1.0 INRODUCTION

The use of Information Technology is well recognised. It has become must for the survival of business houses with the growing information needs. Computer is one of the major components of an Information Technology network. Today, computer technology has permeated every sphere of existence of modern man. From railway reservations to medical diagnosis, from TV programmes to satellite launching, from matchmaking to criminal catching- everywhere, we witness the elegance, sophistication and efficiency possible only with help of computers.

In this block, we will introduce you to the computer hardware technology, how does it work and what is it? In addition we will also discuss some of the terminology closely linked with information Technology and computers. More details on these terms can be obtained from the further readings. We about the basic computer structure, the memory system and characteristics of various memories etc. We will also define the terms such as the main memory, cache memory, magnetic secondary storage and optical memories.

1.1 OBJECTIVES

This unit being the first unit of the block introduces you to the world of computer. At the end of the unit you will be able to:

- define the term computer
- define Von Neumann architecture
- describe key characteristics of memory system.
- distinguish various types of memories
- differentiate various external memories
- define the importance of cache memory.

1.2 WHAT IS A COMPUTER?

Let first define the term computer. Computer is defined in the Oxford dictionary as "An automatic electronic apparatus for making calculations or controlling operations that are expressible in numerical or logical terms".

The definition clearly categorise computer as an electronic apparatus although the initial computers were mechanical and electromechanical, definition is also pointing towards the two major areas of computer application viz. data processing and computer assisted control/operations. Another important confluence of the definition is the fact that the computer can perform only those operations/calculations which can be expressed in Logical or Numerical terms.

The basic function performed by a computer is the execution of a program. Program is a sequence of instructions, which operates on data to perform certain tasks. In modern digital computers data is represented in binary form by using two symbols 0 and 1 which are called binary digits or bits. Computers use eight bits to represent a character internally. This allows up to 2^8=256 different items to be represented uniquely. This collection of eight bits is called a byte. Thus, one byte is used to represent one character internally. One of the most common codes to represent characters in Computers is ASCII (American Standard Code of Information Interchange). Most computers use two bytes or four bytes to represent numbers (positive and negative) internally. Another term that is commonly used in computer is a Word. A word may be defined as a unit of information that a computer can process or transfer at a time. A word must be equal to the number of bits transferred between the central processing unit and the main memory in a single step or it may be defined as the basic unit of storage of integer data in a computer. Normally, a word may be equal to 8, 16, 32, or 64 bits. The terms like 32 bit computer, 64 bit computers etc. basically points to the word size of the computer.

One of the key aspects in program execution is the execution of an instruction. The key questions that can be asked in this respect are (a) how are the instructions supplied to the computer? And (b) how are they interpreted and executed? We will answer these questions along with the discussion on the basic structure of the computer system.

Most of today's computer designs are based on concepts developed by John von Neumann referred to as the Von Neumann architecture. Von Neumann proposed that their should be a unit performing arithmetic and logical operation on the data. This unit is termed as Arithmetic Logic Unit (ALU). A control unit directs the ALU to perform specific arithmetic or logic function on the data. Therefore in such a system, by changing the control signal the desired operation can be performed on data.

But, how can these control signals be supplied? Let us try to answer this from the definition of a program. A program consists of a sequence of steps. Each of these steps, require certain arithmetic of logical or input/output operations to be performed on data. Therefore, each step may require a set of control signals. Is it possible for us to provide a unique code for each set of control signals? Well the answer is Yes. But what do we do with these codes? What about adding a hardware segment that accepts the code and generates control signals? The unit that interprets a code to generate respective control signal is termed as Control unit (CU). Thus, a program now consists of a sequence of codes. This machine is quite flexible, as we only need to provide a new sequence of codes for a new program. Each code is, in effect, an instruction, for the computer. The hardware interprets each of these instructions and generates respective control signals.

The arithmetic Logic Unit (ALU) and the Control Unit (CU) together are termed as the Central Processing Unit (CPU). The CPU is the most important component of a computer's hardware. The ALU performs the arithmetic operations such as addition, subtraction, multiplication and division, and the logical operations such as: "Is A = B? (Where A and B are both numeric or alphanumeric data), "Is a given character equal to M (for male) of F (for female)?" The control unit interprets instructions and produce the respective control signals.

All the arithmetic and logical operations are performed in the CPU in special storage areas called registers. The size of the register is one of the important considerations in determining the processing capabilities of the CPU. Register size refers to the amount of information that can be held in a register at a time for processing. The larger the register size, the faster may be the speed of processing. A CPU's processing power is measured in Million Instructions Per Second (MIPS).

How can the instructions and data be put into the computers? The instruction and data need to be supplied by external environment, therefore, an input module is needed. Main responsibility of input module will be to put the data in the form of signals that can be recognised by the system. Similarly, we need another component that will report the results in proper format and form. This component is called output module. These components are referred together as input/output (I/O) components.

Are these two components sufficient for a working computer? No, because input devices can bring instructions or data only sequentially and a program may not be executed sequentially as jump instructions are normally encountered in programming. In addition, more than one data elements may be required at a time. Therefore, a temporary storage area is needed in a computer to store temporarily the instructions and the data. This component is referred to as memory. It was pointed out by von-Neumann that the same memory can be used for storing data and instructions. In such case the data can be treated as data on which processing can be performed, while instructions can be treated as data which can be used for the generation of control signals.

The memory unit stores all the information in a group of memory cells, also called memory locations, as binary digits (bits). Each memory location has a unique address and can be addressed independently. The contents of the desired memory locations are provided to the central processing unit by referring to the address of the memory location. The amount of information then can be held in the main memory is known as memory capacity. The capacity of the main memory is measured in Kilobytes (KB) or Megabytes (MB). One kilobyte stands for 2^{10} bytes, which is 1024 bytes (or approximately 1000 bytes). A megabyte stands for 2^{10} kilobytes, which is approximately little over one million bytes.

In addition, to transfer the information, the computer system internally needs the interconnections. The most common interconnection structure is the Bus structure. A bus is a set of wires (Lines) which you can visualise on the motherboard of a, computer. It is a shared media. A bus connecting the CPU; memory and I/O components is called a system bus. A system bus may consist of 50 to 100 separate lines.

