Unit 3: Computer Memory (4 Hrs.)

Introduction

Computer memory is a fundamental component that stores data and instructions for the computer's processor to access and execute. Computer memory stores program operations and data while a program is being executed. It plays a crucial role in determining the overall performance and capabilities of a computer system.

There are several types of memory, including registers, cache, RAM, and virtual memory.

Memory Representation

- Bits and Bytes:
  - The basic unit of memory is a bit, which can store a single binary value (0 or 1).
  - A group of 8 bits is called a byte, which is commonly used to represent characters, numbers, and other data types.

    Byte ⇒ 8 bits

    Kilobyte (KB) ⇒ 1024 bytes

    Megabyte (MB) ⇒ 1024 KB

    Gigabyte (GB) ⇒ 1024 MB

    Terabyte (TB) ⇒ 1024 GB

    Petabyte (PB) ⇒ 1024 TB

    Exabyte (EB) ⇒ 1024 PB

    Zettabyte (ZB) ⇒ 1024 EB

    Yottabyte (YB) ⇒ 1024 ZB
- Memory Addresses:
  - Each memory location has a unique address that allows the processor to locate and retrieve data.
  - Addresses are typically expressed in hexadecimal format (base 16)

Example:

Imagine a small memory with a capacity of 16 bytes:

The addresses would range from 0x00 to 0x0F (0 to 15 in decimal).

If a variable named "age" is stored at address 0x05, the CPU would use this address to read or write its value.

- Memory Types:
    - Primary Memory (Main Memory):
        - Directly accessible to the CPU.
        - Volatile (loses data when power is turned off).
        - Commonly used types include:
            - RAM (Random Access Memory)
            - ROM (Read-Only Memory)

Note: While ROM shares some characteristics with primary memory, such as direct CPU access and persistent data storage, its read-only nature distinguish it from the dynamic role of RAM in managing running programs. Therefore, it's more accurate to consider ROM as a **special type of non-volatile memory** *that occupies a unique space between primary and secondary storage.*

-
    - Secondary Memory (Auxiliary Storage):
        - Non-volatile (retains data even without power).
        - Used for long-term storage.
        - Examples include:
            - Hard disk drives (HDDs)
            - Solid-state drives (SSDs)
            - Optical drives (CDs, DVDs)

- Flash drives

Memory Hierarchy

The main purpose of memory hierarchy is to bridge the gap between the speed of the CPU and the slower access times of primary and secondary memory. The lesser the access time, the faster is the speed of memory.

The hierarchical arrangement of memory in computer in such a way that the smallest and fastest memory get placed near to CPU and largest and slowest memory get placed far from CPU is called the ***memory hierarchy***.

