SECONDARY STORAGE TECHNIQUES

Structure **Page No.**

3.0 INTRODUCTION

In the previous units of this block, we have discussed the primary memory system, high speed memories, the memory system of microcomputer, and the input/output interfaces and techniques for a computer. In this unit we will discuss the secondary storage devices such as magnetic tapes, magnetic disks and optical disks, also known as backing storage devices. The main purpose of such a device is that it provides a means of retaining information on a permanent basis. The main discussion provides the characteristics of hard-drives, formatting, drive cache, interfaces, etc. The detailed discussion on storage devices is being presented in the Unit. The storage technologies have moved a dimension from very small storage devices to Huge Giga byte memories. Let us also discuss some of the technological achievements that made such a technology possible.

3.1 OBJECTIVES

Storage is the collection of places where long-term information is kept. At the end of the unit you will be able to:

- describe the characteristics of the different secondary storage drives, i.e., their drive speed, access time, rotation speed, density etc.;
- describe the low-level and high level formatting of a blank disk and also the use of disk partitioning;
- distinguish among the various types of drives, i.e., hard drives , optical drives removable drives and cartridge drive; and
- define different type of disk formats.

3.2 SECONDARY STORAGE SYSTEMS

As discussed in Block 2 Unit 1, there are several limitations of primary memory such as limited capacity, that is, it is not sufficient to store a very large volume of data; and volatility, that is, when the power is turned off the data stored is lost. Thus, the secondary storage system must offer large storage capacities, low cost per bit and medium access times. Magnetic media have been used for such purposes for a long time. Current magnetic data storage devices take the form of floppy disks and hard disks and are used as secondary storage devices. But audio and video media, either in compressed form or uncompressed form, require higher storage capacity than the other media forms and the storage cost for such media is significantly higher.

Optical storage devices offer a higher storage density at a lower cost. CD-ROM can be used as an optical storage device. Many software companies offer both operating system and application software on CD-ROM today. This technology has been the main catalyst for the development of multimedia in computing because it is used in the multimedia external devices such as video recorders and digital recorders (Digital Audio Tape) which can be used for the multimedia systems.

Removable disk, tape cartridges are other forms of secondary storage devices are used for back-up purposes having higher storage density and higher transfer rate.

3.3 HARD DRIVES

The Disks are normally mounted on a disk drive that consists of an arm and a shaft along with the electronic circuitry for read-write of data. The disk rotates along with the shaft. A non-removable disk is permanently mounted on the disk drive. One of the most important examples of a non-removable disk is the hard disk of the PC. The disk is a platter coated with magnetic particles. Early drives were large. Later on, smaller hard (rigid) disk drivers were developed with fixed and removable pack. Each pack held about 30MB of data and became known as the Winchester drive. The storage capacity of today's Winchester disks is usually of the order of a few tens of Megabytes to a few Gigabytes. Most Winchester drives have the following common features:

- the disk and read/write heads are enclosed in a sealed airtight unit;
- the disk(s) spin at a high speed, one such speed may be 7200 revolutions per minute;
- the read/write head do not actually touch the disk surface;
- the disk surface contains a magnetic coating;
- the data on disk surface (platter) are arranged in the series of concentric rings. Each ring is called a track, is subdivided into a number of sectors, each sector holding a specific number of data elements called bytes or characters.
- The smallest unit that can be written to or read from the disk is a sector. The storage capacity of the disk can be determined as the number of tracks, number of sectors, byte per sector and number of read/write heads.

(a) An Open Disk Casing (b) Tracks and Cylinders

Figure 1: The Hard Disk

3.3.1 Characteristics: Drive Speed, Access Time, Rotation Speed

Tracks and Sectors: The disk is divided into concentric rings called tracks. A track is thus one complete rotation of the disk underneath the read/write head. Each track is subdivided into a number of sectors. Each sector contains a specific number of bytes or characters. Typical sector capacities are 128, 256, 512, 1024 and 4096 bytes.

Bad Blocks: The drive maintains an internal table which holds the sectors or tracks which cannot be read or written to because of surface imperfections. This table is called the bad block table and is created when the disk surface is initially scanned during a low-level format.

Sector Interleave: This refers to the numbering of the sectors located in a track. A one to one interleave has sectors numbered sequentially 0,1,2,3,4 etc. The disk drive rotates at a fixed speed 7200 RPM, which means that there is a fixed time interval between sectors. A slow computer can issue a command to read sector 0, storing it in an internal buffer. While it is doing this, the drive makes available sector 1 but the computer is still busy storing sector 0. Thus the computer will now have to wait one full revolution till sector 1 becomes available again. Renumbering the sectors like 0,8,1,9,2,10,3,11 etc., gives a 2:1 interleave. This means that the sectors are alternated, giving the computer slightly more time to store sectors internally than previously.

Drive Speed: The amount of information that can be transferred in or out of the memory in a second is termed as disk drive speed or data transfer rate. The speed of the disk drive depends on two aspects, bandwidth and latency.

- **Bandwidth:** The bandwidth can be measured in bytes per second. The sustained bandwidth is the average data rate during a large transfer, i.e., the number of bytes divided by the transfer time. The effective bandwidth is the overall data rate provided by the drive. The disk drive bandwidth ranges from less than 0.25 megabytes per second to more than 30 megabytes per second.
- **Access latency:** A disk access simply moves the arm to the selected cylinder and waits for the rotational latency, which may take less than 36ms. The latency

depends upon the rotation speed of the disk which may be anywhere from 300 RPM to 7200 RPM. An average latency of a disk system is equal to half the time taken by the disk to rotate once. Hence, the average latency of a disk system whose rotation speed is 7200 RPM will be 0.5 / 7200 minutes = 4.1 ms.

Rotation Speed: This refers to the speed of rotation of the disk. Most hard disks rotate at 7200 RPM (Revolution per Minute). To increase data transfer rates, higher rotation speeds, or multiple read/write heads arranged in parallel or disk arrays are required.

Access Time: The access time is the time required between the requests made for a read or write operation till the time the data are made available or written at the requested location. Normally it is measured for read operation. The access time depends on physical characteristics and access mode used for that device.

The disk access time has two major components:

- **Seek Time:** The seek time is the time for the disk arm to move the heads to the cylinder containing the desired sector.
- **Latency Time:** The latency time is the additional time waiting for the disk to rotate the desired sector to the disk head.

The sums of average seek and latency time is known as the average access time.

3.3.2 Partitioning and Formatting: FAT, Inode

Today the modern PC contains total capacity of approximately 40GB for storage of program and data. Because of this huge capacity, instead of having only one operating system in our PC, partitions are used to provide several separate areas within one disk, each treated as a separate storage device. That is, a disk partition is a sub-division of the disk into one or more areas. Each partition can be used to hold a different operating system. The computer system boots from the active partition and software provided allows the user to select which partition is the active one.

For example, we can run both Windows and Linux operating systems from the same storage of the PC.

A new magnetic disk is just platters of a magnetic recording material. Before a disk can store data, it must be divided into sectors that the disk controller can read and write. This is called low level formatting. Low level formatting fills the disk with a special data structure for each sector, which consists of a header, a data area, and a trailer. The low level formatting is placing track and sector information plus bad block tables and other timing information on the disks. Sector interleave can also be specified at this time.

In any disk system, space at some time in use will become unwanted and hence will be ‘free’ for another application. The operating system allocates disk space on demand by user programs. Generally, space is allocated in units of fixed size called an allocation unit or a cluster, which is a simple multiple of the disk physical sector size, usually 512 bytes. The DOS operating system forms a cluster by combining two or more sectors so that the smallest unit of data access from a disk becomes a cluster, not a sector. Normally, the size of the cluster can range from 2 to 64 sectors per cluster.

High level formatting involves writing directory structures and a map of free and allocated space (FAT or INODE) to the disk. Often this also means transferring the boot file for the operating system onto the hard disks.

FAT and Inode

The FAT maps the usage of data space of the disk. It contains information about the space used by each individual file, the unused disk space and the space that is unusable due to defects in the disk. Since FAT contains vital information, two copies of FAT are stored on the disk, so that in case one gets destroyed, the other can be used. A FAT entry can contain any of the following:

- unused cluster
- reserved cluster
- bad cluster
- last cluster in file
- next cluster number in the file.

The DOS file system maintains a table of pointers called FAT (File allocation table) which consists of an array of 16-bit values. There is one entry in the FAT for each cluster in the file area, i.e., each entry of the FAT (except the two) corresponds to one cluster of disk space. If the value in the FAT entry doesn't mark an unused, reserved or defective cluster, then the cluster corresponding to the FAT entry is part of a file and the value in the FAT entry would indicate the next cluster in the file.

The first two entries (0 & 1) in FAT are reserved for use by the operating system. Therefore, the cluster number 2 corresponds to the first cluster in the data space of the disk. Prior to any data being written on to the disk, the FAT entries are all set to zero indicating a 'free' cluster .The FAT chain for a file ends with the hexadecimal value, i.e., FFFF. The FAT structure can be shown as in Figure 2 below.

Figure 2: FAT structure

Limitation of FAT16: The DOS designers decided to use clusters with at least four sectors in them (thus a cluster size of at least 2KB) for all FAT16 hard disks. That size suffices for any hard disk with less than a 128MB total capacity. The largest logical disk drives that DOS can handle comfortably have capacities up to 2GB. For such a large volume, the cluster size is 32KB. This means that even if a file contains only a single byte of data, writing it to the disk uses one entire 32KB region of the disk, making that area unavailable for any other file's data storage.
The most recent solution to these large-disk problems was introduced by Microsoft in its OSR2 release of Windows 95 and it was named FAT32. The cluster entry for FAT32 uses 32-bit numbers. The minimum size for a FAT32 volume is 512MB. Microsoft has reserved the top four bits of every cluster number in a FAT32 file

allocation table. That means there are only 28-bits for the cluster number, so the maximum cluster number possible is 268,435,456.

In the UNIX system, the information related to all these fields is stored in an Inode table on the disk. For each file, there is an inode entry in the table. Each entry is made up of 64 bytes and contains the relevant details for that file. These details are:

a) Owner of the file
b) Group to which the Owner belongs
c) File type
d) File access permissions
e) Date & time of last access
f) Date & time of last modification
g) Size of the file
h) No. of links
i) Addresses of blocks where the file is physically present.

3.3.3 Drive Cache

Disk cache may be a portion of RAM, sometimes called soft disk cache that is used to speed up the access time on a disk. In the latest technologies such memory can be a part of disk drive itself. Such memory is sometimes called hard disk cache or buffer.