Let us summarise the key features of a von Neumann machine. The hardware of the von Neumann machine consists of:

- A CPU which includes a ALU and CU
- A main memory system
- An Input/output system
- The von Neumann machine used stored program concept, i.e., the program and data are stored in the same memory unit. The computers prior to this idea used to store programs and data on separate memories. Entering and modifying these programs were very, difficult as they were entered manually by setting switches and plugging and unplugging.
- Each location of the main memory of von Neumann machine can be addressed independently.

- Execution of instructions in von Neumann machine is carried out in a sequential fashion (unless explicitly altered by the program itself) from, one instruction to the next.

Figure 1 shows the basic structure of a conventional von Neumann machine.

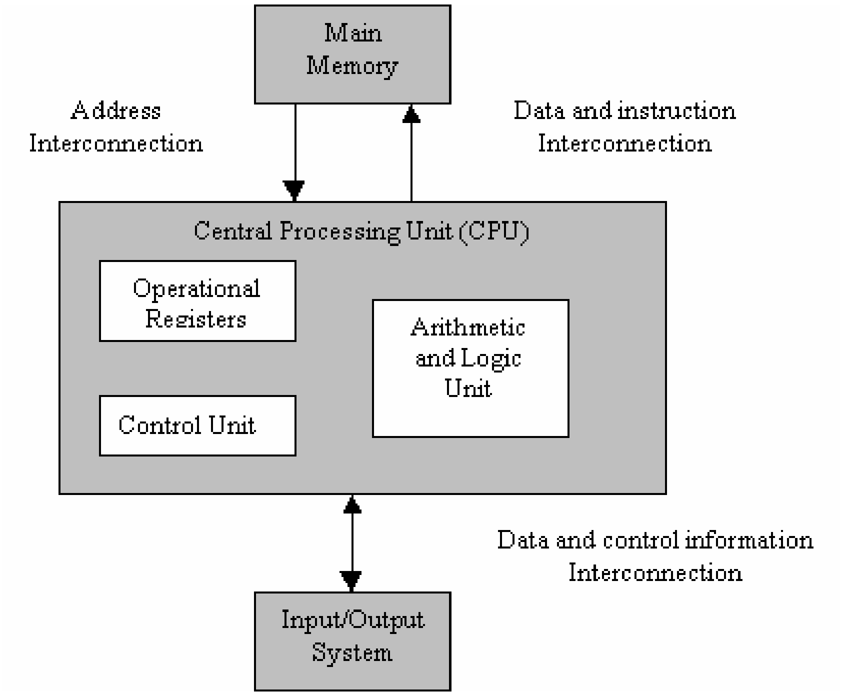

Main Memory
Address Interconnection
Data and instruction Interconnection
Central Processing Unit (CPU)
Operational Registers
Arithmetic and Logic Unit
Control Unit
Data and control information Interconnection
Input/Output System

Figure 1: Structure of a computer

A von Neumann machine has only a single path between the main memory and control unit (CU). This feature/constraint is referred to as von Neumann bottleneck. Several other architectures have been suggested for modern computers.

Check Your Progress 1

State true or false.

1. A byte is equal to 8 bits and can represent a character internally.

True □ False □

2. A word on PC386 is equal to one byte.

True □ False □

3. von Neumann architecture specifies different memory for data and instructions. The memory which stores data is called data memory and the memory that stores instructions is called instruction memory.

True □ False □

4. In von Neumann architecture each bit of memory can be accessed independently.

True □ False □

5. A program is a sequence of instructions designed for achieving a task/goal.

True □ False □

6. One MB is equal to 1024 KB.

True □ False □

1.2.1 The Computer and Integrated Circuit Technology

The era of microelectronics (small electronic) with the invention of Integrated Circuits (ICs) brought a new era of computing. But before we discuss about the relation of computers to integrated circuit technology, let us explore more about the term Integrated Circuit (IC).

In an integrated circuit the components such as transistors, resistors and conductors are fabricated on semiconductor material such as silicon. Thus, a desired circuit can be fabricated in a tiny piece of silicon rather than assembling several discrete components into the same circuit. Hundreds or even thousands of transistors could be fabricated on a single wafer of silicon. In addition, these fabricated transistors can be connected with a process of metalisation to form logic circuits on the same chip they have been produced.

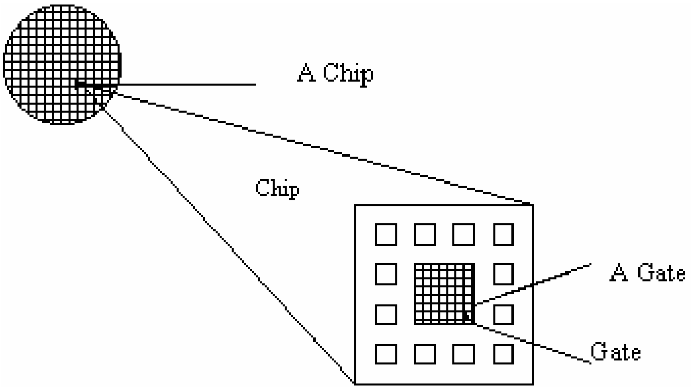
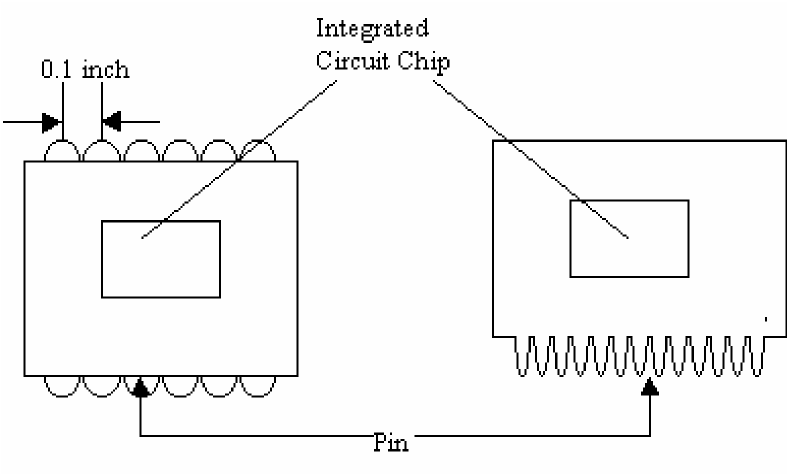


