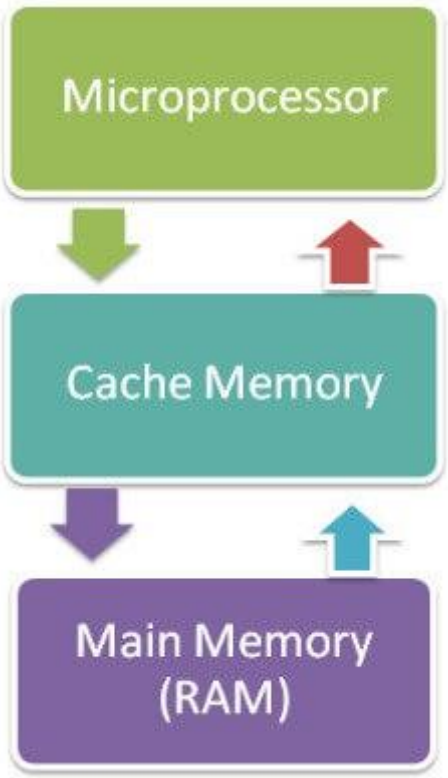
Cache Memory

Cache memory is a high-speed memory, which is small in size but faster than the main memory (RAM). The CPU can access it more quickly than the primary memory. So, it is used to synchronize with high-speed CPU and to improve its performance.



Cache memory can only be accessed by CPU. It can be a reserved part of the main memory or a storage device outside the CPU. It holds the data and programs which are frequently used by the CPU. So, it makes sure that the data is instantly available for CPU whenever the CPU needs this data. In other words, if the CPU finds the required data or instructions in the cache memory, it doesn't need to access the primary memory (RAM). Thus, by acting as a buffer between RAM and CPU, it speeds up the system performance.

Types of Cache Memory:

L1: It is the first level of cache memory, which is called Level 1 cache or L1 cache. In this type of cache memory, a small amount of memory is present inside the CPU itself. If a CPU has four cores (quad core cpu), then each core will have its own level 1 cache. As this memory is present in the CPU, it can work at the same speed as of the CPU. The size of this memory ranges from 2KB to 64 KB. The L1 cache further has two types of caches: Instruction cache, which stores instructions required by the CPU, and the data cache that stores the data required by the CPU.

L2: This cache is known as Level 2 cache or L2 cache. This level 2 cache may be inside the CPU or outside the CPU. All the cores of a CPU can have their own separate level 2 cache, or they can share one L2 cache among themselves. In case it is outside the CPU, it is connected with the CPU with a very high-speed bus. The memory size of this cache is in the range of 256 KB to the 512 KB. In terms of speed, they are slower than the L1 cache.

L3: It is known as Level 3 cache or L3 cache. This cache is not present in all the processors; some high-end processors may have this type of cache. This cache is used to enhance the performance of Level 1 and Level 2 cache. It is located outside the CPU and is shared by all the cores of a CPU. Its memory size ranges from 1 MB to 8 MB. Although it is slower than L1 and L2 cache, it is faster than Random Access Memory (RAM).

The basic operation of a cache memory is as follows:

- When the CPU needs to access memory, the cache is examined. If the word is found in the cache, it is read from the fast memory.
- If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word.
- A block of words one just accessed is then transferred from main memory to cache memory. The block size may vary from one word (the one just accessed) to about 16 words adjacent to the one just accessed.
- The performance of the cache memory is frequently measured in terms of a quantity called **hit ratio**.
- When the CPU refers to memory and finds the word in cache, it is said to produce a **hit**.
- If the word is not found in the cache, it is in main memory and it counts as a **miss**.
- The ratio of the number of hits divided by the total CPU references to memory (hits plus misses) is the hit ratio.