These hard disk caches are more effective, particularly applicable in multiprogramming machines or in disk file servers, but they are expensive, and therefore are smaller. Almost all-modern disk drives include a small amount of internal cache. The cycle time of cache would be about a tenth of the main memory cycle time and its cost about 10 times the cost per byte of main memory.

The disk caching technique can be used to speed up the performance of the disk drive system. A set (cache) of buffers is allocated to hold a number of disk blocks which have been recently accessed. In effect, the cached blocks are in memory copies of the disk blocks. If the data in a cache buffer memory is modified, only the local copy is updated at that time. Hence processing of the data takes place using the cached data avoiding the need to frequently access the disk itself.

The main disadvantage of the system using disk caching is risking loss of updated information in the event of machine failures such as loss of power. For this reason, the system may periodically flush the cache buffer in order to minimize the amount of loss.

The disk drive cache is essentially two-dimensional-all the bits are out in the open.

3.3.4 Hard Drive Interface: IDE, SCSI, EIDE, Ultra DMA and ATA/66

Secondary storage devices need a controller to act as an intermediary between the device and the rest of the computer system. On some computers, the controller is an integral part of the computer's main motherboard. On others, the controller is an expansion board that connects to the system bus by plugging into one of the computer's expansion slots. In order that devices manufactured by independent vendors can be used with different computer manufacturers, it is important that the controllers follow some drive interfacing standard. Following are the commonly used drive interface standards:

- **IDE (Integrated Disk Electronics) Devices**

  IDE devices are connected to the PC motherboard via a 34-wire ribbon cable. The common drive used today for workstations has capacities of 40MB to

1000MB and rotation speed 7200RPM. The controller is embedded on the disk drive itself. It is an interface between the disk controller and an adopter located on the motherboard. It has good access time of 20ms and data transfer rates of about 1Mbps under ideal conditions. Drives are reasonably cheap. The latest version of the IDE specification enables four IDE channels; each one is capable of supporting two IDE devices.

- **SCSI (Small Computer Systems Interface)**

  The other popular way is to attach a disk drive to a PC via a SCSI interface. The common drive choice for servers or high-end workstations with drive capacities ranges from 100MB to 20GB and rotation speed 7200RPM. It is a common I/O interface between the adapter and disk drives or any other peripheral, i.e., CD-ROMs drives, tape drives, printers etc.

SCSI (pronounced “scuzzy”) is an interesting and important variation of the separate device controller idea for each device. It uses a generic device controller (called SCSI controller) on the computer system and allows any device with an SCSI interface to be directly connected to the SCSI bus of the SCSI controller. The SCSI interface of a device contains all circuitry that the device needs to operate with the computer system.

As shown in Figure 3, a SCSI controller connects directly to the computer bus on one side and controls another bus (called SCSI bus) on the other side. Since the SCSI controller is connected to the computer’s bus on one side and to the SCSI bus on the other side, it can communicated with the processor and memory and can also control the devices connected to the SCSI bus. The SCSI bus is a bus designed for connecting devices to a computer in a uniform way.

These drives have fast access time and high data rates but are expensive. One advantage of these drives is that a single SCSI controller can communicate simultaneously with up to seven 16-bit SCSI devices or up to 15 Wide or Ultra-Wide devices. Each device must be assigned a unique SCSI identification between 0 and 7 (or 15).

Figure 3: SCSI Drive Interface Standards

The versions of SCSI:

- The SCSI-1 calls for a cable with 8 data wires plus one for parity.
- The SCSI-2 enables the use of multiple cables to support 16- or even 32-bit data transfers in parallel.

- The SCSI-3 enables the use of multiple cables to support 32- or even 64-bit data transfers in parallel.
- With fast SCSI, it is possible to transfer 40MB of data per second on a single SCSI cable.

- **EIDE (Enhanced IDE)**

  The principle behind the EIDE interface is the same as in the IDE interface but this drive has capacities ranging from 10.2GB to 20.5GB. The rotation speed is 7200RPM. Its feature include 9.5ms access time, a 2MB buffer and support for the Ultra ATA/66 interface for high speed data throughput and greater data integrity.

  Modern EIDE interfaces enable much faster communication. The speed increases due to improvements in the protocol that describes how the clock cycles will be used to address devices and transfer data. The modern EIDE hard drives are Ultra DMA and ATA/66.

- **Ultra DMA or ATA/33 (AT Attachment):** The ATA standard is the formal specification for how IDE and EIDE interfaces are supposed to work with hard drives. The ATA33 enables up to 33.3 million bytes of data to be transferred each second, hence the name ATA33.

- **ATA/66:** The ATA66 enables up to 66.7 millions bytes of data to be transferred each second, hence the name ATA66 doubles the ATA33.

Check Your Progress 1

1. The seek time of a disk is 30ms. It rotates at the rate of 30 rotations per sec. Each track has 300 sectors. What is the access time of the disk?

   ……

   ……

   ……

2. Calculate the number of entries required in the FAT table using the following parameters for an MS-DOS system:

Disk capacity	30MB
Block size	512 bytes
Blocks/cluster	4

   ……

   ……

   ……

3. What are the purposes of using SCSI, EISA, ATA, IDE?

   ……

   ……

   ……

3.4 REMOVABLE DRIVES

A disk drive with removable disks is called a removable drive. A removable disk can be replaced by another similar disk on the same or different computer, thus providing enormous data storage that is not limited by the size of the disk. Examples of removable disks are floppy disks, CDROM, DVDROM, etc.

3.4.1 Floppy Drives

The disks used with a floppy disk drive are small removable disks made up of plastic coated with magnetic recording material. The disk rotates at 360RPM. Floppies can be accessed from both the sides of the disk. Floppy diskette drives are attached to the motherboard via a 34-wire ribbon cable. You can attach zero, one or two floppy drives and how you connect each one determines whether a drive become either A: or B: A typical floppy drive and floppy is as shown in Figure 4.

(a) Floppy Disk **(b) Floppy Drive**

Figure 4: The Floppy Disk and Drive

A floppy is about 0.64 mm thick and is available in diameters 5.25 inch and 3.5 inch. The data are organized in the form of tracks and sectors. The tracks are numbered sequentially inwards, with the outermost being 0. The utility of index hole is that when it comes under a photosenser, the system comes to know that the read/write head is now positioned on the first sector of the current track. The write-protect notch is used to protect the floppy against deletion of recorded data by mistake.

The data in a sector are stored as a series of bits. Once the required sector is found, the average data transfer rate in bytes per second can be computed by the formula:

$$\textbf{Average data transfer rate} = \frac{\text{bytes/sector} * \text{sector/track} * \text{speed in rpm}}{60}$$

Typical values for IBM/PC compatibles are given in the following table:

Size	Capacity	Tracks	Sectors
5.25	360KB	40	9
5.25	1.2MB	80	15
3.5	720KB	40	18

3.5	1.44MB	80	18

3.4.2 CD-ROM and DVD-ROM

Optical disks use Laser Disk Technology, which is the latest, and the most promising technology for high capacity secondary storage. The advent of the compact disk digital audio system, a non-erasable optical disk, paved the way for the development of a new low-cost storage technology. In optical storage devices the information is written using a laser beam. We will discuss here the use of some optical disks such as CD-ROM and DVD-ROM devices that are now becoming increasingly popular in various computer applications.

Figure 5: CD-ROM and DVD-ROM

1. **CD-ROM (Compact Disk Read Only Memory):** This technology has evolved out of the entertainment electronics market where cassette tapes and long playing records are being replaced by CDs. The term CD used for audio records stands for Compact Disk. The disks used for data storage in digital computers are known as CD-ROM, whose diameter is 5.25 inches. It can store around 650MB. Information in CD-ROM is written by creating pits on the disk surface by shining a laser beam. As the disk rotates the laser beam traces out a continuous spiral. The focused beam creates a circular pit of around 0.8-micrometer diameter wherever a 1 is to be written and no pits (also called a land) if a 0 is to be written. Figure 5 shows the CD-ROM & DVD-ROM.

The CD-ROM with pre-recorded information is read by a CD-ROM reader which uses a laser beam for reading. It is rotated by a motor at a speed of 360 RPM. A laser head moves in and out to the specified position. As the disk rotates the head senses pits and land. This is converted to 1s and 0s by the electronic interface and sent to the computer. The disk speed of CD-ROM is indicated by the notation nx, where n is an integer indicating the factor by which the original speed of 150KB/s is to be multiplied. It is connected to a computer by SCSI and IDE interfaces. The major application of CD-ROM is in distributing large text, audio and video. For example, the entire Encyclopedia could be stored in one CD-ROM. A 640MB CD-ROM can store 74 min. of music.

The main **advantages** of CD-ROMs are:

- large storage capacity
- mass replication is inexpensive and fast
- these are removable disks, thus they are suitable for archival storage.

The main **disadvantages** of CD-ROMs are:

- it is read only, therefore, cannot be updated
- access time is longer than that of a magnetic disks.
- very slow as compared to hard disks, i.e., the normal transfer rate is 300 Mbps for double speed drives and 600 Mbps for quadruple speed drives.

2. **DVD-ROM (Digital Versatile Disk Read Only Memory):** DVD-ROM uses the same principle as a CD-ROM for reading and writing. However, a smaller wavelength laser beam is used. The total capacity of DVD-ROM is 8.5GB. In double-sided DVD-ROM two such disks are stuck back to back which allows recording on both sides. This requires the disk to be reversed to read the reverse side. With both side recording and with each side storing 8.5GB the total capacity is 17GB.

In both CD-ROMs and DVD-ROMs, the density of data stored is constant throughout the spiral track. In order to obtain a constant readout rate the disk must rotate faster, near the center and slower at the outer tracks to maintain a constant linear velocity (CLV) between the head and the CD-ROM/DVD-ROM platter. Thus CLV disks are rotated at variable speed. Compare it with the mechamism of constant angular velocity (CAV) in which disk is rotated at a constant speed. Thus, in CAV the density of information storage on outside sectors is low.

The main advantage of having CAV is that individual blocks of data can be accessed at semi-random mode. Thus the head can be moved from its current location to a desired track and one waits for the specific sector to spin under it.

The main disadvantage of CAV disk is that a lot of storage space is wasted, since the longer outer tracks are storing the data only equal to that of the shorter innermost track. Because of this disadvantage, the CAV method is not recommended for use on CD ROMs and DVD-ROMs.

