Chapter 7: Memory Organization

Topics to be covered:

- Memory organization
- Memory Hierarchy
- Main Memory
- External Memory
- Cache Memory

7.1 Memory Organization

- The memory unit is an essential component in any digital computer since ***it is needed for storing programs and data***
- No one technology is optimal in satisfying the memory requirements for a computer system.
- It exhibits widest range of type, technology, organization, performance and cost.
- Not all accumulated information is needed by the CPU at the same time
- Therefore, it is more economical to use low-cost storage devices to serve as a backup for storing the information that is not currently used by CPU

7.1 Memory Organization

- **Some characteristics of Memory Systems**
  - **Location:** Refers to whether it is internal or external to computer
    - **Internal**: Directly accessible by CPU
      - main memory, cache, registers
    - **External:** Accessible by CPU through I/O module
      - magnetic disks, tapes, optical disks
  - **Access method**
    - **Sequential access:** Access to records is made in a specific linear sequence eg. Magnetic tape
    - ***Random access:*** the storage locations can be accessed in any order. e.g. RAM and ROM

7.1 Memory Organization

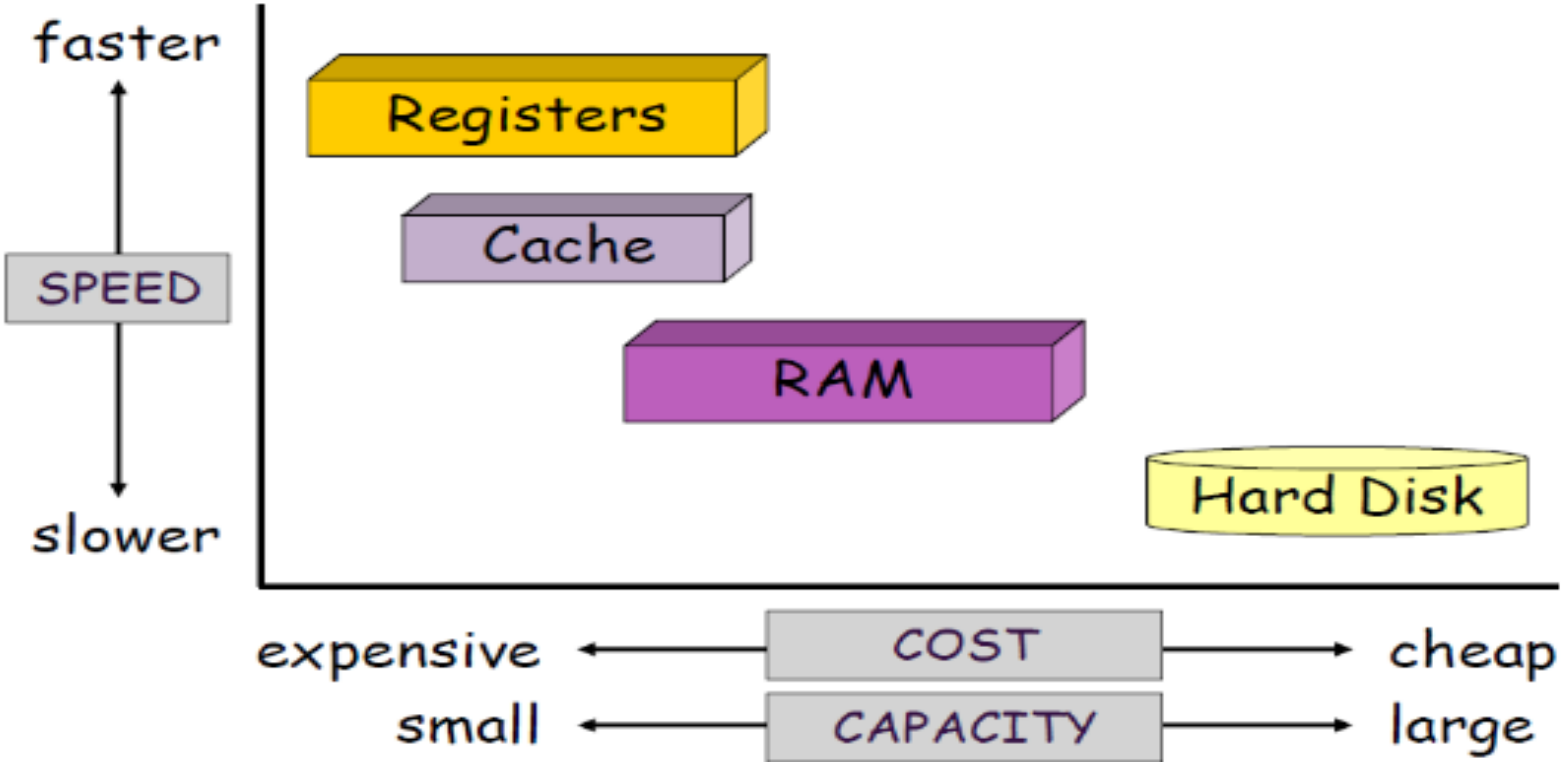
- **Some characteristics of Memory Systems**
  - **Performance:**
    - The average time required to reach a storage location in memory and obtain its contents is called the **access time**
    - **Access time (latency)** is the time between "requesting" data and getting it
    - **The access time = seek time + transfer time**
      - **Seek time:** time required to position the read-write head to a location
      - **Transfer time:** time required to transfer data to or from the device
    - **Transfer rate:** rate at which data can be moved into/out of a memory unit

