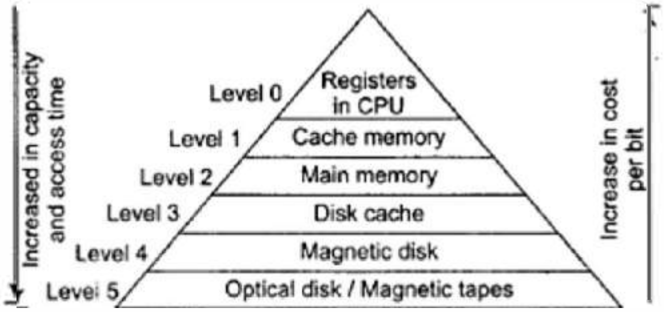
Memory Hierarchy

The memory hierarchy design primarily consists of several storage devices in a computer system. The majority of computers came with extra storage to allow them to run faster than the main memory. The memory unit is used for storing programs and data. It fulfils the need for the storage of information.

The memory hierarchy is important part fo GATE CS syllabus. The additional storage with main memory capacity enhances the performance of the general-purpose computers and makes them efficient. Only those programs and data the processor needs reside in the main memory. Information can be transferred from auxiliary memory to main memory when needed.

Memory Hierarchy Levels

Primary (internal) and secondary (external) memory are the two forms of the designed memory hierarchy. A hierarchical pyramid for computer memory is shown in the diagram below.



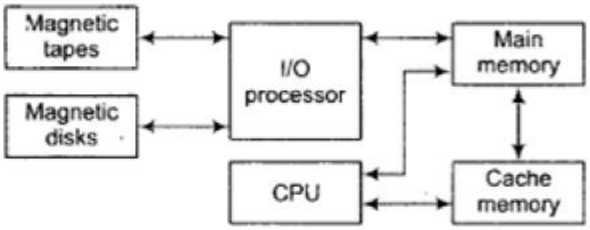
Registers

The register is usually a static RAM or SRAM in the computer's processor that holds the data word, normally 64 or 128 bits. The essential register is the program counter register, which is found in all processors. In addition, a status word register and an accumulator are used by the majority of CPUs. A status word register is used for decision-making, and an accumulator is used to store data like a calculator is used to store numbers. Computers with complex instruction sets typically have many registers for receiving main memory, whereas RISC-based computers have fewer registers.

Cache Memory

Small, fast storage memory is used to improve average access time. Therefore, we can say that cache is a very high-speed memory that rapidly increases processing speed by making current programs and data available to the CPU. The cache memory is followed by main memory as per the memory hierarchy.

The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations.



Memory connection in computer system

Cache Performance

When the processor needs to read or write to a location in the main memory, it first checks whether a copy of that data is in the cache. If so, the processor immediately reads from or writes to the cache.

- Cache hit If the processor immediately reads or writes the data in the cache line.
- Cache miss If the processor does not find the required word in the cache, then a cache miss has occurred.
- Hit ratio Percentage of memory accesses satisfied by the cache.
- Miss ratio = 1– Hit ratio
- Temporal Locality: The word referenced now will likely be referenced again soon. Hence it is wise to keep the currently accessed word handy for a while.
- Spatial Locality: Words near the currently referenced word are likely to be referenced soon. Hence it is wise to prefetch words near the currently referenced word and keep them handy for a while.
- Write through writes the data to memory as well as to the cache.
- Writeback: Don't write to memory now, do it later when this cache block is evicted.

Main Memory

As per the memory hierarchy, main memory is level 2. The main memory refers to the physical memory, one central storage unit in a computer system. The main memory is relatively large and fast memory used to store programs and data during computer operation. The main memory in a general-purpose computer comprises a RAM integrated circuit.

Latency: The latency is the time to transfer a block of data from the main memory or caches.

- As the CPU executes instructions, both the instructions themselves and the data they operate on must be brought into the registers; until the instruction/data is available, the

CPU cannot proceed to execute it and must wait. The latency is, thus, the time the CPU waits to obtain the data.

- The latency of the main memory directly influences the efficiency of the CPU.

Auxiliary Memory

A computer system's auxiliary memory is the lowest-cost, highest-capacity, and slowest-access storage. Auxiliary memory is used to store programs and data that are stored for long periods or are not in use right away. Magnetic tapes and magnetic discs are the most common types of auxiliary memory.

Magnetic Disks

A magnetic disk is a circular plate constructed of metal or plastic coated with magnetized material. Both sides of the disk are often used, and several disks may be stacked on one spindle with reading/write heads available on each surface. magnetic disks are level 4 component in memory hierarchy.

All disks rotate together at high speed. Bits are stored in the magnetized surface in spots along concentric circles called tracks. The tracks are commonly divided into sections called sectors.

Magnetic Tapes

A magnetic tape is a magnetic recording medium made of a thin magnetizable coating on a long, narrow strip of plastic film. These are level 5 in memory hierarchy. Bits are recorded as magnetic spots on the tape along several tracks. Magnetic tapes can be stopped, started to move forward, or in reverse. Read/write heads are mounted in each track so that that data can be recorded and read as a sequence of characters.

