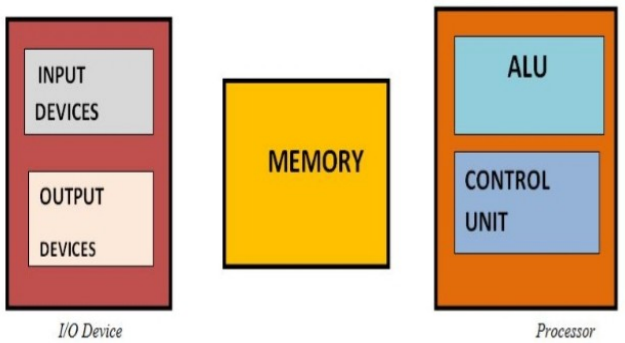
COMPUTER ORGANIZATION

- Introduction to the computer organisation
- Von Neuman architectures
- Computer components
- Interconnection structures
- Bus interconnection

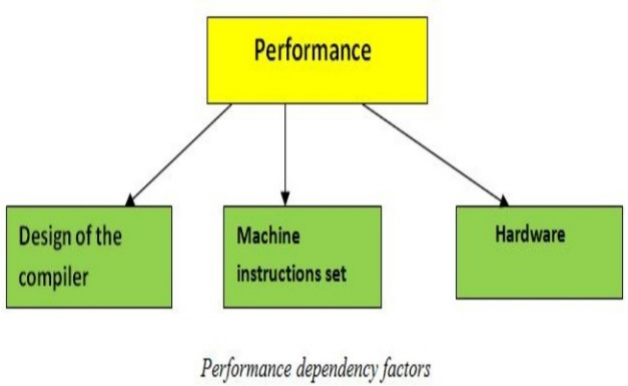
INTRODUCTION TO THE COMPUTER ORGANISATION

Computer organization and architecture mainly focuses on various parts of the computer in order to reduce the execution time of the program, improve the performance of each part. Generally, we tend to think computer organization and computer architecture as same but there is slight difference.

- **DIFFERENCE:**
- **Computer Organization** is study of the system from software point of view and gives overall description of the system and working principles without going into much detail. In other words, it is mainly about the programmer’s or user point of view.
- **Computer Architecture** is study of the system from hardware point of view and emphasis on how the system is implemented. Basically,



Basic Functional Blocks in a system.



Performance dependency factors

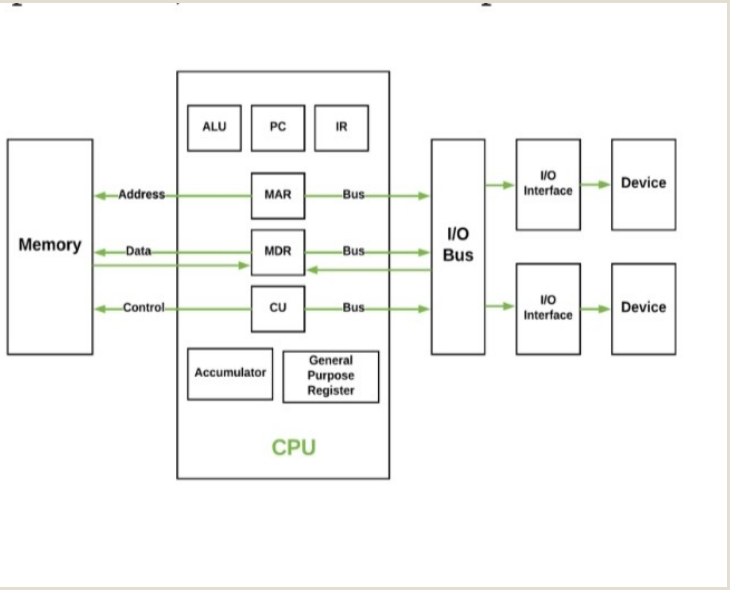
FUNCTIONAL BLOCKS

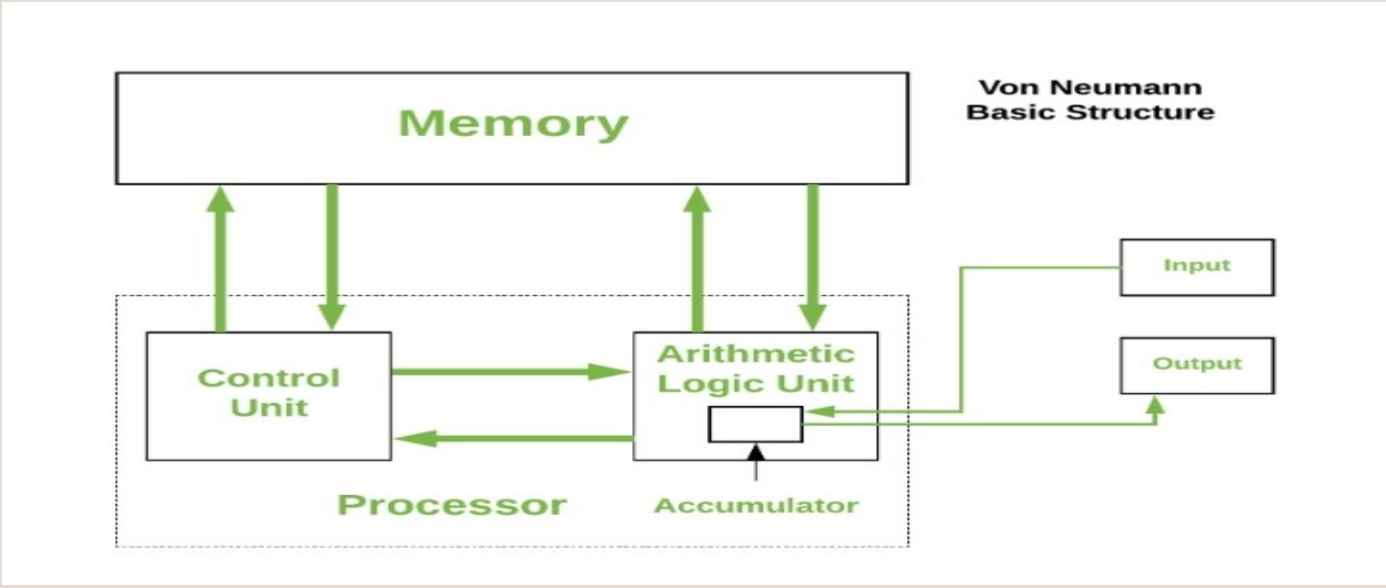
A computer consists of various functional blocks- Input, Output, Memory, arithmetic and logical unit, control units.

- **Input device** provides information in the form of program to the computer and stores it in the **memory**.
- Further the information is fetched from **memory** to the **processer**.
- Inside the **processor**, it is processed by the **ALU**.
- The processed output further passes to **output devices.**
- All these activities are controlled by **control unit**.

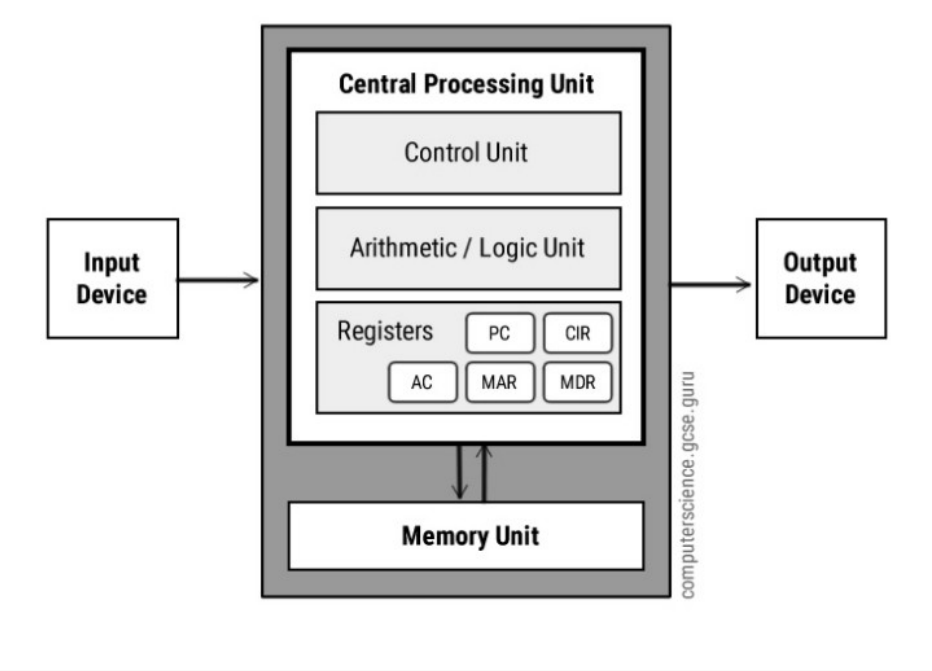
VON NEUMAN ARCHITECTURE

- V**on Neumann Architecture** also known as the *Von Neumann model*, the computer consisted of a CPU, memory and I/O devices. The program is stored in the memory. The CPU fetches an instruction from the memory at a time and executes it.
- Thus, the instructions are executed sequentially which is a slow process. Neumann m/c are called control flow computer because instruction are executed sequentially as controlled by a program counter. To increase the speed, parallel processing of computer have been developed in which serial CPU's are connected in parallel to solve a problem. Even in parallel computers, the basic building blocks are Neumann processors.
- The von Neumann architecture is a design model for a stored-program digital computer that uses a processing unit and a single separate storage structure to hold both instructions and data. It is named after mathematician and early computer scientist John von Neumann. Such a computer implements a universal Turing machine, and the common "referential model" of specifying sequential architectures, in contrast with parallel architectures.



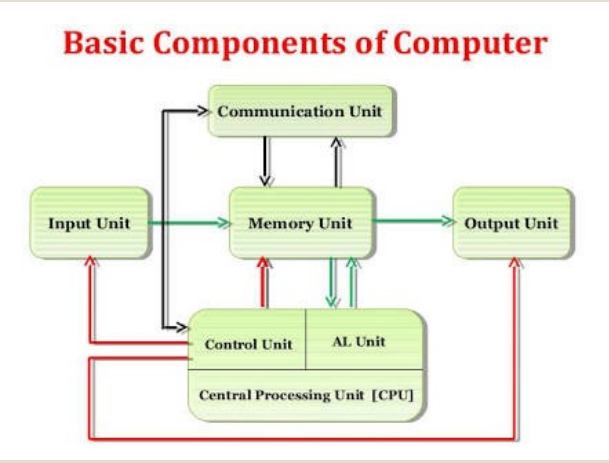


Von Neumann architecture is based on the stored-program computer concept, where instruction data and program data are stored in the same memory. This design is still used in most computers produced today.

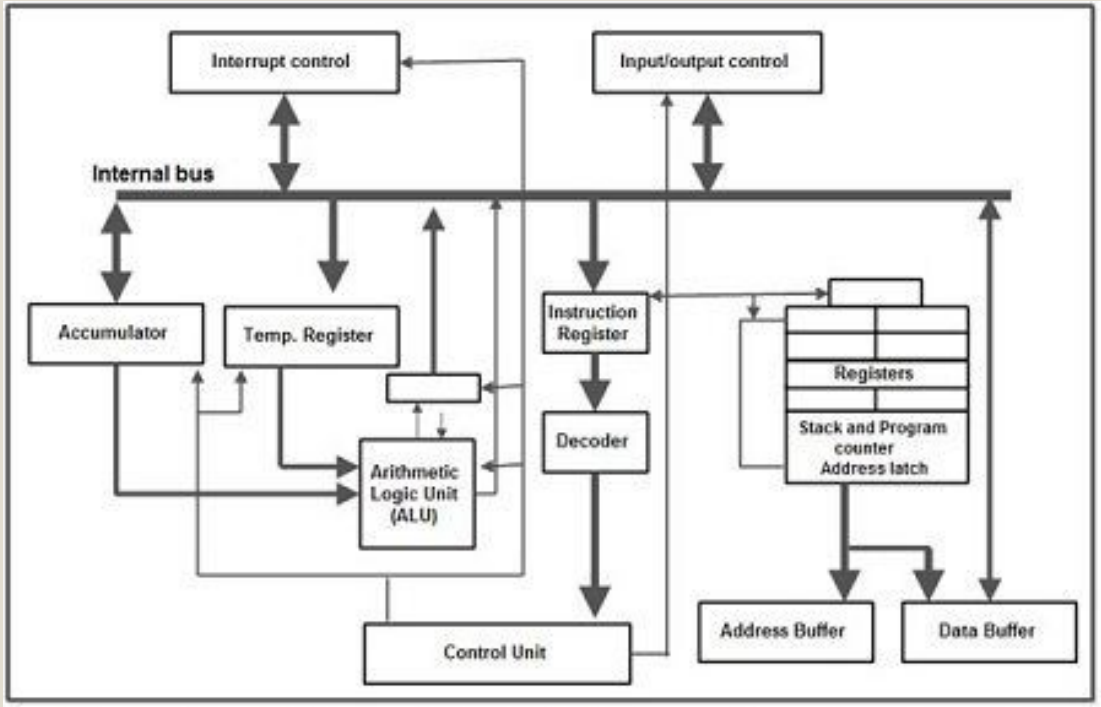


COMPUTER COMPONENTS

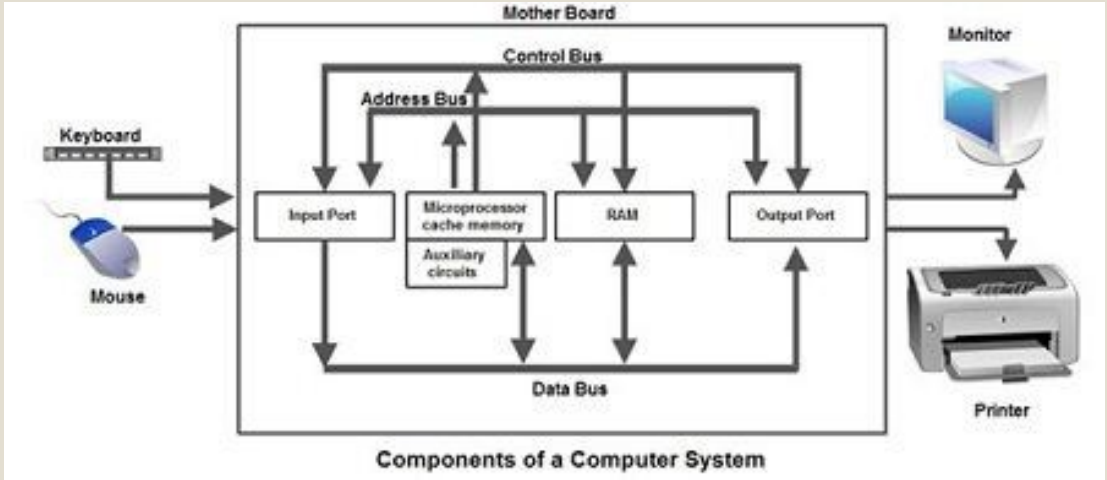
Computers internal architectural design comes in different types and sizes, but the basic structure remains same of all computer systems.



The term '**computer hardware**' or 'computer parts' is used to describe *computer components* that can be seen and touched. The major components of general-purpose computer system are Input Unit, main/internal Memory or Storage Unit, Output Unit, Central Processing unit. The CPU is further includes Arithmetic logic unit (ALU) and control unit (CU). All the units also referred to as "**The functional units**". Devices that are not integral part of CPU referred to as peripherals.



Simplified block diagram of one of the first-generation microprocessors



Components of a Computer System

The below section describe briefly all the computer components in a computer system

Input Unit

Input unit is used for transfers' raw Data and controlsignals into the information processing system by the user before processing and computation. All the input unit devices provide the instructions and data are transformed into binary codes that is the primary memory acceptable format.

Example of Input unit devices: keyboard, mouse, scanner, joystick, MICR, Punched cards, Punched paper tape, Magnetic tape etc.

Memory or Storage Unit

Memory or Storage unit is used for storing Data during before and after processing. The capacity of storage is expressed in terms of Bytes.

The two terms Memory or Storage unit are used interchangeably, so it is important to understand what is the difference between memory and storage

Memory

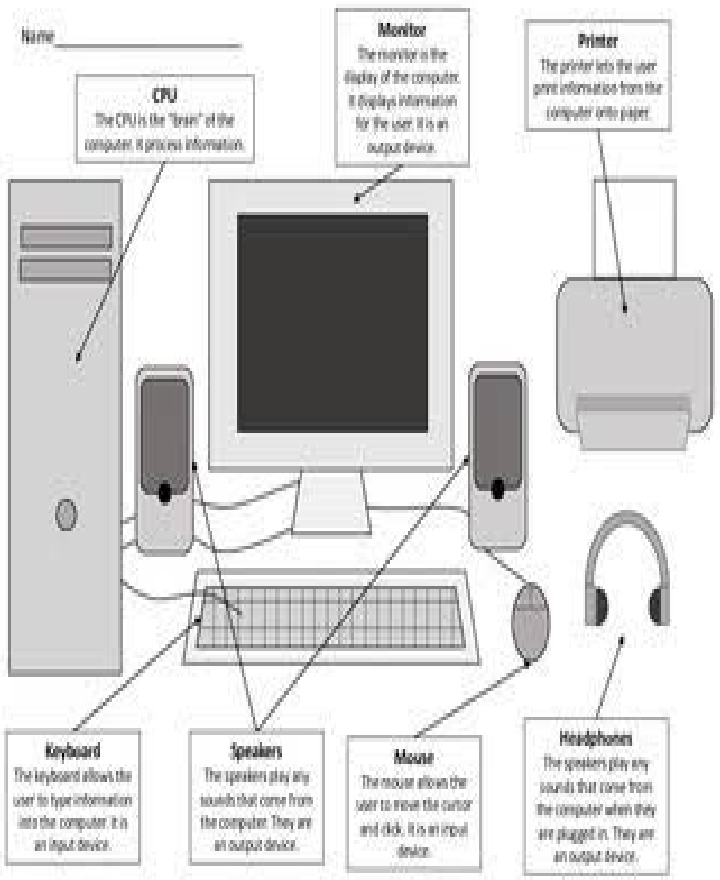
This unit retains temporarily results till further processing, For example, Random Access Memory (RAM).This memory is volatile, which means data is disappears when the power is lost.

