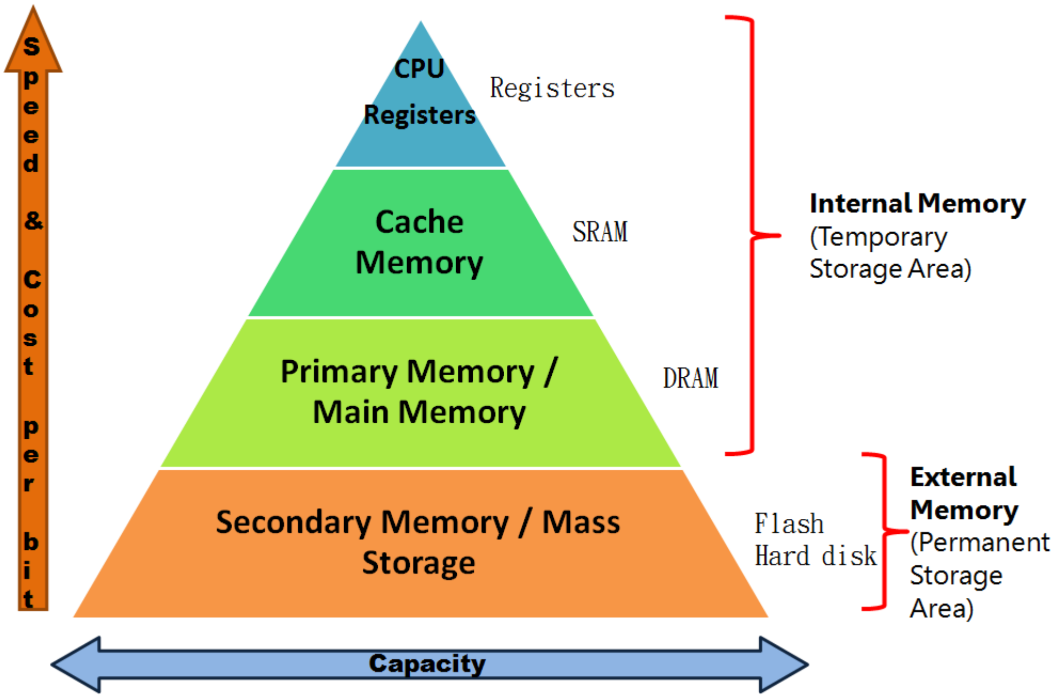
Memory Hierarchy Levels

Memory hierarchy is about arranging different kinds of storage devices in a computer based on their size, cost and access speed. The main purpose is to achieve efficient operations by organizing the memory to reduce access time while speeding up operations.

Since CPU registers are the fastest to read and write, they are placed at the top of the hierarchy. On the other hand, mass storage devices like optical drives and magnetic tapes are the slowest and largest and therefore occupy the last level in the pyramid.



CPU registers A CPU register is a small section of memory in a CPU that can store small amounts of the data required to perform various operations. It loads the resulting data to the main memory and contains the address of the memory location.

Registers are present inside the CPU and therefore have the quickest access time. Since they are the **fastest memory type, they are the most expensive.** They are also the smallest in size, typically measured in kilobytes.

Cache memory Cache memory is required to store segments of programs or chunks of data that are frequently accessed by the processor. When the CPU needs to access program code or data, it first checks the cache memory. If it finds the data, it reads it quickly. If it doesn't, it looks into the main memory to find the required data.

Cache memory is usually smaller in size, typically measured in megabytes (MB). It is implemented using SRAM. Usually, the cache is inside the processor. However, it may also be implemented as a separate integrated circuit (IC).

Primary/main memory The primary memory communicates with the CPU and with the peripheral or auxiliary memory devices through the I/O processor. It is the primary storage unit of a computer system;

It's often referred to as random access memory (RAM) and is implemented using dynamic RAM (DRAM) components. However, main memory may also include read-only memory (ROM).

Any program or data that is not currently required in the main memory is transferred into the auxiliary memory to create space for programs and data that are currently active. **Main memory is less expensive than CPU registers and cache memory, and is also larger in size (typically measured in gigabytes).**

Secondary storage

Secondary storage devices act as backup storage and are much cheaper than the main memory and cache. These memory types are also large in size and generally have capacities of up to 20 terabytes (TB). In secondary storage devices, both faces of a **magnetic disk** are utilized to store programs and data.

Tertiary storage. Tertiary storage devices are usually **magnetic tapes or optical disks**. These devices are typically used to store duplicate or archive copies of data. Also known as auxiliary storage, tertiary memory devices are usually used to store programs and data for the long term or when not required for immediate use.

Tertiary devices are suitable for data archiving and backup. They are the cheapest and slowest memory type; they typically have capacities of 1 TB to 20 TB.

Characteristics of memory hierarchy

The key characteristics of a memory hierarchy include the following:

Capacity Capacity is the volume of information that a memory device can store. As we move down the memory pyramid, the capacity or memory size increases.

Access time Access time is the time interval from when a read/write request is made and when the data actually becomes available. It increases as we move from the top to the bottom of the memory hierarchy. Registers, which are present inside the CPU, have the shortest access time, meaning they are the fastest. At the bottom of the pyramid, magnetic tapes and similar storage devices have the greatest access time.

Performance Without a memory hierarchy, there is a speed gap between CPU registers and the main memory. This increases access time and directly impacts the system's performance. Performance can be improved by reducing the number of levels required to access and manipulate data.

Cost per bit The cost per bit is calculated by dividing the total cost of the memory by the total number of accessed bits. As we move from the top of the memory hierarchy to the bottom, the cost per bit decreases. This is because internal memory is costlier than external memory.