Two main goals to use Memory hierarchy

- To maximize access speed
- To minimize the per bit costs

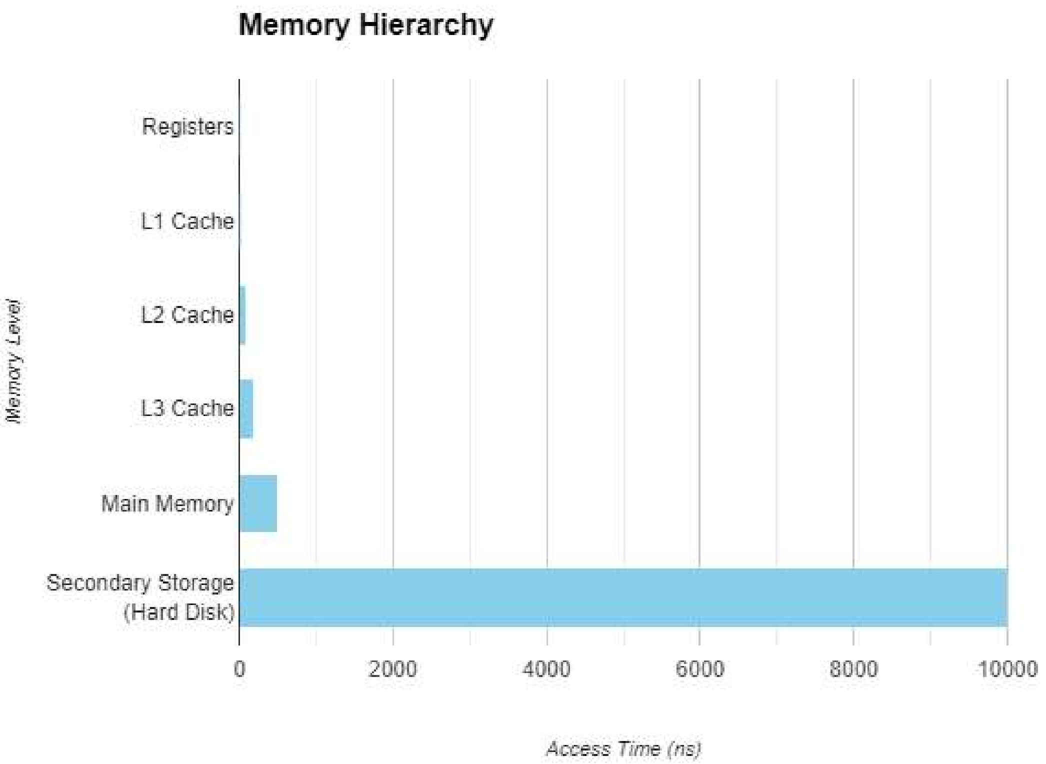

Figure 3: Memory Hierarchy

Level 1: Registers

- Fastest and smallest: Located within the CPU itself.
- Hold temporary data: Frequently used values, intermediate results, and address calculations.
- Size: Few bytes (e.g., 32 or 64) to a few kilobytes. Cost: Most expensive per unit of storage.

Level 2: Cache Memory

- Faster than main memory: Acts as a buffer between CPU and main memory.
- Stores frequently accessed data: Recently used values or instructions anticipated for future use.
- Size: Kilobytes to tens of megabytes.
- Cost: Expensive but less than registers.

Level 3: Main Memory (RAM)

- Primary storage: Holds data and programs currently being used.
- Faster than secondary storage: Accessed directly by the CPU.
- Size: Gigabytes to tens of gigabytes.
- Cost: More affordable than cache memory.

Level 4: Secondary Storage (Hard Drives, SSDs)

- Non-volatile: Retains data even when the computer is off.
- Slower than main memory: Accessed through the main memory by the CPU.
- Larger capacity: Hundreds of gigabytes to terabytes.
- Cost: Most affordable per unit of storage.

Level 5: Tertiary Storage (Optical Discs, Magnetic Tapes)

- Slower than secondary storage: Used for long-term archival or backup.
- Size: Gigabytes to terabytes.
- Cost: Less expensive than secondary storage, but slower access times.

Some important definitions:

- Memory Access Time: The time it takes to locate and retrieve data from memory.
- Memory Capacity: The amount of data that can be stored in memory.
- Memory Latency: The delay between a request for data and its delivery.
- Memory Bandwidth: The rate at which data can be transferred between memory and the CPU.

CPU Registers:

Computers have several storage locations called registers. Registers are the part of the control unit and ALU rather than memory. Registers are a number of small, high speed memory units that hold data and instructions temporarily during processing.

- Location: Built-in within the CPU itself.
- Size: Extremely small, with only a few bytes per register.
- Access Speed: Fastest memory type, with access times in nanoseconds.
- Purpose: Hold data and instructions currently being processed by the CPU.
- Benefits:
    - Provide immediate access to critical data and instructions for the fastest possible execution.
    - Essential for complex calculations and program control flow.
- Drawbacks:
    - Extremely limited size, restricting the amount of data they can hold.
    - Data is lost when the program completes or the CPU switches tasks.