7.2 Memory Hierarchy

- The design constraints on a computer's memory can be summed up by three questions:
  - How much is the capacity? **Storage capacity**
  - How fast? **Speed**
  - How expensive? **Cost**
- A variety of technologies are used to implement memory systems, and across this spectrum of technologies, the following relationships hold:
  - Faster access time, greater cost per bit
  - Greater capacity, smaller cost per bit
  - Greater capacity, slower access time

7.2 Memory Hierarchy

- The total memory of a computer system is organized as the hierarchy of memories as shown below.



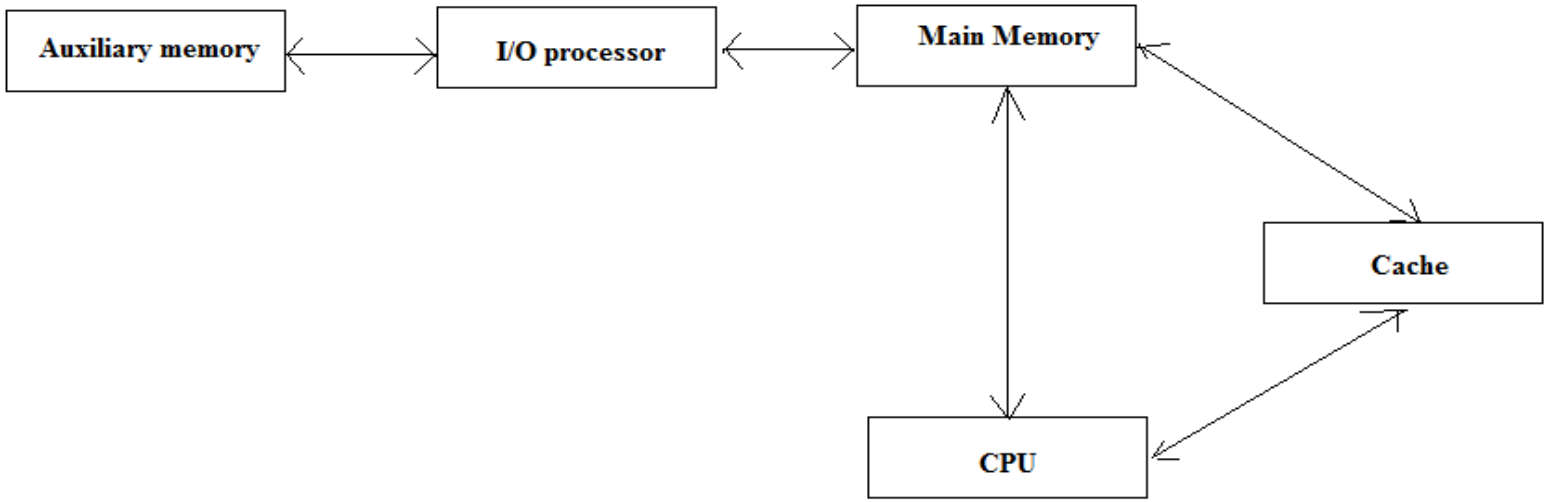
- Economic and performance are the basis for the hierarchy for memory organization