Comparison of CD-ROM and DVD-ROM

Characteristics	CD-ROM	DVD-ROM
Pit length(micron)	0.834	0.4
Track pitch(micron)	1.6	0.74
Laser beam wave-length(nanometer)	635	780

Capacity	
1 layer/1 side 4.7GB	650MB
2 layers/1 side 8.5GB	NO
1 layer/2sides 9.4GB	NO
2 layers/2 sides 17GB	NO
Speed 1x 1.38MB/s	150KB/s

Check Your Progress 2

1. Compare and contrast the difference between CD-ROM and DVD-ROM.

   ……

   ……

   ……

2. List the advantage and disadvantage of CD-ROM.

   ……

   ……

3.5 REMOVABLE STORAGE OPTIONS

3.5.1 Zip, Jaz and Other Cartridge Drives

Zip Drive: Volumes of data, loads of files have definitely increased the onus on today's computer user, and the protection of this data is what bugs each one of us. The Zip drive is a special high-capacity disk drive that uses a 3.5-inch Zip disk which can store 100MB of data. It allows an easy and rapid shift of the data from desktop to laptop. The user can also connect to notebooks running Windows 95/Windows 98/Windows ME via a Zip cable. The latest addition of Zip drive is Iomega's Zip 250MB drive that can store 250MB of data.

Jaz Drive: The Jaz drive is a popular drive with 2GB and unleashes the creativity of professionals in the graphic design and publishing, software development, 3D CAD/CAM, enterprise management systems and entertainment authorizing markets by giving them unlimited space for dynamic digital content. It has an impressive sustained transfer rate of 8.0 MB/s - fast enough to run applications or deliver full-screen, full-motion video. It is compatible with both Windows (95/98/NT 4.0 & 2000) & MAC OS 8.1 through 9.x.

Cartridge Drive: A cartridge is a protective case or covering, used to hold a disk, magnetic tape, a printer ribbon or toner. The contents are sealed inside a plastic container so that they cannot be damaged.

Disk Cartridges: Removable disk cartridges are an alternative to hard disk units as a form of secondary storage. The cartridge normally contains one or two platters enclosed in a hard plastic case that is inserted into the disk drive much like a music tape cassette. The capacity of these cartridges ranges from 5MB to more than 60MB, somewhat lower than hard disk units but still substantially superior to diskettes. They are handy because they give microcomputer users access to amounts of data limited only by the number of cartridges used.

Quarter Inch Cartridge Tapes (QIC Standard): These tape cartridges record information serially in a track with one head. When the end of the tape is reached the tape is rewound and data is recorder on the next track. There are 9 to 30 tracks. Data bits are serial on a track and blocks of around 6000 bytes are written followed by error-correction code to enable correction of data on reading if any error occurs. The density of data is around 16000 bits per inch in modern tapes. The tapes store around 500 MB. The cassette size is 5.25 inch just like a floppy and mounted in a slot provided on the front panel of a computer. The tape read/write speed is around 120 inch/second and data are transferred at the rate of 240KB/s.

These tapes are normally interfaced to a computer using the SCSI standard. The data formats used in these tapes are called QIC standard.

Tape drive	Capacity MB	Transfer rate(KB/S)	read/write speed(KB/S)	Main application
QIC DEC TZK 10	525	240	120	Backup, archiving
QIC DEC TK50	95	62.5	75	-do-
QIC TS/1000	1000	300	66	-do-
DAI DELTLZ06	4000	366	1GB/Hour	-do-

3.5.2 Recordable CDs and DVDs

The optical disk becomes an attractive alternative for backing up information from hard disk drives, and uses large text, audio and video data. The advent of CD-ROM-R and DVD-ROM-R has broken the monopoly of tapes for backing up files from hard disks with enormous storage capacities in GB range.

3.5.3 CD-R vs CD RW

A CD-R disc looks like a CD. Although all pressed CDs are silver, CD-R discs are gold or silver on their label side and a deep green or cyan on the recordable side. The silver/cyan CD-Rs were created because the green dye used in the original CD-R does not reflect the shorter-wavelength red lasers used in new DVD drives. The cyan dye used in the CD-R will allow complete compatibility with DVD drives. The CD-R disc has four layers instead of three for a CD. At the lowest level, the laser light suffices to detect the presence or absence of pits or marks on the recording surface to read the disc. At the higher level, it can actually burn marks into the surface.

CD-RW is relatively new technology, but it has been gaining market share quite rapidly. The drives cost little more than CD-R drives because they can be used to play audio CDs and CD-ROMs as well as playing and recording CD-RW discs. A CD-RW disc contains two more layers than a CD-R. The difference is that the recordable layer is made of a special material, an alloy of several metals.

Iomega Corporation has announced a CD-RW drive, the **Iomega 48*24*48 USB 2.0 external CD-RW drive**. These drive features buffers under run protection, which list user's record safely, even while multitasking. It offers plug-&-play capability with Microsoft Windows & Mac OS operating systems and its digital audio extraction rate (DAE) of 48x allows users to rep or burn a 60-min CD in under 3 min., while maximum drive speed is attainable only with hi-speed USB 2.0 connections.

3.5.4 Tape Backup

Magnetic tapes are used nowadays in computers for the following purposes:

- Backing up data stored in disks. It is necessary to regularly save data stored on disk in another medium so that if accidentally the data on disk are overwritten or if data get corrupted due to hardware failure, the saved data may be written back on the disk.
- Storing processed data for future use. This is called archiving.
- Program and data interchange between organizations.

Comparative Characteristics of Secondary Memories
(Please note that this may change with technology advancements /time)

Memory Type	Average Capacity in byte	Technology	Average time to access a bye	Pemanence of storage	Access mode	Purpose in computer system	Relative cost per byte in units
Hard disk	50 GB	Magnetic surgaces on hard disks	10 msec	Non-volatile	Direct	Large data files and program overflow from main memory	1/100
Floppy disk	10 MB	Magnetic surgaces on hard disks	500 msec	Non-volatile	Direct	Data entry. As input unit	1/1000
Main memory	50 MB	Integrated circuits	20 nsec	Volatile	Random	Program and data	1
Cache memory	0.5 MB	High speed integrated circuits	2 nsec	Non-volatile	Direct	Instructions and data to be immediately used	10
CD-ROM	650 MB	Laser Disk	500 msec	Non-volatile	Direct	Store large text, pictures and audio. Software distribution	1/10000
DVD-ROM	8.5 GB	Laser Disk	500 msec	Non-volatile	Direct	Video files	1/100000
Magnetic tape	5 GB	Long ¼“	25 sec	Non-volatile	Sequential	Historical files. Backup for disk	1/1000

Digital Audio Tape (DAT): The most appropriate tape for backing up data from a disk today is Digital Audio Tape (DAT). It uses a 4mm tape enclosed in a cartridge. It uses a helical scan, read after write recording technique, which provides reliable data recording. The head spins at a high speed while the tape moves. Very high recording densities are obtained. It uses a recording format called Digital Data Storage (DDS), which provides three levels of error correcting code to ensure excellent data integrity. The capacity is up to 4GB with a data transfer speed of 366KB/sec. This tape uses SCSI interface.

Check Your Progress 3

1. What is the difference between CD-R and CD-RW?

……

……

……

2. State True or False:

   (a) Zip drive can be used for storing 10 MB data. T/F

   (b) QIC standard Cartridges have 40 tracks T/F

   (c) DAT is a useful backing store technology T/F

   (d) DVD-ROMs are preferred for data storage over CD-ROMs T/F

   (e) Magnetic tape is faster that CD-ROM. T/F

3.6 SUMMARY

In this unit, we have discussed the characteristics of different secondary storage drives, their drive speed, access time, rotation speed, density, etc. We also describe the low-level and high-level formatting of a blank disk and also the use of disk partitioning. We have also learnt to distinguish among the various types of drives, i.e., hard drives, optical drives, removable drives and cartridge drive, the hard drive interfaces, removable drives and non-removable drives. This unit also described the different types of disk formats. The advanced technologies of optical memories such as CD-ROM, DVD-ROM, CD-R, CD-RW, etc., and backups for storage such as DAT are also discussed in this unit.

3.7 SOLUTIONS /ANSWERS

Check Your Progress 1

1. Access time is ‘seek time’ plus ‘latency time’. Seek time is the time taken by read-write head to get into the right track. Latency time is the time taken by read-write head to position itself in the right sector. Here a track has 300 sectors. So on an average to position in the right word the read-write head should traverse 150 words. Time taken for this will be 150 / (30 × 300) second = 17 ms (approximately). So the access time will be 30 + 17 = 47 ms.

2. Cluster size is 4 ×512 = 2048 bytes
   Number of clusters = 30 ×1,000,000/2048 = 14648 approx.

3. SCSI is a port or rather an I/O Bus that is used for interfacing many devices like disk drives, printers, etc to computer. SCSI interfaces provide data transmission rates up to 80 MBits per second. It is an ANSI standard also. EISA (**E**xtended **I**ndustry **S**tandard **A**rchitecture) is used for connecting peripherals such as mouse etc. ATA (**A**dvanced **T**echnology **A**ttachment) is a disk drive that integrates the controller on the disk drive itself. IDE (**I**ntegrated **D**rive **E**lectronics) is an interface for mass storage devices that integrates the controller into the disk or CD-ROM drive.

Check Your Progress 2

1. A CD-ROM is a non-erasable disk used for storing computer data. The standard uses 12 cm disk and can hold more than 650 MB.
   A DVD-ROM is used for providing digitized compressed representation of video as well as the large volume of digital data. Both 8 and 12 cm diameters are used with a double sided capacity of up to 17GB.

2. The advantages of CD-ROM are:

- Large storage capacity.
- Mass replication is inexpensive and fast.
- These are removable disks, thus they are suitable for archival storage

Check Your Progress 3

1. A CD-R is similar to a CD-ROM but the user can write to the disk only once. A CD-RW is also similar to a CD-ROM but the user can erase and rewrite to the disk multiple times.
2. (a) False (b) False (c) True (d) False (e) False.

UNIT 4 I/O TECHNOLOGY

Structure **Page No.**

4.0 INTRODUCTION

In the previous units you have been exposed to Input/Output interfaces, control and techniques etc. This unit covers Input/Output devices and technologies related to them. The basic aspects covered include:

- The characteristics of the Device.
- How does it function?
- How does it relate with the Main computing unit?

4.1 OBJECTIVES

After going through this unit you will be able to:

- describe the characteristics, types, functioning and interfacing of Keyboards;
- describe the characteristics, technology and working of Mice;
- describe characteristics, technology and working of Video Cards including various parameters, Video Memory, interfaces and Graphic accelerators;
- describe the characteristics, technology and working of Monitors;
- describe the characteristics, technology and working of Liquid Crystal Displays (LCDs), and Video Cameras;
- describe the characteristics, technology and working of Sound Cards;
- describe the characteristics, technology and working of Printers;
- describe the characteristics, technology and working of Modems;
- describe the characteristics, technology and working of Scanners; and
- describe the the purpose of Power Supply and explain SMPS.