Cache Memory:

Cache memory is a high-speed storage component that bridges the speed gap between the CPU and main memory (RAM). It temporarily stores frequently accessed data and instructions, enabling faster retrieval and boosting overall system performance.

- **Location**: Situated between the CPU and main memory (RAM).
- **Size**: Larger than registers, typically ranging from kilobytes to megabytes.
- **Access Speed**: Faster than main memory but slower than registers.
- **Purpose**: Acts as a buffer, storing frequently accessed data and instructions from main memory for quicker retrieval by the CPU.
- **Benefits**:
    - Reduces the need to access slower main memory, boosting overall performance.

    - Improves performance for programs or tasks that access the same data repeatedly.
- **Drawbacks**:
    - Limited size compared to main memory, requiring careful selection of data to store.
    - Data needs to be copied between cache and main memory, adding overhead.

Primary Memory

- **Active Workspace**: The main memory that directly interacts with the CPU, storing actively running programs, data, and operating system instructions.
- **Volatile**: Loses its contents when the computer is powered off.
- **Fast Access**: Offers relatively fast data access speeds, enabling swift program execution.
- **Capacity**: Typically ranges from 4GB to 16GB in modern systems, but can be higher for demanding tasks.
- **Limited Size**: Constrains the number of programs and data that can be held simultaneously.
- The primary memory or main memory of computer is divided into RAM and ROM.

Random Access Memory (RAM)

It allows the computer to store data for immediate manipulation and to keep track of what is currently being processed. It is the place in a computer where the operating system, application programs, and data in current use are kept so that they can be accessed quickly by the computer's processor. RAM is made up of several small storage areas called cell. Each of cells is identified by a number, called address of that particular cell. RAM also refers to read and write memory, that is CPU, can both write data randomly into and read data from RAM.

RAM is volatile memory because the data and instruction will remain there only as the computer has electric power. Every time, when the power is switched on, the system files are load into this memory from the storage device such as a hard disk. Therefore, it is also called the loading memory.

The size of RAM is measured in MB or GB. The size is limited due to its high cost. RAM affects the speed and power of a computer. More the RAM, the better it is. RAM is a microchip implemented using semiconductors.

Different types of RAM

SRAM (Static Random Access Memory)

- **Technology:** Uses latches made of transistors to store each bit of data. No refresh needed.
- **Speed:** Faster than DRAM, offering quick access times for data retrieval and writing.
- **Cost:** More expensive than DRAM due to its more complex circuitry.
- **Density:** Lower than DRAM, meaning it stores less data per unit area.
- **Application:** Used in smaller, performance-critical applications like CPU cache, embedded systems, and high-speed networking equipment.

DRAM (Dynamic Random Access Memory)

- **Technology:** Uses capacitors to store each bit of data. Requires periodic refresh (thousands of times per second) to maintain data integrity.
- **Speed:** Slower than SRAM, but still fast enough for most computing tasks.
- **Cost:** Less expensive than SRAM due to its simpler design.
- **Density:** Higher than SRAM, allowing for larger capacities and lower cost per bit.
- **Application:** Used as the main system memory (RAM) in most computers, where cost and capacity are more important than ultimate speed.

- Difference between SRAM and DRAM

SRAM	DRAM
1. It is made up of transistors and flip flops.	It is made up of with capacitors and few transistors.
2. For single block of memory six transistors are used.	For single block of memory only one transistor is used.
3. It has no charge leakage property so, does not need to be power-refreshed.	It has charge leakage property so, to be refreshed after each read operation.
4. Refreshing circuit is not implemented	Refreshing circuit is implemented
5. It utilizes less power.	It utilizes more power.
6. It is more expensive.	It is less expensive.
7. It is faster than DRAM.	It is slower than SRAM.

8. It has low density.	It has high density.

Read Only Memory (ROM)

ROM acts as the foundation for a computer's initial operation, providing essential instructions and data that remain constant throughout its life life. ROM is also referred to as non-volatile memory because any data stored in ROM will remain there even power is turned off. ROM comes programmed by the manufacturer.