Figure 2: Wafer, Chip and Gate

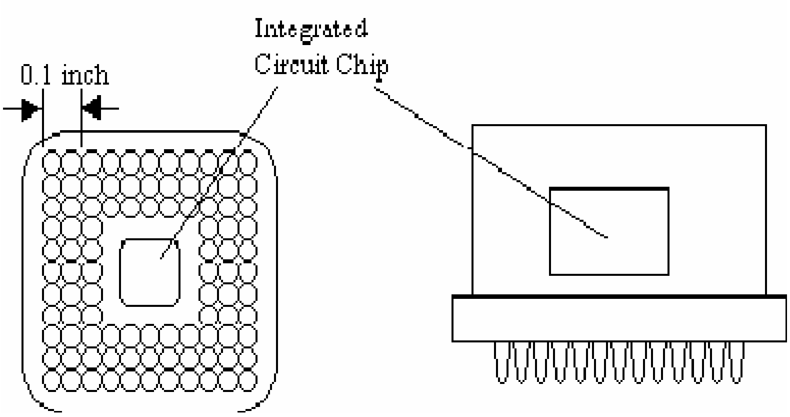
An integrated circuit is constructed on a thin wafer of silicon that is divided into a matrix of small areas (size of the order of a few millimetres squares). An identical circuit pattern is fabricated on each of these areas and the wafer is then broken into chips (Refer figure 2). Each of these chips consists of several gates, a useful logic component, and a number of input and output connection points. Each of these chips then can be packaged separately in a housing to protect it. In addition, this housing provides a number of pins for connecting this chip with other devices or circuits. The pins on these packages can be provided in two ways:

* In two parallel rows with 0.l inch spacing between two adjacent pins in each row. This package is called dual in-line package (DIP) (Refer Figure3 (a)).

* In case, more than hundred pins are required then pin grid array (PGA) where pins are arranged in arrays of rows and columns, with spacing between two adjacent pin of 0. 1 inch (Refer Figure 3(b)).



(a) A 24 pin dual in-line package (DI)
(Contains dual rows which are in-line)



(b) 144-pin Pin Grid Array (PGA) package

Figure 3: Integrated Circuit Packages

Different circuits can be constructed on different wafers. All these packaged circuit chips then can be interconnected on a printed-circuit board to produce several complex electronic circuits such as computers.

Initially, only a few gates were integrated reliably on a chip and then packaged. These initial integration was referred to as small-scale integration (SSI). Later, with the advances in microelectronics technologies the SSI gave way to Medium Scale Integration where 100s of gates were fabricated on a chip. Then came Large Scale Integration (1,000 gates) and very large integration (VLSI 100,000,000 components are expected to be fabricated on a single chip. According to this, expected projections is that in near future almost 10,000,000,000 components will be fabricated on a single chip.

What are advantages of having densely packed Integrated Circuits? These are:

Low cost: The cost of a chip has remained almost constant while the chip density (number of gates per chip) is ever increasing. It implies that the cost of computer logic and memory circuitry has been reducing rapidly.
Greater Operating Speed: More is the density, the closer are the logic or memory elements, which implies shorter electrical paths and hence the higher operating speed.

Smaller computers-better portability

Reduction in power and cooling requirements

Reliability: The integrated circuit interconnections are much more reliable than soldered connections. In addition, densely packed integrated circuits enable fewer inter-chip connections. Thus, the computers are more reliable.

One of the major milestones in this technology was the very large scale integration (VLSI) where thousands of transistors can be integrated on a single chip. The main impact of VLSI was that, it was possible to produce a complete CPU or main memory or other similar devices on a single IC chip. Let us discuss some of the important breakthroughs of VLSI technologies.

Semiconductor Memories

Initially the IC technology was used for constructing processor, but soon it was realised that same technology can be used for construction of memory. The first memory chip was constructed in 1970 and could hold 256 bits. Although the cost of this chip was high, but gradually the cost of semiconductor memory is coming down. The memory capacity per chip has increased for e.g. 1k, 4K, 16K, 64K 256K and 1M bits.

Microprocessors

Keeping pace with electronics as more and more component were fabricated on single chip, fewer chips were needed to construct a single processor. Intel in 1971 achieved the breakthrough of putting all the components on a single chip. The single chip processor is known as a microprocessor. The Intel 4004 was the first microprocessor. It was a primitive microprocessor designed for a specific application. Intel 8080 that came in 1974 was the first general purpose microprocessor. It was an 8 bits microprocessor. Motorola is another manufacturer in this area. At present 32 and 64 bit general purpose microprocessors are already in the market. For example, Intel Pentium is a 32 bit processor, similarly Motorola's 68000 is a 32 bit microprocessor. P6 that is announced by Intel 1995 can process 64 bit data at a time. Figure 4 shows the Intel family of microprocessors.