7.2 Memory Hierarchy

- As one goes down the hierarchy, the following occurs:
  - Cost per bit Decrease (cheaper)
  - Increasing storage capacity
  - Increasing access time (slower in speed)
  - Decreasing frequency of access of the memory by the processor
- The memory unit that directly communicate with CPU is called the ***main memory***
- Devices that provide backup storage are called ***auxiliary memory***
- The memory hierarchy system consists of all storage devices employed in a computer system from the slow by high-capacity **auxiliary memory** to a relatively faster **main memory**, to an even smaller and faster **cache memory**

7.2 Memory Hierarchy

- The main memory occupies a central position by being able to communicate directly with the CPU and with auxiliary memory devices through an I/O processor
- A special very-high-speed memory called **cache** is used to **increase the speed of processing** by making current programs and data available to the CPU at a rapid rate



7.2 Memory Hierarchy

- CPU logic is usually faster than main memory access time, with the result that processing speed is limited primarily by the speed of main memory
- The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations
- The typical access time ratio between cache and main memory is about 1 to 7
- Auxiliary memory access time is usually 1000 times that of main memory

7.3 Main Memory

- The memory unit that communicates directly with CPU is called ***main memory/Primary memory***.
- Most of the main memory in a general purpose computer is made up of RAM integrated circuits chips, but a portion of the memory may be constructed with ROM chips

- RAM– Random Access memory or read/write memory
  - Integrated RAM are available in two possible operating modes, S*tatic and Dynamic*
- ROM– Read Only memory

7.3 Main Memory

Random-Access Memory (RAM)

- Static RAM (SRAM)
  - Each cell stores bit with a six-transistor circuit.
  - Retains value indefinitely, as long as it is kept powered.
  - Relatively insensitive to disturbances such as electrical noise.
  - Faster and more expensive than DRAM and used for cache memory
- Dynamic RAM (DRAM)
  - Each cell stores bit with a capacitor and transistor.
  - Loses its stored information in a very short time (a few milliseconds) even though the power supply is on
    - Value must be refreshed every 10-100 ms.
  - Sensitive to disturbances.
  - Slower and cheaper than SRAM and used for main memory

7.3 Main Memory

Read Only Memory (ROM)

- ROM is used for storing programs that are **PERMENTLY** resident in the computer
- The ROM portion of main memory is needed for storing an initial program called a bootstrap loader.
- The ***bootstrap loader***((BIOS) is a program whose function is to start the computer software operating when power is turned on.
- Nonvolatile memory: The contents of ROM remain unchanged after power is turned off.

