**Memory Hierarchy**

* Memory is an essential component in any digital computer which is need for storing programs and data.
* A computer run more efficiently if it has additional storage beyond the capacity of main memory.
* The memory unit that communicates directly with the CPU is called the main memory.
* Devices that provide backup storage are called auxiliary memory.
  + Example : Magnetic discs , Magnetic tapes…etc.
* Only program and data needed by the processor reside in the main memory. And the remaining are stored in auxiliary memory.
* The memory hierarchy system consists of all storage devices employed in a computer system.

CPU

Cache Memory

Main Memory

I/O Processor

Magnetic Discs

Magnetic Tapes

Fig: Memory Hierarchy in a Computer System

* Main memory able to communicate directly with the CPU and with auxiliary memory devices through I/O processor.
* A special very high speed memory called a cache is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.
* The cache memory is employed in computer system to compensate for the speed differential between the main memory access time and processor logic.
* The cache is used for storing segments of program currently being executed in the CPU and temporary data frequently need in the present calculation.
* The main goal of using a memory hierarchy to obtain the highest possible average access speed while minimizing the total cost of the entire memory system.
* The part of the computer system that supervises the flow of information between auxiliary memory and main memory is called the memory management system.
* So memory hierarchy contains:
  + Cache Memory
  + Primary Memory: RAM, ROM.
  + Auxiliary Memory: Magnetic Discs, Magnetic Tapes, Optical Semiconductor devices…etc

**Cache Memory**

* A special very high speed memory called a cache is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.
* The cache memory is employed in computer systems to compensate for the speed differential between main memory access time and processor logic.
* The CPU logic is usually faster than main memory access time, with the result that processing sped is limited primarily by the speed of main memory.
* A technique used to compensate for the mismatch in operating speeds is to employ an extremely fast, small cache between the CPU and main memory whose access time is close to processor logic clock cycle time.
* The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations.
* By making programs and data available at a rapid rate, it is possible to increase the performance rate of the computer.
* The cache organization is concerned with the transfer of information between main memory and CPU.