**Module-4**

**Memory**

A memory cell is capable of storing 1-bit of information. A number of memory cells are organized in the form of a matrix to form the memory chip. Each row of cells constitutes a memory word, and all cells of a row are connected to a common line which is referred as word line. An address decoder is used to drive the word line. At a particular instant, one word line is enabled depending on the address present in the address bus. Additional circuitry is required to access the memory cells based on address generated by the CPU; this extra circuitry is called memory interfacing circuitry.

Register, cache memory and main memory are internal memory. Magnetic Disk, removable media etc. are external memory. Internal memories are semiconductor memory. Semiconductor memories are categorized as volatile memory and non-volatile memory. Static RAM (SRAM) and Dynamic RAM (DRAM) are examples of volatile memory. PROM, EPROM, EEPROM etc. are instances of non-volatile memory.

Analysis of large set of programs has shown that many instructions in each of a few localized areas of the program are repeatedly executed, while the remainder of the program is accessed relatively less. This phenomenon is referred to as locality of reference. If it can be arranged to have the active segments of a program in a fast memory (based on locality of reference), then the total execution time can be significantly reduced. The faster memory that is inserted between CPU and Main Memory is termed as Cache memory. When a Read request is received from the CPU, the contents of a block of main memory words containing the word specified are transferred into the cache. When any of the words in this block is referenced by the program subsequently, its contents are read directly from the cache; this is called cache hit. On the other hand if the word specified is not present in the cache it is called cache miss and the corresponding block is loaded from the main memory to the cache.

The cache memory can store a number of such blocks at any given time. The correspondence between the Main Memory Blocks and those in the cache is specified by means of a mapping function. This mapping function is used to transfer the block from main memory to cache memory. Three basic mapping functions are available viz. direct mapping, Associative mapping, Block-set-associative mapping.

###### **Generally speaking, a full code does not fit into the free blocks of a cache at a given time. So at some point of time the contents of certain cache blocks are needed to be replaced with new ones.** Such an overwritten block is called the *victimized block* and the block to be victimized is decided by the replacement algorithm employed. There are a number of replacement algorithms like First In First Out (FIFO), Least Recently Used (LRU), Least Frequently Used (LFU) etc.