7.4 Auxiliary Memory/External Memory

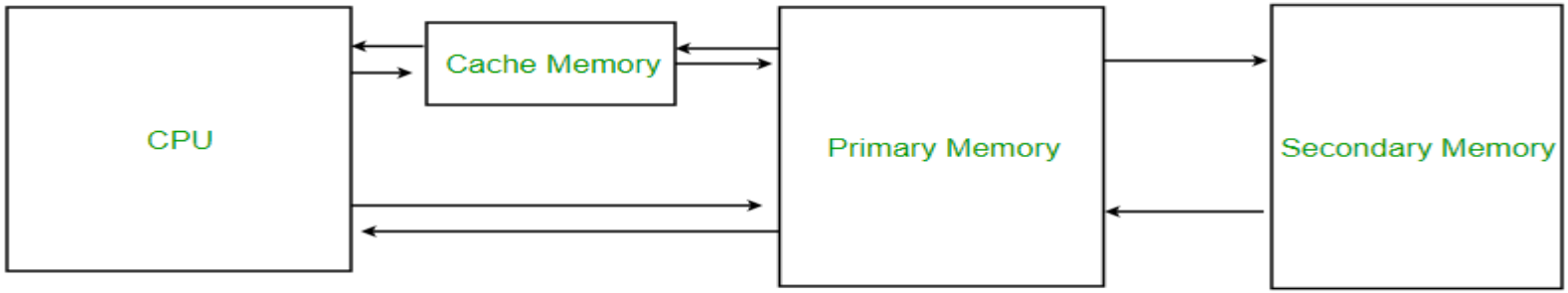
- Storage devices that provide backup storage are called **auxiliary/external memory/secondary memory.**
  - Used to store programs and data which are not needed immediately by the CPU, large data files, backup data
  - Not urgently needed data are stored here.
- It is non volatile
- Common used secondary memory includes: Magnetic disks, Optical disks, Magnetic tapes, etc.

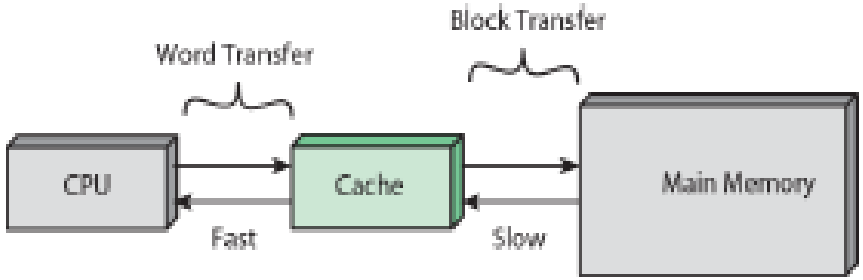
7.5 Cache memory

- **Cache Memory** is a special very high-speed memory.
- It is used to speed up and synchronize with high-speed CPU.
- Cache memory is costlier than main memory or secondary memory but more economical than CPU registers.
- Cache memory is an extremely fast memory type that acts as a buffer between RAM and the CPU.
- It holds frequently requested data and instructions so that they are immediately available to the CPU when needed.
- Cache memory is used to reduce the average time to access data from the Main memory.
  - The basic characteristic of cache memory is its fast access time
- The cache is a smaller and faster memory that stores copies of the data from frequently used main memory locations.

7.5 Cache memory

- Cache memory is placed between the CPU and the main memory.





7.5 Cache memory

- **Cache Performance:** When the processor needs to read or write a location in main memory, it first checks for a corresponding entry in the cache.
  - If the processor finds that the memory location is in the cache, a **cache hit** has occurred and data is read from the cache.
  - If the processor **does not** find the memory location in the cache, a **cache miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from main memory, then the request is fulfilled from the contents of the cache.
- When the CPU refers to memory and finds the word in cache, it is said to produce a **hit**.
- If the word is not found in the cache, it is in main memory and it counts as a **miss**.

7.5 Cache memory

- The performance of cache memory is frequently measured in terms of a quantity called **Hit ratio**.
  - **Hit ratio = hit / (hit + miss) = no. of hits/total accesses**
- **Cache Mapping:** There are three different types of mapping used for the purpose of cache memory which is as follows: Direct mapping, associative mapping, and set-Associative mapping. (you can read more about their detail)

Chapter 8: Input Output Organization

- Topics to be covered:
  - Peripheral Devices
  - I/O Interface
  - Modes of I/O Data Transfer
    - **Programmed I/O**
    - **Interrupt driven I/O**
    - **Direct Memory Access (DMA)**
  - I/O Processor

Peripheral Devices

- The **I/O subsystem** of a computer provides an efficient mode of communication between the central system and the outside environment. It handles all the input/output operations of the computer system.

Peripheral Devices:

- Input or output devices that are connected to computer are called peripheral devices.
- These devices are designed to read information into or out of the memory unit upon command from the CPU and are considered to be the part of computer system. These devices are also called peripherals.
- For example: Keyboards, display units and printers are common peripheral devices.