The ROM memory chip stores the Basic Input Output System (BIOS). BIOS provide the processor with the information required to boot the system. It provides the system with the settings and resources that are available on the system.

While traditional ROM is unchangeable, advancements have led to various types that offer different levels of re-programmability for specific applications.

Types of ROM

- **PROM (Programmable Read-Only Memory):** Can be programmed once using special equipment called PROM programmer or burner. Once the PROM has been programmed, the information written is permanent and cannot be erased or deleted, so PROM is the ROM that can be written once. The PROM chip is often called 1 TP i.e. one-time programmable chip because we cannot convert a 0 back to 1.
- **EPROM (Erasable Programmable Read-Only Memory):** Can be erased using ultraviolet light and then reprogrammed. It is reconfigured using EPROM programmer. To erase the content stored in EPROM, one need to remove the chip from the system. In this ROM selective programming cannot be done.
- **EEPROM (Electrically Erasable Programmable Read-Only Memory):** Can be electronically erased and reprogrammed multiple times. EEPROM chip need not be taken out of the computer or electronic device of which it is part when new program or data needs to be written on it. Selective programming can be done using to an EEPROM chip. The user can alter the value of certain cells without needing to erase the programming on other cells. Erasing can be done byte by byte in EEPROM.
- **Flash ROM:** Also called flash BIOS or flash memory is a special type of EEPROM commonly used in storage devices like USB drives and solid-state drives (SSDs). Many modern PCs have their BIOS stored on a flash memory chip so that it can be easily updated, if necessary, such BIOS is sometimes called flash BIOS.

Secondary Memory

- **Long-Term Storage**: Provides non-volatile storage for data that needs to be preserved even when the computer is powered off.
- **Slower Access**: Access speeds are slower than RAM, but capacities are much larger.
- **Types**:
    - Hard Disk Drives (HDDs): Traditional magnetic storage, offering large capacities at a lower cost.
    - Solid State Drives (SSDs): Faster, more reliable, and compact, but generally more expensive per gigabyte.
    - External Storage: Flash drives, memory cards, and optical drives (CDs, DVDs) for portability and backup.

The difference between primary and secondary memory is as follows:

Primary Memory	**Secondary Memory**
i. Primary memory is the memory that is directly accessed by the CPU to store and retrieve information.	i. Secondary Memory is not accessible directly by the CPU.
ii. Primary memory is accessed using address and data buses by the CPU.	ii. Secondary memory is accessed using input/output channels.
iii. It stores data and program under execution.	iii. It stores the data and program which is currently in non-executing stage.
iv. It is temporary and volatile except ROM.	iv. It is permanent and non-volatile.
v. It is fast and expensive.	v. It is slow and very inexpensive as compared to primary memory.
vi. It has smaller storage capacity.	vi. Its size is large
vii. It is also known as internal or main memory.	vii. It is also known as external or auxiliary or backup storage.
viii. E.g. RAM, ROM etc.	viii. E.g. Magnetic disk, optical disk, Tape etc.

Access Types of Storage Devices

- **Sequential Access**: Data is accessed in a linear sequence, requiring reading through preceding data to reach the desired information. The data access method is less expensive than other methods because it uses magnetic tape, which is cheaper than disks. The disadvantage is that the searching for data is slower.
    - Examples: Tape drives, older magnetic storage devices.

  Working of Sequential Access:

    - Similar to reading a book from page to page.
    - To reach a specific point, you must go through all preceding pages.
    - Often used for archival or backup purposes where access speed is less critical.

- **Direct Access (Random Access):** Data can be accessed directly, regardless of its physical location, enabling quick retrieval of any desired piece of information by locating the data's address. In this method, information is viewed as a numbered sequence of blocks and there is no restriction on the order of reading or writing in direct access method. It is ideal for application such as airline reservation systems or computer-based directory assistance operation.