Storage

The storage or "secondary storage" is used for retain digital data after processing for permanently. For example hard drive. The Storage is non-volatile in nature. CPU does not access directly to secondary storage memories, instead they accessed via input-output unit. The contents of secondary storage memories are first transferred to the main memory (RAM) and then CPU access it

Output Unit

Output Unit receives information from the CPU and then delivers it the external storage or device in the soft or hard processed form. The devices which are used to display output to the user are called output devices. The Monitor or printer is common output device.



Central Processing Unit

The main chip in a computer is the microprocessor chip, which is also known as the CPU (central processing unit). The CPU is mounted on a printed circuit board called the main board or mother board. This chip is considered to be the controlling chip of a computer system since it controls the activities of other chips as well as outside devices connected to the computer, such as monitor and printer. In addition, it can also perform logical and computational tasks. Microprocessors work on a parallel system. Figure shows a typical structure of one of the first-generation microprocessors. The recent ones possess greater complexity, although the basic design concept has not changed much.

Arithmetic logic unit (ALU)

Arithmetic Logical Unit is used for processing data after inputting data is stored into primary unit. The major operations of Arithmetic Logical Unit are addition, subtraction, multiplication, division, logic and comparison.

Control unit (CU)

It is like a supervisor, that checks ordaining operations or check sequence in which instructions are executed

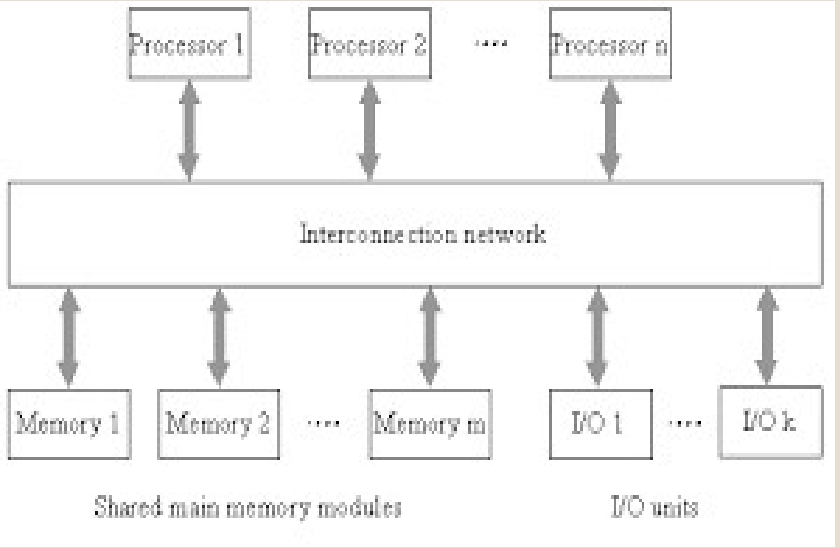
INTERCONNECTION STRUCTURES

A computer consists of a set of components or modules of three basic types (processor, memory, I/O) that communicate with each other. In effect, a computer is a network of basic modules. Thus, there must be paths for connecting the modules. The collection of paths connecting the various modules is called the interconnection structure. The design of this structure will depend on the exchanges that must be made among modules.
Figure below suggests the types of exchanges that are needed by indicating the major forms of input and output for each module type.

.

• **Memory**: Typically, a memory module will consist of N words of equal length. Each word is assigned a unique numerical address (0, 1, . . . , N – 1). A word of data can be read from or written into the memory.The nature of the operation is indicated by read and write control signals. The location for the operation is specified by an address.

• **I/O module**: From an internal (to the computer system) point of view, I/O is functionally similar to memory. There are two operations, read and write. Further, an I/O module may control more than one external device. We can refer to each of the interfaces to an external device as a port and give each a unique address (e.g., 0, 1,…, M – 1). In addition, there are external data paths for the input and output of data with an external device. Finally, an I/O module may be able to send interrupt



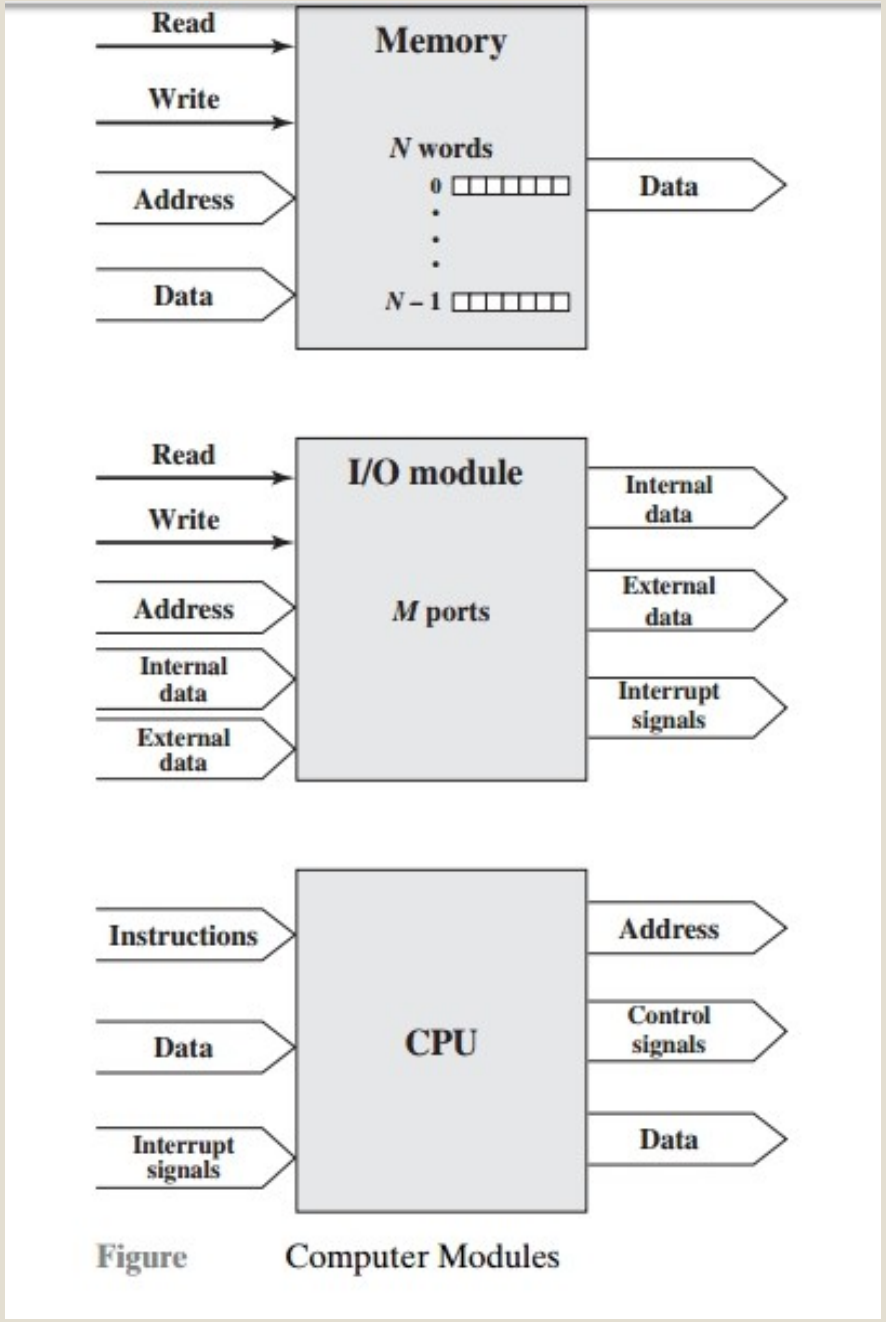


Figure Computer Modules

• **Processor**: The processor reads in instructions and data, writes out data after processing, and uses control signals to control the overall operation of the system. It also receives interrupt signals. The preceding list defines the data to be exchanged. The interconnection structure must support the following types of transfers:

• Memory to processor: The processor reads an instruction or a unit of data
from memory.

• Processor to memory: The processor writes a unit of data to memory.

• I/O to processor:The processor reads data from an I/O device via an I/O module.

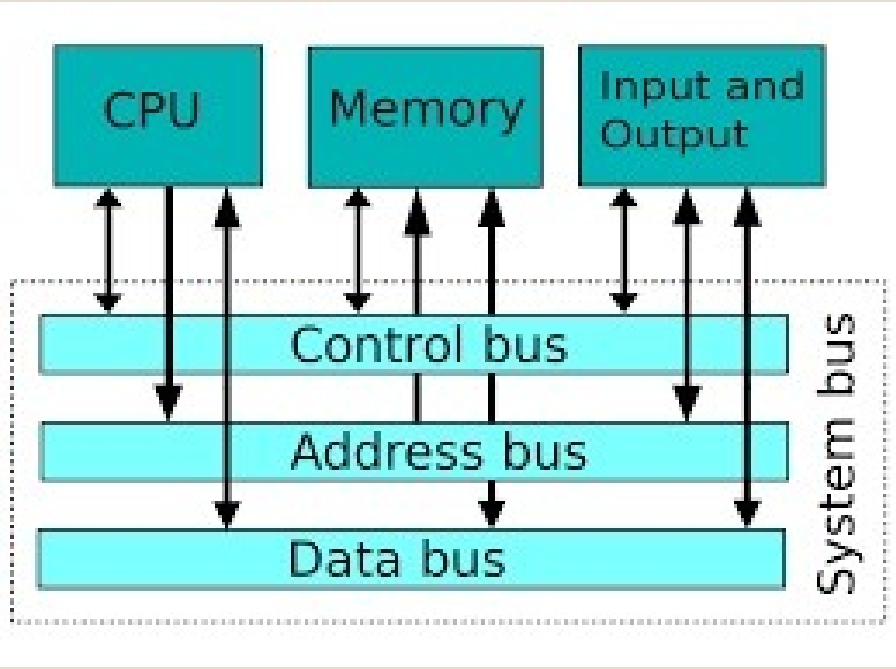
• Processor to I/O: The processor sends data to the I/O device.

• I/O to or from memory: For these two cases, an I/O module is allowed to exchange data directly with memory, without going through the processor, using direct memory access (DMA).

Though a number of interconnection structures have been tried. By far the most common is the bus and various multiple-bus structures

BUS INTERCONNECTION

- The system bus is a pathway composed of cables and connectors used to carry data between a computer microprocessor and the main memory. The bus provides a communication path for the data and control signals moving between the major components of the computer system. The system bus works by combining the functions of the three main buses: namely, the data, address and control buses. Each of the three buses has its separate characteristics and responsibilities.

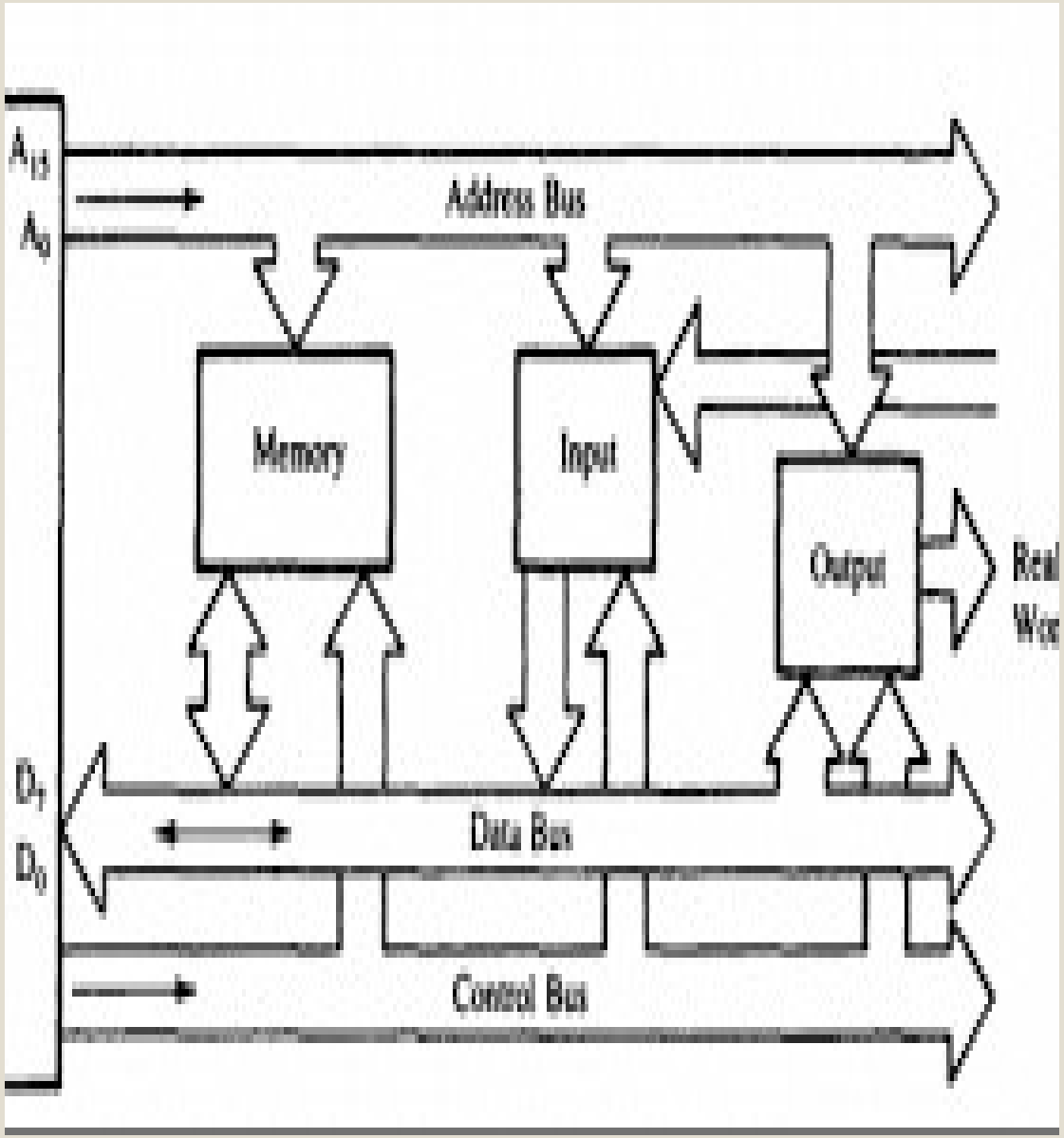


•The system bus connects the CPU with the main memory and, in some systems, with the level 2 (L2) cache. Other buses, such as the IO buses, branch off from the system bus to provide a communication channel between the CPU and the other peripherals.

The system bus combines the functions of the three main buses, which are as follows:

The control bus carries the control, timing and coordination signals to manage the various functions across the system.
•The address bus is used to specify memory locations for the data being transferred.
•The data bus, which is a bidirectional path, carries the actual data between the processor, the memory and the peripherals.
The design of the system bus varies from system to system and can be specific to a particular computer design or may be based on an industry standard. One advantage of using the industry

System bus characteristics are dependent on the needs of the processor, the speed, and the word length of the data and instructions. The size of a bus, also known as its width, determines how much data can be transferred at a time and indicates the number of available wires. A 32-bit bus, for example, refers to 32 parallel wires or connectors that can simultaneously transmit 32 bits.



The design and dimensions of the system bus are based on the specific processor technology of the motherboard. This, in effect, affects the speed of the motherboard, with faster system buses requiring that the

INPUT-OUTPUT ORGANIZATION

UNIT 2
Input output organization:

- I/O interface
- interrupt driven I/O,
- Priority
- interrupt,
- DMA,Input output organization:
- I/O interface models of transfer,
- int
- I/O processor and serial communication,
- Synchronous data transfer ,
- Asynchronous data transfer,
- strobe control,
- handshaking,
- PCI, working mechanism of Peripherals:
- Keyboard,
- Mouse,
- Scanners ,
- Video Display,
- Touch Sreen panel etc.(features and principles)

Input - Output Interface

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 central processing unit.

