Hierarchical memory system

By

Smita Mande

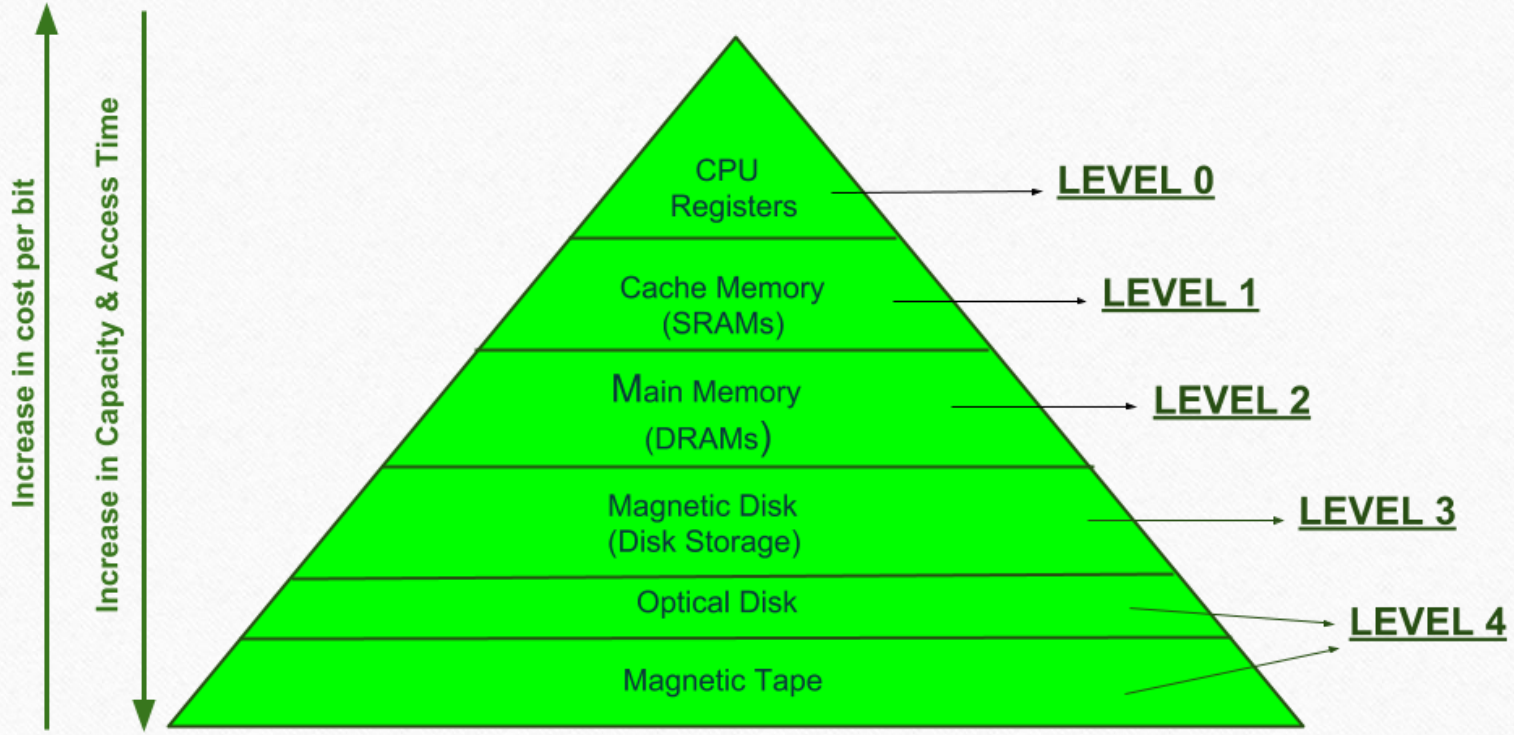
Hierarchical Memory System

In the Computer System Design, Memory Hierarchy is an enhancement to organize the memory such that it can minimize the access time. The Memory Hierarchy was developed based on a program behavior known as locality of reference.

This Memory Hierarchy Design is divided into 2 main types:

1.External Memory or Secondary Memory – Comprising of Magnetic Disk, Optical Disk, Magnetic Tape i.e. peripheral storage devices which are accessible by the processor via I/O Module.

2.Internal Memory or Primary Memory – Comprising of Main Memory, Cache Memory & CPU registers. This is directly accessible by the processor.



