Memory hierarchy

The Computer memory hierarchy looks like a pyramid structure which is used to describe the differences among **memory types**. It separates the computer storage based on hierarchy.

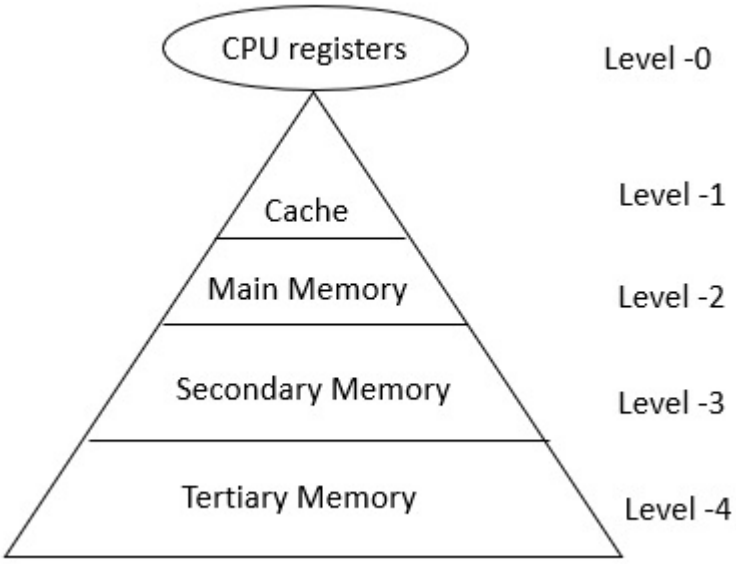
Level 0: CPU registers

Level 1: Cache memory

Level 2: Main memory or primary memory

Level 3: Magnetic disks or secondary memory

Level 4: Optical disks or magnetic types or tertiary Memory



In Memory Hierarchy the cost of memory, capacity is inversely proportional to speed. Here the devices are arranged in a manner Fast to slow, that is form register to Tertiary memory.

Let us discuss each level in detail:

Level-0 − Registers

The registers are present inside the **CPU**. As they are present inside the CPU, they have least access time. Registers are most expensive and smallest in size generally in kilobytes. They are implemented by using Flip-Flops.

Level-1 − Cache

Cache memory is used to store the segments of a program that are frequently accessed by the processor. It is expensive and smaller in size generally in Megabytes and is implemented by using static **RAM**.

Level-2 − Primary or Main Memory

It directly communicates with the CPU and with auxiliary memory devices through an I/O processor. **Main memory** is less expensive than cache memory and larger in size generally in Gigabytes. This memory is implemented by using **dynamic RAM**.

Level-3 − Secondary storage

Secondary storage devices like **Magnetic Disk** are present at level 3. They are used as backup storage. They are cheaper than main memory and larger in size generally in a few TB.

Level-4 − Tertiary storage

Tertiary storage devices like magnetic tape are present at level 4. They are used to store removable files and are the cheapest and largest in size (1-20 TB).

Let us see the memory levels in terms of size, access time, bandwidth.

Level	**Register**	**Cache**	**Primary memory**	**Secondary memory**
Bandwidth	4k to 32k MB/sec	800 to 5k MB/sec	400 to 2k MB/sec	4 to 32 MB/sec
Size	Less than 1KB	Less than 4MB	Less than 2 GB	Greater than 2 GB
Access time	2 to 5nsec	3 to 10 nsec	80 to 400 nsec	5ms
Managed by	Compiler	Hardware	Operating system	OS or user

Why memory Hierarchy is used in systems?

Memory hierarchy is arranging different kinds of storage present on a computing device based on speed of access. At the very top, the highest performing storage is CPU registers which are the fastest to read and write to. Next is cache memory followed by conventional DRAM memory, followed by disk storage with different levels of performance including SSD, optical and magnetic disk drives.

Different Types of RAM (Random Access Memory)

Memory plays an important component in determining the performance and efficiency of a system. In between various types of memory, Random Access Memory (RAM) stands out as a necessary component that enables computers to process and store data temporarily. In this article, we will explore the world of RAM, exploring its definition, types, and characteristics, as well as its significance in modern computing.

What is RAM?

Random Access Memory, is a type of computer memory that allows data to be read and written randomly, meaning that the computer can access any location in the memory directly rather than having to read the data in a specific order. This makes RAM an essential component of a computer system, as it enables the CPU to access data quickly and efficiently.

RAM is volatile in nature, which means if the power goes off, the stored information is lost. RAM is used to store the data that is currently processed by the CPU. Most of the programs and data that are modifiable are stored in RAM.

The block diagram of the RAM chip is given below:

Types of RAM?

Mainly RAM have 2types

- SRAM (Static RAM)
- DRAM (Dynamic RAM)

What is SRAM?