I/O BUS and Interface Module It defines the typical link between the processor and several peripherals. The I/O Bus consists of data lines, address lines and control lines. The I/O bus from the processor is attached to all peripherals interface. To communicate with a particular device, the processor places a device address on address lines. Each Interface decodes the address and control received from the I/O bus, interprets them for peripherals and provides signals for the peripheral controller. It is also synchronizes the data flow and supervises the transfer between peripheral and processor. Each peripheral has its own controller. For example, the printer controller controls the paper motion, the print timing The control lines are referred as I/O command. The commands are as following: Control command- A control command is issued to activate the peripheral and to inform it what to do. Status command- A status command is used to test various status conditions in the interface and the peripheral. Data Output command- A data output command causes the interface to respond by transferring data from the bus into one of its registers. Data Input command- The data input command is the opposite of the data output. In this case the interface receives on item of data from the peripheral and places it in its buffer register. I/O Versus Memory Bus

Interrupt driven I/O

Interrupt driven I/O is an alternative scheme dealing with I/O. Interrupt I/O is a way of controlling input/output activity whereby a peripheral or terminal that needs to make or receive a data transfer sends a signal. This will cause a program interrupt to be set. At a time appropriate to the priority level of the I/O interrupt. Relative to the total interrupt system, the processors enter an interrupt service routine. The function of the routine will depend upon the system of interrupt levels and priorities that is implemented in the processor. The interrupt technique requires more complex hardware and software, but makes far more efficient use of the computer's time and capacities.

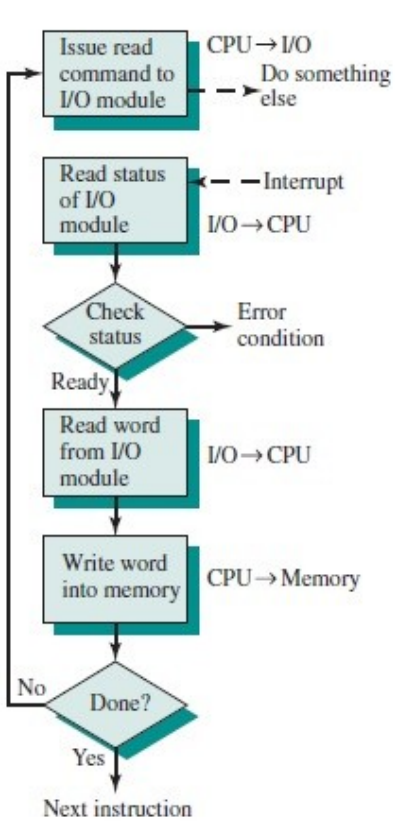
- For **input**, the device interrupts the CPU when new data has arrived and is ready to be retrieved by the system processor. The actual actions to perform depend on whether the device uses I/O ports or memory mapping.

  For **output**, the device delivers an interrupt either when it is ready to accept new data or to acknowledge a successful data transfer. Memory-mapped and DMA-capable devices usually generate interrupts to tell the system they are done with the buffer

- ***Advantages & Disadvantages of Interrupt Drive I/O***
- **Advantages**
- \- fast
- \- efficient

- **Disadvantages**
- - can be tricky to write if using a low level language
- - can be tough to get various pieces to work well together
- - usually done by the hardware manuf

Interrupt Processing



PRIORITY INTERRUPT

- In a typical application, a number of I/O devices are attached to computer, with each device being able to originate an interrupt request, so to provide services to device which initiate interrupt request, the task of interrupt system is to identify the source(device) of interrupt and then provide services to them.

- But, in most cases there is a possibility that several sources will request service simultaneously.So, in this case, the interrupt system must also need to decide which device to service first.But, these simple interrupt system are not able for that, so, another system known as Priority interrupt system is provided.

- Priority Interrupt are systems, that establishes a Priority over the various sources(interrupt devices) to determine which condition is to be serviced first when two or more requests arrive simultaneously.This system may also determine which condition are permitted to interrupt to the computer while another interrupt is being serviced.

Usually, in Priority Systems, higher-priority interrupt levels are served first, as if they delayed or interrupted, could have serious consequences.And the devices with high-speed transfer such as magnetic disks are given high-priority, and slow devices such as keyboards receives low-priority.

Establishing Priority of Simultaneous Interrupt:

The priority of simultaneous interrupts can be established either by software method or hardware.

The software method which gives priority to simultaneous interrupt is:
Polling
And the hardware method which gives priority to simultaneous interrupt is:
Daisy-Chaining Priority
Now, we will explore to each one of them one by one.

1. **Polling:**

Polling is the software method of establishing priority of simultaneous interrupt.In this method, when the processor detects an interrupt, it branches to an interrupt service routine whose job is to pull each I/O module to determine which module caused the interrupt.

The poll could be in the form of separate command line(e.g., Test I/O).In this case, the processor raises the Test I/O and places the address of particular I/O module on the address line.If it has interrupt that is, if interrupt is identified in it.

And, it is the order in which they are tested i.e., the order in which they appear on address line(Service Routine) determine the priority of each interrupt.As while testing, highest priority source(devices) are tested first then lower-priority devices.

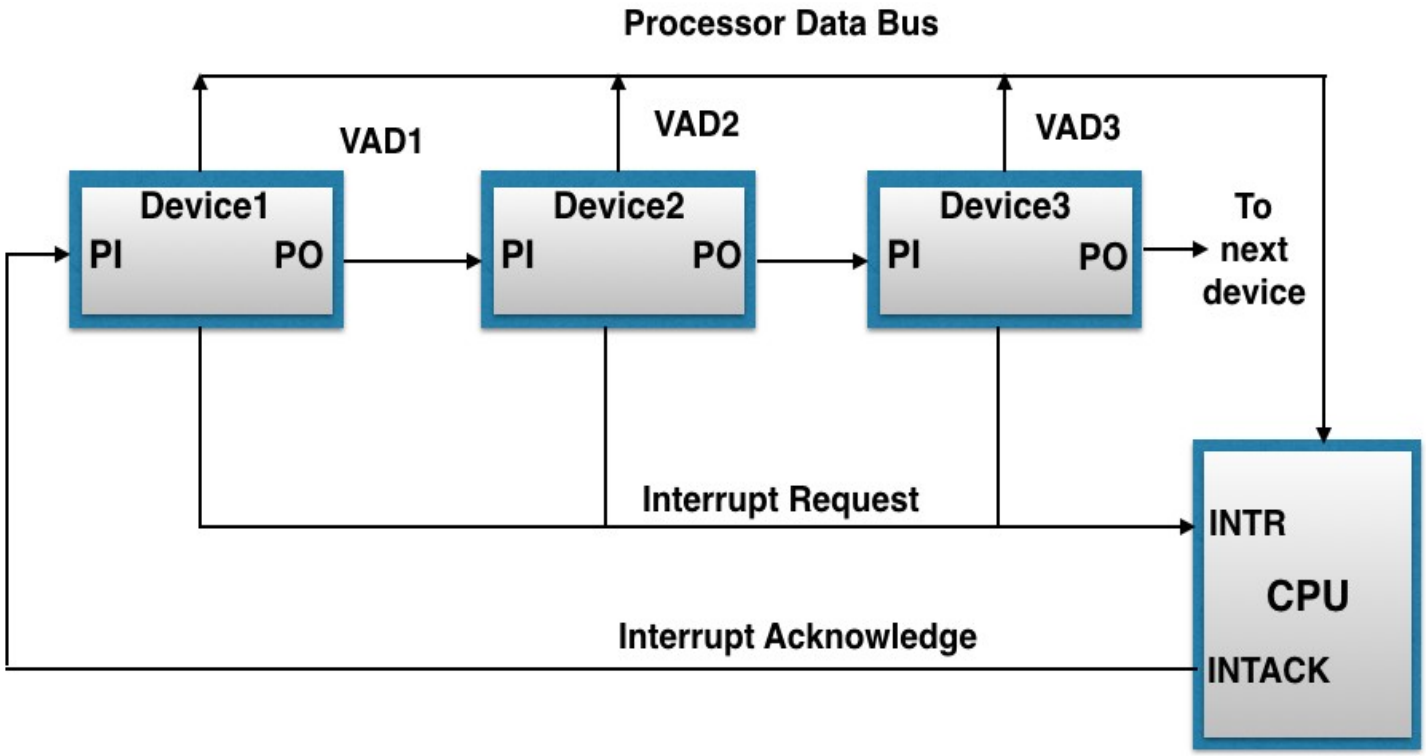
This is very simple method of establishing priority on simultaneous interrupt.But the disadvantage opolling is that it is very time consuming.

2. **Daisy-Chaining Priority:**

The Daisy–Chaining method of establishing priority on interrupt sources uses the hardware i.e., it is the hardware means of establishing priority.

In this method, all the device, whether they are interrupt sources or not, connected in a serial manner.Means the device with highest priority is placed in the first position, which is followed by lowest priority device.And all device share a common interrupt request line, and the interrupt acknowledge line is daisy chained through the modules.

The figure shown below, this method of connection with three devices and the CPU.



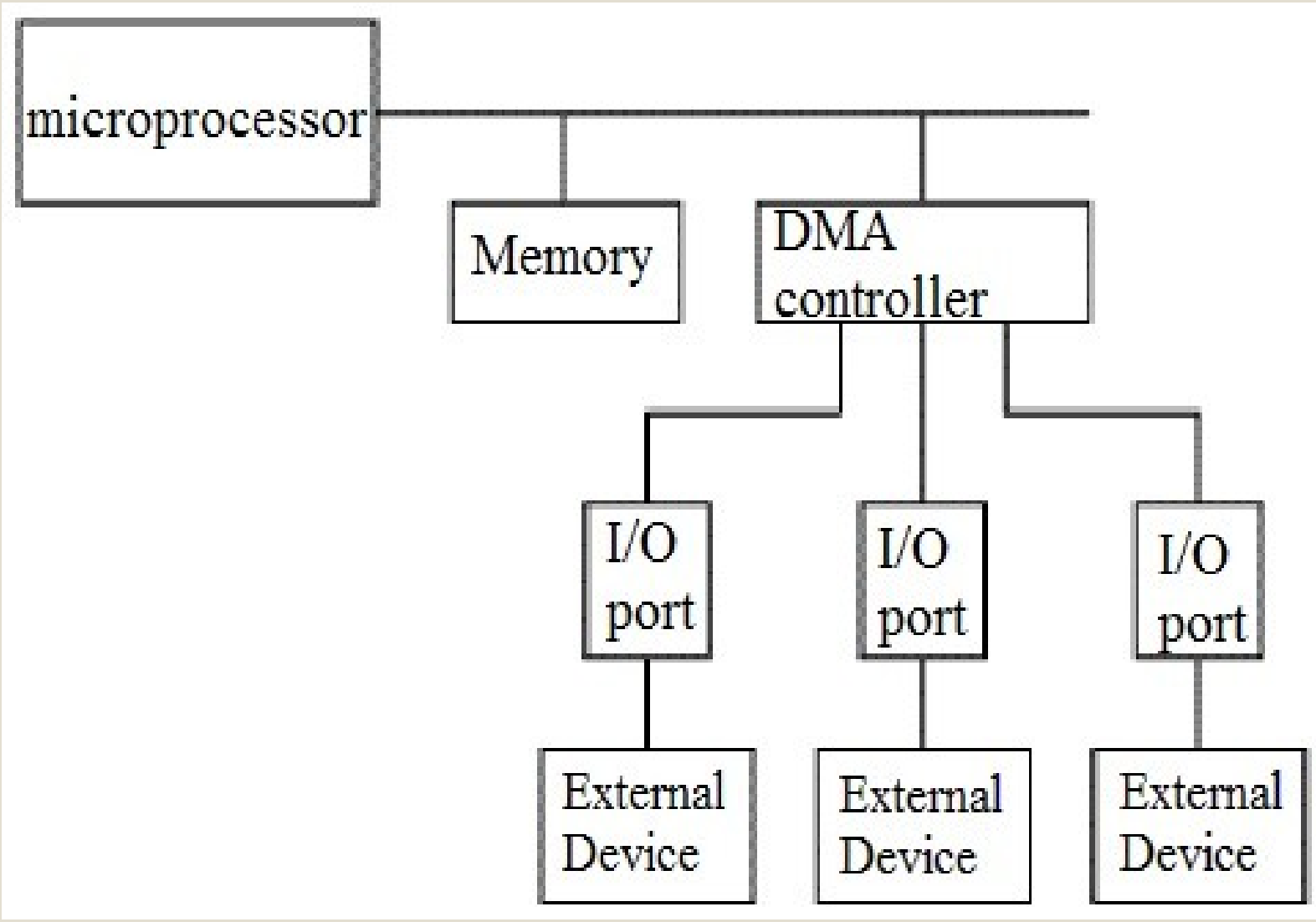
It works as follows:

When any device raise an interrupt, the interrupt request line goes activated, the processor when sense it, it sends out an interrupt acknowledge which is first received by device1.If device1 does not need service, i.e., processor checks, whether the device has pending interrupt or initiate interrupt request, if the result is no, then the signal is passed to device2 by placing 1 in the PO(Priority Out) of device1.And if device need service then service is given to them by placing first 0 in the PO of 1 device1, which indicate the next-lower-priority device that acknowledge signal has been blocked.And device that have processor responds by inserting its own interrupt vector address(VAD) into the data bus for the CPU to use during interrupt cycle.

In this way, it gave services to interrupt source according to their priority.And thus, we can say that, it is the order of device in chain that determine the priority of interrupt sources.

DMA

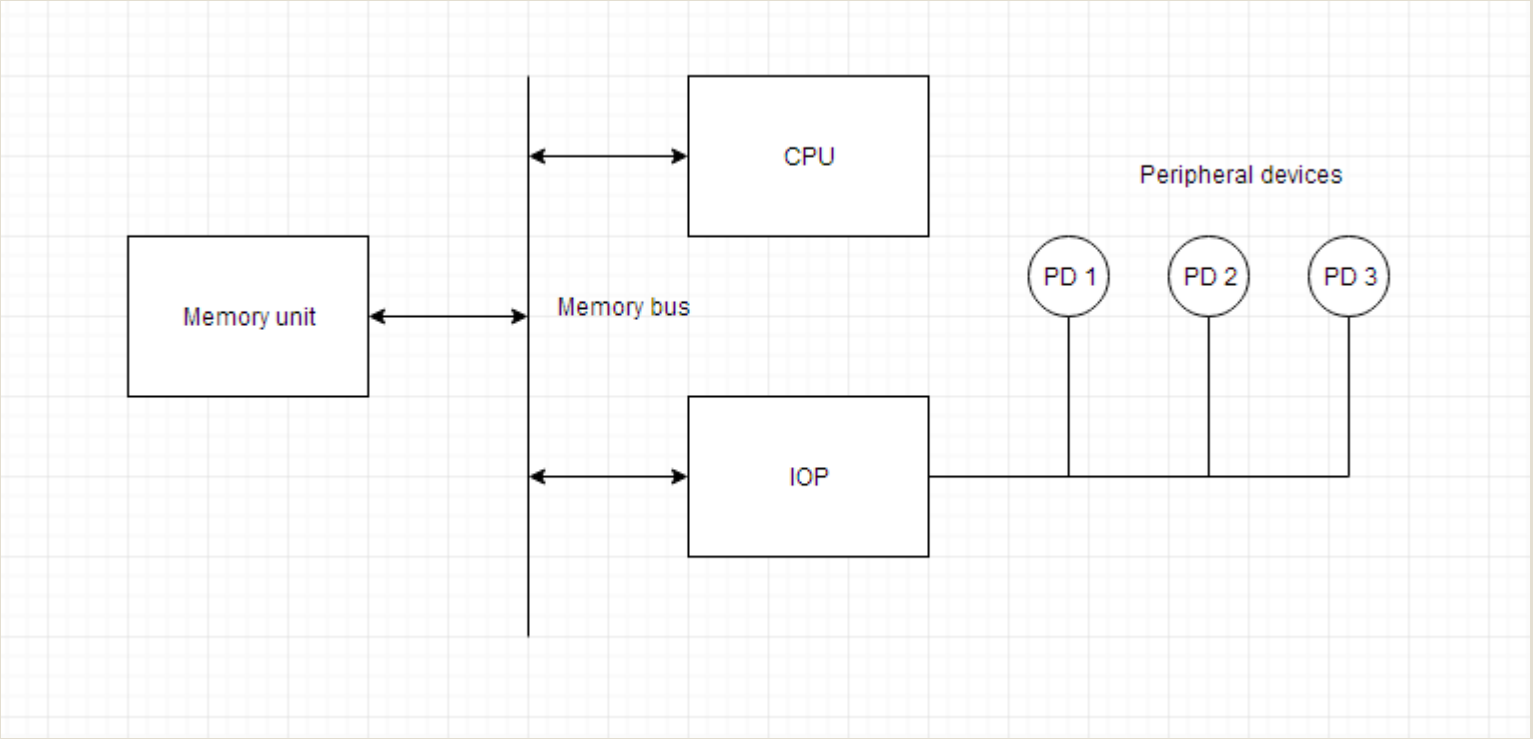
- Stands for "Direct Memory Access." DMA is a method of transferring data from the computer's RAM to another part of the computer without processing it using the CPU. While most data that is input or output from your computer is processed by the CPU, some data does not require processing, or can be processed by another device. In these situations, DMA can save processing time and is a more efficient way to move data from the computer's memory to other devices.
- For example, a sound card may need to access data stored in the computer's RAM, but since it can process the data itself, it may use DMA to bypass the CPU. Video cards that support DMA can also access the system memory and process graphics without needing the CPU. Ultra DMA hard drives use DMA to transfer data faster than previous hard drives that required the data to first be run through the CPU.


microprocessor
Memory
DMA controller
I/O port
I/O port
I/O port
External Device
External Device
External Device

I/O Processor and serial communication

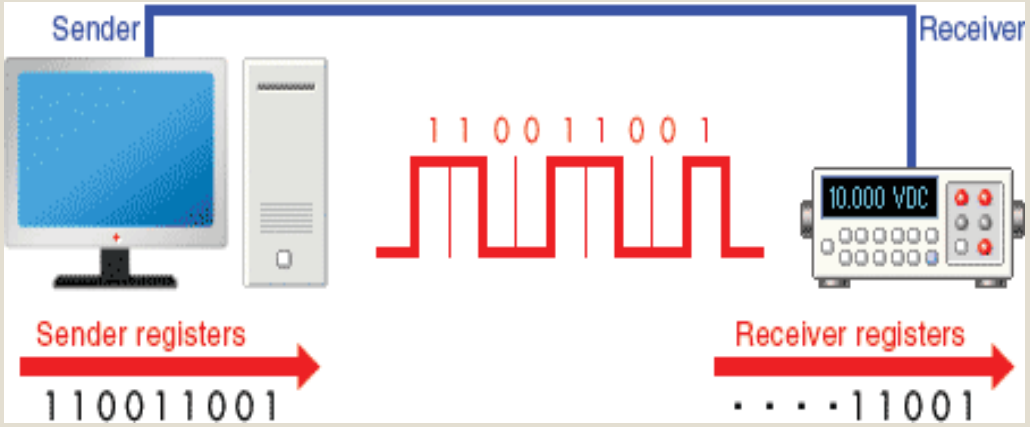
- An input-output processor (IOP) is a processor with direct memory access capability. In this, the computer system is divided into a memory unit and number of processors.
- Each IOP controls and manage the input-output tasks. The IOP is similar to CPU except that it handles only the details of I/O processing. The IOP can fetch and execute its own instructions. These IOP instructions are designed to manage I/O transfers only.

