CACHE MEMORY

Cache memory is a small-sized type of volatile computer memory that provides high-speed data access to a processor and stores frequently used computer programs, applications and data.

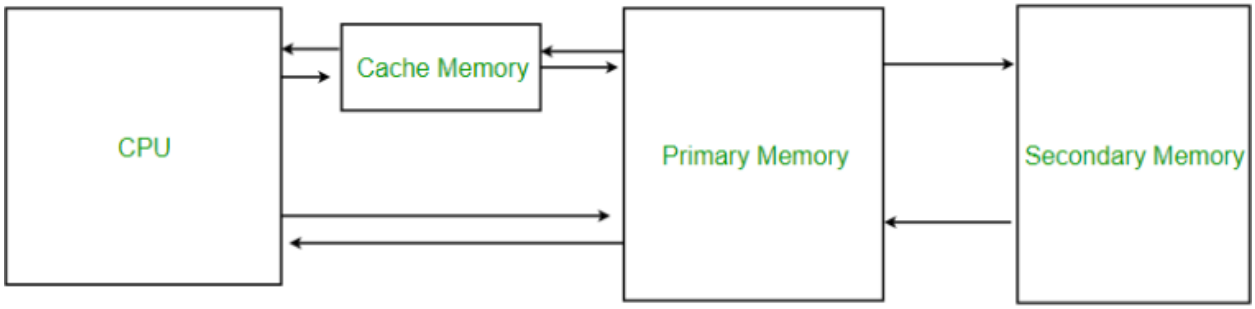
A temporary storage of memory, cache makes data retrieving easier and more efficient.
It is the fastest memory in a computer, and is typically integrated onto the motherboard and directly embedded in the processor or main random access memory.

Cache memory provides faster data storage and access by storing instances of programs and data routinely accessed by the processor. Thus, when a processor requests data that already has an instance in the cache memory, it does not need to go to the main memory or the hard disk to fetch the data.

The processor checks whether a corresponding entry is available in the cache every time it needs to read or write a location, thus reducing the time required to access information from the main memory.

Whenever the processor accesses data for the first time, a copy is made into the cache.

When that data is accessed again, if a copy is available in the cache, that copy is accessed first so the speed and efficiency is increased. If it’s not available, then larger, more distant, and slower memories are accessed (such as the RAM or the hard disk).



Levels of Memory

- **Level 1 or Register:** It is a type of memory in which data is stored and accepted that are immediately stored in the CPU. The most commonly used register is Accumulator, Program counter, Address Register, etc.
- **Level 2 or Cache memory:** It is the fastest memory that has faster access time where data is temporarily stored for faster access.
- **Level 3 or Main Memory:** It is the memory on which the computer works currently. It is small in size and once power is off data no longer stays in this memory.
- **Level 4 or Secondary Memory:** It is external memory that is not as fast as the main memory but data stays permanently in this memory.

Cache Performance:

When the processor needs to read or write a location in the main memory, it first checks for a corresponding entry in the cache.

- If the processor finds that the memory location is in the cache, a **Cache Hit** has occurred and data is read from the cache.
- If the processor does not find the memory location in the cache, a **Cache Miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from the main memory, then the request is fulfilled from the contents of the cache.

The performance of cache memory is frequently measured in terms of a quantity called **Hit ratio.**

Hit Ratio(H) = hit / (hit + miss) = no. of hits/total accesses
Miss Ratio = miss / (hit + miss) = no. of miss/total accesses = 1 - hit ratio(H)

CACHE ACCESS TIME(CACHE HIT TIME): time required to access the word from cache.

MISS PENALTY: The time required to fetch the required block from main memory.

MP= Cache Access Time + Main mem access time

AVG ACCESS TIME OF CPU: Hit Ratio X Cache access time + (1-hit ratio) x miss penalty

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.

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.

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.

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).

How does cache memory work with CPU?

When CPU needs the data, first of all, it looks inside the L1 cache. If it does not find anything in L1, it looks inside the L2 cache. If again, it does not find the data in L2 cache, it looks into the L3 cache. If data is found in the cache memory, then it is known as a cache hit. On the contrary, if data is not found inside the cache, it is called a cache miss.