  Based on access, magnetic tapes are sequential access devices, and magnetic disks, optical disks and magneto-optical disks are direct access devices.

    - Examples: Hard disk drives (HDDs), solid-state drives (SSDs), flash drives, optical drives (CDs, DVDs).

  Working of Direct Access:

    - Analogous to finding a song on a CD or a specific chapter in a book using a table of contents.
    - You can jump directly to the desired location without reading through everything beforehand.
    - Essential for modern computing where quick data retrieval is paramount.

Magnetic Tape

- **Storage Medium**: Reel of plastic coated with magnetic particles.
- **Access Type**: Sequential, which means that the tape needs to rewind or move forward to the location where the requested data is positioned in the magnetic tape.
- **Working Mechanism:** Magnetic tape is made up of Mylar plastic coated with iron oxide on only one side of tape. Data is stored in frames across the width of the tape. It stores information by converting electrical signals into patterns of magnetism.

  The working process includes **Recording** and **Playback:**

  Recording:

  - The tape, coated with tiny magnetic particles, is pulled past a recording head.
  - The electrical signal (e.g., audio, data) is fed to the head, creating a varying magnetic field.
  - This field aligns the magnetic particles on the tape in a specific pattern, mimicking the original signal.

  Playback:

  - The magnetized tape is passed over a playback head.
  - The varying magnetic patterns on the tape induce a corresponding electrical signal in the head.
  - This electrical signal is then amplified and converted back into its original form (e.g., sound played through speakers, data sent to a computer).

- **Different type of magnetic tape found**:

  - Half inch tape with 60 MB to 400 MB data storage capacity.
  - Quarter inch tape with 40 MB to 5 GB data storage capacity.
  - 4 mm DAT tape with 2 GB to 24 GB data.

- **Advantages**:

  - High Capacity: Can store massive amounts of data (terabytes) at a very low cost per bit of storage and no complicated software is required for file handling.
  - Tapes can be erased and reused many times.
  - Compact size and light weight

    - Durable: Tapes can last for decades under proper conditions.
    - Archiving: Ideal for long-term data storage and backups.
- **Disadvantages**:
    - Slow Access: Retrieving specific data takes a long time due to sequential access.
    - Vulnerable to Wear: Tapes can degrade over time and are susceptible to physical damage.
    - Not for Everyday Use: Not efficient for frequent data access due to slow speeds.

Magnetic Disk

- **Storage Medium**: Circular metal plate/plotters coated with magnetic material such as iron oxide or ferrous oxide on the both side which can be magnetized. Information is recorded on the disk surface in the form of invisible tiny magnetic spots. The presence of a magnetized spot represents a 1 bit and its absence represents a 0 bit.
- **Access Type**: Direct. Data can be accessed randomly from magnetic disk.
- **Floppy disk, hard disk** and **zip disk** are the different types of magnetic disks.
- **Advantages**:
    - Faster Access: Data can be retrieved much faster than with magnetic tape due to direct access of data.
    - Widely Used: Suitable for both on-line and off-line storage of data. Standard storage format for personal computers and servers.
    - Very large amount of data can be stored in a small storage space.
    - Can be erased and reused many times.
    - Relatively Durable: Disks can last for several years, although they are susceptible to physical damage.
- **Disadvantages**:
    - Cost of magnetic disks storage is more expensive than the cost of magnetic tapes.
    - Prone to Fragmentation: Continuous data writes can fragment files, slowing down access further.

- Magnetic disks must be stored in dust free environment in order to protect them from crashing down.

- **Working mechanism of a magnetic disk.**

  Components:

  - **Platters:** Flat, circular disks coated with a magnetic material.
  - **Read/Write Head:** A device that moves across the platters, reading and writing data.
  - **Disk Arm:** Arm that positions the read/write head over the desired location on the platter.
  - **Tracks:** Concentric circles on the platter surface, storing data.
  - **Sectors:** Smaller segments within a track, storing portions of data.