The SRAM memories consist of circuits capable of retaining the stored information as long as the power is applied. That means this type of memory requires constant power. SRAM memories are used to build Cache Memory.

SRAM Memory Cell

Static memories(SRAM) are memories that consist of circuits capable of retaining their state as long as power is on. Thus this type of memory is called **volatile memory**.

What is DRAM?

DRAM stores the binary information in the form of electric charges applied to capacitors. The stored information on the capacitors tends to lose over a period of time and thus the capacitors must be periodically recharged to retain their usage. DRAM requires refresh time. The main memory is generally made up of DRAM chips.

DRAM Memory Cell

Though SRAM is very fast, it is expensive because of its every cell requires several transistors. Relatively less expensive RAM is DRAM, due to the use of one transistor and one capacitor in each cell. Information is stored in a DRAM cell in the form of a charge on a capacitor and this charge needs to be periodically recharged.

Understanding the different types of RAM is crucial for grasping how memory works in computers. RAM comes in various forms, including SRAM and DRAM, each serving different purposes within a computer system.

Types of DRAM

There are mainly 5 types of DRAM.

- **Asynchronous DRAM (ADRAM):** The DRAM described above is the asynchronous type of DRAM. The timing of the memory device is controlled asynchronously. A specialized memory controller circuit generates the necessary control signals to control the timing. The CPU must take into account the delay in the response of the memory.
- **Synchronous DRAM (SDRAM):** These RAM chips' access speed is directly synchronized with the CPU's clock. For this, the memory chips remain ready for operation when the CPU expects them to be ready. These memories operate

at the CPU-memory bus without imposing wait states. SDRAM is commercially available as modules incorporating multiple SDRAM chips and forming the required capacity for the modules.

- **Double-Data-Rate SDRAM (DDR SDRAM):** This faster version of SDRAM performs its operations on both edges of the clock signal; whereas a standard SDRAM performs its operations on the rising edge of the clock signal. Since they transfer data on both edges of the clock, the data transfer rate is doubled. To access the data at a high rate, the memory cells are organized into two groups. Each group is accessed separately.
- **Rambus DRAM (RDRAM):** The RDRAM provides a very high data transfer rate over a narrow CPU-memory bus. It uses various speedup mechanisms, like synchronous memory interface, caching inside the DRAM chips and very fast signal timing. The Rambus data bus width is 8 or 9 bits.
- **Cache DRAM (CDRAM):** This memory is a special type of DRAM memory with an on-chip cache memory (SRAM) that acts as a high-speed buffer for the main DRAM.

Difference Between SRAM and DRAM

The below table lists some of the differences between SRAM and DRAM.

SRAM	**DRAM**
SRAM stands for Static Random Access Memory.	DRAM stands for Dynamic Random Access Memory.
Uses a flip-flop circuit to store data	Uses a capacitor and a transistor to store data

SRAM	DRAM
SRAM has a lower access time, so it is faster compared to DRAM.	DRAM has a higher access time, so it is slower than SRAM.
SRAM has long data life.	DRAM has short data life.
SRAM has a storage capacity of 1 MB to 16 MB in most cases.	DRAM, which is often found in tablets and smartphones, has a capacity of 1 GB to 2 GB
SRAM is costlier than DRAM.	DRAM costs less compared to SRAM.
SRAM provides faster speed of data read/write.	DRAM provides slower speed of data read/write.
SRAM requires a constant power supply, which means this type of memory consumes more power.	DRAM offers reduced power consumption due to the fact that the information is stored in the capacitor.
Good choice for applications that may be exposed to extreme temperatures.	Not suitable for such applications.
Due to complex internal circuitry, less storage is available compared to the	Due to the small internal circuitry in the one-bit memory cell of DRAM, a large storage capacity is available.

SRAM	DRAM
same physical size of a DRAM memory chip.	
SRAM has low packaging capacity.	DRAM has a high packaging density.
SRAM is used in cache memories.	DRAM is used in main memories.
SRAM does not require refresh time.	DRAM requires periodic refresh time.
SRAMs are used as cache memory in computer and other computing devices.	DRAMs are used as main memory in computer systems.

Computer Memory

Memory is the electronic storage space where a computer keeps the instructions and data it needs to access quickly. It's the place where information is stored for immediate use. Memory is an important component of a computer, as without it, the system wouldn't operate correctly. The computer's operating system (OS), hardware, and software all rely on memory to function properly.

Computer memory functions similarly to the human brain, storing data, information, and instructions. It acts as a storage unit or device where data to be processed and the instructions necessary for processing are kept. Both input and output data can be stored in memory.

How Computer Memory Communicates With the CPU ?