4.2 KEYBOARD

The keyboard is the main input device for your computer. It is a fast and accurate device. The multiple character keys allow you to send data to your computer as a stream of characters in a serial manner. The keyboard is one device which can be used in public spaces or offices where privacy is not ensured. The keyboard is efficient in jobs like data entry. The keyboard is one device which shall stay on for years to come, probably even after powerful voice-based input devices have been developed.

The precursor of the keyboard was the mechanical typewriter, hence it has inherited many of the properties of the typewriter.

The Keys

A full size keyboard has the distance between the centres of the keycaps (keys) as 19mm (0.75in).The keycaps have a top of about 0.5in (12.5in) which is shaped as a sort of dish to help you place your finger. Most designs have the keys curved in a concave cylindrical shape on the top.

4.2.1 Keyboard Layout

A keyboard layout is the arrangement of the array of keys across the keyboard. There is one keyboard layout that anybody who has worked on a standard keyboard or typewriter is familiar with; that layout is QWERTY. However, there are other less popular layouts also.

QWERTY

q,w,e,r,t,y are the first six letters of the top row of the alphabets of the QWERTY layout. The QWERTY arrangement was given by Sholes, the inventor of the typewriter. The first typewriter that Sholes created had an alphabetic layout of keys. However, very soon Sholes designed QWERTY as a superior arrangement though he gave no record of how he came upon this arrangement.

QWERTY-based keyboards

Besides the standard alphabet keys having the QWERTY arrangement, a computer keyboard also consists of the control (alt, Del, Ctrl etc. keys), the function keys (F1, F2 .. etc.), the numerical keypad etc.

PC 83-key and AT 84-key Keyboards

The PC 83-key was the earliest keyboard offered by IBM with its first Personal Computers (PC). This had 83 keys. Later IBM added one more key with its PC AT computer keyboards to make it a 84-key keyboard. The special feature of these keyboards was that they had function keys in two columns on the left side of the keyboard.

101-key Enhanced Keyboard

With its newer range of PCs IBM introduced the 101-key Enhanced/Advanced keyboard. This keyboard is the basic keyboard behind modern QWERTY keyboards. This has the function keys aligned in a separate row at the top of the PC, to

Figure 1: IBM 101-key Keyboard layout

correspond to the function keys shown by many software on the monitor. However, this has also been criticised at times for having a small enter key and function keys on the top! ! ! .

Windows 104-key keyboard

This is enhancements of the 101-key keyboard with special keys for Windows functions and popup. Individual vendors sometimes make changes to the basic keyboard design, for example by having a larger enter key.

Dvorak-Dealey keyboard

This was one keyboard layout designed to be a challenger to the QWERTY layout. This was designed by August Dvorak and William Dealey after much scientific research in 1936. This layout tries to make typing faster. The basic strategy it tries to incorporate is called *hand alteration*. *Hand alteration* implies that if you press one key with the left hand, the next key is likely to be pressed by the right hand, thus speeding up typing (assuming you type with both hands).

Figure 2: Dvorak keyboard layout

However, the Dvorak has not been able to compete with QWERTY and almost all systems now come with QWERTY 101-key or 104-key based keyboards. Still, there may be a possibility of designing new keyboards for specific areas, say, for Indian scripts.

4.2.2 Keyboard Touch

When using a keyboard, the most important factor is the feel of the keyboard, i.e., how typing feels on that particular keyboard. The keyboard must respond properly to your keypress. This not only means that keys must go down when pressed and then come up but also that there must be a certain feedback to your fingers when a key gets activated. This is necessary for you to develop faith in the keyboard and allow fast, reliable typing.

Linear travel or linear touch keyboards increase resistance linearly with the travel of the key. Therefore, you have to press harder as the key goes lower. There can be audible feedback as a click and visual feedback as the appearance of a character on screen letting you know when a key gets activated. Better keyboards provide tactile feedback (to your fingers) but suddenly reducing resistance when the key gets actuated. This is called an over-center feel. Such keyboards are best for quick touch typing. These were implemented by using springs earlier but now they are usually elastic rubber domes. Keyboards also differ in whether they 'click' or not (soundless), on the force required and the key travel distance to actuate a key. The choice is usually an issue of personal liking. Laptops usually have short travel keys to save space which is at a premium in laptops.

4.2.3 Keyboard Technology

Each key of a keyboard is like an electric switch changing the flow of electricity in some way. There are two main types — capacitor-based and contact-based keyboards.

Capacitor-Based Keyboards

These keyboards are based on the concept of Capacitance. A simple capacitor consists of a pair of conductive plates having opposite charges and separated by an insulator. This arrangement generates a field between the plates proportional to the closeness of the plates. Changing the distance between the plates causes current to flow. Capacitive keyboards have etched circuit boards, with tin and nickel-plated copper pads acting as capacitors under each key (a key is technically called a station). Each key press presses a small metal-plastic circle down causing electric flow. These keyboards work

well but have the drawback that they follow an indirect approach though they have a longer life than contact-based keyboards. These keyboards were introduced by IBM.

Contact-Based Keyboards

Contact-based keyboards use switches directly. Though they have a comparatively shorter life, they are the most preferred kind nowadays due to their lower cost. Three such kinds of keyboards have been used in PCs:

1. **Mechanical Switches:** These keyboards use traditional switches with the metal contacts directly touching each other. Springs and other parts are used to control positioning of the keycaps and give the right feel. Overall, this design is not suited to PC keyboards.

2. **Rubber Dome:** In rubber dome keyboards, both contact and positioning are controlled by a puckered sheet of *elastomer*, which is a stretchy, rubber-like synthetic material. This sheet is moulded to have a dimple or dome in each keycap. The dome houses a tab of carbon or other conductive material which serves as a contact. When a key is pressed, the dome presses down to touch another contact and complete the circuit. The elastomer then pushes the key back. This is the most popular PC keyboard design since the domes are inexpensive and proper design can give the keyboards an excellent feel.

3. **Membrane:** These are similiar to rubber domes except that they use thin plastic sheets (membranes) with conductive traces on them. The contacts are in the form of dimples which are plucked together when a key is pressed. This design is often used in calculators and printer keyboards due to their low cost and trouble-free life. However, since its contacts require only a slight travel to actuate, it makes for a poor computer keyboard.

Scan Codes

A scan code is the code generated by a microprocessor in the keyboard when a key is pressed and is unique to the key struck. When this code is received by the computer it issues an interrupt and looks up the scan code table in the BIOS and finds out which keys have been pressed and in what combination. Special memory locations called *status bytes* tell the status of the locking and toggle keys, e.g., Caps lock etc. Each keypress generates two different scan codes — one on key-push down called Make code, another on its popping back called Break code.This two-key technique allows the computer to tell when a key is held pressed down, e.g., the ALT key while pressing another key, say, CTRL-ALT-DEL.

OPERATOR

— means dash which is longer.
- means hypen which is shorter.

There are three standards for scan codes: Mode1 (83-key keyboard PC, PC-XT), Mode2 (84-key AT keyboard), Mode3 (101-key keyboard onwards). In Mode1 Make and Break codes are both single bytes but different for the same key. In Mode2 and Mode3, Make code is a single byte and Break code is two bytes (byte F0(Hex) + the make code).

Interfacing

The keyboard uses a special I/O port that is like a serial port but does not explicitly follow the RS-232 serial port standard. Instead of multiple data and handshaking signals as in RS-232, the keyboard uses only two signals, through which it manages a bi-directional interface with its own set of commands.

Using its elaborate handshaking mechanism, the keyboard and the PC send commands and data to each other. The USB keyboards work differently by using the USB coding and protocol.

Table 1: Some Scan Codes

	Mode1			Mode2	Mode 3
Key	KeyNo.	Make	Break	Make	Break
A	31	1E	9E	1C	F0 1C
0	11	0B	8B	45	F0 45
Enter	43	1C	9C	5A	F0 5A
Left Shift	44	2A	AA	12	F0 12
F1	112	3B	BB	07	F0 07

Connections

5-pin DIN connector: This is the connector of the conventional keyboard having 5 pins (2 IN, 2 OUT and one ground pin), used for synchronization and transfer.

PS/2 connector (PS/2 keyboards): These were introduced with IBM's PS/2 computers and hence are called PS/2 connectors. They have 6-pins but in fact their wiring is simply a rearrangement of the 5-pin DIN connector. This connector is smaller in size and quite popular nowadays. Due to the similiar wiring, a 5-pin DIN can easily be connected to a PS/2 connector via a simple adapter.

Ergonomic Keyboards

Ergonomics is the study of the environment, conditions and efficiency of workers[1] . Ergonomics suggests that the keyboard was not designed with human beings in mind. Indeed, continuous typing can be hazardous to health. This can lead to pain or some ailments like the Carpal Tunnel Syndrome.

For normal typing on a keyboard, you have to place your hands apart, bending them at the wrists and hold this position for a long time. You also have to bend your wrist vertically especially if you elevate your keyboard using the little feet behind the keyboards. This stresses the wrist ligaments and squeezes the nerves running into the hand through the Carpal tunnel, through the wrist bones.

To reduce the stress, keyboards called ergonomic keyboards have been designed. These split the keyboard into two and angle the two halves so as to keep the wrists straight. To reduce vertical stress, many keyboards also provide extended wrist rests. For those who indulge in heavy, regular typing, it is recommended that they use more ergonomics based keyboards and follow ergonomic advice in all aspects of their workplace.

4.3 MOUSE

The idea of the Mouse was developed by Douglas C. Engelbart of Stanford Research institute, and the first Mouse was developed by Xerox corporation. Mouse itself is a device which gives you a pointer on screen and a method of selection of commands through buttons on the top. A single button is usually sufficient (as in Mouse with Apple Macintosh machines) but Mice come with upto 3 buttons.

Types of Mice

Mice can be classified on the basis of the numbers of buttons, position sensing technology or the type of Interface:

[1]Oxford Advanced Learner's Dictionary

Sensing Technology

The Mice can be Mechanical or Optical.

Mechanical Mice have a ball made from rough rubbery material, the rotation of which effects sensors that are perpendicular to each other. Thus, the motion of the ball along the two axes is detected and reflected as the motion of the pointer on the screen.

Optical Mice can detect movement without any moving parts like a ball. The typical optical Mouse used to have a pair of LEDs (Light Emitting Diodes) and photo-detectors in each axis and its own Mousepad on which it is slided. However, due to the maintenance needs of the Mousepad, this was not very successful. Recently, optical Mice have made a comeback since they can now operate without a Mousepad.

