Memory Organization

6.1. BASIC MEMORY CHARACTERISTICS

The most modern computers are designed using the Von-Neumann Model, which is centered on Main Memory. The user programs and user data that perform the processing are stored initially on auxiliary memory devices from there they are to be brought to main memory by operating system. We know that memory is logically organized as a linear array of locations or addresses, with addresses from 0 to the maximum memory size the processor can address. This address is normally assigned by a programmer in OCTAL or Hexadecimal format.

The various different types of memory devices and technology used for fabrication for these devices for better performance and how each is part of the memory hierarchy system. We then look at cache memory (a special high-speed memory) and a method that utilizes memory to its fullest by means of virtual **memory** implemented via paging technique. The goal of Memory Hierarchy is to **obtain the highest** possible access speed while minimizing the total cost of the memory system.

Location of the Device

(*i*) CPU registers–we call sometime internal memory.

(*ii*) Cache memory–between CPU and Main memory or On chip Cache.

(*iii*) Main Memory–SRAM or DRAM Devices.

(*iv*) Secondary Storage (external or auxiliary device).

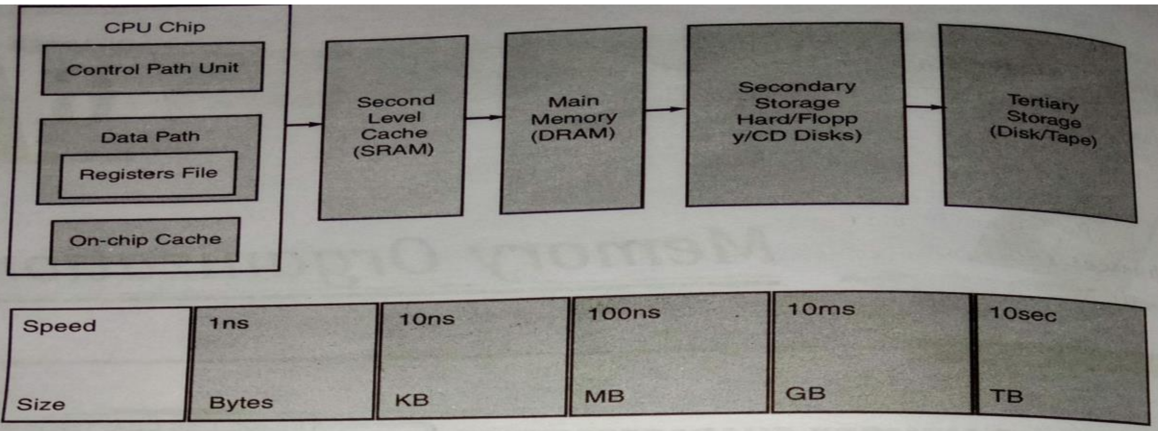

Fig. 6.1. CPU and Memory Interaction

Capacity of the Device—Unit of Data Transfer

The number of data lines into and out of the memory module. Internally it is usually governed by data bus width. Externally it is usually a block which is much larger than a word.

Access Methods For The Device

Sequential Access

We start at the beginning and read through in order and access time depends on location of data and previous location (tape)

Direct Access

Individual blocks have unique address in memory and Access is by jumping to vicinity plus sequential search. Access time depends on location and previous location.

Random Access

Individual addresses identify locations exactly. Access time is independent of location or previous access (RAM).

Associative

Data is located by a comparison with contents of a portion of the store. Access time is independent of location or previous access e.g. cache. The words are retrieved based on a portion of its contents rather than its address and constant access time.

Memory Device Performance

Access Time : The time it takes from the instant that an address is presented to the memory by CPU to the instant that data have been made available for use on data bus.

Memory Cycle Time

Time may be required for the memory to "recover" before next access. Cycle time is access + recover.

Data Transfer Rate or Memory Bandwidth

Rate at which data can be transferred into or out of a memory unit. Random access : 1/(Cycle time).

Non-random : TN = TA + (N/R)

TA = access time
N = number of bits
R = transfer rate (bps)

Physical Characteristics of a Device

(i) Volatility
(ii) Power Consumption
(iii) Erasable
(iv) Density
(v) Cost Per Bit Storage

Device Organisation and Trade-off for Performance

(i) Physical arrangement of bits into words.
(ii) Not always obvious e.g. interleaved unit.

Tradeoff among three main characteristics of memory (cost, size, access time)

(i) Smaller access time, greater cost per bit.
(ii) Greater capacity, smaller cost per bit.
(iii) Greater capacity, greater access time.

Different types of memory are required at different level of operation in a computer system. A classification is discussed further.

6.2. MEMORY HIERARCHY

The widening speed gap between CPU and main memory needs to be filled. Modern processor operations take of the order of 1 ns, while memory access requires 10s or even 100s of ns. The second parameter is memory bandwidth, which limits the instruction execution rate. Each instruction executed involves at least one memory access. Hence, a few to 100s of MIPS is the best that can be achieved. A fast buffer memory can help bridge the CPU-memory gap. The fastest memories are expensive and thus not very large.

One of the most important considerations in understanding the performance capabilities of a modern processor is the memory hierarchy. Unfortunately, as we have seen, not all memory is created equal, and some types are far less efficient and thus cheaper than others. To deal with this disparity, today's computer systems use a combination of memory types to provide the best performance at the best cost. This approach is called hierarchical memory. As a rule, the faster memory is, the more expensive it is per bit of storage. By using a hierarchy of memories, each

with different access speeds and storage capacities, a computer system can exhibit performance above what would be possible without a combination of the various types. The base types that normally constitute the hierarchical memory system include registers, cache, main memory, and secondary memory.

At the bottom of the hierarchy are the relativity slow magnetic tapes used to store removable files. Magnetic disk used as back up storage. The main memory occupies a central position by being able to communicate directly with the CPU and with auxiliary memory device through an input/output processor when program not residing in main memory are needed by the CPU. They are in bought in from auxiliary memory. Program not currently needed in main memory are transferred into auxiliary memory to provided space for currently used programs and data.

As we move from top to bottom in hierarchy.

(i) Access Speed will Decrease
(ii) Storage Capacity will Increase
(iii) Cost Per Bit will Decrease

While the I/O processor manages data transfer between auxiliary memory and main memory, the cache organization is concerned with the transfer of information between main memory and CPU. Thus each is involved with a different level in memory hierarchy system. The reason for having two or three level of memory hierarchy is economics. As the storage capacity of the memory increase, the **cost per bit** for storing binary information decrease and the access time of the memory become longer.