Block Diagram Of I/O Processor



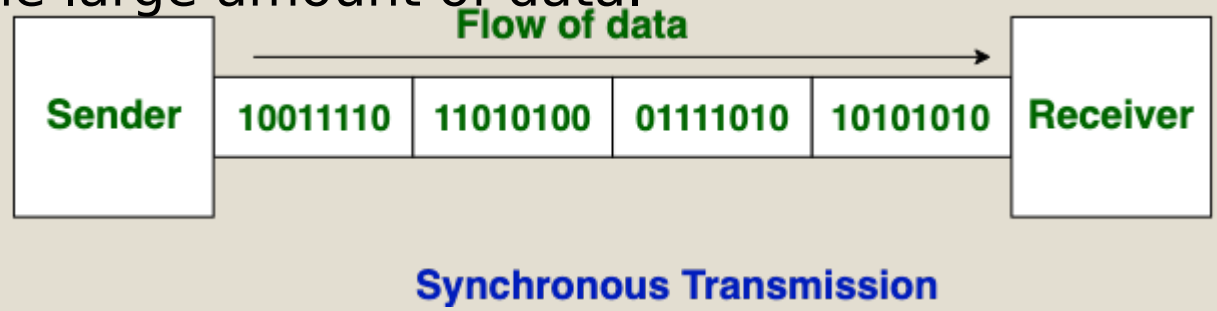
Serial Communication

- Serial communication is a communication method that uses one or two transmission lines to send and receive data, and that data is continuously sent and received one bit at a time.Since it allows for connections with few signal wires, one of its merits is its ability to hold down on wiring material and relaying equipment costs.



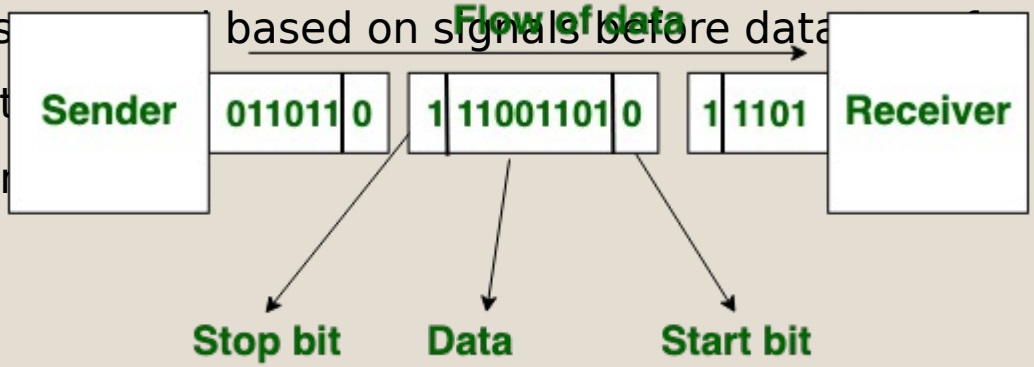
Synchronous data transfer

- In Synchronous Transmission, data is sent in form of blocks or frames. This transmission is the full duplex type. Between sender and receiver the synchronization is compulsory. In Synchronous transmission, There is no gap present between data. It is more efficient and more reliable than asynchronous transmission to transfer the large amount of data.



Asynchronous data transfer

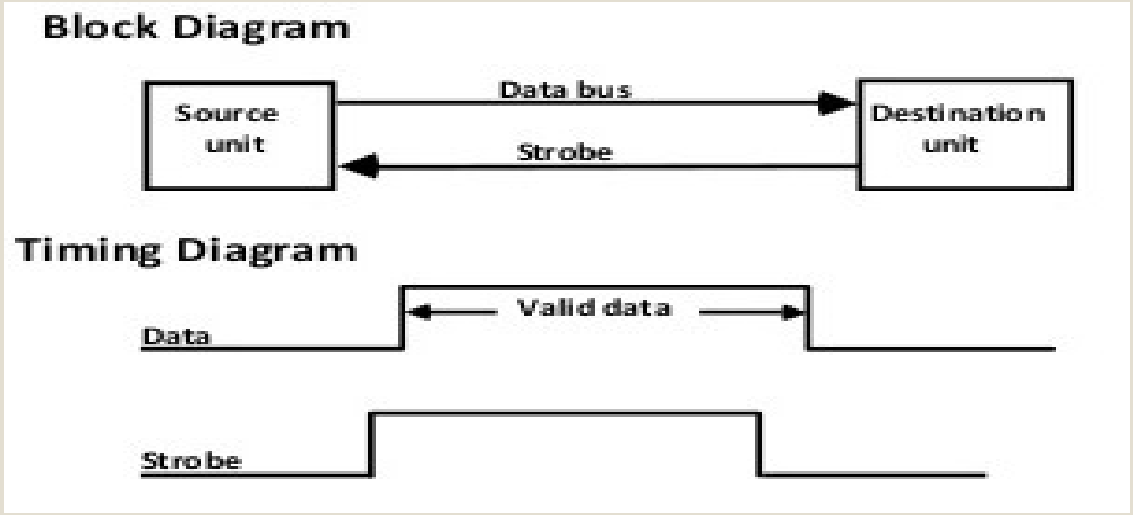
- This Scheme is used when speed of I/O devices do not match with microprocessor, and
- timing characteristics of I/O devices is not predictable. In this method, process initiates the
- device and check its status. As a result, CPU has to wait till I/O device is ready to transfer
- data. When device is ready CPU issues instruction for I/O transfer. In this method two types
- of techniques based on signals before data
- i.Strobe Cont
- ii. Handshaki



Asynchronous Transmission

Strobe Control

- The strobe control method of Asynchronous data transfer employs a single control line to
- time each transfer. The strobe may be activated by either the source or the destination unit.
- Data Transfer Initiated by Source Unit:



Source initiated strobe for data transfer

In the block diagram fig. 1 the data bus carries the binary information from source to
destination unit. Typically, the bus has multiple lines to transfer an entire byte or word. The
strobe is a single line that informs the destination unit when a valid data word is available.
The timing diagram fig. 2 the source unit first places the data on the data
bus. The information on the data bus
Data Transfer Initiated by Destination Unit:
In this method, the destination unit activates the strobe pulse, to informing the source to
provide the data. The source will respond by placing the requested binary information on the
data bus.
The data must be valid and remain in the bus long enough for the destination
unit to accept it. When accepted the destination unit then disables the strobe and the source
unit removes the data from the bus.

Handshaking

- The handshaking method solves the problem of strobe method by introducing a second
- control signal that provides a reply to the unit that initiates the transfer.
- Principle of Handshaking:
- The basic principle of the two-wire handshaking method of data transfer is as follow:
- One control line is in the same direction as the data flows in the bus from the source to
- destination. It is used by source unit to inform the destination unit whether there a valid data
- in the bus. The other control line is in the other direction from the destination to the source. It
- is used by the destination unit to inform the source whether it can accept the data. The
- sequence of control during the transfer depends on the unit that initiates the transfer.
- Source Initiated Transfer using Handshaking:
- The sequence of events shows four possible states that the system can be at any given time.

PCI

- Stands for "Peripheral Component Interconnect." PCI is a hardware bus used
  internal components to a desktop computer. For example, a PCI card can be i
  slot on a motherboard, providing additional I/O ports on the back of a comput
- The PCI architecture, also known as "conventional PCI," was designed by Inte
  in 1992. Many desktop PCs from the early 1990s to the mid 2000s had room
  cards. Each card required an open slot on the motherboard and a removable
  of the system unit. Adding PCI cards was an easy way to upgrade a computer
  add a better video card, faster wired or wireless networking, or add new port
- The original 32-bit, 33 MHz PCI standard supported data transfer rates of 133
  second. An upgraded 64-bit, 66 MHz standard was created a few years later a
  much faster data transfer rates up to 533 MHz. In 1998, IBM, HP, and Compa
  (or "PCI eXtended"), which was backwards compatible with PCI. The 133 MHz
  supported data transfer rates up to 1064 MHz.
- Both PCI and PCI-X were superseded by PCI Express, which was introduced in

WORKING MECHANISM OF PERIPHERALS

Working Principle of a Keyboard:-

Inside the keyboard, there are metallic

- plate, circuit board and processor, which are responsible for transferring
- information from the keyboard to the computer. Depending upon the
- working principle, there are two main types of keys, namely, capacitive and
- hard-contact. Let's discuss in brief about the functioning of capacitive and
- hard contact key.When a capacitive key is pressed, the metal plunger applies
- a gentle pressure to the circuit board. The pressure is identified by the
- computer and the circuit flow is initiated, resulting in the transfer of
- information from the circuit to the currently installed software.
- The key identifying to computer is identified using a keyboard driver and
- finding the preferred key called
- source code.

WORKING MECHANISM OF MOUSE

- • Working of a mouse---
- > With most of the system you will find mechanical mouse.The primary mechanical part of a
- mouse is a ball on the bottom of the mouse. There are these little wheels which turn/rotate
- when the ball moves against them. The wheels are monitored electronically. When they trun
- or rotate they transmit how much they have turned to the computer. Out of these three wheels

- the two wheels perpendicular to each other are used for tracking the motion on X-axis and Y-
- axis. The third one just balances the two.

- When the mouse is moved on a flat surface the roller ball moves in the locking ring. When the
- mouse is positioned on the desktop the actuators register the mouse balls movement in X-axis
- and Y-axis direction. The sensors attached to it generate a series of pulses representing
- movement on both axis. The pulse generated are in same ratio as the mouse movement i.e.
- More pulse mean more movement.
- Normally a mouse is used along with a mouse pad. Place the mouse pad on a flat surface and
- place the mouse on it. Move the mouse pad and the pointer moves in the direction of the
- movement of mouse.

Various terms related to the use of mouse

- > Click
- > Double click
- > Drag
- 1. When the left button of mouse is pressed and released quickly the
  say '
- clicking the mouse '.
- 2. The double clicking is used to initiate some action on the selected
  it selects

WORKING OF SCANNER

- A scanner is a device that is used for producing an exact digital image
- replica of a photo, text written in paper, or even an object. This digital
- image can be saved as a file to your computer and can be used to
- alter/enhance the image or apply it to the web. The most comonly
- used scanner is the flatbed scanner, in which you keep the object on
- top of the glass window. The scanned output will be obtained in your
- computer. The image and text are obtained exactly through the
- process of optical character recognition [OCR].
- Handheld scanners use the same basic technology as a flatbed scanner, but rely on the user to move them
- instead of a motorized belt. This type of scanner typically does not provide good image quality. However, it can
- be useful for quickly capturing text.
- • Drum scanners are used by the publishing industry to capture incredibly detailed images. They use a
- technology called a photomultiplier tube (PMT). In PMT, the document to be scanned is mounted on a glass
- cylinder. At the center of the cylinder is a sensor that splits light bounced from the document into three beams.

Each beam is sent through a color filter into a photomultiplier tube where the light is changed into an electrical

signal.

- The basic principle of a scanner is to analyze an image and process it in some way. Image and text capture
  - (optical character recognition or OCR) allow you to save information to a file on your computer.

WORKING OF VIDEO DISPLAY

- Video display device means an electronic device with an output surface that
- displays, or is capable of displaying, moving graphical images or a visual
- representation of image sequences or pictures, showing a number of quickly
- changing images on a screen in fast succession to create the illusion of motion,
- including, if applicable, a device that is an integral part of the display, in that it
- cannot be easily removed from the display by the consumer, that produces the
- moving image on the screen. A video display device may use, but is not limited to, a
- cathode ray tube (CRT), liquid crystal display (LCD), gas plasma, digital light
- processing, or other image projection technology.
- device means a printer or a unit capable of presenting images
- electronically on a screen, with a video display greater than four inches
- when measured diagonally, that are viewed by the user, and includes
- televisions, computer monitors, laptop computers, cathode ray tubes,
- plasma displays, liquid crystal displays, rear and front enclosed projection
- devices, and other similar displays that may be developed.

WORKING OF TOUCH SCRE PANNEL

- Different kinds of touchscreen work in different ways. Some can sense
- only one finger at a time and get extremely confused if you try to press
- in two places at once. Others can easily detect and distinguish more
- than one key press at once. These are some of the main
- technologies:
- • Resistive
- • Resistive touchscreens (currently the most popular technology) work a
- bit like "transparent keyboards" overlaid on top of the screen. There's
- a flexible upper layer of conducting polyester plastic bondedd to a
- rigid lower layer of conducting glass and separated by an insulating
- membrane. When you press on the screen, you force the polyester to
- touch the glass and complete a circuit—just like pressing the key on a
- keyboard. A chip inside the screen figures out the coordinates of the
- place you touched.

Capacitive

• These screens are made from multiple layers of glass. The inner layer conducts electricity and so does the outer layer, so effectively the screen behaves like two electrical conductors separated by an insulator—in other words, a capacitor.When you bring your finger up to the screen, you alter the
electrical field by a certain amount that varies according to where your hand
is. Capacitive screens can be touched in more than one place at once. Unlike most other types of touchscreen, they don't work if you touch them with a plastic stylus (because the plastic is an insulator and stops your hand
from affecting the electric field).

Infrared

• Just like the magic eye beams in an intruder alarm, an infrared touchscreen
uses a grid pattern of LEDs and light-detector photocells arranged on opposite sides of the screen. The LEDs shine infrared light in front of the screen—a bit like an invisible spider's web. If you touch the screen at a certain point, you interrupt two or more beams. A microchip inside the screen
can calculate where you touched by seeing which beams you interrupted.