Interface

Mouse is usually a serial device connected to a serial port(RS232), but these connections can themselves take various forms:

Serial Mouse

Mice that use the standard serial port are called "serial". Since Serial ports 1 and 4 (COM1, COM4 under DOS, /dev/ttyS0 and /dev/ttyS3 under Unix/GNU-Linux systems) and ports 2 and 3 (COM2, COM3 or /dev/ttyS1/dev/ttyS2) share the same interrupts respectively, one should be careful not to attach the mouse so that it shares the interrupt with another device in operation like a modem.

Bus Mouse

These Mice have a dedicated Mouse card and port to connect to. Recently, USB mouse has become popular.

Proprietary

Mouse ports specific to some PCs e.g., IBM's PS/2 and some Compaq computers.

Mouse Protocols

The mouse protocol is the digital code to which the signal from the mouse gets converted. There are four major protocols: Microsoft, Mouse Systems Corporation(MSC), Logitech and IBM. Most mice available do support at least the Microsoft protocol or its emulation.

Resolution versus Accuracy

Resolution of mouse is given in CPI(Counts per Inch) i.e. the number of signals per inch of travel. This means the mouse will move faster on the screen but it also means that it will be more difficult to control the accuracy.

Check Your Progress 1

1. Discuss the merits and demerits of Dvorak-Dealey keyboard vs. QWERTY keyboard.

   .....

   .....

   .....

   .....

   .....

2. Why is keyboard touch important? What kind of touch would you prefer and which kind of keyboard will give that touch?

.....

.....

.....

.....

.....

3. What precautions should be taken while attaching a Serial Mouse?

.....

.....

.....

.....

.....

4. You enter 'a' as left-shift + 'A' ? What will be the scan-code generated in Mode-3 by the keyboard?

a) 2A1E9EAA b) 1CF01C
c) 121CF01CF012 d) 1CF01C5AF05A

4.4 VIDEO CARDS

Before discussing in detail video hardware, let us have a brief overview of graphic display technology. The purpose of your graphic display system is to display bit-mapped graphics on your monitor. The image displayed on your system thus consists of small dots called pixels (short for 'picture elements') and your video system contains a description of each of these dots in the memory. At any moment, the display memory contains the exact bit-map representation of your screen image and what is coming next. This is like a time-slice of what you see on your monitor. Therefore, display memory is also called a framebuffer. These frames are read dozens of times a second and sent in a serial manner through a cable to the monitor. The monitor recieves the train of data and displays it on the screen. This happens by a scanning raster movement from up to down one row at a time. A CRT (Cathode Ray Tube) based monitor will light its small phosphor dots according to this raster movement. In this respect, it is like a television, which is also a CRT based device.

Figure 3: Raster Display

The more the number of dots, i.e., the higher the resolution of the image, the sharper the picture is. The richness of the image is also dependant on the number of colours (or gray levels for a monochrome display) displayed by the system. The higher the number of colours, the more is the information required for each dot. Hence, the amount of memory (framebuffer) required by a system is directly dependent on the resolution and colour depth required.

4.4.1 Resolution

Resolution is the parameter that defines the possible sharpness or clarity of a video image. Resolution is defined as the number of pixels that make up an image. These pixels are then spread across the width and height of the monitor. Resolution is independent of the physical characteristics of the monitor. The image is generated without considering the ultimate screen it is to be displayed upon. Hence, the unit of resolution is the number of pixels, not the number of pixels per inch. For example, a standard VGA native graphic display mode has a resolution of 640 pixels horizontally by 480 pixels vertically. Higher resolutions mean the image can be sharper because it contains more pixels.

The actual on-screen sharpness is given as dots-per-inch, and this depends on both the resolution and the size of the image. For the same resolution, an image will be sharper on a smaller screen, i.e., an image which may look sharp on a 15" monitor may be a little jagged on a 17” display.

4.4.2 Colour Depth

It is clear that an image consists of an array of pixels. If we tell which pixels are ‘on’ and which are ‘off’ to the monitor, it should be able to display the image as a pure black and white image. But what about Colour and Contrast? Clearly, if only a single bit is assigned to a pixel, we cannot give any additional quality to the image. It will look like a black and white line drawing. Such a system is typically called a two-colour system. Such black and white picture can be converted to gray levels by assigning more bits, e.g., with two bits we can get the following levels: White, Light Gray, Dark Gray and Black.
To add colour to an image, we have to store colour of the pixel with each pixel.This is usually stored as intensity measures of the primary light colours — Red, Green and Blue. That means we have to assign more than 1 bit to describe a pixel. Hence, 1 bit per pixel implies 2 colours or 2 gray-levels, 2 bits per pixel 4 colours or 4 gray-levels and so *n* bits per pixel means a display of colours or gray-levels is possible.

Colour Depth (or the number of *Colour Planes*) is the number of bits assigned to each pixel to code colour information in it. These are also called *Colour Planes* because each bit of a pixel represents a specific colour and the bit at the same position on every pixel represents the same colour. Hence, the bits at the same position can be thought of as forming a plane of a particular colour shade and these planes piled on top of each other give the final colour at each point. Thus, if each pixel is described by 3 bits, one each for red, green and blue colour, then, there are 3 *Colour Planes* (one each for red, green and blue) and 6 colour planes if there are 6 bits — see Figure 4.

Figure 4: Colour Planes

What Colour depths are practically used?

Practically, the number of colours are an exponential power of 2, since for *Colour Depth n*, colours can be displayed. The most popular colour modes are given in Table 2.

	Colour Mode	Depth(bits/pixel)
1	Monochrome	1
2	16-Colours	4
3	256-Colours	8
4	High Color	16
5	True Color	24

Table 2: Major Colour Depths

The Bad News

The bad news is that most monitors can only display upto a maximum of 262,144 colours (= i.e. 18 bits/pixel *Colour Depth*). The other bad news is that the human eye can only perceive a few million colours at the most. So, even if you had lots of bits per pixel and very advanced display systems, it would be useless. Maybe, this is good news rather than bad news for the hardware developer!

This also implies that 24-bit colour bit-depth is the practical upper limit. Hence, this depth is also called true colour because with this depth the system stores more colours than can ever be seen by the human eye and, hence, it is a true colour representation of the image. *Though, 24-bit colour or true colour systems have more colour than possibly useful, they are convenient for designers because they assign 1 byte of storage for each of the three additive primary colours (red, green and blue)*. Some new systems even have 32 bits per pixel. Why? Actually, the additional bits are not used to hold colours but something called an *Alpha Channel*. This 8-bit *Alpha Channel* stores special effect information for the image.

Why are all resolutions in the ratio of 43? The answer you'll find in a later section.

4.4.3 Video Memory

As stated before, video memory is also called framebuffer because it buffers video frames to be displayed. The quality of a video display depends a lot on how quickly can the framebuffer be accessed and be updated by the video system. In early video systems, video memory was just a fixed area of the system RAM. Later, there was video RAM which came with the video cards themselves and could be increased by putting additional video RAM under the UMA (Unified Memory Architecture). Video RAM is again part of the system RAM. UMA is what you get in the modern low-cost motherboards with on-board video and sound cards etc.

The amount of video memory required is dependant on the resolution and colour-depth required of the system. Let us see how to calculate the amount of video memory required. The video memory required is simply the resolution (i.e., the total number of pixels) multiplied by the Colour Depth. Let us do the calculations for a standard VGA graphics screen (640 ×480) using 16 colours.

Total number of Pixels = 640 ×480 =		307200	
Colour Depth (16-colours)	=	4	bits
Total minimum Memory	=	1,228,800	bits
Total minimum memory (in bytes)	=	153,600	bytes
		153 KB	

Minimum Video RAM required and available = 256 KB.

Therefore, 16-colour VGA needs at least 153,600 bytes of memory but memory is only available in exponential powers of 2, hence, the next highest available memory is = 256 KB.

What is a good resolution? Actually, it depends on your hardware. So, it is the maximum your hardware can allow you. However, one odd-looking resolution which has become popular is 1152 × 864 pixels. Can you judge why this should be so? (Hint: Think of this resolution at 8-bit colour).

If you can't wait any longer, here is the answer: 1152 × 864 is nearly one million pixels. Since 8-bit colour depth means 8 million bits or 1 MB. This is the highest resolution you can get in 1 MB video memory at 8-bit colour depth, plus this still leaves you square pixels (in the ratio 4: 3) to allow easy programming.

The above calculations hold good for only two-dimensionsal display systems. This is because 3-D systems require much more memory because of techniques such as "Double Buffering" and "Z-Buffering".

4.4.4 Refresh Rates

A special circuit called the Video Controller scans the video memory one row at a time and reads data value at each address sending the data out in a serial data stream. This data is displayed by a process called *Scanning* where the electron beam is swept across the screen one-line-at-a-time and left-to-right. This is controlled by a vertical and a horizontal field generated by electromagnets — one moving the beam horizontally and another vertically.
The rate at which horizontal sweeps take place is called horizontal frequency or horizontal refresh rate and the rate at which vertical sweeps take place are called vertical frequency or vertical refresh rate or simply refresh rate or frame rate. The term frame rate is used because actually one vertical sweep means display of a single frame. Since each frame contains several hundred rows, horizontal frequency is hundreds of times higher than vertical frequency. Therefore, the unit of horizontal frequecny is KHz and that of vertical frequency is Hz.

The most important thing is maintaining the same frequencies between the Video system and monitor. The monitor must support these refresh rates, hence the supported refresh rates are given with the manual of the monitor. More about this topic will be discussed in the section on Monitors.

4.4.5 Graphic Accelerators and 3-D Accelerators

A Graphic Accelerator is actually a chip, in fact the most important chip in your video card. The Graphic Accelerator is actually the modern development of a much older technology called the *Graphic Co-Processor*. The accelerator chip is actually a chip that has built-in video functions. These functions execute the algorithms for image construction and rendering. It does a lot of work which would otherwise have to be done by the microprocessor. Hence, the accelerator chip is actually optional but very important for good graphics performance.

The graphic accelerator determines whether your system can show 3-D graphics, how quickly your system displays a drop-down menu, how good is your video playback, etc. It determines the amount and kind of memory in the framebuffer and also the resolution your PC can display.

The first major graphic accelerators were made by the S3 corporation. Modern Graphic accelerators have internal registers at least 64-bit wide to work on at least 2 pixels at a time. They can use the standard Dynamic RAM (DRAM) or the more expensive but faster dual-ported Video RAM (VRAM). They support at least the standard resolutions up to 1024 × 768 pixels. They often use RAMDACs for colour support giving full 24-bit or 32-bit colour support. A RAMDAC (Random Access

Memory Digital-to-Analog Converter) is a microchip that converts digital image data into the analog data needed by a computer display. However, the higher the resolution required, the higher is the speed at which the chip has to function. So, for a resolution of 1280×1024, the chip operates at 100 MHz. At the cutting edge of technology, chips now run even as fast as 180 or 200 Mhz.