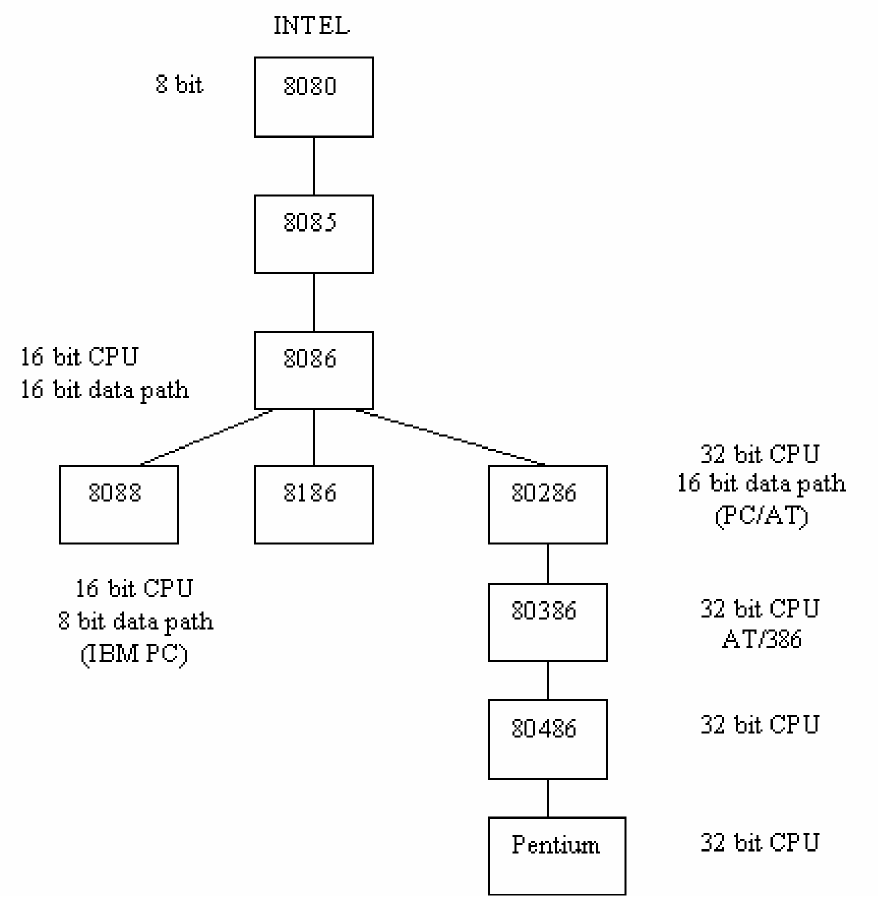


Figure 4: Intel Microprocessor Families

The VLSI technology is still evolving and more and more powerful microprocessor and more storage space now is being put in a single chip.

1.2.2 Classification of Computers

One question which we have still not answered is- Is there any classification of computers? Well for quite sometime computers have been classified under three main classes. These are:

Microcomputers

Minicomputers

Mainframes

Although with development in technology the distinction between these is becoming blurred. Yet it is important to classify them as it is sometimes useful to differentiate the key elements and architecture among the classes.

Microcomputers

A microcomputer's CPU is a microprocessor. The first microcomputers were built around 8-bit microprocessor chips. What do we mean by an 8-bit chip? It means that the chip can retrieve instructions/data from storage, manipulate, and process an 8-bit data at a time or we can say that the chip has a built-in 8-bit data Transfer path. An improvement on 8-bit chip technology was seen in early 1980s, when a series of 16-bit chips namely 8086 and 8088 were introduced by Intel Corporation, each one with an advancement over the other. 8088 is a 8/16 bit chip i.e. an 8-bit path is used to move data between chip and primary storage (external path), at a time, but processing is done within the chip using a 16 bit path (internal path) at a time. 8086 is a 16/16 bit chip i.e. the internal and external paths both are 16 bit wide. Both these chips can support a primary storage capacity of upto 1-megabyte (MB).

Most of the popular microcomputers are developed around Intel's chips, while most of the minis and, superminis are built around Motorola's 68000 series chips. With the advancement of display and VLSI technology now a microcomputer is available in very small size. Some of these are laptops, notebook computers etc. Most of these are of the size of a small notebook but equivalent capacity of an older mainframe.

Minicomputer

The term minicomputer originated in 1960s when it was realised that many computing tasks do not require an expensive contemporary mainframe computers but can be solved by a small, inexpensive computer. Initial minicomputers were 8 bit and 12 bit machines but by 1970s almost all minicomputers were 16 bit machines. The 16 bit minicomputers have the advantage of large instruction set and address field; and efficient storage and handling of text, in comparison to lower bit machines. Thus, 16 bit minicomputer was more powerful machine which could be used in variety of applications and could support business applications alongwith the scientific applications.

With the advancement in technology the speed, memory size and other characteristics developed and the minicomputer was then used for various stand alone or dedicated applications. The minicomputer was then used as a multi-user system, which can be used by various users at the same time. Gradually the architectural requirement of minicomputers grew and a 32-bit minicomputer, which was called supermini, was introduced. The supermini had more peripheral devices, large memory and could support more users working simultaneously on the computer in comparison to previous minicomputers.

Mainframes

Mainframe computers are generally 32-bit machines or on the higher side. These are suited to big organisations, to manage high volume applications. Few of the popular mainframe series are IBM, HP, etc. Mainframes are also used as central host computers in distributed systems. Libraries of applications programs developed for mainframe computers are much large than those of the micro or minicomputers because of their evolution over several decades as families of computing. All these factors and many more make the mainframe computers indispensable even with the popularity of microcomputers.

Supercomputer

The upper ends of the state of the art mainframe machine are the supercomputer. These are amongst the fastest machines in terms of processing speed and use multiprocessing techniques, were a number of processors are used to solve a problem. Lately ranges of parallel computing products, which are multiprocessors sharing common buses, have been in use in combination with the mainframe supercomputers. The supercomputers are reaching upto speeds as well over 25000 million arithmetic operations per second. India also has its indigenous supercomputer.

Supercomputers are mainly being used for weather forecasting, computational fluid dynamics, remote sensing, image processing, biomedical applications, etc. In India, we have one such mainframe supercomputer system- CRAY XMP-14, which is at present, being used by Meterological Department.

Check Your Progress 2

1. What is a general-purpose machine?

....

....

2. Define the following terms briefly:

(i) Microprocessors

(ii) Laptop

(iii) Supercomputer

....

....

1.3 MEMORY SYSTEM

Memory in a computer system is required for storage and subsequent retrieval of the instructions and data. A computer system uses variety of devices for storing the instructions and data which are required for its operations. Normally we classify the information to be stored on computer in two basic categories: Data and the Instructions.