Computer memory communicates with the CPU through a structured system of electronic pathways and controllers, enabling the CPU to fetch and store data rapidly and efficiently. Here's a detailed breakdown:

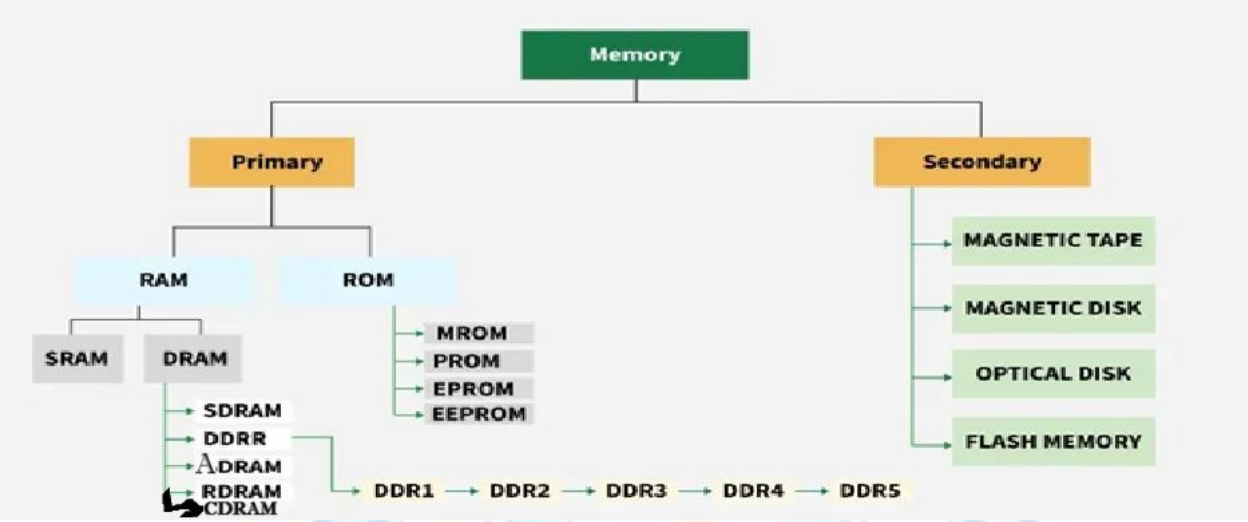
- **SystemBusStructure:**
  The main channel for communication between the CPU and memory is the *system bus*, which is a collection of three types of buses:
    - **Data Bus:** Transfers the actual data between CPU and memory.
    - **Address Bus:** Carries the memory address that specifies where data should be read from or written to.
    - **Control Bus:** Sends signals that coordinate and control the activity, such as indicating read or write operations.
- **MemoryController:**
  Communication is orchestrated by a *memory controller*, which manages the flow of data and ensures that signals between the CPU and memory are synchronized. In older systems, this controller was located on the motherboard; in modern computers, it's typically integrated into the CPU for greater speed and efficiency

Communication Process:

1. When the CPU needs to access data or instructions in memory, it places the address of the required memory location on the address bus.
2. The CPU sends a control signal (read or write command) on the control bus.
3. If reading, the memory controller retrieves the data from the specified address and sends it back to the CPU via the data bus. If writing, the CPU sends data over the data bus to be stored at the designated memory location.

4. This process is repeated billions of times per second during computing operations, forming the backbone of the *fetch-decode-execute* cycle used to run programs

Types of Computer Memory



In general, computer memory is divided into three types:

- Primary memory
- Secondary memory
- Cache memory

Now we discuss each type of memory one by one in detail:

1. Primary Memory

It is also known as the main memory of the computer system. It is used to store data and programs, or instructions during computer operations. It uses semiconductor technology and hence is commonly called semiconductor memory. Primary memory is of two types:

RAM (Random Access Memory):

It is a volatile memory. Volatile memory stores information based on the power supply. If the power supply fails/ interrupted/stopped, all the data and information on this memory will be lost. RAM is used for booting up or starting the computer. It temporarily stores programs/data which has to be executed by the processor. RAM is of two types:

- **S RAM (Static RAM):**S RAM uses transistors and the circuits of this memory are capable of retaining their state as long as the power is applied. This memory consists of the number of flip flops with each flip flop storing 1 bit. It has less access time and hence, it is faster.
- **D RAM (Dynamic RAM):**D RAM uses capacitors and transistors and stores the data as a charge on the capacitors. They contain thousands of memory cells. It needs refreshing of charge on capacitor after a few milliseconds. This memory is slower than S RAM.