The touchscreen on Sony Reader ebooks (like the one pictured in our top

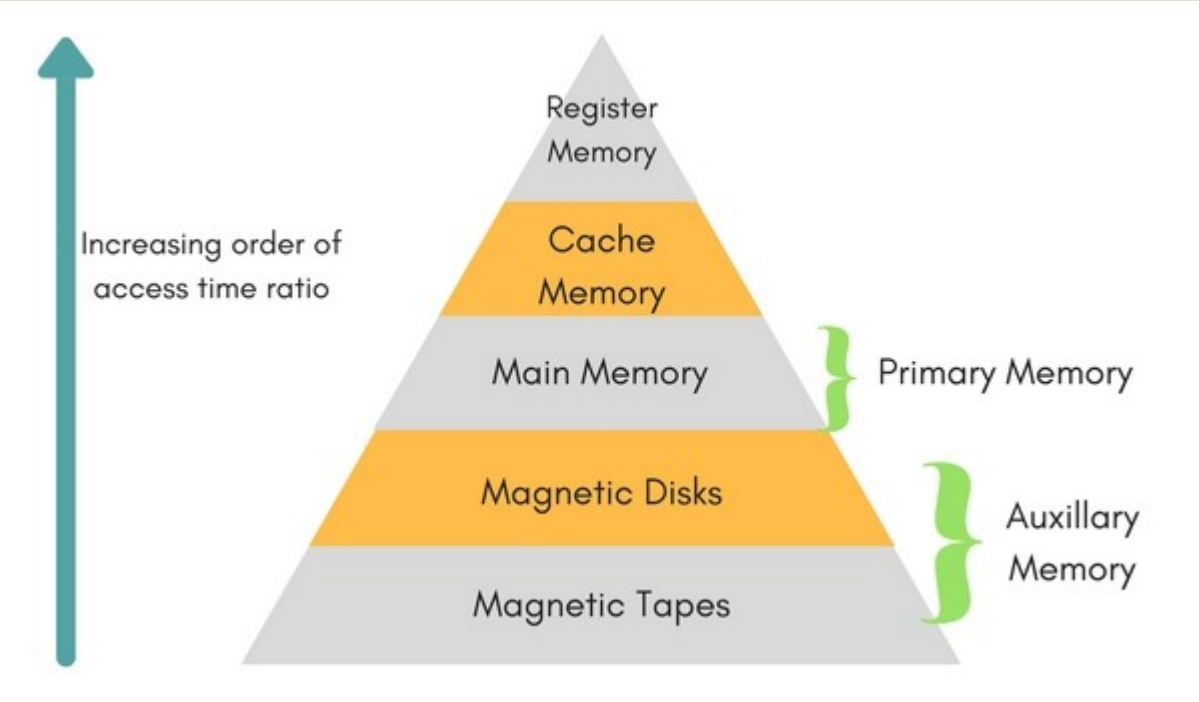
Surface Acoustic Wave

- Surprisingly, this touchscreen technology detects your fingers using
- Sound instead of light. Ultrasonic sound waves (too high pitched for humans to hear) are generated at the edges of the screen and reflected back and forth across its surface. When you touch the screen, you interrupt the sound beams and absorb some of their energy. The screen's microchip controller figures out from this where exactly you touched the screen.

Memory Organisation

- A memory unit is the collection of storage units or devices together. The memory unit stores the binary information in the form of bits. Generally, memory/storage is classified into 2 categories:
- **Volatile Memory**: This loses its data, when power is switched off.
- **Non-Volatile Memory**: This is a permanent storage and does not lose any data when power is switched off.

Memory Hierarchy



Internal Memory

Internal memory typically refers to main memory (RAM), but may also refer to ROM and flash memory. In either case, internal memory generally refers to chips rather than disks or tapes.

In a computer, all of the storage spaces that are accessible by a processor without the use of the computer input-output Internal memory usually includes several types of storage, such as main storage, cache memory, and special registers, all of which can be directly accessed by the processor.

- Primary storage (or main memory or internal memory), often referred to simply as memory, is the only one directly accessible to the CPU. The CPU continuously reads instructions stored there and executes them as required. Any data actively operated on is also stored there in uniform manner.

Types of Internal Memory

RAM (Random Access Memory)

Random access memory, or RAM, is memory storage on a computer that holds data while the computer is running so that it can be accessed quickly by the processor. RAM holds the operating system, application programs and data that is currently being used.

RAM data is much faster to read than data stored on the hard disk. RAM is stored in microchips and contains much less data than the hard disk. RAM can never run out of memory, but the processor must overwrite old data if the RAM is filled, which results in slower computer function. Any file stored in RAM can be accessed directly if the user knows the row and column where the data is stored.

Random access memory is used to store temporary but necessary information on a computer for quick access by open programs or applications.

RAM, is a volatile yet fast type of memory used in computers. RAM is more expensive to incorporate.

RAM allows reading and writing (electrically) of data at the byte level

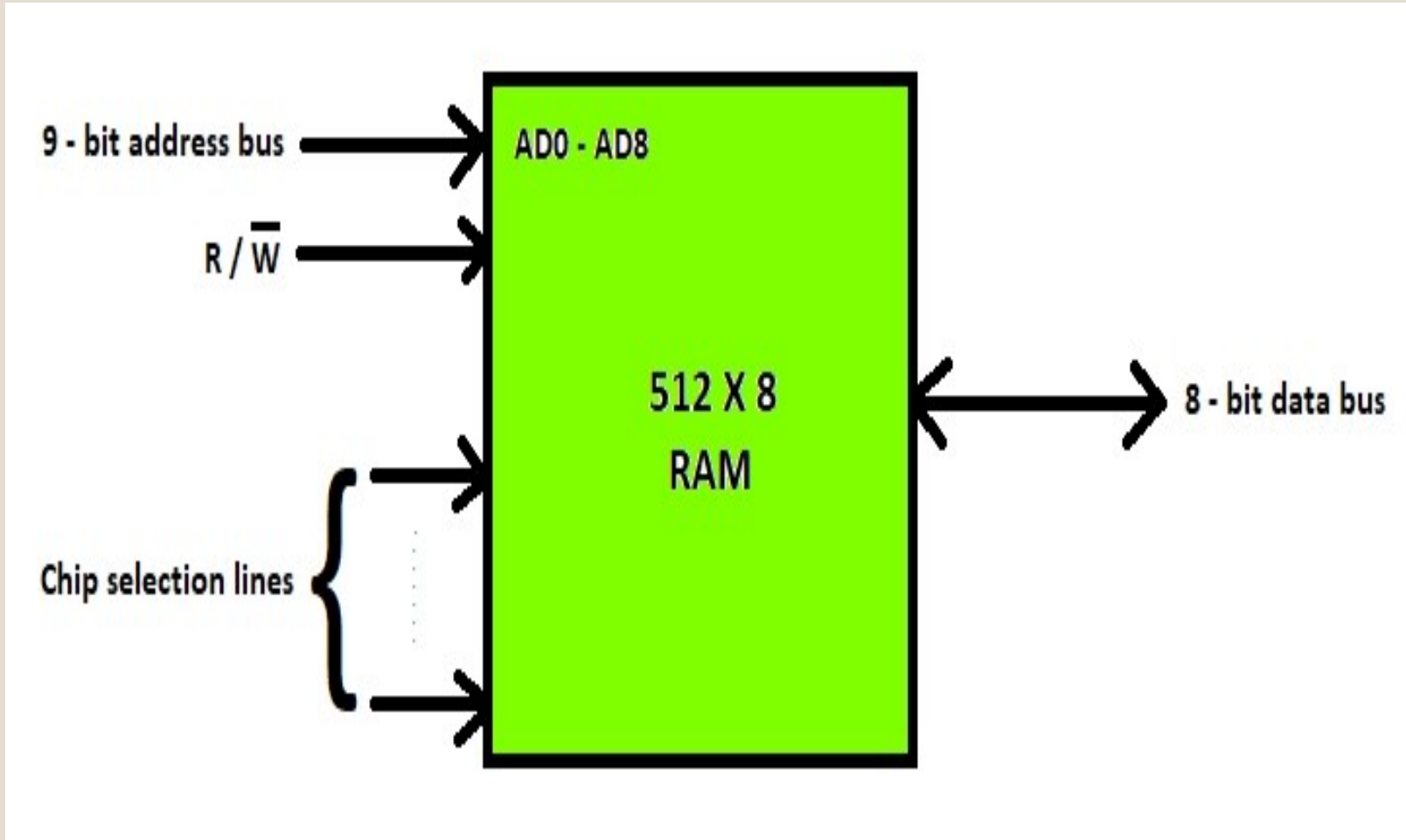
- RAM is the Volatile memory.

Types of RAM

Static RAM

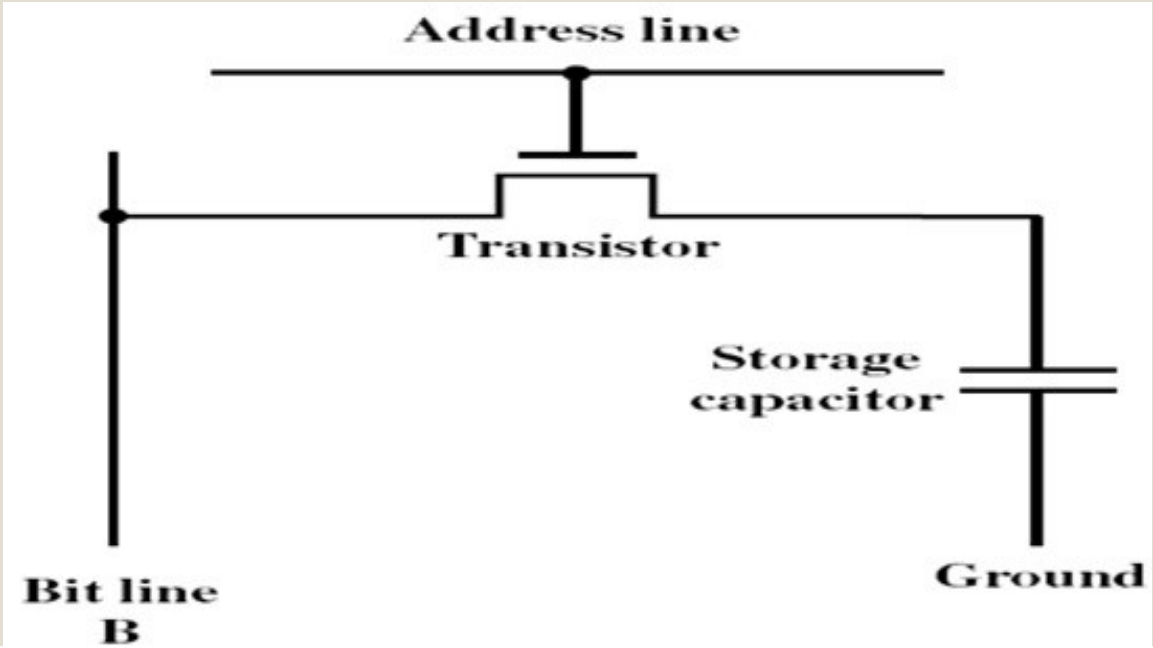
Static RAM stores a bit of information in a flip-flop. Static RAM is usually used for applications that do not require large capacity RAM memory.

Static(RAM) is a memory technology based on flip-flops. SRAM has an access time of 2 – 10 nanoseconds. All of main memory can be viewed as fabricated from SRAM, although such a memory would be unrealistically expensive


9 - bit address bus
R / W
Chip selection lines
AD0 - AD8
512 X 8
RAM
8 - bit data bus

Dynamic RAM

Dynamic RAM data store one bit of information as a payload. Dynamic RAM using a substrate capacitance gate MOS transistors as memory cells shut. To keep dynamic RAM stored data remains intact, the data should be refreshed again by reading and re-write the data into memory. Dynamic RAM is used for applications that require large RAM capacity, for example in a personal computer.



- EDO (Extended Data-output) and SD (**Synchronous Dynamic Random Access Memory**) are type of Dynamic RAM.
- Dynamic RAM (DRAM) is a memory technology based on capacitors
- Dynamic RAM is cheaper than static RAM and can be packed more densely on a computer chip
- DRAM has an access time in the order of 60 – 100 nanoseconds, slower than SRAM.

ROM (Read only Memory)

Sometimes can be erased for reprogramming, but might have odd requirements such as UV light or erasure only at the block level.

- Data are written into a ROM when it is manufactured.
- ROM is mask programmed by the manufacturer in the factory with the contents ordered by the customers.
- The contents are fixed by metal masks used during chip fabrication.
- Once programmed, the contents cannot be erased.
- Even a single bit wrongly programmed the ROM chip is useless

Applications

- Used to store control programs such as micro program.
- Character generation, code conversion
- Permanent storage – nonvolatile
- Microprogramming
- Library subroutines
- Systems programs (BIOS)
- Function tables
- Embedded system code

Types of ROM

- PROM (Programmable Read Only Memory)
- EPROM (Erasable Programmable Read Only Memory)
- EEPROM (Electrically Erasable Programable Read Only Memory)

If the content is determined by the vendor ROM, PROM sold empty and can then be filled with a program by the user. Having completed the program, fill PROM cannot be removed.

PROM (Programmable Read Only Memory)

- PROM is a field programmable device.
- The customer buy a blank PROM and store desired data using PROM programmer(burner).
- Programmability achieved by inserting a fuse at point P.
- Before programmed, the memory contains all 0s
- The user can insert 1 by burning out the fuse in the particular cell using high current pulse.
- The PROM chip can be programmed only once and its contents cannot be erased.
- PROM are flexible, faster and less expensive because they can be programmed directly by the user.

EPROM (Erasable Programmable Read Only Memory)

A rewritable chip that holds its contents without power. Previous data can be erased and new data can be inserted

EPROM chips are written on an external programming device before being placed on the circuit board. Capable of retaining stored information for a long time.

Eraser contd., requires breakup the charges trapped in the transistors of memory cell.[this is done by break the chip to ultraviolet light].

This reason EPROM packaged with transparent window.

Disadvantages: Entire EPROM is erased as a whole and selective erasing is not possible.

Should be removed from the chip for reprogramming.

- Unlike the PROM, EPROM contents can be deleted after being programmed. Elimination is done by using ultraviolet light.

EEPROM (Electrically Erasable Programable Read Only Memory)

- EEPROM can store data permanently, but its contents can still be erased electrically through the program. One type EEPROM is Flash Memory. Flash Memory commonly used in digital cameras, video game consoles, and the BIOS chip.
- It can be both programmed and erased electrically (flashed back to Zero).
- They do not need to removed when the chip content erasure.
- Also, erase selected content in the chip.
- Erasing and programming dynamically without removing the EEPROM from the circuit.

Disadvantages

- Different voltages are required for erasing, reading and writing the data.

External Memory

- **External memory** typically refers to storage in an **external** hard drive or on the Internet. The main "**memory**" in the **computer** is the **computer** work-space, not its storage facility.
- External memory which is sometimes called *backing store* or *secondary memory*, allows the permanent storage of large quantities of data. Some method of magnetic recording on magnetic disks or tapes is most commonly used.
- The capacity of external memory is high, usually measured in hundreds of megabytes or even in **gigabytes**
- The most common form of external memory is a **hard disc** which is permanently installed in the computer and will typically have a capacity of hundreds of megabytes

Types of External Memory

- Magnetic Tapes
- Hard disk
- Magnetic Disk
  - Optical Drives (CD-R/W, CD-ROM)

Cache Memory

Cache memory, also called CPU memory, is high-speed static random access memory (SRAM) that a computer microprocessor can access more quickly than it can access regular random access memory (RAM). This memory is typically integrated directly into the CPU chip or placed on a separate chip that has separate bus interconnect with the CPU. The purpose of cache memory is to store program instructions and data that are used repeatedly in the operation of programs or information that the CPU is likely to need next. The computer processor can access this information quickly from the cache rather than having to get it from computer’s main memory. Fast access to these instructions increases the overall speed of the program.

- As the microprocessor processes data, it looks first in the cache memory. If it finds the instructions or data it's looking for there from a previous reading of data, it does not have to perform a more time-consuming reading of data from larger main memory or other data storage devices. Cache memory is responsible for speeding up computer operations and processing.
- Once they have been opened and operated for a time, most programs use few of a computer's resources. That's because frequently re-referenced instructions tend to be cached. This is why system performance measurements for computers with slower processors but larger caches can be faster than those for computers with faster processors but less cache space.

Cache memory mapping

Caching configurations continue to evolve, but cache memory traditionally works under three different configurations:

- **Direct mapped cache** has each block mapped to exactly one cache memory location. Conceptually, direct mapped cache is like rows in a table with three columns: the data block or cache line that contains the actual data fetched and stored, a tag with all or part of the address of the data that was fetched, and a flag bit that shows the presence in the row entry of a valid bit of data.

- **Fully associative cache mapping** is similar to direct mapping in structure but allows a block to be mapped to any cache location rather than to a prespecified cache memory location as is the case with direct mapping.
- **Set associative cache mapping** can be viewed as a compromise between direct mapping and fully associative mapping in which each block is mapped to a subset of cache locations. It is sometimes called *N-way set associative mapping*, which provides for a location in main memory to be cached to any of "N" locations in the L1 cache.

Virtual Memory

- Virtual memory is a memory management capability of an operating system (OS) that uses hardware and software to allow a computer to compensate for physical memory shortages by temporarily transferring data from random access memory (RAM) to disk storage. Virtual address space is increased using active memory in RAM and inactive memory in hard disk drives (HDDs) to form contiguous addresses that hold both the application and its data.

- Virtual memory was developed at a time when physical memory -- the installed RAM -- was expensive. Computers have a finite amount of RAM, so memory can run out, especially when multiple programs run at the same time. A system using virtual memory uses a section of the hard drive to emulate RAM. With virtual memory, a system can load larger programs or multiple programs running at the same time, allowing each one to operate as if it has infinite memory and without having to purchase more RAM.
- While copying virtual memory into physical memory, the OS divides memory into pagefiles or swap files with a fixed number of addresses. Each page is stored on a disk and when the page is needed, the OS copies it from the disk to main memory and translates the virtual addresses into real addresses.

Pros of using virtual memory

Among the primary benefits of virtual memory is its ability to handle twice as many addresses as main memory. It uses software to consume more memory by using the HDD as temporary storage while MMUs translate virtual memory addresses to physical addresses via the CPU. Programs use virtual addresses to store instructions and data; when a program is executed, the virtual addresses are converted into actual memory addresses.