MEMORY HIERARCHY DESIGN

Characteristics of Hierarchical memory system

1. **Capacity:** It is the global volume of information the memory can store. As we move from top to bottom in the Hierarchy, the capacity increases.
2. **Access Time:** It is the time interval between the read/write request and the availability of the data. As we move from top to bottom in the Hierarchy, the access time increases.
3. **Performance:** Earlier when the computer system was designed without Memory Hierarchy design, the speed gap increases between the CPU registers and Main Memory due to large difference in access time. This results in lower performance of the system and thus, enhancement was required. This enhancement was made in the form of Memory Hierarchy Design because of which the performance of the system increases. One of the most significant ways to increase system performance is minimizing how far down the memory hierarchy one has to go to manipulate data.
4. **Cost per bit:** As we move from bottom to top in the Hierarchy, the cost per bit increases i.e. Internal Memory is costlier than External Memory.

Principle of Locality of Reference

- Locality of reference refers to the tendency of the computer program to access the same set of memory locations for a particular time period. The property of Locality of Reference is mainly shown by loops and subroutine calls in a program.
- On an abstract level there are two types of localities which are as follows −

1. Temporal locality
2. Spatial locality

Temporal locality

This type of optimization includes bringing in the frequently accessed memory references to a nearby memory location for a short duration of time so that the future accesses are much faster.

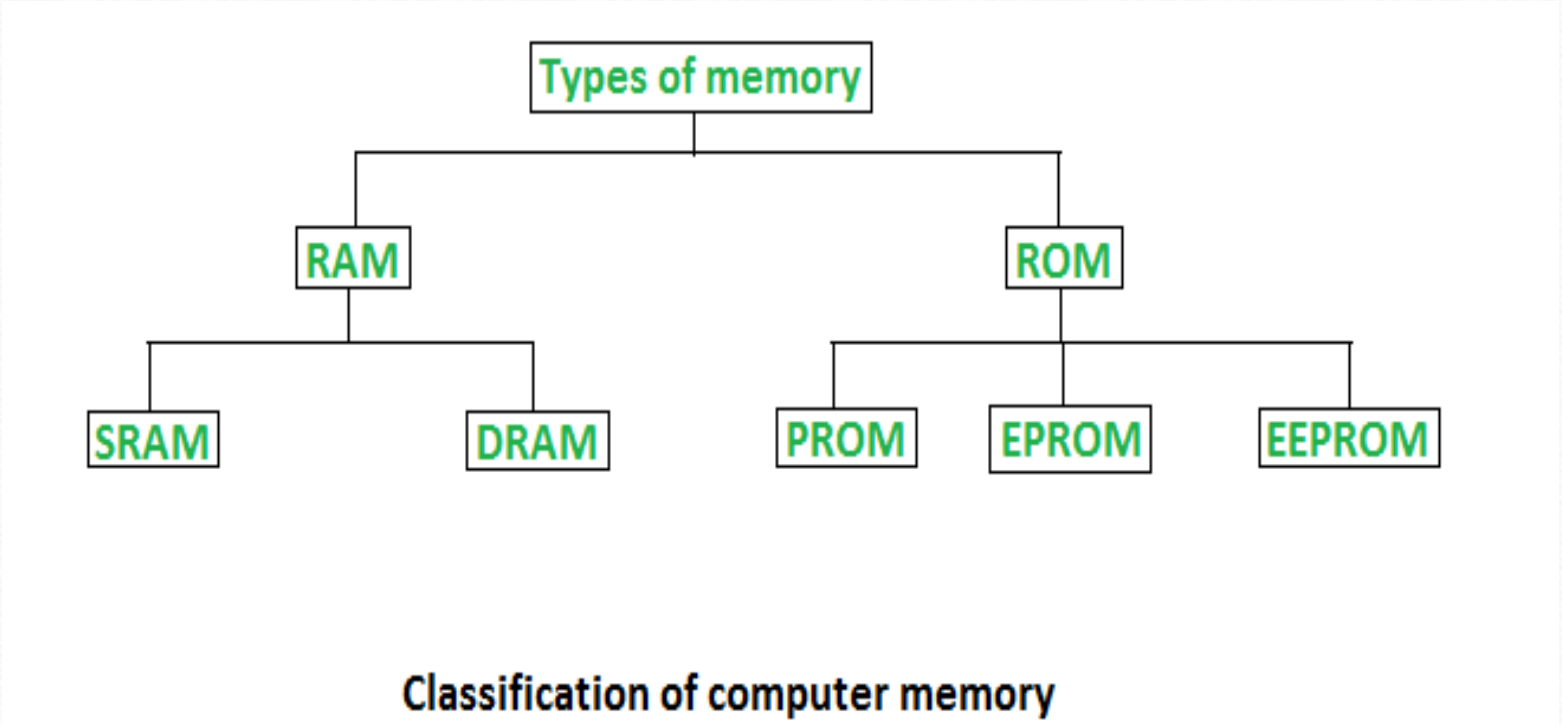
For example, if in an instruction set we have a variable declared that is being accessed very frequently we bring in that variable in a memory register which is the nearest in memory hierarchy for faster access.