"The storage device 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. But unfortunately, faster memory technology is more expensive. In addition, fast memories require power supply till the information need to be stored. Both these things are not very convenient, but on the other hand the memories with less cost have very high access time, that is the time taken by CPU to access a location in the memory in high, which will result in slower operation of the CPU. Thus, the cost versus access time anomaly has lead to a hierarchy of memory where we supplement fast memories with larger, cheaper, slower memories. These memory units may have very different physical and operational characteristics, therefore, making the memory system very diverse in type, cost, organisation, technology and performance. This memory hierarchy will be fruitful if the frequencies of access to slower memories are significantly less than the faster memories.

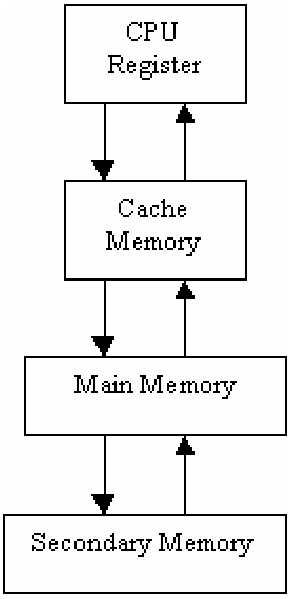


Figure 5: The Memory Hierarchy

Thus, a memory system can be considered to consist of three groups of memories. These are:

(a) **Internal Processor Memories:** These consist of the small set of high speed registers which are internal to a processor and are used as temporary locations where actual processing is done.

(b) **Primary Memory or Main Memory:** It is a large memory, which is fast but not as fast as internal processor memory. This memory is accessed directly by the processor. It is mainly based on integrated circuits.

(c) **Secondary Memory/Auxiliary Memory/Backing Store:** Auxiliary memory in fact is much larger in size than main memory but is slower than main memory. It normally stores system programs (programs which are used by system to perform various operational functions), other instructions, programs and data files. Secondary memory can also be used as an overflow memory in case the main memory capacity has been exceeded. (How? The answer is not supplied in the block. You need to refer to further readings to get this answer). Secondary memories cannot be accessed directly by a processor. First the information of these memories is transferred to the main memory and then the information can be accessed as the information of main memory.

There is another kind of memory, which is increasingly being used in modern computers and this is called Cache memory. It is logically positioned between the internal memory (registers) and main memory. It stores or catches some of the content of the main memory, which is currently in use of the processor.

Before discussing more about these memories let us first discuss the technological terms commonly used in defining and accessing the memory.

1.4 CHARACTERISTICS TERMS FOR VARIOUS MEMORY DEVICES

The following terms are most commonly used for identifying comparative behaviour of various memory devices and technologies.

Storage Capacity: It is a representative of the size of the memory. The capacity of internal memory and main memory can be expressed in terms of number of words or bytes. The storage capacity of external memory is normally measured in terms of bytes.

Access Modes: A memory is considered to consist of various memory locations. The information from these memory locations can be accessed in the following ways:

- **Random Access Memory (RAM):** It is the mode in which and memory location can be accessed in any order in the same amount of time. Ferrite and Semiconductor memories, which generally constitute main memory, are of this nature. The storage locations can be accessed independently and there exist separate access mechanism for each location.

- **Sequential Access**: On the other hand we have memories which can be accessed in a pre-defined sequence for example, the songs stored on a cassette can be accessed only one by one. The example of sequential access memory is Magnetic Tape. Here the access mechanism need to be shared among different locations. Thus, either the location or the read/write head or both should be moved to access the desired location.

- **Direct Access:** In certain cases the information is neither accessed randomly nor in sequence but something in between. In direct access, a separate read/write head exist for a track and on a track the information can be accessed serially. This semi-random mode of operations exists in magnetic disks.

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

Permanence of Storage:

Some memories loose information over a period of time. For example, there can be some memories where the stored data bit value 1 looses its strength to become 0 over a period of time. These kinds of memories require refreshing. The memories, which require refreshing, are termed as dynamic memories. In contrast, the memories, which do not require refreshing, are called static memories. Another factor, which can destroy the contents, is the presence and absence of electricity. The memories, which looses their content on failure of power are termed as **volatile** memories, those, which do not are called **non-volatile.** Magnetic memories are non-volatile and semi-conductor main memories are volatile in nature.

Physical Characteristics:

In this respect the memory devices can be categorised into four main categories viz., electronic, magnetic, mechanical and optical. One of the requirements for a storage device is that it should exhibit two well-defined physical states, such that 0 and 1 can be represented in those two states. The Data transfer rate of the memory depends on the how quickly the state can be recognised and altered. The following table lists some of the memory technologies along with their physical and other important characteristics.

Technology	**Access time (in seconds)**	**Access Mode**	**Performance of Storage**	**Physical nature of storage medium**	**Average cost (Rs/bit) (Approx.)**
Semiconductor memories	10^{-8}	Random	Volatile	Electronic	10^{-2}
Magnetic disk	10^{-2}	Direct	Non-volatile	Magnetic	10^{-5}
Magnetic tape	10^{-1}	Sequential	Non-volatile	Magnetic	10^{-5}
Compact disk ROM	Approx. 10^{-1}	Direct	Non-volatile	Optical	10^{-7}

Figure 6: Characteristics of some memory technologies

The physical size of memories should be small and it must consume less power. Higher power consumption may result in more costly equipment for internal cooling of computer. The storage devices, which require mechanical motion e.g. hard disks are more prone to failure rather than the semiconductors memories which are totally electronic in nature. Very high speed semiconductor memories are also prone to failure as technology is moving towards its limits.

Check Your Progress 3

1. State true or false.

(a) Memory hierarchy is built in computer system, as the main memory can not store very large data.

True □ False □

(b) The secondary memory is slower than that of main memory but has a larger capacity.

True □ False □

(c) In Random access Memory any memory location can be accessed independently.

True □ False □

2. What are the differences in:

(a) Volatile versus Non-volatile memory

....

....

(b) Static versus dynamic memories

....

....

1.5 MAIN MEMORY OR PRIMARY STORAGE

Primary memory consists of semiconductor memory chips and is used to store the data and programs currently in use. Each storage element of memory is directly (randomly) accessible and can be examined and modified without affecting other cells and hence primary memory is also called Random Access Memory (RAM). Main memory stores a variety of critical information required for processing by the CPU. How does it store the information? Please answer it yourself.