What is a 3-D Accelerator?

3-D Accelerator is no magic technology. It is simply an accelerator chip that has built-in ability to carry out the mathematics and the algorithms required for 3-D image generation and rendering. A 3-D imaging is simply an illusion, a projection of 3-D reality on a 2-D screen. These are generated by projection and perspective effects, depth and lighting effects, transparency effects and techniques such as Ray-Tracing (Tracing the path of light rays emitting from a light source), Z-buffering (a buffer storing the Z-axis positions) and Double-Buffering (two buffers instead of one).

4.4.6 Video Card Interfaces

A video interface is the link of the video system to the rest of the PC. To enhance video performance, there is sought to be an intimate connection between the microprocessor and the video system, especially the framebuffer. In modern displays, only in the UMA system is the framebuffer actually a part of the main memory; in the rest the connection is through a bus, which may be PCI or AGP. Let us briefly discuss these interfaces:

PCI

PCI stands for Peripheral Connect Interface. It is the revolutionary high speed expansion bus introduced by Intel. With the growing importance of video, video cards were shifted to PCI from slower interfaces like ISA. The PCI standard has now developed into the even more powerful AGP.

AGP

AGP stands for Advanced (or Accelerated) Graphics Port. It is a connector standard describing a high speed bus connection between the PC video system, the microprocessor and the main memory. It is an advancement of the PCI interface.

AGP uses concepts such as pipelining to allow powerful 3-D graphic accelerators to function when used in conjuction with fast processors. AGP uses three powerful innovations to achieve its performance:

- Pipelined Memory: The use of Pipelining eliminates wait states allowing faster operation.
- Seperate Address and Data Lines.
- High speeds through a special 2X mode that allows running AGP at 133 MHz instead of the default 66 MHz.

Through AGP, the video board has a direct connection to the microprocessor as a dedicated high speed interface for video. The system uses DMA (Direct Memory Access) to move data between main memory and framebuffer. The accelerator chip uses the main memory for execution of high level functions like those used in 3-D rendering.

UMA

UMA stands for Unified Memory Architecture. It is an architecture which reduces the cost of PC construction. In this, a part of the main memory is actually used as framebuffer. Hence, it eliminates the use of a bus for video processing. Therefore, it is less costly. Though it is not supposed to perform as well as AGP etc., in some

cases it may give a better performance than the bus-based systems. It is the interface used nowadays in low-cost motherboards.

Figure 5: AGP Video Architecute and its working

4.5 MONITORS

A Monitor is the television like box connected to your computer and giving you a vision into the mind of your PC. It shows what your computer is thinking. It has a display which is technically defined as the image-producing device, i.e., the screen one sees and a circuitry that converts the signals from your computer (or similiar devices) into the proper form for display.

Monitors are or were just like television sets except that television sets have a tuner or demodulator circuit to convert the signals. However, now monitors have branched beyond television. They have greater sharpness and colour purity and operate at higher frequencies.

Generally, when you go to purchase a monitor from the market, you see the following specifications: The maximum Resolution, the Horizontal and Vertical Frequencies supported, the tube size and the connectors to the monitor. There are many vendors on the market like Samsung, LG, Sony etc. Home users generally go in for monitors of size 17", 15" or 14" . Monitors are also available as the traditional curved screens, flat screens or LCD. The technology behind Monitors and the above specifications are discussed ahead.

4.5.1 Cathode Ray Tubes

Cathode ray tube is the major technology on which monitors and televisions have been based. CRT is a partially evacuated glass tube filled with inert gas at low pressure. A specially designed Cathode (negatively charged electrode) shoots beams of electrons at high speed towards an anode (positively charged electrode) which impinges on the screen which is coated with small phosphor coated dots of the three primary colours. This cathode is also called an Electron Gun. In fact, there can be three seperate guns for the three colours (Red, Green and Blue) or one gun for all three.

Four factors influence the quality of image of the monitor:

1. **The Phosphor coating** : This affects the colour and the persistence (The period the effect of a single hit on a dot lasts).
2. **The Cathode (Electron Gun)** : The sharpness of the image depends on the good functioning of this gun.

3. **Shadow Mask/ Aperture Grill** : This determines the resolution of the screen in colour monitors.
4. The Screen, glare and lighting of the monitor.

4.5.2 Shadow Mask

The Shadow Mask is a metal sheet which has fine perforations (holes) in it and is located a short distance before the phosphor coated screen. The Phosphor dots and the holes in the shadow mask are so arranged that the beams from a particular gun will strike the dots of that colour only. The dots of the other two colours are in the shadow. In an attempt to overcome some shortcomings of Shadow masks due to their round holes, Sony introduced Aperture grills (in their Trinitron technology) which are slots in an array of vertically arranged wires.

Figure 6: Shadow Mask and Aperture

4.5.3 Dot Pitch

Dot Pitch of a CRT is the distance between phosphor dots of the same colour. In Trinitron screens, the term Slot Pitch is used instead of Dot Pitch — this is the distance between two slots of the same colour. Dot Pitch is a very important parameter of monitor quality. For a particular resolution, you can get the minimum dot pitch required by dividing the physical screen size by the number of pixels. Therefore, for smaller screens, you require finer Dot Pitch.

4.5.4 Monitor Resolutions

We have discussed about resolutions and vertical and horizontal refresh rates in the section on Video Cards. Let us refer to them from the monitor point of view. So, we have the following definitions (from the manual of a monitor available in the market):

Horizontal Frequency: The time to scan one line connecting the right edge to the left edge of the screen horizontally is called the Horizontal cycle and the inverse number of the Horizontal cycle is called Horizontal Frequency. The unit is KHz (KiloHertz).

Vertical Frequency: Like a Flouroscent lamp, the screen has to repeat the same image many times per second to display an image to the user. The frequency of this repetition is called Vertical Frequency or Refresh Rate.

If the resolution generated by the video card and the monitor resolution is properly matched, you get a good quality display. However, the actual resolution achieved is a physical quality of the monitor. In colour systems, the resolution is limited by Convergence (Do the beam of the 3 colours converge exactly on the same dot?) and the Dot Pitch. In monochrome monitors, the resolution is only limited by the highest frequency signals the monitor can handle.

4.5.5 DPI

DPI (Dots Per Inch) is a measure for the actual sharpness of the onscreen image. This depends on both the resolution and the size of the image. Practical experience shows that a smaller screen has a sharper image at the same resolution than does a larger screen. This is because it will require more dots per inch to display the same number of pixels. A 15-inch monitor is 12-inches horizontally. A 10-inch monitor is 8 inches horizontally. To display a VGA image (640□480) the 15-inch monitor will require 53DPI and the 10-inch monitor 80 DPI.

4.5.6 Interlacing

Interlacing is a technique in which instead of scanning the image one-line-at-a-time it is scanned alterenately, i.e., alternate lines are scanned at each pass. This achieves a doubling of the frame rate with the same amount of signal input. Interlacing is used to keep bandwidth (amount of signal) down. Presently, only the 8514/A display adapters use interlacing. Since Interlaced displays have been reported to be more flickery, with better technology available, most monitors are non-interlaced now.

4.5.7 Bandwidth

Bandwidth is the amount of signal the monitor can handle and it is rated in MegaHertz. This is the most commonly quoted specification of a monitor. The Bandwidth should be enough to address each pixel plus synchronizing signals.

Check Your Progress 2

1. Redraw Figure 4 showing Colour-Planes for a true-colour system.

   .....

   .....

   .....

   .....

2. What is a FrameBuffer? Discuss the placement of the FrameBuffer w.r.t. to the different Video Card interfaces.

   .....

   .....

   .....

   .....

3. What is the difference between Shadow Mask and Dot Pitch for Trinitron and non-Trinitron monitors?

   .....

   .....

   .....

4. How much Video-RAM would you require for a high-colour (16-bits) Colour-Depth at 1024□768 resolution? What would be the size of the corresponding single memory chip you would get from the market?

   a) 900KB, 1MB b) 1.6 MB, 4MB
   c) 12.6MB, 16MB d)7.6MB, 8MB

5. There is an image of resolution 1024X768. It has to be displayed on a 15-inch monitor (12-inch horizontal, 9-inch Vertical display). What is the minimum Dot-pitch required for this image? (minimum here means the largest useful dot pitch).

   a) 1.4×10^{-4} inches b) 2.8×10^{-4} inches
   c) 1.4×10^{4} inches d) 1.2×10^{-2} inches

4.6 LIQUID CRYSTAL DISPLAYS (LCD)

LCDs are the screens of choice for portable computers and lightweight screens. They consume very little electricity and have advanced technologically to quite good resolutions and colour support. They were developed by the company RCA in the 1960s. LCDs function simply by blocking available light so as to render display patterns.

LCDs can be of the following types:

1. **Reflective LCDs:** Display is generated by selectively blocking reflected light.
2. **Backlit LCDs :** Display is due to a light source behind LCD panel.
3. **Edgelit LCDs :** Display is due to a light source adjacent to the LCD panel.

LCD Technology

The technology behind LCD is called Nematic Technology because the molecules of the liquid crystals used are nematic i.e. rod-shaped. This liquid is sandwiched between two thin plastic membranes. These crystals have the special property that they can change the polarity and the bend of the light and this can be controlled by grooves in the plastic and by applying electric current.

Passive Matrix

In a passive matrix arrangement, the LCD panel has a grid of horizontal and vertical conductors and each pixel is located at an intersection. When a current is recieved by the pixel, it becomes dark. This is the technology which is more commonly used.

Active Matrix

This is called TFT (Thin Film Transistor) technology. In this there is a transistor at every pixel acting as a relay, receiving a small amount and making it much higher to activate the pixel. Since the amount is smaller, it can travel faster and hence response times are much faster. However, TFTs are much more difficult to fabricate and are costlier.

4.7 DIGITAL CAMERA

A Digital camera is a camera that captures and stores still images and video (Digital Video Cameras) as digital data instead of on photographic film. The first digital cameras became available in the early 1990s. Since the images are in digital form they can be later fed to a computer or printed on a printer.

Like a conventional camera, a digital camera has a series of lenses that focus light to create an image of a scene. But instead of this light hitting a piece of film, the camera focuses it on to a semiconductor device that records light electronically. An in-built computer then breaks this electronic information down into digital data.

This semiconductor device is called an Image sensor and converts light into electrical charges. There are two main kinds of Image sensors: CCD and CMOS. CCD stands for Charge coupled devices and is the more popular and more powerful kind of sensor. CMOS stands for Complementary Metal oxide semiconductor and this kind of technology is now only used in some lower end cameras. While CMOS sensors may improve and become more popular in the future, they probably won't replace CCD sensors in higher-end digital cameras.