Secondary Storage

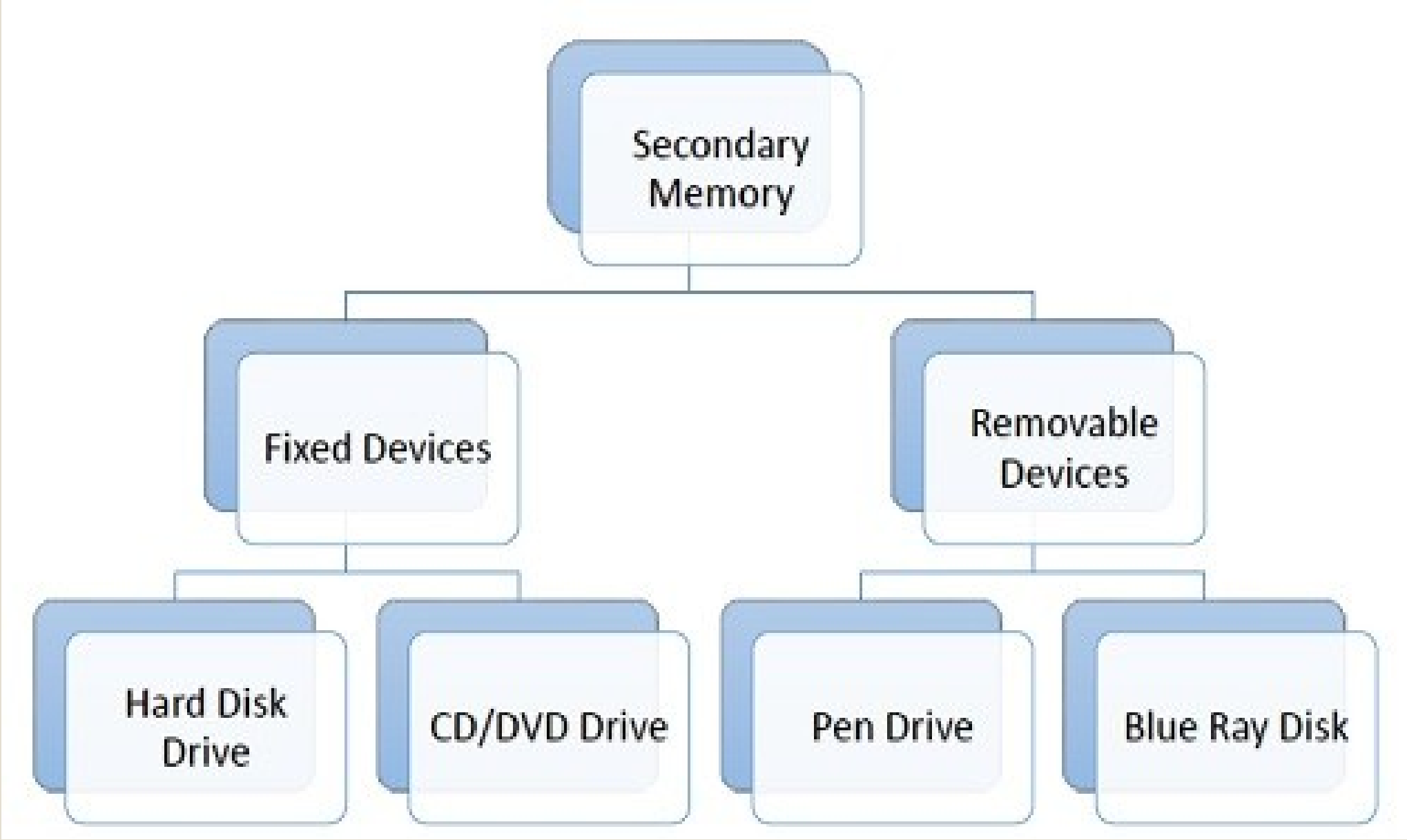
Secondary memory (or secondary storage) is the slowest and cheapest form of memory. It cannot be processed directly by the CPU. It must first be copied into primary storage (also known as RAM).

Characteristics of Secondary Memory

It is non-volatile, i.e. It retains data when power is switched off

It is large capacities to the tune of terabytes

- It is cheaper as compared to primary memory


Secondary Memory
Fixed Devices
Removable Devices
Hard Disk Drive
CD/DVD Drive
Pen Drive
Blue Ray Disk

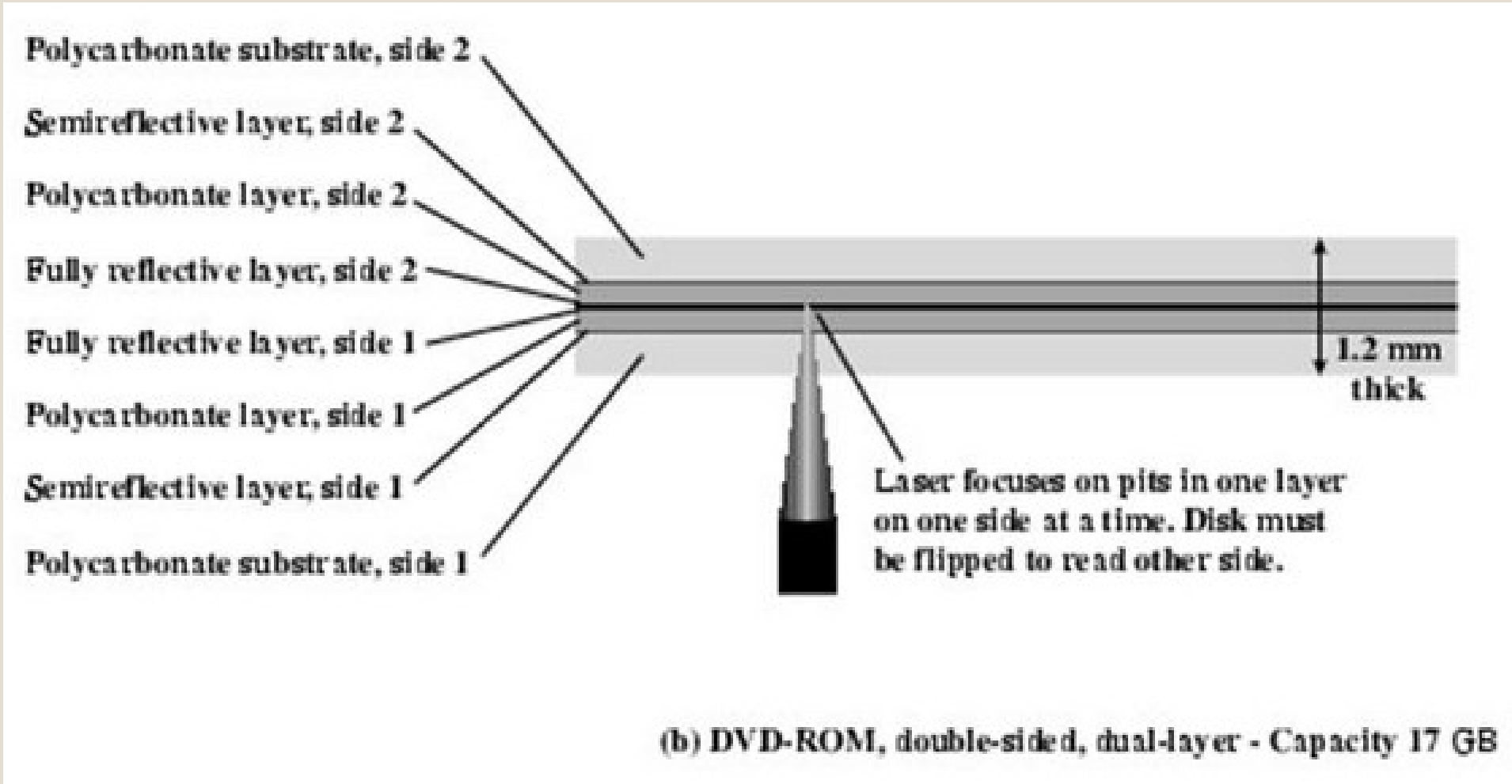
CD Drive

CD stands for Compact Disk. CDs are circular disks that use optical rays, usually lasers, to read and write data. They are very cheap as you can get 700 MB of storage space for less than a dollar. CDs are inserted in CD drives built into CPU cabinet. They are portable as you can eject the drive, remove the CD and carry it with you. There are three types of CDs −

- **CD-ROM (Compact Disk – Read Only Memory)** − The data on these CDs are recorded by the manufacturer. Proprietary Software, audio or video are released on CD-ROMs.
- **CD-R (Compact Disk – Recordable)** − Data can be written by the user once on the CD-R. It cannot be deleted or modified later.
- **CD-RW (Compact Disk – Rewritable)** − Data can be written and deleted on these optical disks again and again.

DVD Drive

- DVD stands for Digital Video Display. DVD are optical devices that can store 15 times the data held by CDs. They are usually used to store rich multimedia files that need high storage capacity. DVDs also come in three varieties – read only, recordable and rewritable.

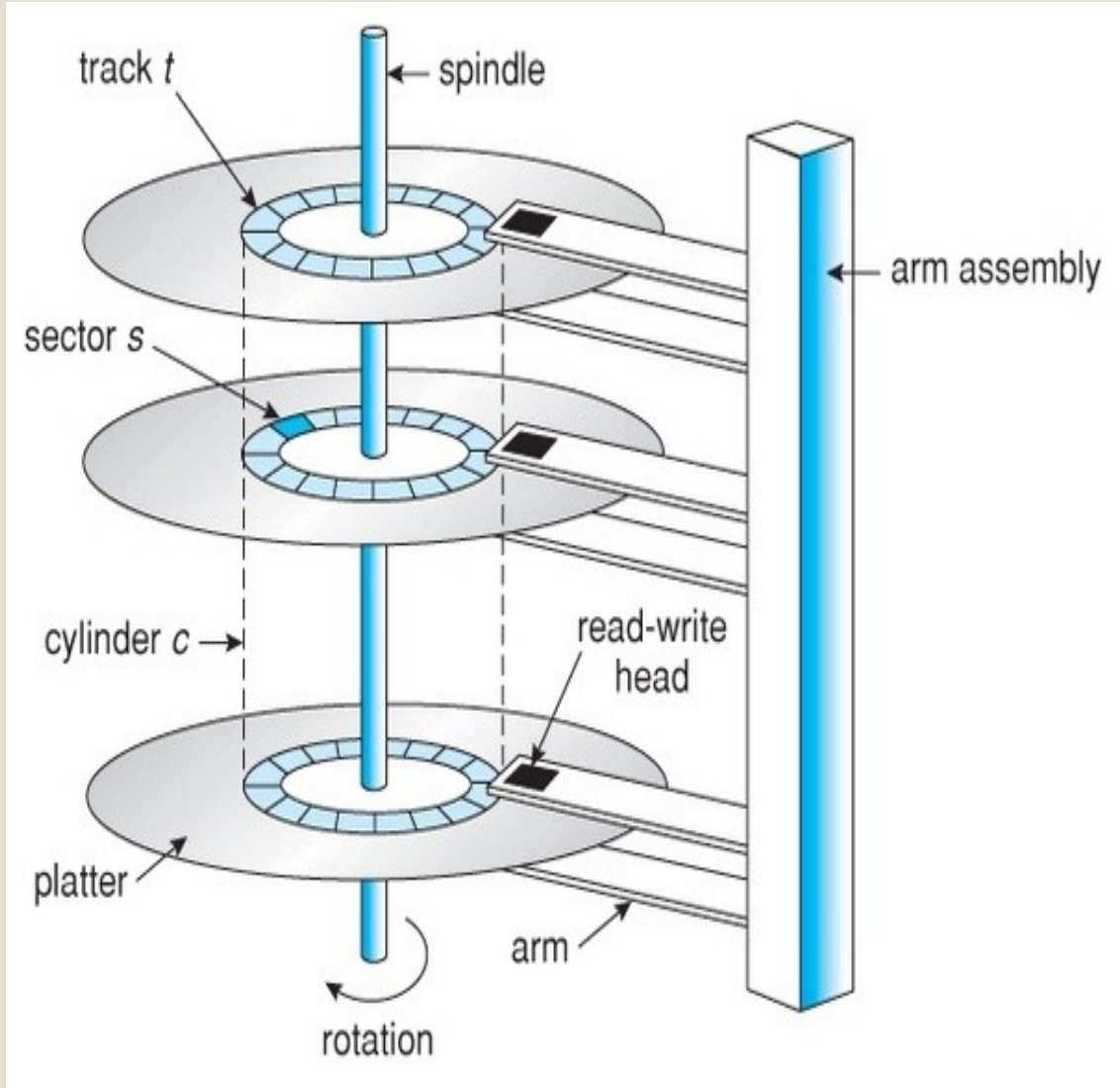


(b) DVD-ROM, double-sided, dual-layer - Capacity 17 GB

Magnetic Disk

A magnetic disk primarily consists of a rotating magnetic surface and a mechanical arm that moves over it. The mechanical arm is used to read from and write to the disk. The data on a magnetic disk is read and written using a magnetization process. Data is organized on the disk in the form of tracks and sectors, where tracks are the circular divisions of the disk. Tracks are further divided into sectors that contain blocks of data. All read and write operations on the magnetic disk are performed on the sectors.

- Magnetic disks have traditionally been used as primary storage in computers. With the advent of solid-state drives (SSDs), magnetic disks are no longer considered the only option, but are still commonly used.


track t
spindle
arm assembly
sector s
cylinder c
read-write
head
platter
arm
rotation

Magnetic Tape

Originally, magnetic tape was designed to record sound. In computing, it holds binary data. In recent years, magnetic tape devices have become more scarce with the emergence of digital imaging and audiovisual media storage.

- Magnetic tape was used in many of the larger and less complex mainframe computers that predated today’s personal computers (PC).

- Uses the same reading and recording techniques as disk systems.
- Medium is flexible polyester tape coated with magnetizable material.
- Data on the tape are structured as a number of parallel tracks running lengthwise.
- Data are laid out as a sequence of bits along each track.
- Data are read and written in contiguous blocks called physical records.
- Blocks on the tape are separated by gaps referred to as inter-record gaps

CPU Organisation

- General Resistor Organisation
- Stack organisation and Accumulator types
- Instruction Format
- Addressing Modes

Introduction

- The part of computer that performs the bulk of data-processing operations is called central processing unit and is referred to as the CPU. The CPU is made up of three major parts.

1. The register set stores intermediate data used during the execution of the instructions.
2. The arithmetic logic unit(ALU) performs the required microoperations for executing the instructions.
3. The control unit supervises the transfer of information among the registers and instruct the ALU as to which operation to perform.

- The CPU performs a variety of functions dictated by the type of instruction that are incorporated in the computer.

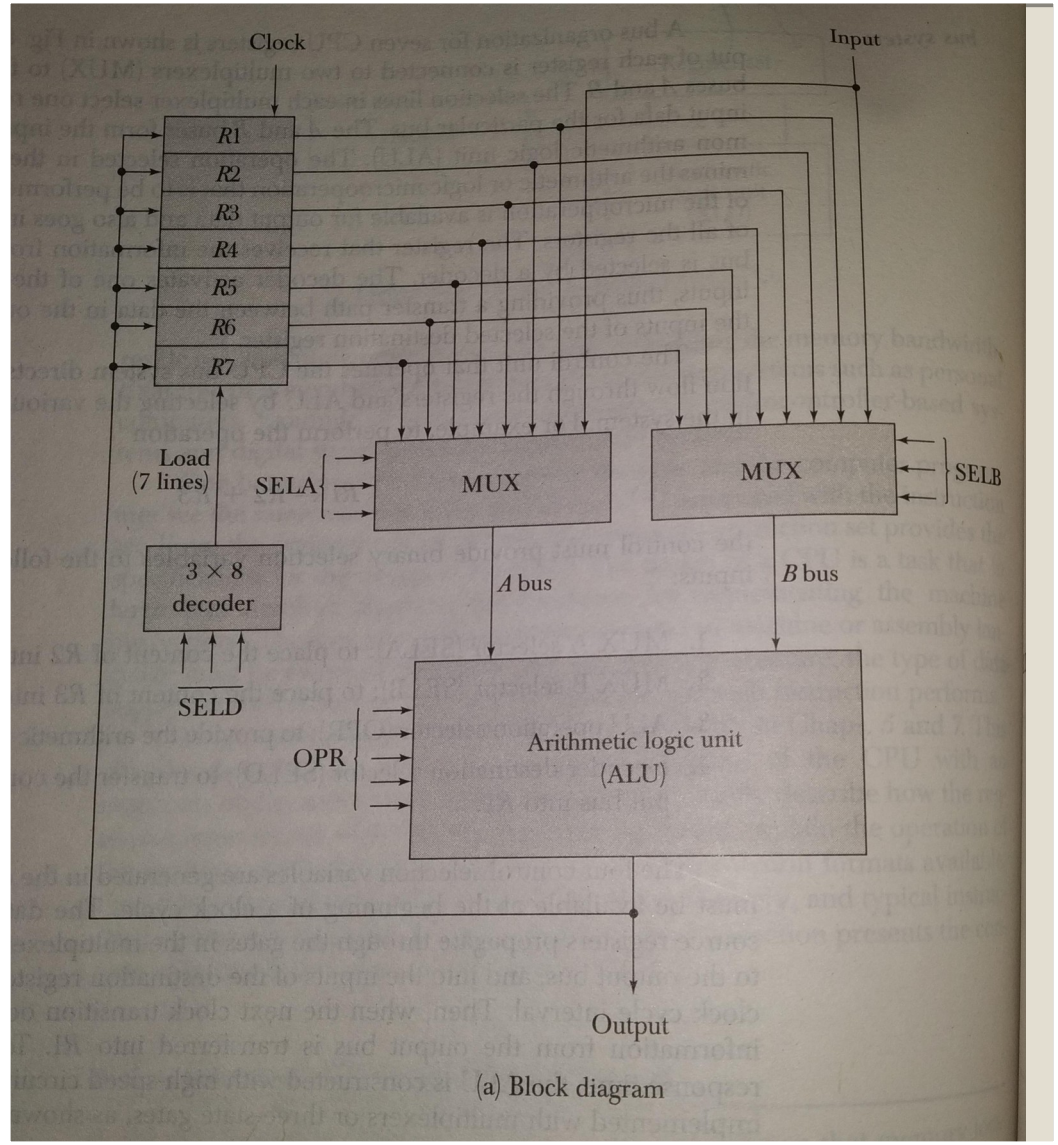
- This includes the instruction formats, addressing modes, the instruction set, and the general organisation of the CPU registers leading to two computer architectures as reduced instruction set computer(RISC) and complex instruction set computer(CISC).