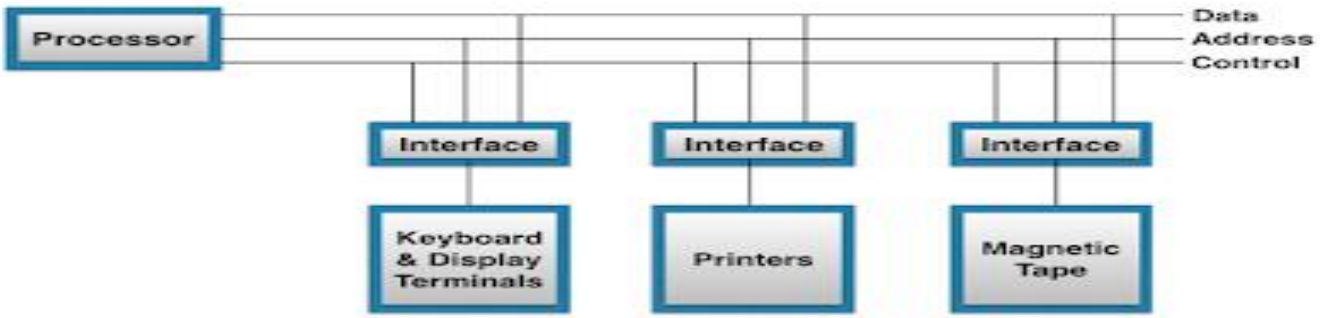
Peripheral Devices

There are three types of peripherals:

1. **Input peripherals** :
   - Allows user input, from the outside world to the computer. Example: Keyboard, Mouse etc.
2. **Output peripherals**:
   - Allows information output, from the computer to the outside world. Example: Printer, Monitor etc
3. **Input-Output peripherals**:
   - Allows both input(from outside world to computer) as well as, output(from computer to the outside world).
     - Example: Touch screen etc.

Input-Output Interface

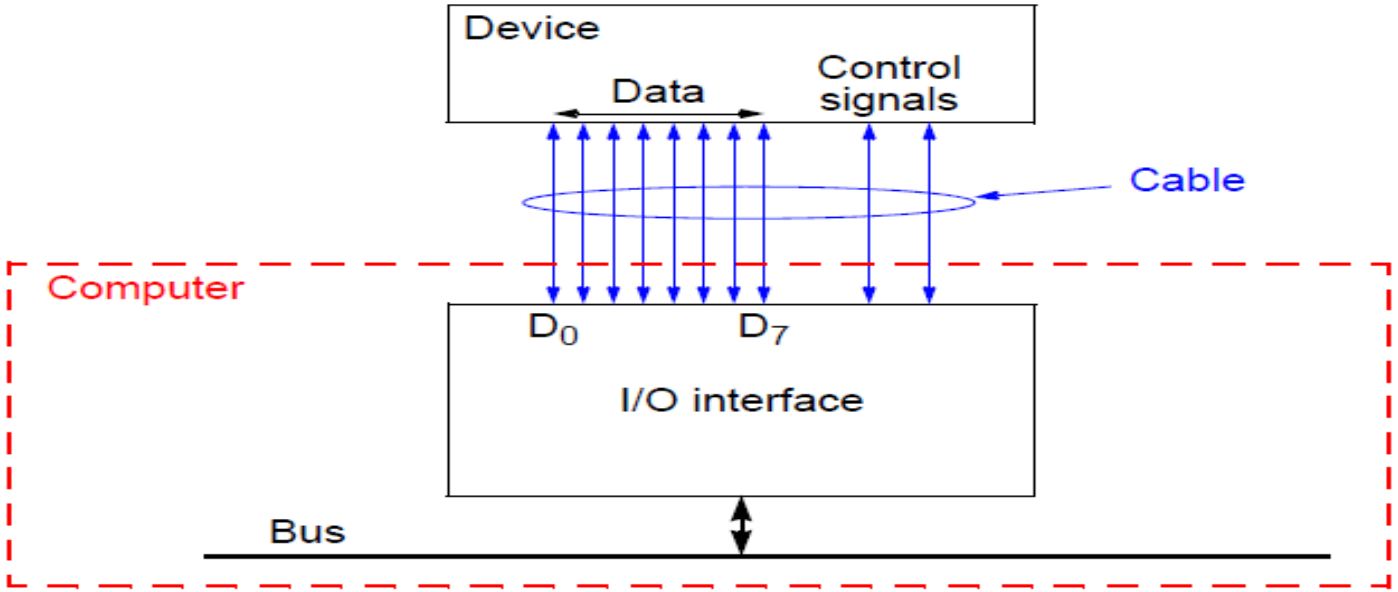
- Peripherals connected to a computer need special communication links for interfacing with CPU.
- In computer system, there are special hardware components between the CPU and peripherals to control or manage the input-output transfers.
- These components are called **input-output interface units** because they provide communication links between processor bus and peripherals.
- They provide a method for transferring information between internal system and input-output devices.