In brief, the CCD is a collection of tiny light-sensitive diodes called photosites, which convert photons (light) into electrons (electrical charge). Each photosite is proportionally sensitive to light – the brighter the light that hits a single photosite, the greater the electrical charge that will accumulate at that site.

A digital Camera is also characterised by its resolution (like monitors and printers) which is measured in pixels. The higher the resolution, the more detail is available in an image.

Mobile Cameras

Mobile cameras are typically low-resolution Digtial cameras integrated into the mobile set. The photographs are typically only good enough to show on the low resolution mobile screen. They have become quite popular devices now and the photographs taken can be used for MMS messages or uploading to a Computer.

4.8 SOUND CARDS

Multimedia has become a very important part of today's PC. The home user wants to watch movies and hear songs. The Software developer hacking away at her computer wants to have the computer playing MP3 or OGG (The latest Free Sound format standard) in the background. Thus, the sound system is a very important part of the system.

As you must have read in your high school physics, sound is a longitudinal wave travelling in a medium, usually air in the case of music. Sound can be encoded into electrical form using electrical signals which encode sound strengths. This is called analog audio. This analog audio is converted to digital audio, which is conversion of those signals into bits and bytes through the process called Sampling. In Sampling, analog 'samples' are taken at regular intervals and the amplitude (Voltage) of these samples is encoded to bits. These sounds are manipulated by your PCs microprocessor etc. To play back these digital audio sounds, the data are sent to the Sound card which converts them to analog audio, which is played back through speakers.

The Sound card (The card is often directly built into motherboards nowadays) is a board that has digital to analog sound converter, amplifier, etc., circuitry to play sound and to connect the PC to various audio sources.
A sound card may support the following functions:

1. Convert digital sound to analog form using digital-to-analog converter to play back the sound.
2. May record sound to play back later with analog-to-digital converter.
3. May have built-in Synthesizers to create new sounds.
4. May use various input sources (Microphone, CD, etc.) and mixer circuits to play these sounds together.
5. Amplifiers to amplify the sound signals to nicely audible levels.

Sound cards are described by the Compatibility, Connections and Quality.

Compatibility: Sound cards must be compatible at both hardware and software levels with industry standards. Most software, especially games, require sound cards to be compatible with the two main industry standards: AdLib (A Basic standard) and Sound Blaster (an advanced standard developed by Creative Labs).

Connections: Sound cards should have connections to allow various functions. One of the most important is the MIDI port (MIDI stands for Musical Instrument Device Interface). MIDI port allows you to create music directly with your PC using the Sound Cards synthesizer circuit and even attach a Piano keyboard to your PC.

Quality: Sound Cards vary widely in terms of the quality they give. This ranges from the frequency range support, digital quality and noise control.

4.9 PRINTERS

Printers are devices that put ink on paper in a controlled manner. They manually produce readable text or photographic images. Printers have gone through a large transition in technology. They are still available in a wide range of technology and prices from the dot matrix printer to Inkjet printers to Laser Printers.

4.9.1 Classification of Printers

Printers can be classified on the following bases:

a) Impact: Impact printers print by the impact of hammers on the ribbon (e.g., Dot-Matrix Printers) whereas non-impact printers use other means (e.g., Inkjet, Laser).

b) Character formation: Character formed by Bit-maps or by dots.

c) Output : The quantity of output processed at a time: Serial, Line or Page Printers.

Actually, there are many specifications one has to keep in mind while purchasing a printer. Some of these are Compatibiltiy with other hardware, in-built Memory, maximum supported memory, actual technology, Printer resolution (Colour, BW), PostScript support, output type, Printer speed, Media capacity, Weight, Height and Width of the Printer.

Let us discuss some of these parameters that characterize printers :

4.9.2 Print Resolution

Print Resolution is the detail that a printer can give determined by how many dots a printer can put per inch of paper. Thus, the unit of resolution is *Dots per inch*. This is applicable to both impact and non-impact printer though the actual quality will depend on the technology of the printer.

The required resolution to a great extent determines the quality of the output and the time taken to print it. There is a tradeoff between quality and time. Lower resolution means faster printing and low quality. High resolution means slower printing of a higher quality. There are three readymade resolution modes: draft, near letter quality (NLQ) and letter quality. Draft gives the lower resolution print and letter quality higher resolution. In Inkjet and Laser Printers, the highest mode is often called 'best' quality print.

4.9.3 Print Speed

The speed at which a printer prints is often an important issue. However, the printer has to take a certain time to print. Printing time increases with higher resolution and coloured images. To aid printing, all operating systems have spooling software that accumulates print data and sends it at the speed that the printer can print it.
The measure of speed depends on whether the printer is a **Line Printer** or **Page**

Printer. Let us understand these:

Line Printer: Line Printer processes and prints one line of text at a time.

Page Printer: A page printer processes and prints one full page at a time.

Actually, it rasterizes the full image of the page in its memory and then prints it as one line of dots at a time. For a line printer, the speed is measured in *characters per second* (cps) whereas for page printing, it is *pages per minute* (ppm). Hence, Dot Matrix usually have speeds given in *cps* whereas Lasers have speed in *ppm*. The actual speed may vary from the rating speed given by the manufacturer because, as expected, the printer chooses the more favourable values.

4.9.4 Print Quality

Print quality depends on various factors but ultimately the quality depends on the design of the printer and its mechanical construction.

DotMatrix/InkJet Printers

Three main issues determine the quality of characters produced by DotMatrix/InkJet Printers: - Number of dots in the matrix of each character, the size of the dots and the addressability of the Printer. Denser matrix and smaller dots make better characters.

Addressability is the accuracy with which a dot can be produced (e.g., 1/120 inch means printer can put a dot with 1/120 inch of the required dot). Minimum dot matrix used by general dot matrix printers is 9×9 dots, 18-pin and 24-pin printers use 12×24 to 24×24 matrices. Inkjets may even give up to 72×120 dots. Quality of output also depends on the paper used. If the ink of an Inkjet printer gets absorbed by the paper, it spreads and spoils the resolution.

Laser Printer

Laser Printers are page printers. For print quality, they also face the same addressability issues as DMP/InkJet Printers. However, some other techniques are possible to use for better quality here.

One of these is ReT(Resolution Enhancement Technology) introduced by Hewlett-Packard. It prints better at the same resolution by changing the size of the dots at character edges and diagonal lines reducing jagged edges.

A very important requirement for Laser Printers to print at high quality is Memory. Memory increases as a square of resolution, i.e., the Dot density, i.e., the dpi. Therefore, if 3.5 MB is required for a 600 dpi page, approximately 14 MB is required for 1200 dpi. You need even more memory for colour.

For efficient text printing, the Laser printer stores the page image as ASCII characters and fonts and prints them with low memory usage. At higher resolutions, the quality of print toner also becomes important since the resolution is limited by the size of toner particles.

4.9.5 Colour Management

There are three primary colours in pigments - *Red, Yellow and Blue*. There are two ways to produce more colours:

Physical Mixing: Physically mix colours to make a new colour. This is difficult for printers because their colours are quick drying and so colours to be mixed must be applied simultaneously.

Optical Mixing: Mixing to give the illusion of a new colour. This can be done in ways:

- Apply colours one upon another. This is done using inks which are somewhat transparent, as modern inks are.
- Applying dots of different colours so close to one another that the human eye cannot distinguish the difference. This is the theory behind **Dithering**.

3 or 4 colour Printing?

For good printing, printers do not use RBY, instead they use CMYK (Cyan instead of Blue, Magenta instead of Red, Yellow, and a separate Black). A separate Black is required since the 3 colours mixed to produce a black (which is called Composite Black) is often not satisfactory.

What is Dithering?

CMYK gives only 8 colours (C, M, Y K, Violet= C + M, Orange= M + Y, Green = C + Y, and the colour of the paper itself!). What about other colours? For these, the technique of Dithering is used. Dithering is a method in which instead of being a single colour dot, it is a small matrix of a number of different colour dots. Such pixels are called **Super-pixels**. The dots of a given colour in a Super-pixel decide the intensity of that colour. The problem with dithering is that it reduces the resolution of the image since more dots are taken by a single pixel now.

Monitors versus Printer

Monitor screens and Printers use different colour technologies. The monitor uses RGB and the Printer CMYK. So, how does one know that the colour that is seen is going to be printed. This is where the Printer driver becomes very important, and where many computer models and graphic oriented machine score. For long, a claim to fame of the Apple Macintosh machines has been its very good correspondence between print and screen colours.

4.10 MODEMS

A Modem is one device that most computer users who have surfed the Internet are aware of. A modem is required because though most of the telecommunications have become digital, most telephone connections at the user end are still the analog POTS (Plain Old Telephone Systems/Sets/Service). However, the computer is a digital device and hence another device is needed which can convert the digital signals to analog signals and vice-versa. Such a device is the Modem.
Modem stands for Modulator/Demodulator. Modulation is the process which puts digital information on to the analog circuit by modifying a constant wave (signal) called the Carrier. This is what happens when you press a button to connect to the Internet or to a web site. Demodulation is the reverse process, which derived the digital signal from the modulated wave. This is what happens when you receive data from a website which then gets displayed by your browser.

Discussion of modulation techniques is out of scope here (you can refer to your course on Computer Networks).

Modems are available as the following types:

1. **Internal Modems:** Internal Modems plug into expansion slots in your PC. Internal Modems are cheap and efficient. Internal Modems are bus-specific and hence may not fit universally.
2. **External Modems:** Modems externally connected to PC through a serial or parallel port and into a telephone line at the other end. They can usually connect to any computer with the right port and have a range of indicators for troubleshooting.
3. **Pocket Modems:** Small external Modems used with notebook PCs.
4. **PC-Card Modems:** PC and Modems are read with PCMCIA slots found in notebooks. They are like external Modems which fit into an internal slot. Thus, they give the advantage of both external and internal modems but are more expensive.

Modems come according to CCITT/ITU standards, e.g., V.32, V.32bis, V.42 etc.

Modem Language

Modems understand a set of instructions called *Hayes Command Set* or the *AT Command Set*. These commands are used to communicate with the Modem. Sometimes, when you are in trouble setting up your Modem, it is useful to know some basic commands, e.g., ATDT 17776 will dial the number 17776 across a Tone Phone and ATDP 17776 to the number 17776 if it is a Pulse phone.

4.11 SCANNERS

A Scanner is a device that allows you to capture drawings or photographs or text from tangible sources (paper, slides etc.) into electronic form. Scanners work by detecting differences in brightness of reflections from an image or object using light sensors. These light sensors are arranged in an array across the whole width that is scannable. This packing determines the resolution and details that can be scanned.