Memory Hierarchical Design

There are various factors such as typical size, bandwidth, access time, etc which are important as per the design. Check out the Memory Hierarchical design shown below:

Level Name	**Registers**	**Cache Memory**	**Main Memory**	**Secondary memory**
Typical Size	<1KB	<16MB	<16GB	>100GB
Implementation	Customized Multiport	SRAM (Flip-Flops)	DRAM (Capacitor)	Magnetic devices
Bandwidth (MB/s)	(20,000-1,00,000)	(5,000-10,000)	(1,000-5,000)	(20-150)

Access Time (ns)	(0.25-0.5)	(0.5-2.5)	(80-250)	5000000
Managed By	Compiler	Hardware	OS	OS
Backed By	Cache Memory	Main Memory	Secondary Memory	Compact Disk

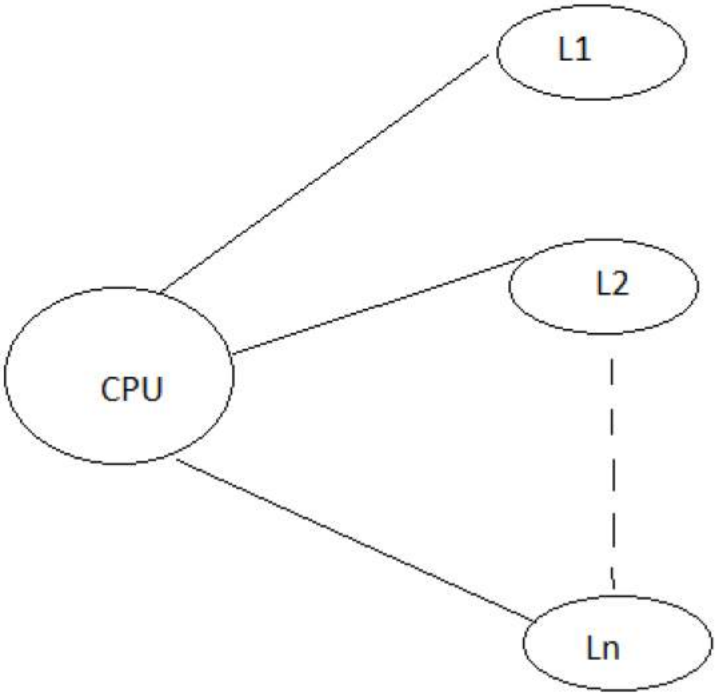
Memory Organization

Based on the style of accessing the memory, it can be classified into two types:

- Simultaneous Memory Access Organization.
- Hierarchical Memory Access Organization.

Simultaneous Access Memory Organization

- In this memory Organization, the CPU is directly connected to all the levels of memory. Still, access is allowed in a sequence, i.e., Whenever there is a miss in level 1 memory, data can be accessed directly from level 2 without copying it into level 1.
- Levels are designed so that the lower the level, the higher the performance, the lower the access time, and the higher the level, the lower the performance, and the higher the access time.
- The hit ratio of the last level must be 1 since it represents the whole system.

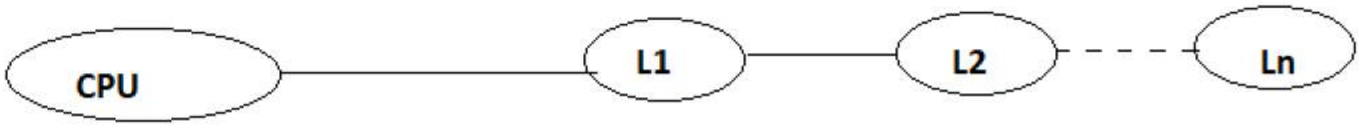


Average Memory Access time= H1T1+ (1-H1)H2T2+ (1-H1)(1-H2)H3T3+.....

H1, H2, and H3.. are the hit ratio for level 1, level 2, Level 3, and so on.

Hierarchical Memory Access Organization

- In this memory organization, the CPU always accesses the data from the first memory level. Whenever there is a miss in level 1, data is transferred from a higher to a lower memory level. After the data transfer, the CPU access the data from the lower level.
- It is a better approach if the spatial locality is needed in the program.



Average Access time= H1T1+ (1-H1)H2(T2+ T1)+ (1-H1)(1-H2)H3(T3+ T2+T1) +...

Characteristics of Memory Hierarchy

The following are the main characteristics of memory hierarchy:

Performance

Initially, computer systems were designed without a memory hierarchy. The speed gap between the main memory and CPU registers grew due to the large differential in access time, resulting in lower system performance. As a result, enhancement was required. Because of the system's increased performance, this was enhanced in the memory hierarchy model.

Ability

The total quantity of data the memory hierarchy can store is its capability because its capacity grows as we move from top to bottom.

Cost per bit

When we move from the bottom to the top of the memory hierarchy, the cost of each bit increases, implying that internal memory is more expensive than external memory.

Access Time

In the memory hierarchy, the access time is the time delay between data availability and requests to read or write because the access time increases as we move from the top to the bottom of the memory hierarchy.

Advantages of Memory Hierarchy

Memory hierarchy is necessary. Here are a few advantages of memory hierarchy:

- Memory distribution is easy and cost-effective.
- External destruction is removed.
- Data can be spread all over.
- Allows for pre-paging and demand paging.
- Swapping will be a lot easier.