General Register Organisation

- When a large number of registers are included in the CPU, it is most efficient to connect them through a common bus system. The registers communicate with each other not only for direct data transfers, but also while performing various microoperations. Hence it is necessary to provide a common unit that can perform all the arithmetic, logic, and shift microoperations in the processor.

1. BUS SYSTEM

- A bus organisation for seven CPU registers is shown. The output of each register is connected to two multiplexers(MUX) to form the two buses A and B. The selection lines in each mutliplexer select one register or the input data for the particular bus.



(a) Block diagram

- The A and B buses from the inputs to a common arithmetic logical unit(ALU). The operation selected in the ALU determines the arithmetic or logic microoperation that is to be performed. The result of the microoperation is available for output data and also goes into the inputs of all the resisters. The resister that receives the information from the output bus is selected by a decoder. The decoder activates one of the register load inputs of the selected destination register.

CONTROL WORD

- There are 14 binary selectioninputs in the unit, and their combined value specifies a *control word .* The 14-bit control word is defined in the fig. b. It consists of four fields. Three fields contain three bit each , and one has five bits. The three bit of SELA select a source register for the A input of the ALU. the three bits of SELD select a destination registor for the B input of the ALU. The three bits of SELD select a destination register using the decoder and its seven load outputs. The five bits of OPR select one of the operations in the ALU. The 14- bit control word when applied to the selection inputs specify a particular microoperation.

ALU

- The ALU provides arithmetic and logic operations. In addition, the CPU must provide shift operations. The shifter may be placed in the input of the ALU to provide a preshift capability, or at the output of the ALU to provide postshifting capability. In some cases, the shift operations are included with the ALU. An arithmetic logic and shift unit was designed in Sec. 4-7. The functiton table for this ALU is listed in Table 4-8. The encoding of the ALU operations for the CPU is taken from Sec.4-7 and is specified in Table 8-2. The OPR field has five bits and each operation is designated with a symbolic name.

STACK ORGANIZATION

- A useful feature that is included in the CPU of most computers is a stack or last-in, first-out (LIFO) list. A stack is a storage device that stores information in such a manner that the item stored last is the first item retrieved. The operation of a stack can be compared to a stack of trays. The last tray placed on top of the stack is the first to be taken off.
- The stack in digital computers is essentially a memory unit with an address register that can count only (after an inital value is loaded into it). The register that holds the address for the stack is called a stack pointer (SP) because its value always points at the top item in the stack. Contrary to a stack of trays where the tray itself may be taken out or inserted, the physical registers of a stack are always available for reading or writing. It is the content of the word that is inserted or deleted.

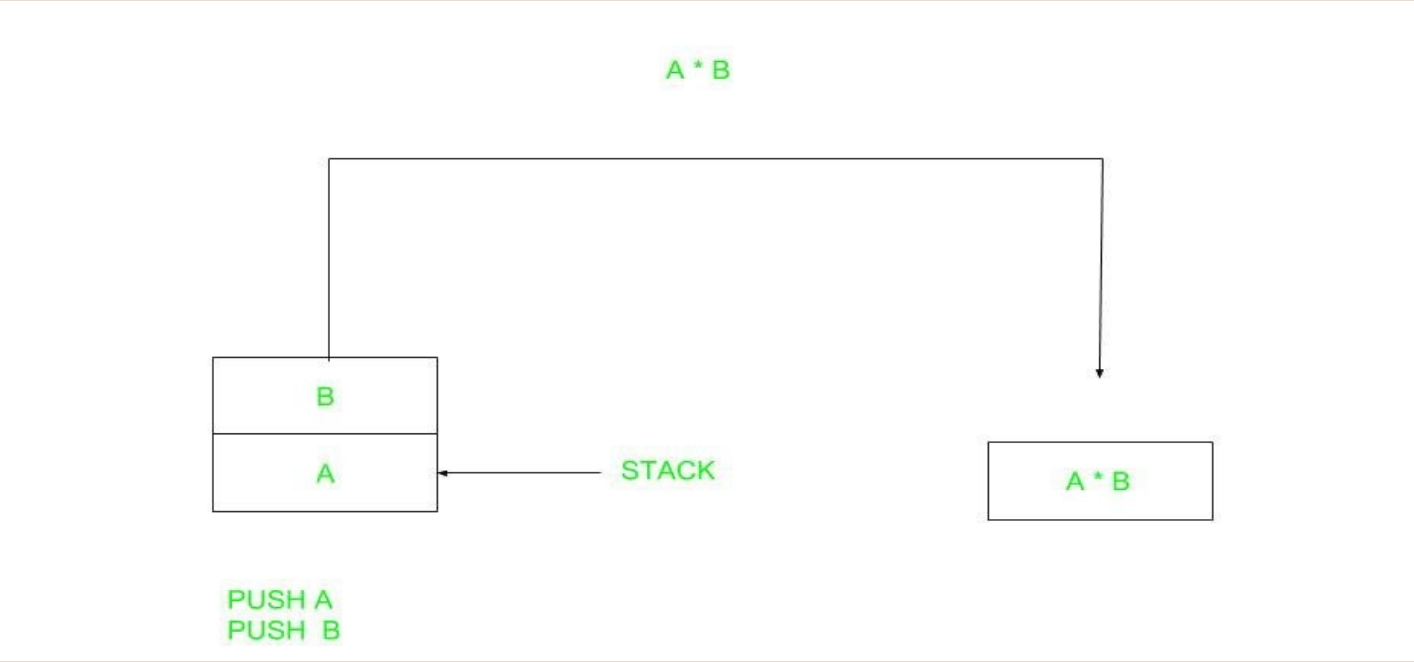
The two operations of a stack are the insertion and deletion of items. The operation of insertion is called push (or push-down) because it can be thought of as the result of pushing a new item on top. The operation of deletion is called pop (or pop-up) because it can be thought of as the result of removing one item so that the stack pops up. However, nothing is pushed or popped in a computer stack.' The operations are simulated by incrementing or decrementing the stack pointer register

ACCUMULATOR TYPE ORGANISATION

- The computers, present in the early days of computer history, had accumulator based CPUs. In this type of CPU organization, the accumulator register is used implicitly for processing all instructions of a program and store the results into the accumulator. The instruction format that is used by this CPU Organisation is **One address field**. Due to this the CPU is known as **One Address Machine**.
- The main points about Single Accumulator based CPU Organisation are:
- In this CPU Organization, the first ALU operand is always stored into the Accumulator and the second operand is present either in Registers or in the Memory.
- Accumulator is the default address thus after data manipulation the results are stored into the accumulator.
- One address instruction is used in this type of organization.

Instruction Formats

- **Zero Address Instructions –**



A stack based computer do not use address field in instruction.To evaluate a expression first it is converted to revere Polish Notation i.e. Post fix Notation.
Expression: X = (A+B)*(C+D) Postfixed : X = AB+CD+* TOP means top of stack M[X] is any memory location.

PUSH	A	TOP = A
PUSH	B	TOP = B
ADD		TOP = A+B
PUSH	C	TOP = C
PUSH	D	TOP = D
ADD		TOP = C+D
MUL		TOP = (C+D)*(A+B)
POP	X	M[X] = TOP

ONE ADDRESS ACCUMULATOR

- **One Address Instructions –**
  This use a implied ACCUMULATOR register for data manipulation.One operand is in accumulator and other is in register or memory location.Implied means that the CPU already know that one operand is in accumulator so there is no need to specify it.

LOAD	A	AC = M[A]
ADD	B	AC = AC + M[B]
STORE	T	M[T] = AC
LOAD	C	AC = M[C]
ADD	D	AC = AC + M[D]
MUL	T	AC = AC * M[T]
STORE	X	M[X] = AC

Two Address Instructions –
This is common in commercial computers.Here two address can be specified in the instruction.Unlike earlier in one address instruction the result was stored in accumulator here result cab be stored at different location rather than just accumulator, but require more number of bit to represent address.

MOV	R1, A	R1 = M[A]
ADD	R1, B	R1 = R1 + M[B]
MOV	R2, C	R2 = C
ADD	R2, D	R2 = R2 + D
MUL	R1, R2	R1 = R1 * R2
MOV	X, R1	M[X] = R1

Three Address Instructions –
This has three address field to specify a register or a memory location. Program created are much short in size but number of bits per instruction increase. These instructions make creation of program much easier but it does not mean that program will run much faster because now instruction only contain more information but each micro operation (changing content of register, loading address in address bus etc.) will be performed in one cycle only.

ADD	R1, A, B	R1 = M[A] + M[B]
ADD	R2, C, D	R2 = M[C] + M[D]
MUL	X, R1, R2	M[X] = R1 * R2

- The A and B buses from the inputs to a common arithmetic logical unit(ALU). The operation selected in the ALU determines the arithmetic or logic microoperation that is to be performed. The result of the microoperation is available for output data and also goes into the inputs of all the resisters. The resister that receives the information from the output bus is selected by a decoder. The decoder activates one of the register load inputs of the selected destination register.

CONTROL WORD

- There are 14 binary selectioninputs in the unit, and their combined value specifies a *control word .* The 14-bit control word is defined in the fig. b. It consists of four fields. Three fields contain three bit each , and one has five bits. The three bit of SELA select a source register for the A input of the ALU. the three bits of SELD select a destination registor for the B input of the ALU. The three bits of SELD select a destination register using the decoder and its seven load outputs. The five bits of OPR select one of the operations in the ALU. The 14- bit control word when applied to the selection inputs specify a particular microoperation.

ALU

- The ALU provides arithmetic and logic operations. In addition, the CPU must provide shift operations. The shifter may be placed in the input of the ALU to provide a preshift capability, or at the output of the ALU to provide postshifting capability. In some cases, the shift operations are included with the ALU. An arithmetic logic and shift unit was designed in Sec. 4-7. The functiton table for this ALU is listed in Table 4-8. The encoding of the ALU operations for the CPU is taken from Sec.4-7 and is specified in Table 8-2. The OPR field has five bits and each operation is designated with a symbolic name.

STACK ORGANIZATION

- A useful feature that is included in the CPU of most computers is a stack or last-in, first-out (LIFO) list. A stack is a storage device that stores information in such a manner that the item stored last is the first item retrieved. The operation of a stack can be compared to a stack of trays. The last tray placed on top of the stack is the first to be taken off.
- The stack in digital computers is essentially a memory unit with an address register that can count only (after an inital value is loaded into it). The register that holds the address for the stack is called a stack pointer (SP) because its value always points at the top item in the stack. Contrary to a stack of trays where the tray itself may be taken out or inserted, the physical registers of a stack are always available for reading or writing. It is the content of the word that is inserted or deleted.

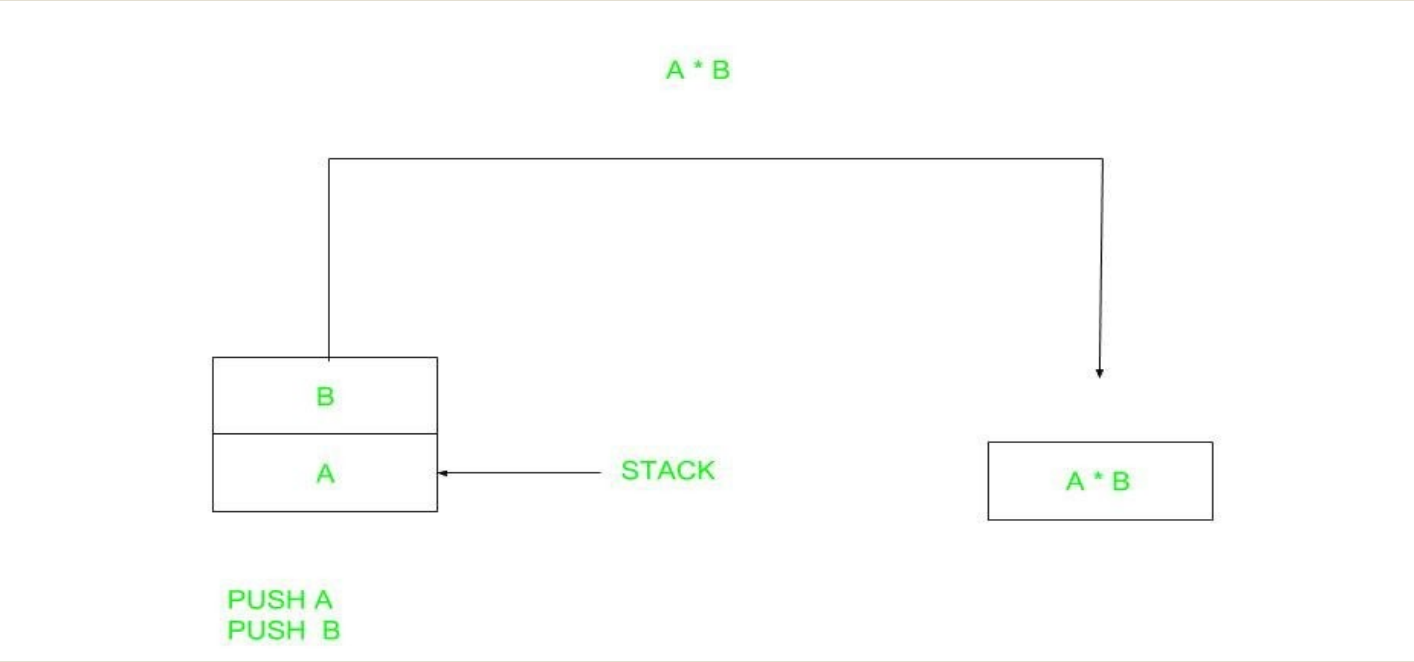
The two operations of a stack are the insertion and deletion of items. The operation of insertion is called push (or push-down) because it can be thought of as the result of pushing a new item on top. The operation of deletion is called pop (or pop-up) because it can be thought of as the result of removing one item so that the stack pops up. However, nothing is pushed or popped in a computer stack.' The operations are simulated by incrementing or decrementing the stack pointer register

ACCUMULATOR TYPE ORGANISATION

- The computers, present in the early days of computer history, had accumulator based CPUs. In this type of CPU organization, the accumulator register is used implicitly for processing all instructions of a program and store the results into the accumulator. The instruction format that is used by this CPU Organisation is **One address field**. Due to this the CPU is known as **One Address Machine**.
- The main points about Single Accumulator based CPU Organisation are:
- In this CPU Organization, the first ALU operand is always stored into the Accumulator and the second operand is present either in Registers or in the Memory.
- Accumulator is the default address thus after data manipulation the results are stored into the accumulator.
- One address instruction is used in this type of organization.

Instruction Formats

- **Zero Address Instructions –**



A stack based computer do not use address field in instruction.To evaluate a expression first it is converted to revere Polish Notation i.e. Post fix Notation.
Expression: X = (A+B)*(C+D) Postfixed : X = AB+CD+* TOP means top of stack M[X] is any memory location.

PUSH	A	TOP = A
PUSH	B	TOP = B
ADD		TOP = A+B
PUSH	C	TOP = C
PUSH	D	TOP = D
ADD		TOP = C+D
MUL		TOP = (C+D)*(A+B)
POP	X	M[X] = TOP

ONE ADDRESS ACCUMULATOR

- **One Address Instructions –**
  This use a implied ACCUMULATOR register for data manipulation.One operand is in accumulator and other is in register or memory location.Implied means that the CPU already know that one operand is in accumulator so there is no need to specify it.

LOAD	A	AC = M[A]
ADD	B	AC = AC + M[B]
STORE	T	M[T] = AC
LOAD	C	AC = M[C]
ADD	D	AC = AC + M[D]
MUL	T	AC = AC * M[T]
STORE	X	M[X] = AC

Two Address Instructions –
This is common in commercial computers.Here two address can be specified in the instruction.Unlike earlier in one address instruction the result was stored in accumulator here result cab be stored at different location rather than just accumulator, but require more number of bit to represent address.