Scanners come in various types: *Drum Scanners, Flatbed Scanners, Hand Scanners and Video Scanners*. Drum Scanners use a rotating drum to scan loose paper sheets. Flatbed scanners have movable sensors to scan images placed on a flat glass tray.

These are the most expensive kind. Hand held Scanners are the cheapest and most portable.

They are useful for many applications but are small in size and need good hand control for high quality scanning. Video Scanners use Video technology and Video cameras instead of Scanning technology. Potentially, they can give high resolutions, scanners in the economical range give poor resolutions.

Figure 7: Scanners

When you buy a scanner, there are many factors that can be looked at: Compatibility of the Scanner with your Computer, The Technology (Depth, Resolution), the media types supported for scanning, How media can be loaded, Media size supported, Interfaces supported, physical dimensions, style and ease of use of the scanner.
One exciting application of Scanners is Optical Recognition of Characters (*OCR*). OCR software tries to recognise characters from their shapes and write out the scanned text as a text file. Though this technology is steadingly improving, it is still not completely reliable especially w.r.t. Indian scripts. However, it can be very useful to digitize the ancient texts written in Indian scripts.

Scanning technology is also everpresent nowadays in Bar-Code readers and MICR (Magnetic Ink Character Recognition) cheques. This technology is very useful for automating data at source of origin, thereby avoiding problems like inaccuracies in data entry, etc.

4.11.1 Resolution

Optical Resolution

Optical resolution or hardware resolution is the mechanical limit on resolution of the Scanner. For scanning, the sensor has to advance after each line it scans. The smallness of this advancement step gives the resolution of the Scanner. Typically, Scanners may be available with mechanical resolutions of 300, 600, 1200 or 2400 dpi. Some special scanners even scan at 10,000 dpi.

Interpolated Resolution

Each Scanner is accompanied by a software. This software can increase the apparent resolution of the scan by a technique called Interpolation. By this technique, additional dots are interpolated (added) between existing dots. This gives a higher resolution and smoother picture but without adding any additional information. The added dots will however lead to larger file sizes.

4.11.2 Dynamic Range/Colour Depth

Dynamic Range is the number of colours a colour scan or the number of grays a monochrome scanner can differentiate. The dynamic range is usually given as bit-depth or colour depth. This is simply the number of bits to distinguish the colours. Most scanners can do 256(8-bit), 1024(10-bit) or 4096(12-bit) for each primary colour. This adds up to and is advertised as 24-bit, 30-bit and 36-bit colour scanners. Actually though, to utilise the Colour Depth, the image under scanning must be properly focused upon and properly illuminated by the scanner.

Since the minimum colour range useful for human vision is 24-bits, more bits may seem useful. However, extra bits of scanning give you firm control for filtering the image colour to your requirements.

4.11.3 Size and Speed

Before actual scanning, a quick, low resolution scan called pre-scan is made to preview the image and select scanning area. After this only does the actual scan take place. Early colour scanners used to take three passes for a scan - one pass for each colour. Now, Scanners use just one pass and use photodetectors to detect the colours. Then, they operate as fast as monochrome Scanners. However, other issues are also involved.

High resolution scans of large images result in large file sizes. These can slow down processing since they need Hard Disk I/O for virtual memory. Hence, for large scans, it is necessary to have higher RAM in your PC.

4.11.4 Scanning Tips

- Do not scan at more resolution than required. This saves both time and Disk Space.
- Usually, it is not useful to scan at more than the optical resolution since it adds no new information. Interpolation can be done later with Image processing softwares.
- If scanning photographs for Printers, it is enough to scan at one-third the resolution of printing, since Printers usually use Super-Pixels (Dithering) for printing. Only for other kind of Printers, like continuous tone Printers, do you need to scan at the Printer resolution for best quality.
- For images to be seen only at the Computer Monitor, you may need to only scan so that the image size in pixels is the same as display resolution. That is, Scan resolution = Height of image in pixels divided by the screen size in inches. This may be surprisingly small.

4.12 POWER SUPPLY

Computer operate electronically — either through power supply obtained from your electric plug or batteries as in the case of portable computers. However, the current coming through your electric line is too strong for the delicate computer circuits. Also, electricity is supplied as AC but the computer uses DC. Thus, power supply is that equipment which takes AC from electrical supply and converts it to DC to supply to computer circuits. Early power supplies were linear power supplies and they worked by simply blocking one cycle of the AC current. They were superseded by the SMPS.

SMPS (Switched Mode Power Supply)

SMPS is the unit into which the electric supply from the mains is attached to your PC and this supplies DC to the internal circuits. It is more efficient, less expensive and more complex than linear supplies.

SMPS works in the following way: The electric supply received is sent to a component called **triac** which shifts it from 50 Hz to a much higher frequency (almost 20.000 Hz). At the same time, using a technique called **Pulse Width modulation**, the pulse is varied to the needs of the computer circuit. Shorter pulses give lower output voltage. A transformer then reduces back the voltage to the correct levels and rectifiers and filters generate the pure DC current.
SMPS has two main advantages: They generates less heat since they waste less power, and use less expensive transformers and circuits since they operate at higher frequencies.

The power requirement of a PC depends on the motherboard and the peripherals in your computer. Still, in modern PCs, your requirement may not be more than 150-200 Watts.

Check Your Progress 3

1. In what ways does a digital camera differ from a conventional camera?

   .....

   .....

   .....

   .....

2. Explain the term Resolution and how it applies to Monitors, Cameras, Printers, Scanners etc.

   .....

   .....

   .....

   .....

3. Explain the process of Colour management in Printers.

   .....

   .....

   .....

   .....

   .....

4. Compare Laptops using passive matrix and TFT technology. Which are cheaper in price?

   .....

   .....

   .....

   .....

   .....

5. To connect my modem to an ISP, I have to dial to the Pulse phone number 26176661, what Hayes Command set command would I give:

   a) ATDT 1777626176661 b) ATDT 26176661

   c) ATDP 1777626176661 d) ATDP 26176661

6. It is given that to print a 300dpi page on a particular Laser Printer, 2MB of memory is required. How much memory is required for a 600dpi page?

   a) 2MB b) 4 MB

   c) 8MB d) 16MB

7. Which of the following are Sound storage format standards?

   a) MIDI b) WAVE

   c) MP3 d) OGG

4.13 SUMMARY

In this unit, we discussed various Input/Output devices. We have covered the input devices Keyboard, Mouse and Scanner. Various types of Keyboards, Keyboard layouts (QWERTY, Dvorak) and technologies have been discussed. Various types of mice and their operation have been discussed. Different types of Scanners, the underlying technology and use in applications like OCR have been discussed.

The output devices discussed are Monitor, LCD and Printer. The technologies and specifications behind Monitors, LCD and Printers have been discussed. Colour management has also been discussed. Video cards, which control the display on monitors from the CPU and their system of display have been discussed with their characteristics like depth, resolution and memory. Modem is a communication device and thereby an I/O device. Its functioning has been discussed. The Power supply, and especially, the SMPS, which is actually input of electric power for the computing unit, has also been discussed.

4.14 SOLUTIONS/ ANSWERS

Check Your Progress 1

Question 4. Ans. (c)

Check Your Progress 1

1. The main Merit of Dvorak-Dealey keyboard is the scientific design using Hand alteration. However, since it came much later than QWERTY it did not become popular, as QWERTY was already well established.

2. Keyboard touch gives you a feedback mechanism. This tells you when you have pressed a key enough and invloluntarlily allows faster typing. The preferred touch is an individual choice but the best feedback is provided with an overcenter feel with a `click' sound. The most suitable touch is given by Rubber Dome keyboards. (refer text for details).

3. Besides the standard precautions while attaching Hardware, one has to take precaution regarding interrupt conflict for serial devices., since Serial ports share their interrupts. (refer text for details).

Check Your Progress 2

1. A true-colour system has a depth of 24 bits per pixel. This means that 8 bits each are assigned to R,G and B i.e. there are 8 Colour Planes. Hence, in figure-4 replace ` n' by 8 to draw the new figure.

2. Framebuffer is another name for the Display Memory. This is like a time-slice of what you see on your monitor. Discuss how framebuffer is handled differently in early display systems, PCI, AGP and UMA. (refer text for details).

3. Shadow Mask: Trinitron uses Aperture Grills instead of Shadow Mask, for the same purpose.
Dot Pitch: Similiarly, instead of Dot Pitch, there is Slot Pitch.
explain the terms Shadow Mask, Aperture Grill, Dot Pitch and Slot Pitch (refer text).

4. Ans. (b) $1024 \times 768 \times 2$Bytes = 1.6MB. RAM is/was available as 1MB, 4MB, 16MB etc.
5. Ans. (a) Total screen size = $12 \times 9 = 108$ inches. image size $= 1024 \times 768 = 786432$ pixels. divide 108 inches by 786432.

Check Your Progress 3

1. In a digital camera, photos are stored in digital format. Instead of film, these cameras use Semiconductor devices, called image sensors. There are many other differences regarding quality, resolution etc.

2. Resolution is a generic term the parameter that defines the possible sharpness or clarity of something i.e. how clearly that thing can be resolved. This applies especially to images. See in what different ways it is used for Monitors, Cameras, Printers, Scanners and even Mice.

3. It tells about physical mixing, optical mixing and RGB and CMYK schemes. The rechnique of dithering is used for rich colour quality. Colours also differ on monitors and printers. To maintain similarity is also an important issue.

4. Compare Laptops made using passive matrix and TFT technology. Which are cheaper in price?

   In a Passive matrix arrangement, the LCD has a grid of horizontal and vertical conductors. Each pixel is located at an intersection. When a current is recieved by the pixel, it becomes dark whereas in Active Matrix, also called TFT (Thin Film Transistor) technology, each pixel is active, working as a relay. Hence, it needs less power and gives better quality display. Passive matrix LCDs are cheaper but now, TFT LCDs are also economically available. (find out the latest from the market).

5. Ans. (d) ATDP 26176661.

6. Ans. (c) 8MB. The memory requirement increases as a square of the resolution(dpi), so an increase of two times in the dpi leads to an increase of four times in the memory requirement.

7. Ans. All of them.

References:

1) http: //whatis.techtarget.com/.

2) http: //www.epanorama.net/links/pc/index.htm.

3) http: //www.howstuffworks.com/.

4) Mark Minasi. *The Complete PC Upgrade and Maintenance Guide*. BPB Publications, New Delhi, 2002.

5) Winn L. Rosch. *Hardware Bible*. Techmedia, New Delhi, 1997.

SOCIAS-IGNOU/P.O.10.5T/MAY, 2004