CACHE MEMORY

Analysis of a large number of typical programs has shown that most of their execution time is spent on a few main row lines in which a number of instructions are executed repeatedly. These instructions may constitute a simple loop, nested loops or few procedure that repeatedly call each other. The main observation is that many instructions in a few localized are as of the program are repeatedly executed and that the remainder of the program is accessed relatively infrequently. This phenomenan is referred to as locality of reference.

If the active segments of a program can be placed in a fast memory, then the total execution time can be significantly reduced, such a memory is referred as a cache memory which is in served between the CPU and the main memory as shown in fig.1

Fig.1 cache memory between main memory & cpu.

Two Level memory Hierarchy: We will adopt the terms Primary level for the smaller, faster memory and the secondary level for larger, slower memory, we will also allow cache to be a primary level with slower semiconductor memory as the corresponding secondary level. At a different point in the hierarchy, the same S.C memory could be the primary level with disk as the secondary level.

Primary and Secondary addresses:-

A two level hierarchy and its addressing are illustrated in fig.2. A system address is applied to the memory management unit (MMU) that handles the mapping function for the particular pair in the hierarchy. If the MMU finds the address in the Primary level, it provides Primary address, which selects the item from the Primary memory. This translation must be fast, because every time memory is accessed, the system address must be translated. The translation may fail to produce a Primary address because the requested items is not found, so that information can be retrieved from the secondary level and transferred to the Primary level.

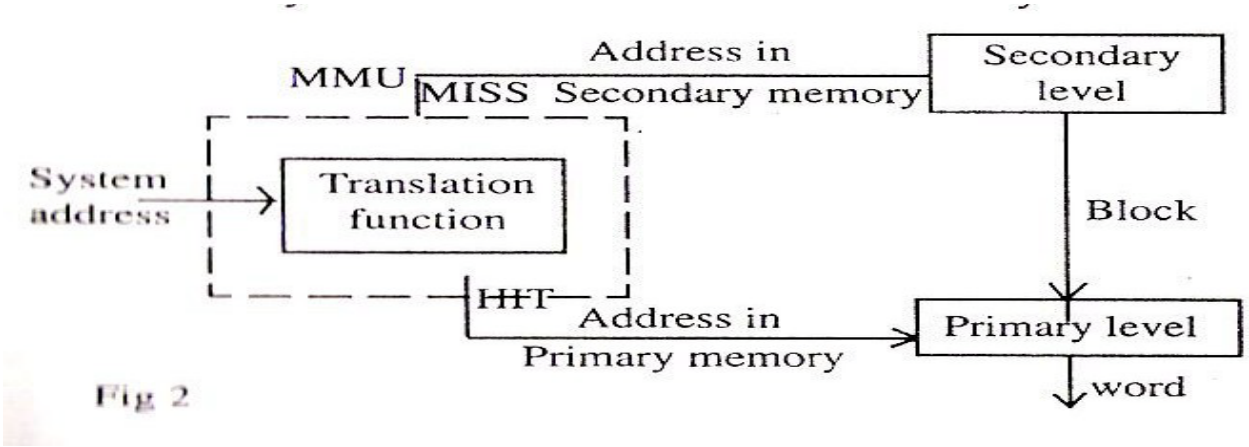

Fig 2

Hits and Misses:- Successful translation of reference into Primary address is called a hit, and failure is a miss. The hit ratio is (1-miss ratio). If t_p is the Primary memory access time and t_s is the secondary access time, the average access time for the two level hierarchy is

$$t_a = h\, t_p + (1-h)t_s$$

Source : http://elearningatria.files.wordpress.com/2013/10/cse-iv-computer-organization-10cs46-notes.pdf