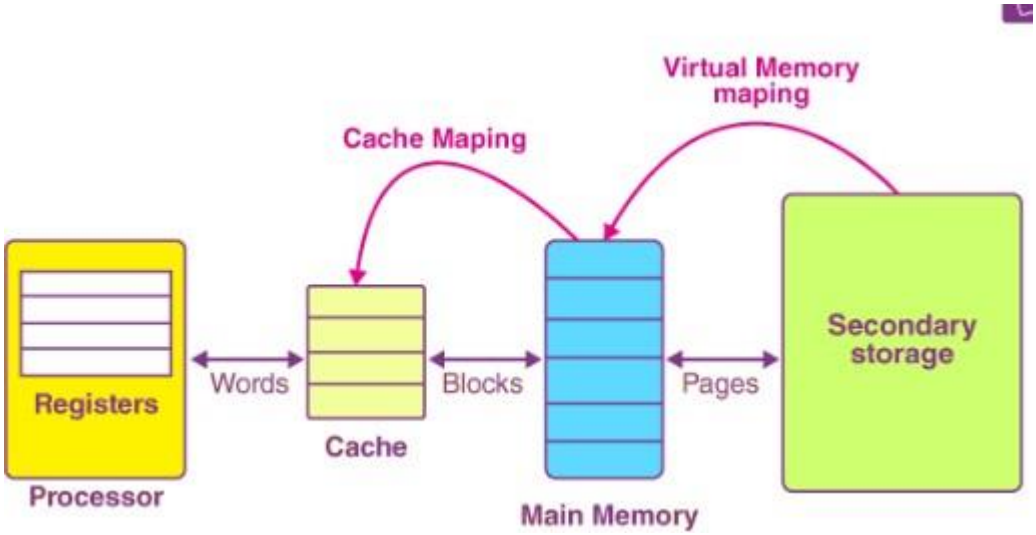
If data is not available in any of the cache memories, it looks inside the Random Access Memory (RAM). If RAM also does not have the data, then it will get that data from the Hard Disk Drive.

So, when a computer is started for the first time, or an application is opened for the first time, data is not available in cache memory or in RAM. In this case, the CPU gets the data directly from the hard disk drive.

Process of Cache Mapping

The process of cache mapping helps us define how a certain block that is present in the main memory gets mapped to the memory of a cache in the case of any cache miss.

In simpler words, cache mapping refers to a technique using which we bring the main memory into the cache memory.

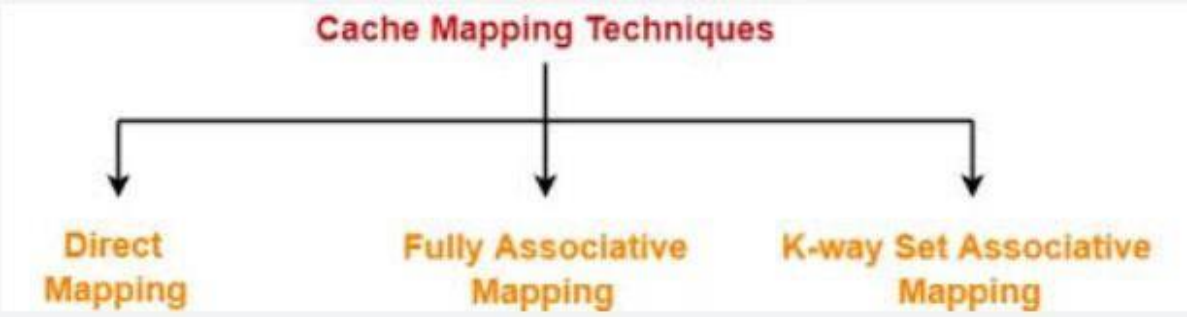


Important Note:

4. The main memory gets divided into multiple partitions of equalsize, known as the **frames or blocks**.

5. The cache memory is actually divided into various partitions of thesame sizes as that of the blocks, known as **lines**.

6. The main memory block is copied simply to the cache during the process of cache mapping, and this block isn't brought at all from themain memory.



f-List of Reference Material

1. Computer Organization & Architecture (10th ed)William Stallings

2. Computer Organization & Design

(5th ed) David A. Patterson &
JOHN L. Hennessy

3. Computer Organization & Design- RISC V ed
   David A. Patterson & JOHN L. Hennessy