ROM (Read Only Memory):

It is a non-volatile memory. Non-volatile memory stores information even when there is a power supply failed/ interrupted/stopped. ROM is used to store information that is used to operate the system. As its name refers to read-only memory, we can only read the programs and data that are stored on it. It contains some electronic fuses that can be programmed for a piece of specific information. The information is stored in the ROM in binary format. It is also known as permanent memory. ROM is of four types:

- **MROM(Masked ROM):** Hard-wired devices with a pre-programmed collection of data or instructions were the first ROMs. Masked ROMs are a type of low-cost ROM that works in this way.
- **PROM (Programmable Read Only Memory):** This read-only memory is modifiable once by the user. The user purchases a blank PROM and uses

a PROM program to put the required contents into the PROM. Its content can't be erased once written.

- **EPROM (Erasable Programmable Read Only Memory):**EPROM is an extension to PROM where you can erase the content of ROM by exposing it to Ultraviolet rays for nearly 40 minutes.
- **EEPROM (Electrically Erasable Programmable Read Only Memory):** Here the written contents can be erased electrically. You can delete and reprogram EEPROM up to 10,000 times. Erasing and programming take very little time, i.e., nearly 4 -10 ms(milliseconds). Any area in an EEPROM can be wiped and programmed selectively.

2. Secondary Memory

It is also known as auxiliary memory and backup memory. It is a non-volatile memory and used to store a large amount of data or information. The data or information stored in secondary memory is permanent, and it is slower than primary memory. A CPU cannot access secondary memory directly. The data/information from the auxiliary memory is first transferred to the main memory, and then the CPU can access it.

Characteristics of Secondary Memory

- It is a slow memory but reusable.
- It is a reliable and non-volatile memory.
- It is cheaper than primary memory.
- The storage capacity of secondary memory is large.
- A computer system can run without secondary memory.
- In secondary memory, data is stored permanently even when the power is off.

Types of Secondary Memory

1. Magnetic Tapes: Magnetic tape is a long, narrow strip of plastic film with a thin, magnetic coating on it that is used for magnetic recording. Bits are recorded

on tape as magnetic patches called RECORDS that run along many tracks. Typically, 7 or 9 bits are recorded concurrently. Each track has one read/write head, which allows data to be recorded and read as a sequence of characters. It can be stopped, started moving forward or backwards or rewound.

2. Magnetic Disks: A magnetic disk is a circular metal or a plastic plate and these plates are coated with magnetic material. The disc is used on both sides. Bits are stored in magnetized surfaces in locations called tracks that run in concentric rings. Sectors are typically used to break tracks into pieces.

Hard discs are discs that are permanently attached and cannot be removed by a single user.

3. Optical Disks: It's a laser-based storage medium that can be written to and read. It is reasonably priced and has a long lifespan. The optical disc can be taken out of the computer by occasional users.

Types of Optical Disks

CD - ROM

- It's called a compact disk. Only read from memory.
- Information is written to the disc by using a controlled laser beam to burn pits on the disc surface.
- It has a highly reflecting surface, which is usually aluminium.
- The diameter of the disc is 5.25 inches.
- 16000 tracks per inch is the track density.
- The capacity of a CD-ROM is 600 MB, with each sector storing 2048 bytes of data.
- The data transfer rate is about 4800KB/sec. & the new access time is around 80 milliseconds.

WORM-(WRITE ONCE READ MANY)

- A user can only write data once.

- The information is written on the disc using a laser beam.
- It is possible to read the written data as many times as desired.
- They keep lasting records of information but access time is high.
- It is possible to rewrite updated or new data to another part of the disc.
- Data that has already been written cannot be changed.
- Usual size - 5.25 inch or 3.5 inch diameter.
- The usual capacity of a 5.25-inch disk is 650 MB,5.2GB etc.

DVDs

The term "DVD" stands for "Digital Versatile/Video Disc," and there are two sorts of DVDs:

- DVDR (writable)
- DVDRW (Re-Writable)
- **DVD-ROMS (Digital Versatile Discs)**: These are read-only memory (ROM) discs that can be used in a variety of ways. When compared to CD-ROMs, they can store a lot more data. It has a thick polycarbonate plastic layer that serves as a foundation for the other layers. It's an optical memory that can read and write data.
- **DVD-R**: DVD-R is a writable optical disc that can be used just once. It's a DVD that can be recorded. It's a lot like WORM. DVD-ROMs have capacities ranging from 4.7 to 17 GB. The capacity of 3.5 inch disk is 1.3 GB.

3. Cache Memory

Cache Memory is a type of high-speed semiconductor memory that can help the CPU run faster. Between the CPU and the main memory, it serves as a buffer. It is used to store the data and programs that the CPU uses the most frequently.

Advantages of Cache Memory

- It is faster than the main memory.
- When compared to the main memory, it takes less time to access it.
- It keeps the programs that can be run in a short amount of time.
- It stores data for temporary use.

Disadvantages of Cache Memory

- Because of the semiconductors used, it is very expensive.
- The size of the cache (amount of data it can store) is usually small.