Spatial locality

This type of optimization assumes that if a memory location has been accessed it is highly likely that a nearby/consecutive memory location will be accessed as well and hence we bring in the nearby memory references too in a nearby memory location for faster access.

For example, traversal of a one-dimensional array in any instruction set will benefit from this optimization.

Main Memory Organization



Classification of computer memory

Random Access Memory (RAM)

- It is also called read-write memory or the main memory or the primary memory.
- The programs and data that the CPU requires during the execution of a program are stored in this memory.
- It is a volatile memory as the data is lost when the power is turned off.
- RAM is further classified into two types- SRAM (Static Random Access Memory) and DRAM (Dynamic Random Access Memory).

Read-Only Memory (ROM)

- Stores crucial information essential to operate the system, like the program essential to boot the computer.
- It is non-volatile.
- Always retains its data.
- Used in embedded systems or where the programming needs no change.
- Used in calculators and peripheral devices.
- ROM is further classified into four types- MROM, PROM, EPROM, and EEPROM.

DRAM	SRAM
1. Constructed of tiny capacitors that leak electricity.	1.Constructed of circuits similar to D flip-flops.
2.Requires a recharge every few milliseconds to maintain its data.	2.Holds its contents as long as power is available.
3.Inexpensive.	3.Expensive.
4. Slower than SRAM.	4. Faster than DRAM.
5. Can store many bits per chip.	5. Can not store many bits per chip.
6. Uses less power.	6.Uses more power.
7.Generates less heat.	7.Generates more heat.
8. Used for main memory.	8. Used for cache.

Difference between SRAM and DRAM

Types of Read-Only Memory (ROM)

- **PROM (Programmable read-only memory)** – It can be programmed by the user. Once programmed, the data and instructions in it cannot be changed.

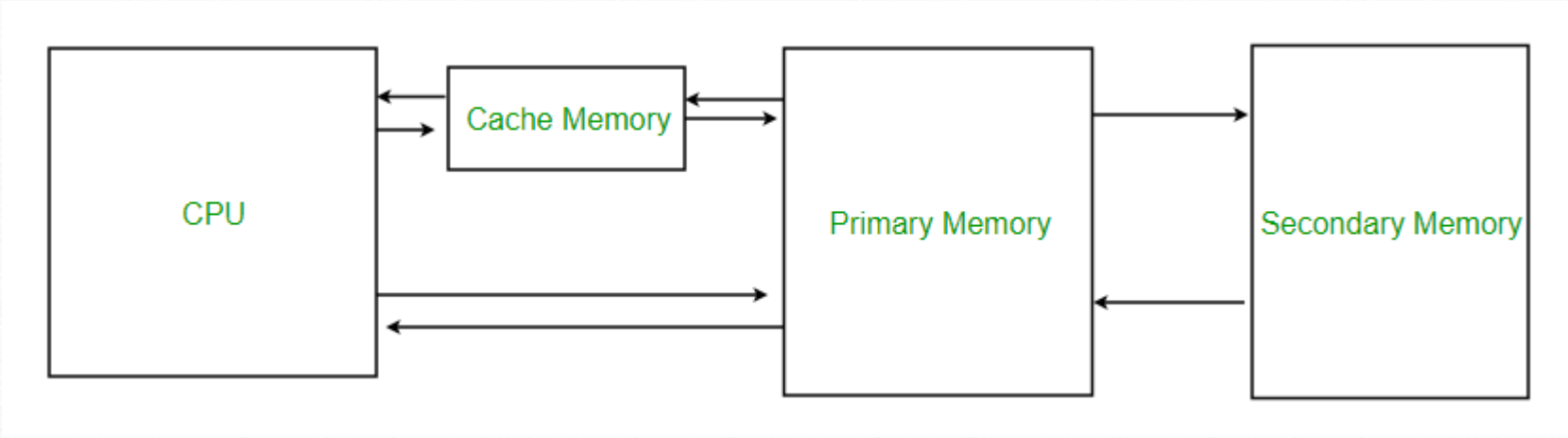
- **EPROM (Erasable Programmable read-only memory)** – It can be reprogrammed. To erase data from it, expose it to ultraviolet light. To reprogram it, erase all the previous data.