The memory unit stores all the information in memory cells also called memory locations, in binary digits. Each memory location has a unique address. The contents of the desired memory locations are provided to the central processing unit by referring to the address of the memory location. The amount of information that can be held in the main memory is known as memory capacity. The capacity of the main memory is measured in kilobytes (KB) or Megabytes (MB).

All modern computers use semiconductor memory as its main memory. Semiconductor memory is known as Random Access Memory (RAM) since any part of the memory can be accessed for reading and writing. Another part of main memory is Read Only Memory (ROM). ROMs (Read Only Memories) 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. The ROMs can be used in storing programs provided by the manufacturer of computer for basic operations. ROMs are non-volatile in nature and need not be loaded in a secondary storage device. ROMs are fabricated in large number in a way where there is no room for even a single error.

ROMs can be written only at the time of manufacture. Another similar memory is PROM. PROMs are also non-volatile and can be programmed only once by a special write device hence the name Programmable ROM (PROM). The writing process in PROM can be performed electrically by the supplier or the customer. 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 (in ROMs at the time manufacture and PROMs at any time later also), but in both the cases once whatever is written on, cannot be changed. But what about a case where you read mostly but write only very few times. This lead to the concept of Read mostly memories and the best examples of these are EPROMs (Erasable PROMS) and EEPROMs (Electrically erasable PROMS). The EPROMs can be 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 bring them to same initial state. This erasure is a time consuming process. 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 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. Figure 7 summarise the features of these read only and read mostly memories.

Memory Type	**Write Time**	**Order of Read Time**	**Number of Cycles allowed**
ROM	Once at the time of manufacture	Nano seconds	ONE
PROM	Hours	Nano seconds	ONE
EPROM	Minutes (including time of ensure)	Nano Seconds	HUNDREDS
EEPROM	Milliseconds	Nano seconds	THOUSANDS

Common features
* Non-destructive
* Long data life
* Non-volatile

Figure 7: Features of Read Only and Read Mostly Memories

1.6 EXTERNAL/AUXILIARY MEMORY

As discussed earlier the cost of RAM is very high and the semiconductor RAMs are mostly volatile, therefore, it is highly likely that a secondary cheap media should be used which should show some sort of permanence of storage and should be relatively inexpensive. The magnetic material was found to be inexpensive and quite long lasting material. Therefore, became an ideal choice to do so. Magnetic tape and magnetic disks are commonly used as storage media. With the advancements in the optical technology now the optical disks are trying to make inroads as one of the major external memory. We will discuss about the characteristics of these memories in this section.

1.6.1 Magnetic Disk

A magnetic disk is a circular platter of plastic, which is coated with magnetised material. One of the key components of a magnetic disk is a conducting coil named as Head which performs the job of reading and writing on the magnetic surface. The head remains stationary while the disk rotates below it for reading or writing operation.

Data Organisation and Format

The head of disk is a small coil and reads or writes on the position of the disk rotating below it, therefore, the data is stored in concentric set of rings (refer figure 4). These are called tracks. The width of a track is equal to the width of the head. To minimise the interference of magnetic fields and to minimise the errors of misalignment of head, the adjacent tracks are separated by inter track gaps. As we go towards the outer tracks the size of a track increase but to simplify electronics same number of bits are stored on each track.

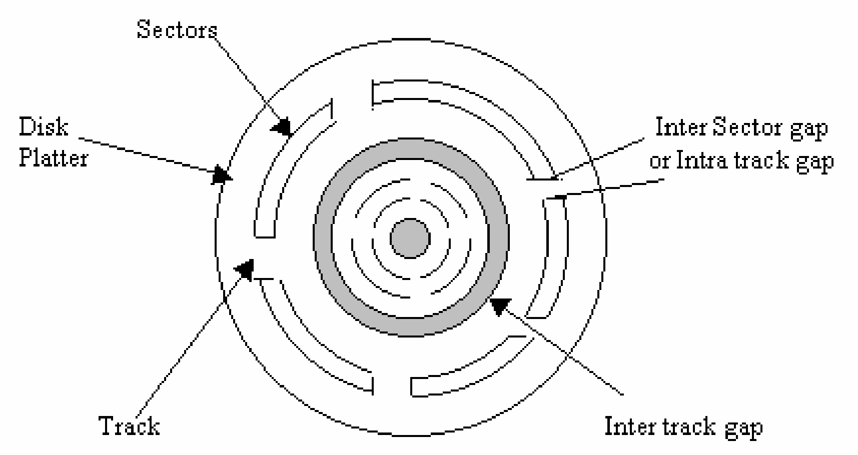


Figure 8: **Logical layout of Magnetic Disk**

The data is transferred from and to the disks in blocks. Block is a section of disk data and is normally equal to a sector. A track is divided into 10-100 sectors and these sectors should be either fixed or variable length sectors. Two adjacent sectors are separated by intra-track gaps. This helps in reducing the precision requirements of sectors. To identify the sector position normally there may be a starting point on a track or a starting and end point of each sector.

Floppy disks

A floppy disk is made of a flexible thin sheet of plastic material with a magnetic coating and grooves arranged in concentric circles with tracks. Floppy disk becomes a convenient recording medium to transport information from one location to another. Disk is removable from the reading device attached to the computer and therefore provides unlimited storage capacity. The floppy disks of today are available in two sizes 5.25 inches and 3.5 inches and their capacity ranges from 360 KB to 1.44 MB per disk.

1.6.2 Winchester Disk

This is a sealed rigid magnetic oxide medium disk, which typically holds 10 MB to 10 GB of data. Winchester disks are not removable from the drive and since they are sealed dust and other contaminations, which are likely in a floppy disk, are minimised. These provide substantially faster data access compared to floppy disk and provide very large data storage for on-line retrieval.

Sides: The magnetic coating if applied to both the sides of the platter is called as double sided disks. The data can be recorded on either side of these disks. Some inexpensive disks were initially single sided.

