CACHE EVOLUTION

Current Cache Design

What is a Cache?

The cache is a very high speed, expensive piece of memory, which is used to speed up the memory retrieval process. Due to it's higher cost, the CPU comes with a relatively small amount of cache compared with the main memory. Without cache memory, every time the CPU requests for data, it would send the request to the main memory which would then be sent back across the system bus to the CPU. This is a slow process. The idea of introducing cache is that this extremely fast memory would store data that is frequently accessed and if possible, the data that is around it. This is to achieve the quickest possible response time to the CPU.

Types of Cache Memory

- **Memory Cache**: A memory cache, sometimes called a cache store or RAM cache, is a portion of memory made of high-speed static RAM (SRAM) instead of the slower and cheaper dynamic RAM (DRAM) used for main memory. Memory caching is effective because most programs access the same data or instructions over and over. By keeping as much of this information as possible in SRAM, the computer avoids accessing the slower DRAM.

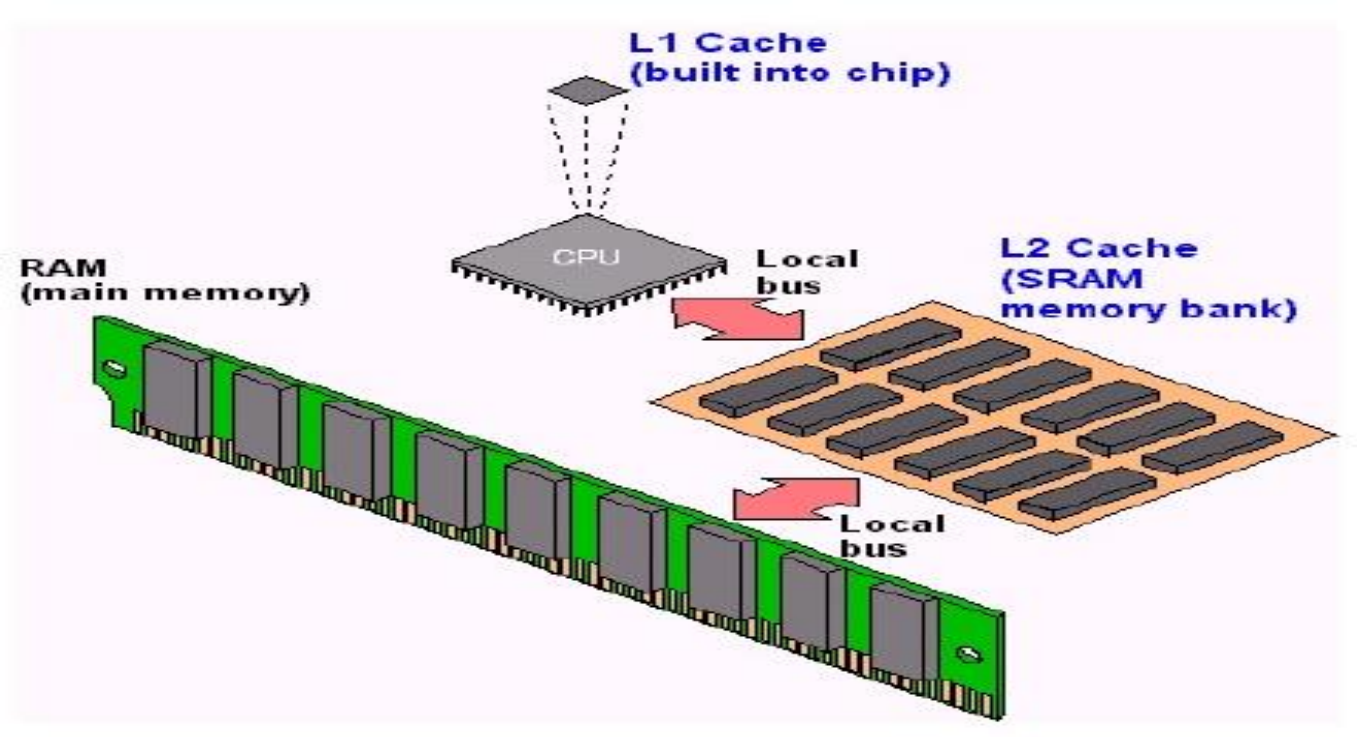
- **Disk Cache**: Disk caching works under the same principle as memory caching, but instead of using high-speed SRAM, a disk cache uses conventional main memory. The most recently accessed data from the disk (as well as adjacent sectors) is stored in a memory buffer. When a program needs to access data from the disk, it first checks the disk cache to see if the data is there. Disk caching can dramatically improve the performance of applications, because accessing a byte of data in RAM can be thousands of times faster than accessing a byte on a hard disk.

Levels of Cache: Cache memory is categorized in levels based on it's closeness and accessibility to the microprocessor. There are three levels of a cache.

- **Level 1(L1) Cache**: This cache is inbuilt in the processor and is made of SRAM(Static RAM) Each time the processor requests information from memory, the cache controller on the chip uses special circuitry to first check if the memory data is already in the cache. If it is present, then the system is spared from time consuming access to the main memory. In a typical CPU, primary cache ranges in size from 8 to 64 KB, with larger amounts on the newer processors. This type of Cache Memory is very fast because it runs at the speed of the processor since it is integrated into it.

- **Level 2(L2) Cache**: The L2 cache is larger but slower in speed than L1 cache. It is used to see recent accesses that is not picked by L1 cache and is usually 64 to 2 MB in size. A L2 cache is also found on the CPU. If L1 and L2 cache are used together, then the missing information that is not present in L1 cache can be retrieved quickly from the L2 cache. Like L1 caches, L2 caches are composed of SRAM but they are much larger. L2 is usually a separate static RAM (SRAM) chip and it is placed between the CPU & DRAM(Main Memory)

Level 3(L3) Cache: L3 Cache memory is an enhanced form of memory present on the motherboard of the computer. It is an extra cache built into the motherboard between the processor and main memory to speed up the processing operations. It reduces the time gap between request and retrieving of the data and instructions much more quickly than a main memory. L3 cache are being used with processors nowadays, having more than 3 MB of storage in it.



Thank You

G.Mahidhar.