- **EEPROM (Electrically erasable programmable read-only memory)** – The data can be erased by applying an electric field, with no need for ultraviolet light. We can erase only portions of the chip.

- **MROM(Mask ROM)** – Mask ROM is a kind of read-only memory, that is masked off at the time of production. Like other types of ROM, mask ROM cannot enable the user to change the data stored in it.

Cache memory Organization

- **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 disk 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 cache is a smaller and faster memory that stores copies of the data from frequently used main memory locations.


CPU
Cache Memory
Primary Memory
Secondary 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.
- 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

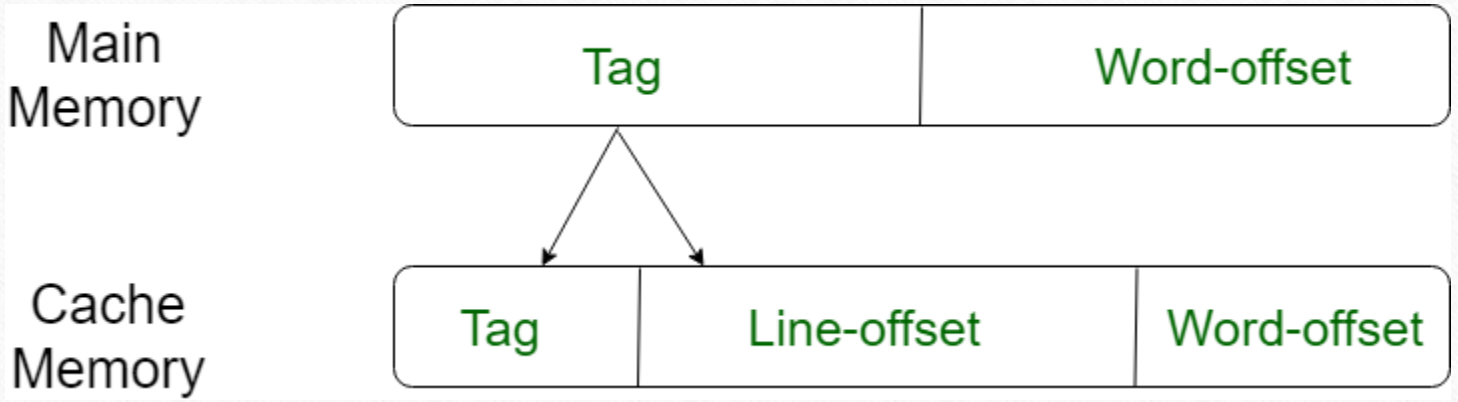
- We can improve Cache performance using higher cache block size, reduce miss rate, and reduce the time to hit in the cache.

Mapping Functions

- The correspondence between main memory blocks and cache is specified by a **mapping function.**
  - **Direct Mapping Technique**
  - **Associative Mapping Technique**
    - **Fully Associative**
    - **Set Associative**

1. Direct Mapping

- Each block of main memory maps to only one cache line
  - i.e. if a block is in cache, it must be in one specific place
- Address is in two parts
- Least Significant w bits identify unique word
- Most Significant s bits specify one memory block
- The MSBs are split into a cache line field or block field r and a tag of s-r (most significant)


Main Memory
Tag
Word-offset
Cache Memory
Tag
Line-offset
Word-offset

Direct Mapping Address Structure

Tag s - r	Block Field r	Word w
5	7	4

- The address is divided into 3 fields
- The lower order 4 bits select one of the 16 words in a block.
- The second field known as **block field** is used to distinguish a block from other blocks. (Since 2^7=128)
- The third field is **tag field**. Used to store high-order 5 bits of memory address of block. These tag bits are used to identify which of the 32 – blocks or pages (2^5 = 32) that are mapped into the cache.

Direct Mapping Cache Organization