  Process:

  1. **Positioning:**
     - The disk spins at high speed (60-150 times per second).
     - The disk arm moves the read/write head to the desired track (seek time).
     - The head waits for the correct sector to rotate beneath it (latency time).
  2. **Reading/Writing:**
     - **Reading:** The head detects the magnetic patterns on the sector, translating them into electrical signals and then into usable data (e.g., documents, music).
     - **Writing:** The head generates magnetic patterns on the sector, storing data based on electrical signals received.
  3. **Data Transfer:** The actual transfer of data between the disk and the computer.

  Overall Access Time: The sum of seek time, latency time, and data transfer rate determines the total time it takes to access data on the disk.

Optical Disk

- **Storage Medium**: Plastic disc with reflective layer and protective coating. Data is stored as microscopic pits and lands.
  The pits are the tiny reflective bumps that are created with the laser beam. The lands are flat areas separating the pits. The land reflects the laser light, which is read as binary digit

1. A pit absorbs or scatters light, which is read as binary digit 0. A high-powered laser beam reads the pits and low powered laser beam reads the data from the disk.

Optical disks can store large amount of data, up to 25 GB, in a small space. Commonly used optical disks store 600–700 MB of data. The access time for an optical disk range from 100 to 200 m s.

- **Access Type**: Direct. Data can be accessed directly at any location on the disc.
- **Advantages**:
  - The per bit cost of storage for optical disks is very low because of their low cost and enormous storage density.
  - Durable: Resistant to dust, scratches, and magnetic fields.
  - Due to their compact size and light weight, optical disks are portable, easy to handle and store.
  - Read-Only or Read-Write: Available in both write-once and rewritable formats.
  - Optical disks are more reliable storage medium than magnetic tape or magnetic disks.
- **Disadvantages**:
  - Limited Capacity: Compared to tape and hard drives, optical disks have lower capacities.
  - Slower Access: Read and write speeds are slower than magnetic disks and SSDs.
  - Optical disk requires a complicated drive mechanism.
- **Examples of Optical Disks**
  - CD-ROM
  - DVD-ROM
  - Recordable Optical Disk: CD-R, CD-RW, DVD-R

Magneto-Optical Disk

- **Storage Medium**: Combines magnetic and optical recording technologies. Magneto-optical disks use laser beam to read data and magnetic field to write data to disk cartridge. The surface of the cartridge contains tiny embedded magnets. These are optical disks where data can be written, erased and rewritten.
- **Access Type**: Direct. Data can be accessed directly at any location on the disc.
- **Advantages**:

  - High Capacity: Offers similar capacities to hard drives.
  - Rewritable: Can be erased and rewritten multiple times.
  - Durable: Resistant to dust, scratches, and magnetic fields.
- **Disadvantages**:
  - Lower Performance: Read and write speeds are slower than hard drives and SSDs.
  - Costly: Traditionally more expensive than other storage options.
  - Less Common: Less widely used and supported compared to other storage technologies.

How the Computer uses its memory

1. **Computer Startup Launch**:
   - When the computer is turned on it checks its hardware (using data from ROM).
   - The BIOS, loaded from ROM, provides basic information about the system. The operating system is then loaded from the hard drive into RAM for faster access by the CPU.
2. **Program Launch**:
   - When you launch a program, it's initially loaded from secondary storage (HDD/SSD) into RAM.
3. **Active Data and Instructions**:
   - RAM becomes the primary workspace for the program, holding its data, instructions, and operating system components.
4. **Data Access**:
   - The CPU constantly accesses data and instructions from RAM for processing.
5. **Cache Optimization**:
   - Frequently used data and instructions are copied from RAM to the cache for even faster access (reducing RAM access).
6. **Data Swap and Saving**:
   - When needed, data can be swapped between RAM and secondary storage to make room for other programs or tasks.

- Saving files involves writing data from RAM back to secondary storage for permanent storage.