MOV	R1, A	R1 = M[A]
ADD	R1, B	R1 = R1 + M[B]
MOV	R2, C	R2 = C
ADD	R2, D	R2 = R2 + D
MUL	R1, R2	R1 = R1 * R2
MOV	X, R1	M[X] = R1

Three Address Instructions –
This has three address field to specify a register or a memory location. Program created are much short in size but number of bits per instruction increase. These instructions make creation of program much easier but it does not mean that program will run much faster because now instruction only contain more information but each micro operation (changing content of register, loading address in address bus etc.) will be performed in one cycle only.

ADD	R1, A, B	R1 = M[A] + M[B]
ADD	R2, C, D	R2 = M[C] + M[D]
MUL	X, R1, R2	M[X] = R1 * R2

CONTROL UNIT

Contents

- Instruction word format
- Fetch and execution cycle
- Sequence of operation of control registers
- Control of arithmetic operations
- Microprogramming concepts

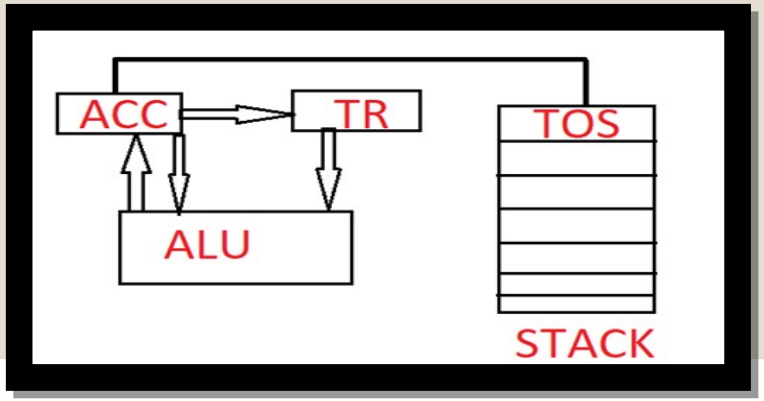
Instruction word format

- **Instruction format** describes the internal structures (layout design) of the bits of an instruction, in terms of its constituent parts.
- An **Instruction format** must include an opcode, and address is dependent on an availability of particular operands.
- The format can be implicit or explicit which will indicate the addressing mode for each operand.

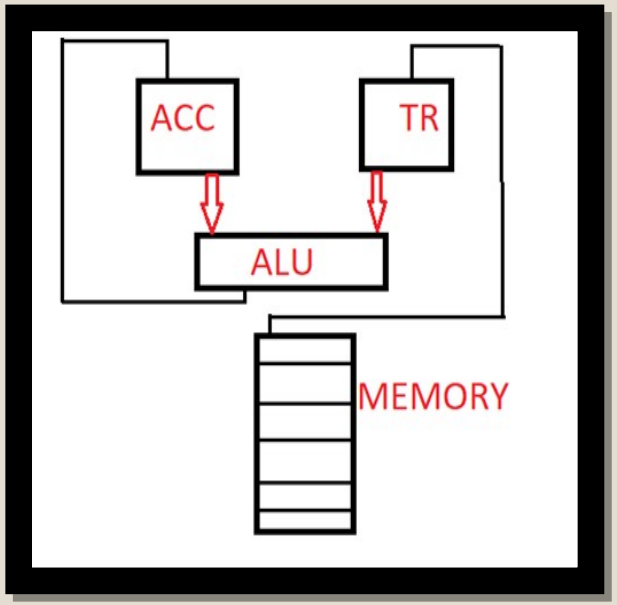
- Designing of an **Instruction format** is very complex. As we know a computer uses a variety of instructional. There are many designing issues which affect the instructional design, some of them are given are below:
  - **Instruction length:** It is a most basic issue of the format design. A longer will be the instruction it means more time is needed to fetch the instruction.
  - **Memory size:** If larger memory range is to be addressed then more bits will be required in the address field.
  - **Memory organization:** If the system supports the virtual memory then memory range which needs to be addressed by the instruction, is larger than the physical memory.
  - **Memory transfer length:** Instruction length should be equal to the data bus length or it should be multiple of it.
- **Instruction formats** are classified into 5 types based on the type of the CPU organization. CPU organization is divided into three types based on the availability of the ALU operands

- **STACK CPU**
- In this organization, ALU operands are performed only on a stack data. This means that both of the ALU operations are always required in the stack. The same stack is also used as the destination. In the stack, we can perform insert and deletion operation at only one end which is called as the top of a stack. So in this format, there is no need of address because in this TOS becomes the default location.
-

  In this organization, only the ALU operands are zero address operation whereas data transfer instructions are not a zero address instruction. The computable instruction format of STACK CPU is **Zero Address Instruction Format**



- **Accumulator CPU**
- In this organization, one of the ALU operands is always present in the accumulator. The same accumulator is also used as the destination. Another ALU operand is present either in the register or in memory. In processor design, only one accumulator is present so it becomes the default location.
- The computable instruction format of Accumulator CPU is **One Address Instruction Format**.



General Register CPU

- Based on the number of the registers possible in the processors, the architecture is divided into two types:
- Register-Memory references CPU
- Register-Register references CPU

- **Register-Memory Reference CPU**
- In this architecture, processors support less number of registers. Therefore register file size is small. In this organization, the first ALU operand is always required in the register. The same register can also be used as the destination. The second ALU operand is present either in a register or in memory. The computable instruction format of the register to memory reference CPU is **Two Address Instruction Format**.

- **Register-Register Reference CPU**
- In this architecture, processors support number of registers, therefore, register file size is large. In this organization, ALU operands are performed only on a registers data that means both of the ALU operands are required in the register. Due to more number of register present in the CPU, the separate register is used to store the result. The computable instruction format of Register-Register Reference CPU is Three Address Instruction Format.

- **Four Address instruction format**
- This format contains the 4 different address fields with an opcode. Since PC is used as the mandatory register in the CPU design which is used to hold the next instruction address. So four instruction format is not in the use.

Fetch and execution cycle

- The fetch execute cycle is the basic operation (instruction) cycle of a computer (also known as the fetch decode execute cycle).
- During the fetch execute cycle, the computer retrieves a program instruction from its memory. It then establishes and carries out the actions that are required for that instruction.

- A CPU has the following components:
- **Control Unit** – controls all parts of the computer system. It manages the four basic operations of the Fetch Execute Cycle as follows:
  - **Fetch** – gets the next program command from the computer’s memory
  - **Decode** – deciphers what the program is telling the computer to do
  - **Execute** – carries out the requested action
  - **Store** – saves the results to a Register or Memory
- **Arithmetic Logic Unit (ALU)** – performs arithmetic and logical operations
- **Register** – saves the most frequently used instructions and data

- Here’s a summary of the fetch – decode – execute cycle:
- The processor reviews the program counter to see which command to execute next.
- The program counter gives an address value in the memory of where the next command is.
- The processor fetches the command value from the memory location.
- Once the command has been fetched, it needs to be decoded and executed. For example, this could include taking one value, putting it into the Arithmetic Logic Unit (ALU), then taking a different value from a register and adding the two together.
- Once this has been completed, the processor returns to the program counter to find the next command.
- This cycle is replicated until the program stops.
- The **Execute Cycle** is the only step useful to the end user, everything else is required to make the execute cycle happen, as it performs the function of the command. The ALU is utilised if the command involves arithmetic or logical operations

Sequence of operation of control registers

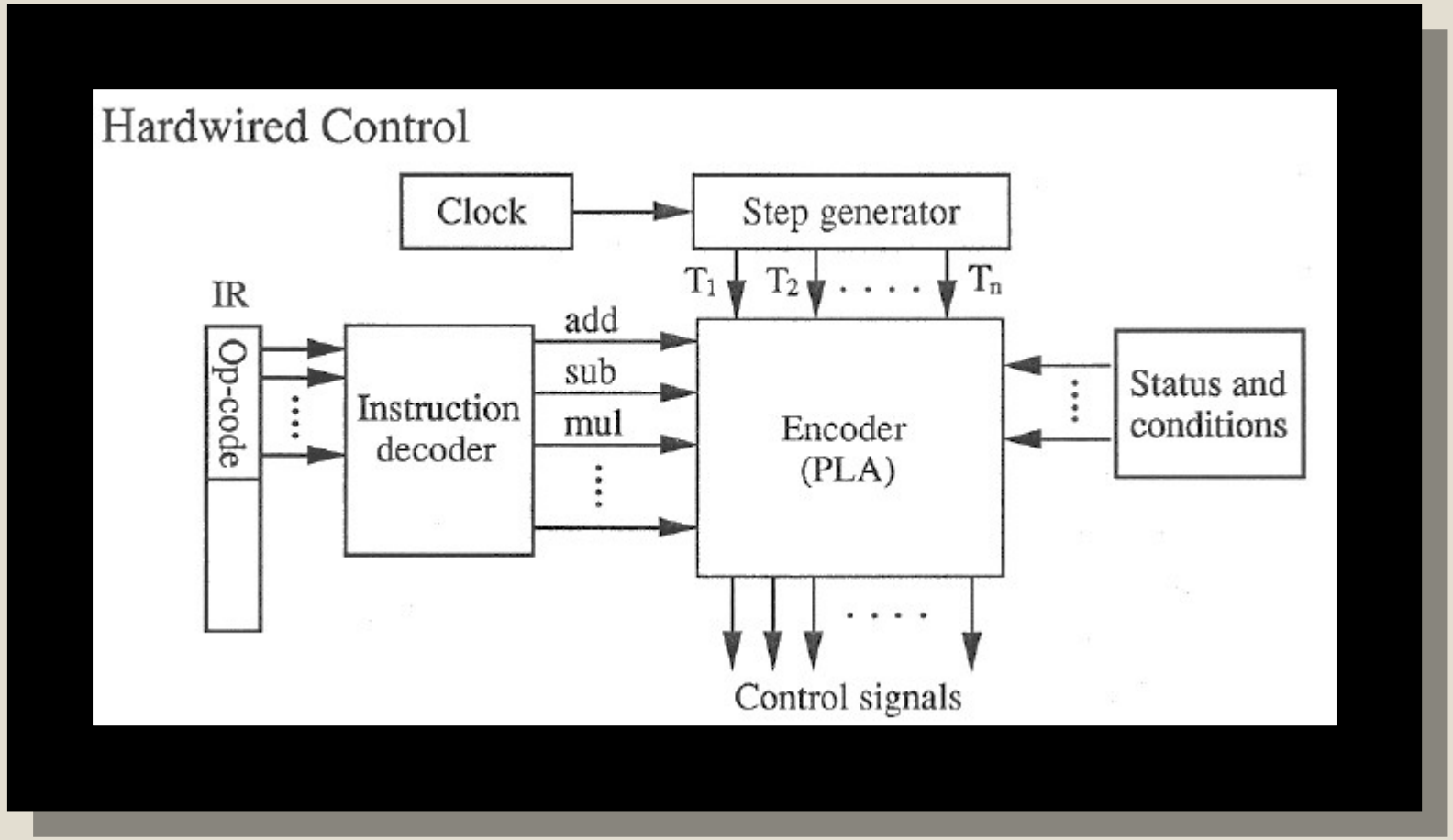
- register in the control unit of the CPU that is used to keep track of the address of the current or next instruction. Typically, the program counter is advanced to the next instruction, and then the current instruction is executed. Also known as a "sequence control register" and the "instruction pointer."

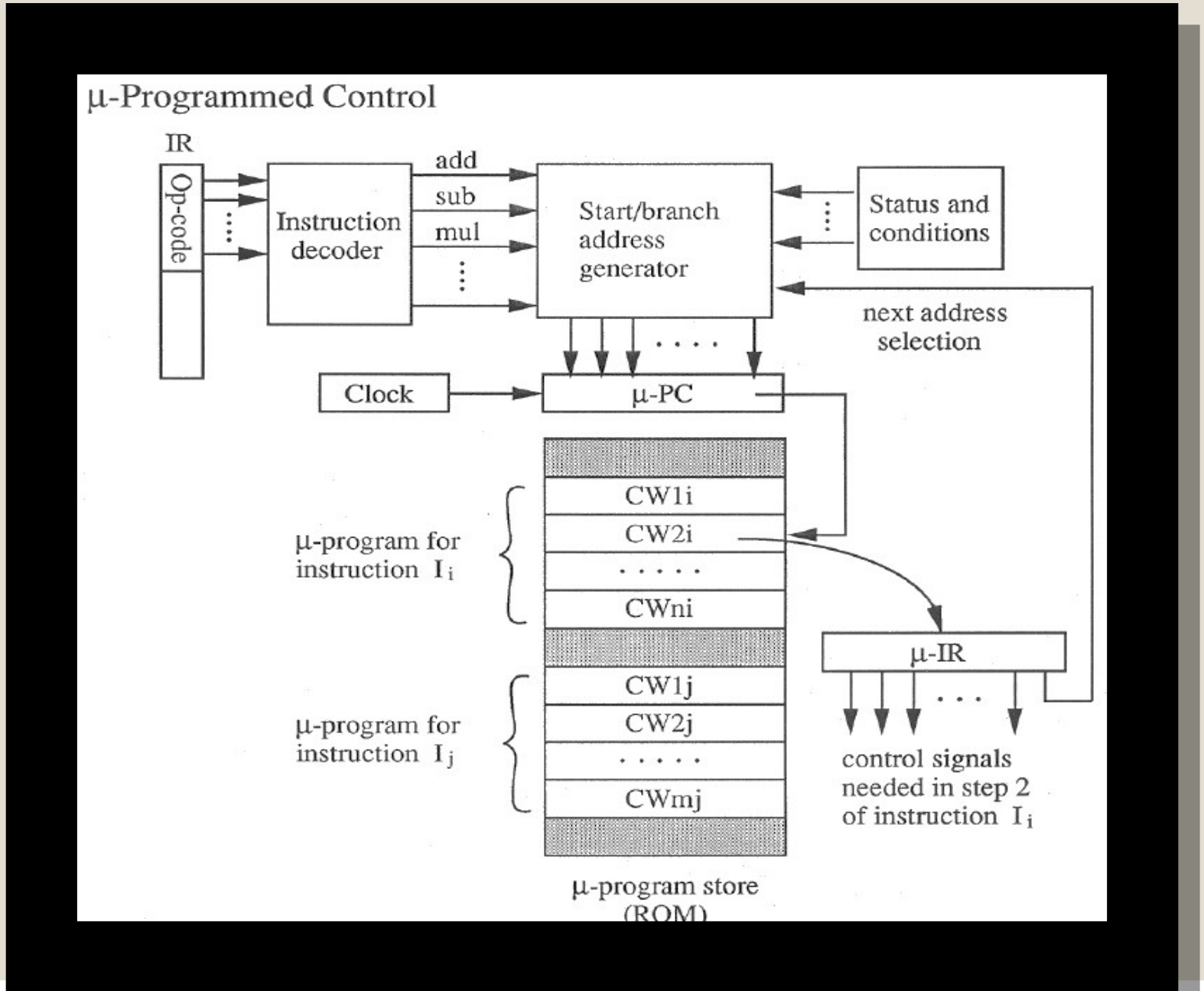
Control of arithmetic operations

- An **arithmetic logic unit** (**ALU**) is a combinational digital electronic circuit that performs arithmetic and bitwise operations on integer binary numbers. This is in contrast to a floating-point unit (FPU), which operates on floating poin numbers. An ALU is a fundamental building block of many types of computing circuits, including the central processing unit (CPU) of computers, FPUs, and graphics processing units(GPUs). A single CPU, FPU or GPU may contain multiple ALUs.
- The inputs to an ALU are the data to be operated on, called operands, and a code indicating the operation to be performed; the ALU's output is the result of the performed operation. In many designs, the ALU also has status inputs or outputs, or both, which convey information about a previous operation or the current operation, respectively, between the ALU and external status registers

Microprogramming concepts

- **Basic Concepts of Microprogramming:**
- **Control word (CW):**A word with each bit for one of the control signals. Each step of the instruction execution is represented by a control word with all of the bits corresponding to the control signals needed for the step set to one.
- **Microinstruction:**Each step in a sequence of steps in the execution of a certain machine instruction is considered as a *microinstruction*, and it is represented by a control word. All of the bits corresponding to the control signals that need to be asserted in this step are set to 1, and all others are set to 0 (*horizontal organization*).
- **Microprogram:**Composed of a sequence of microinstructions corresponding to the sequence of steps in the execution of a given machine instruction.
- **Microprogramming:**The method of generating the control signals by properly setting the individual bits in a control word of a step.


Hardwired Control
Clock
Step generator
T1
T2
Tn
IR
Op-code
Instruction decoder
add
sub
mul
Encoder (PLA)
Status and conditions
Control signals


μ-Programmed Control
IR
Op-code
Instruction decoder
add
sub
mul
Start/branch address generator
Status and conditions
next address selection
Clock
μ-PC
CW1i
CW2i
CWni
μ-program for instruction I_i
CW1j
CW2j
CWmj
μ-program for instruction I_j
μ-IR
control signals needed in step 2 of instruction I_i
μ-program store (ROM)