Platters: Some disks have single platter e.g. floppy disks while some disks have multiple platters which are stacked vertically, normally at a distance of an inch. This is known as disk pack. In disk pack one additional term cylinder is defined which is the ring of all co-centric tracks (Figure 9). A disk pack can contain multiple heads mounted with the same arm.

Access time on Disk

Disk operates in semi-random mode of operation and normally is referenced block wise. The data access time on 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 electronic circuit to select the require head while in movable head disk it is the time required to move the head to a particular track.

Latency time: 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.

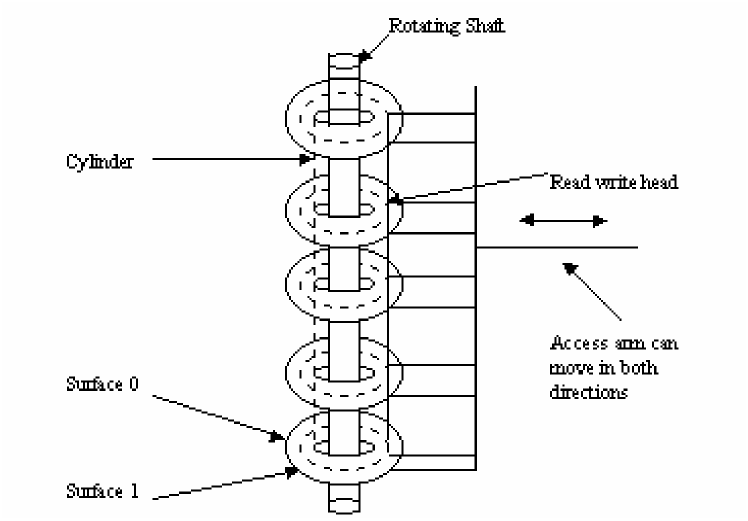


Figure 9: The Disk Pack

In addition to these two times the time taken to read block of word 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 sum of seek time and latency time. Since, access time to disks is large, therefore, it is advisable to read a sizable portion of data in a single go and that is why the disks are referenced block wise.

1.6.3 Magnetic Tape

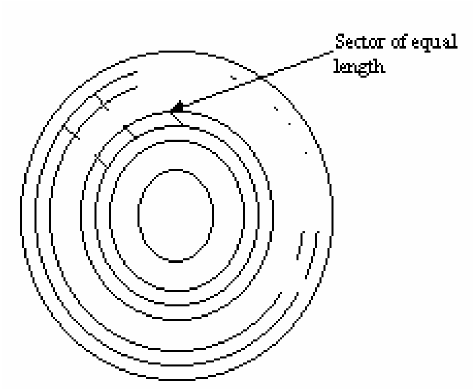
Magnetic tapes are mounted on reels or a cartridge or a cassette of tape to store large volumes or backup data. These are cheaper and since these are removable from the drive, they provide unlimited storage capacity. Information retrieval from tapes is sequential and not random. These are not suitable for on-line retrieval of data, since sequential searching will take long time. These are convenient for archival storage, or for backup. The tapes are one of the earliest storage devices. They are low cost, low speed, portable and are still widely used because of their low cost.

1.6.4 Optical Memories

Optical memories are alternate mass storage devices with huge capacity. The advent of 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 laser beam. These devices which are memories can store large amount of data. We will discuss here three optical memory devices, which are now becoming increasingly popular in various computer applications.

CD-ROM

The CD-ROM (Compact disk read-only memory) is a direct extension of audio CD. CD-ROM players are more rugged and have error-correction facility. This ensures proper data transfer from CD-ROM to the main memory of the computer. CD-ROM is written into during the process of manufacture by a high power laser beam. Information is retrieved from a CD-ROM using a low power laser, which ingenerate in an optical disk drive unit. In CD-ROMs the information is stored evenly across the disk in segments of the same size. Therefore, in CD-ROMs data stored on a track increases as we go towards outer surface of disk. Thus, the CD-ROMs are rotated at variable speeds for the reading process.



Identification of first address is 0 minute
0 second
0 sector

Figure 11 : A CD-ROM's disk layout

Figure 10 indicates the layout used for CD-ROMS. As discussed earlier, the data is stored sequentially along a spiral track. In this disk random access becomes more difficult because locating a desired address involves first moving the head to the specific area then adjusting the rotation speed and then reading the address, and then to find and access the specific sector.

CD-ROMs are very good for distributing large amount of information **or** data to large number of users. The three main advantages of CD-ROMs are:

- Large data/information storage capacity
- Mass replication is inexpensive and fast
- These are removable disks, thus, are suitable for archival storage

The disadvantages of CD-ROMs are:

- It is read-only, therefore, cannot be updated
- Access time is longer than that of a magnetic disk

WORM

In certain applications only few copies of compact disks are to be made which makes the CD-ROMs production economically unviable. For such cases write-once, read-many CD has been developed. WORM disks are prepared in such a way that they can be written only once subsequently by a laser beam of modest intensity. The disk controller of WORM is more expensive than that of CD-ROM. WORM uses sector structures same as that of magnetic disks. High power laser first prepares the WORM disk. A CD writer can write them into once.

Erasable Optical Disk

The most recent development in optical disks is the erasable optical disk. The data in these disks can be changed repeatedly as the case with any magnetic disk. A feasible technology that has proved commercially feasible for erasable optical disk is the magneto-optical system. In such systems, a laser beam is used along with a magnetic field to read or write the information on a disk which is coated with a magnetic material.

The erasable optical disk is a true secondary storage device (unlike CD-ROMs and WORM). The main advantages of erasable optical disk over magnetic disk are:

- The capacity of an erasable disk is very high in comparison to magnetic disk. For example, a 5.1/4 inch optical disk can store around 650 Mbytes of data while the Winchester disks normally can store a maximum capacity of 512 MB.
- The erasable optical disks are portable while a Winchester disk is not.
- The erasable optical disks are highly reliable and have a longer life.
- Erasable optical disk also uses format that makes semi-random access feasible.

The only disadvantage of this disk is the high cost. This disadvantage will disappear in near future.

Check Your Progress 4

1. Compare and contrast RAM and ROM.

   ....

   ....