Connection of I/O Bus to I/O Device

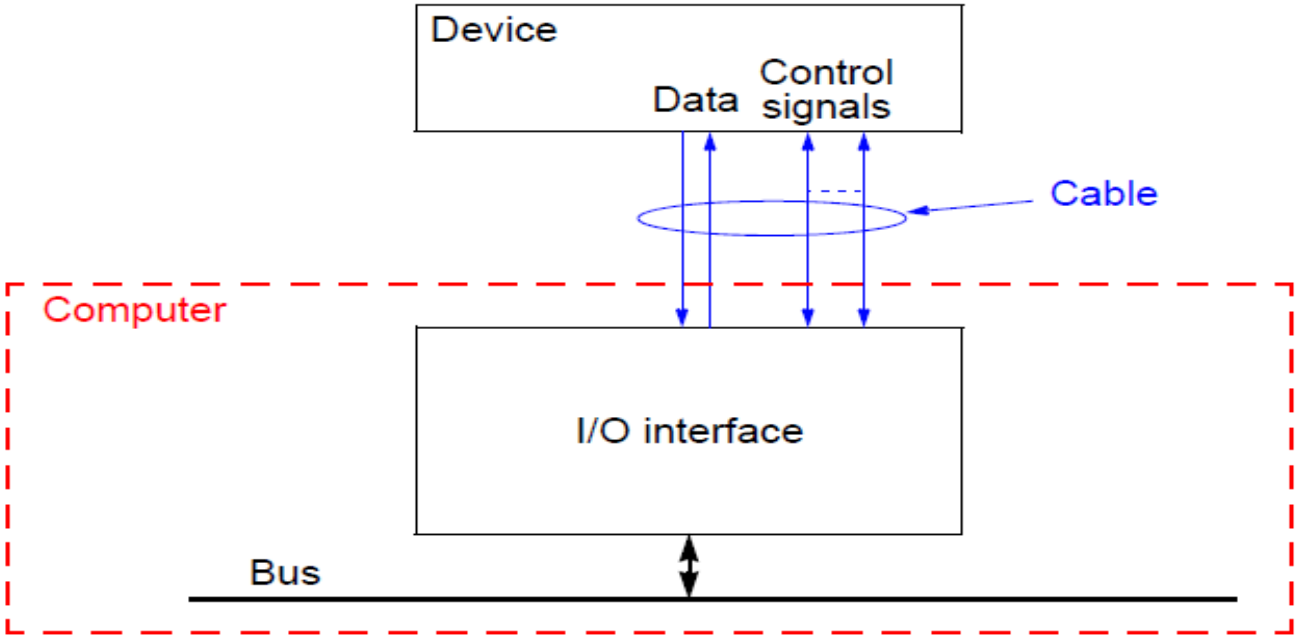
Input-Output Interface

- The transfer of data between two units may be done in parallel or serial.
- **Parallel Interface**
  - Data is transmitted with one wire assigned to each bit of the data: (an n bit message must be transmitted through n separate paths)
  - In parallel data transmission, each bit of the message has its own path and the total message is transmitted at the same time.
  - **Parallel transmission is faster but requires many wires.**