2. What is the importance of read mostly memories?

   ....

   ....

3. What is the head of a disk?

   ……………………………………………………………………………………………………

   ……………………………………………………………………………………………………

4. Match the following pairs

(i)	Variable rotation speed	(a)	Magnetic tape
(ii)	Low cost, low speed devices	(b)	Floppy disks
(iii)	Double sided double density	(c)	CD-ROM

1.7 HIGH SPEED MEMORIES

The Need: Why the high speeds memories? Is the main memory not a high-speed memory? The answer to second question is definitely "No", but why so, well for this we have to go to the fundamentals of semiconductor technology that is beyond the scope of the unit then if the memories are slower then how slow are they? It has been found that the access of main memories is slower than the speed of the processor. Since each instruction require several memory accesses therefore, faster memories will be of tremendous help in increasing the overall throughput of a computer.

There are four possible ways to increase the memory speed.

a) Decrease the memory access time, use a faster but expensive technology for main memory probably it will be feasible after few years.

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. This is termed as memory interleaving.

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

d) Use associative addressing in place of random access.

In this section we will only discuss about one of the most popular technology the cache memory. You can refer to further reading for most details on other terms mentioned in this section.

Cache Memory

These are small fast memories placed between the processor and the main memory. Caches although are fast yet are very extensive memories and are used in only small sizes. For example caches of sizes 64 K, 128K, 256 KB, etc. are normally used in typical PC-486 and Pentium based PCs while they can have 1 to 64 MB RAMs or even more. Thus, small cache memories are intended to provide fast speed of memory retrieval without sacrificing the size of memory (because of main memory size). If we have such a small size of fast memory how could it be advantageous in increasing the overall speed of memory reference? The answer lies in the principles of locality, which says that if a particular memory location is accessed at a time then it is highly likely that its near by locations will be accessed in the near future. Cache contains a copy of certain portions of main memory. The memory read or writes operation is first checked with cache and if the desired location data is available in cache then used by the CPU directly. Otherwise, a block of words is read from main memory to cache and CPU uses the word from cache. Since cache has limited space, so for this incoming block a portion called a slot need to be vacated in Cache. The contents of this vacating block is written back to the main memory at the position it belongs to. The reason of bringing a block of words to cache is once again locality of reference. We expect that next few addresses will be close to this address and, therefore, the block of word is transferred from main memory to cache. Thus, for the word, which is not in cache, access time is slightly more than the access time for main memory without cache. But, because of locality of references, cache performs better. For example, if memory read cycle takes 100 nos and a cache read cycle takes 20 nos, then for four continuous references (first one brings the main memory content to cache and next three from cache).

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

Thus, the closer are the reference, better is the performance of cache and that is why structured code is considered to be a good programming practice, since it provides maximum possible locality.

Check Your Progress 5

1. State True or False.

(a) High speed memories are needed to bridge the gap of speed between I/O device and memory

True □ False □

(b) Cache memory increases load on main memory.

True □ False □

(c) The principle of locality says that all the references to data have to be to the same memory location.

True □ False □

1.8 SUMMARY

This completes our discussion on the introductory concepts of computers. Some very elementary terms have not been defined here, you have to study these of your own. The von Neumann architecture discussed in the unit is not the only architecture but many new architectures have come up which you will find in later courses.

In addition, we have taken a complete view of the memory system of computer system along with the various technologies. The unit has outline the importance of memory system, the memory hierarchy, the main memory, the secondary memories and its technologies and the high speed memories. We have also discussed the key characteristics of these memories and the technologies which are used for constructing these memories. here are several other concepts such as virtual memory and for more details on the memory system a student can go through the further readings.

A course in an area of computer must be supplemented by further readings to keep your knowledge, upto date, as the computer world is changing with leaps and bounds. In addition to further readings the students are advised to study several Indian Journals oh computers to enhance his knowledge.

1.9 MODEL ANSWERS

Check Your Progress 1

1. True 2. False 3. False 4. False 5. True 6. True

Check Your Progress 2

Ans. 1. A machine, which can be used for variety of applications and is not modelled only for specific applications. Von Neumann machines are general purpose machines since they can be programmed for any general application, while a microprocessor based control systems are not general purpose machines as they are specifically modelled as control systems.

Ans. 2. (i) Microprocessor is a complete processor constructed on a single chip using VLSI technology. Intel and Motorola are two popular Microprocessors.

(ii) A very small, battery operated microcomputer, which uses liquid crystal display technology and very handy to carry.

(iii) The upper end of the state-of-art mainframe machines. Very powerful computational capabilities, which are mainly suitable for very large processing requirements such as weather forecasting.

Check Your Progress 3

1. (a) False

(b) True

(c) True

2.

(a) Volatile memory	**Non-volatile memory**
(i) Power is required all the time	(i) Power not required once data is stored
(ii) Temporary storage	(ii) Permanent storage
(iii) Normally faster than non-volatile types	(iii) Slower than volatile memories

(b) Static memory	**Dynamic memory**
(i) Looses its signal 1 becomes 0	(i) Does not loose signal strength

(ii)	Periodic refreshing of memory is needed	(ii)	No refreshing
(iii)	Expensive	(ii)	Cheap
(iv)	Used as Cache	(iv)	Used as main memory

Check Your Progress 4

1.

RAM	ROM
♦ Read-write memory	Only can be read
♦ Semi-Conductor memories	Semi-Conductor memories
♦ Volatile	Non-Volatile
♦ Can be used by user programs and system programs	cannot be used for user programs

2. They are better than ROMs and PROMs for rewriting purposes, however, can be used in similar ways.

3. Head is conducting coil in a disk which performs the job of reading or writing or data on a disk using magnetic effects of electricity.

4. (i) - (c), (ii) - (a), (iii) - (b)

Check Your Progress 5

1. (a) False

(b) False

(c) False

1.10 FURTHER READINGS

1. Stallings, William, Computer Organisation and Architecture, Third edition, Maxwell Macmillan International Editions.
2. Mano, M. Morris, Computer System Architecture and Organisation, Second Edition, McGraw-Hill International Editions, 1988.