Input-Output Interface

- **Serial Interface**
  - In serial data transmission, each bit in the message is sent in sequence one at a time.
  - Data transmitted along one wire (for each direction).
  - Bits of the data are sent one after the other.
  - **Serial transmission is slower but less expensive since it requires less data wires.**



Input-Output Interface

- The Input/output Interface is required because there are exists many **differences between the CPU and each peripheral** while transferring information. Some of the major differences are:

  1. Peripherals are electromechanical and electromagnetic devices and their manner of operation is different from the operation of CPU.
  2. The data transfer rate of peripherals is usually **slower** than the transfer rate of CPU, and consequently a synchronization mechanism is needed.
  3. Data codes and formats in peripherals differ from the word format in the CPU and Memory.
  4. The operating modes of peripherals are differ from each other and each must be controlled so as not to disturb the operation of other peripherals connected to CPU.

Modes of Transfer (Input/output Mechanisms)

- Binary information received from an external device is usually stored in memory for later processing.
- Information transferred from the central computer into an external device originates in the memory.
- The CPU merely executes the I/O instructions and may accept the data temporarily, but the ultimate source or destination is the memory unit.
- Data transfer between the central computer and I/O devices may be handled in a variety of modes.
- Some modes use the CPU as an intermediate path; others transfer the data directly to and from the memory unit.

Modes of Transfer (Input/output Mechanisms)

- The CPU must have a way to pass information to and from an I/O device.
- **There are three approaches (Modes of I/O Data Transfer) available to communicate the CPU with I/O devices.**
  - **Programmed I/O**
  - **Interrupt driven I/O**
  - **Direct Memory Access (DMA)**
- Data transfer to and from the peripherals may be handled in one of the above three possible modes.

Modes of Transfer (Input/output Mechanisms)

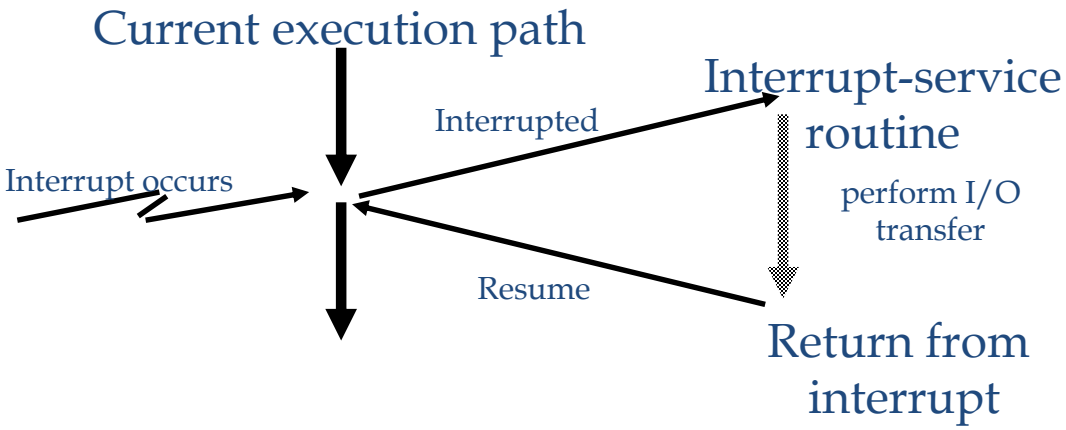
Programmed I/O: Special Instruction I/O

- This uses CPU instructions that are specifically made for controlling I/O devices.
- These instructions typically allow data to be sent to an I/O device or read from an I/O device.
  - Each data item transfer is initiated by the instruction in the program.
- Usually the program controls data transfer to and from CPU and peripheral.
- Transferring data under programmed I/O requires **constant monitoring of the peripherals by the CPU**.

Modes of Transfer (Input/output Mechanisms)

Interrupt driven I/O

- When an I/O device is ready to send (receive) data to (from) the CPU, **it signals (or interrupts) the CPU for its attention**.
- After receiving the interrupt signal, the CPU stops the task which it is processing and service the I/O transfer and then returns back to its previous processing task.
  - As soon as the CPU finishes the current instruction, it transfers its execution to an interrupt-service routine which responds to the external interrupt.



Interrupts

- Interrupt is the mechanism by which the processor is made to transfer control from its current program execution to another program having **higher priority**.
- **Interrupt** is the method of creating a temporary halt during program execution and allows peripheral devices to access the microprocessor.
- **An *interrupt* is generated by a signal from hardware or software, and it may occur at random times during the execution of a program.**
  - An interrupt break the sequence of operation
  - An interrupt cause a temporary halt in the execution of program
- It allows an application program to be suspended, in order that a variety of interrupt conditions can be serviced and later resumed.

Modes of Transfer (Input/output Mechanisms)

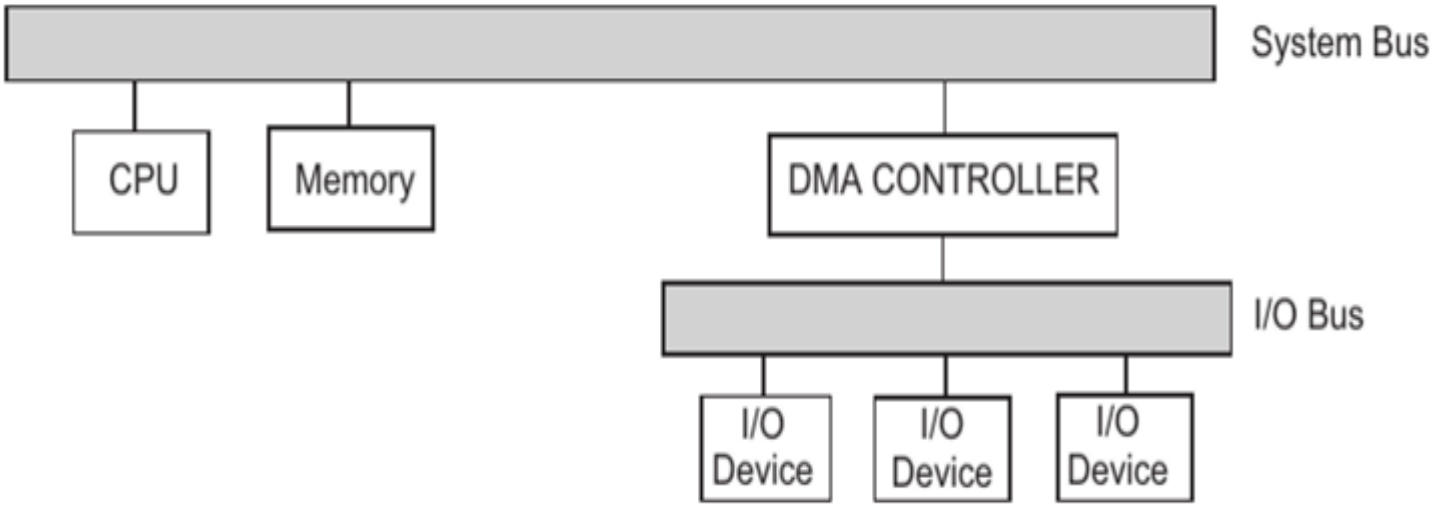
Direct Memory Access (DMA)

- The data transfer from I/O devices to the memory or from the memory to I/O devices **through the accumulator is a time consuming process.** For this situation, the **DMA** technique is preferred.
- **I/O device exchanges data directly with memory**
- It is a technique of **transferring data between memory and I/O devices without CPU intervention**.
  - CPU sets up transfer with DMA controller; then transaction occurs without CPU

Modes of Transfer (Input/output Mechanisms)

Direct Memory Access (DMA)

- Independent dedicated I/O processors (*smart DMA controllers*) are used in computer systems to communicate with I/O devices.



Input Output Processor (IOP)

- An 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.