**Part B Question.no.2**

**Cache**

A cache is a smaller, faster memory, located closer to a processor core, which stores copies data from frequently used main memory location. A cache represents a layered structure of cache memory within a computer system crafted to boost data retrieval speed and enhance the overall efficiency of the system Many CPUs have multiple of cache levels (L1, L2, L3 and rarely even L4) with different instruction-specific and data-specific. Cache is a component in the computer that stores data so that future requests for the data can be served faster.

**Definition of L1, L2 and L3 cache**

L1 Cache (Level 1 Cache)

A L1 cache is a memory cache which is directly built through the microprocessor, and used for storing the microprocessor’s newly entered data and information. It is also called as the primary cache. It is also fastest memory that is present in a computer. In terms, of priority of access, the L1 cache has the data the CPU is most likely to need while completing a certain tasks.

L2 Cache (Level 2 Cache)

A L2 cache is a CPU cache memory which is located outside and separated from the microprocessor chip core, although it is found on the same processor chip package. It is also called secondary cache. L2 is larger than L1.

L3 Cache (Level 3 Cache)

A L3 cache is a specialized cache that is used by the CPU and it is typically built onto the motherboard and, and in certain case of special processors, within the CPU module itself. L3 cache works together with L1 and L2 cache to improve computer performance by deducting the time it takes to fetch and implement the data. It is located at the main memory.

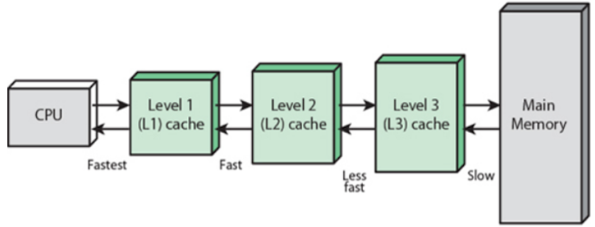


Figure: Cache Memory

**Function of L1, L2, and L3 Cache**

|  |  |  |
| --- | --- | --- |
| L1 Cache | L2 Cache | L3 Cache |
| 1. L1 cache is fast access to frequently used data. | 1. Supplemental cache Storage. | 1. Shared cache s pace. |
| 2. It also minimize memory access latency. | 2. Intermediate data-access. | 2. Inter-Core communication optimization. |
| 3. It enhanced CPU performance. | 3. Reduced main memory access. | 3. Improved scalability. |

**Capacity**

**L1 cache:** Typically L1 is 16KB to 128KB per core.

**L2 cache:** Commonly 256-512 KB, Sometimes high as 1 MB

**L3 cache:** Typically L3 is 2 MB to 32 MB or more, shared among CPU cores in multi-core processors

The cache also exploit the idea of locality, which includes two aspects

1. Temporal locality: Temporal locality means that all those instructions in the caches which access the same set of memory locations within the small time period, such that the data likely remains fetched easily and takes no time in searching for the same instruction.
2. Spatial locality: Spatial locality means that all those instruction which access to memory location and to each other in address space, and refers to the used data instructions which are relatively close in storage.

**→** Cache memory is an indispensable part of modern CPU’s. Although the differences between cache L1, L2 and L3 seems to be complex at first sight, they help to maximize the efficiency and performance of a computer. A meticulously planned architecture improves system responsiveness in practical scenarios by optimizing data access and reduced latency in memory retrievals.

Cache plays a vital role in computer architecture, enhancing the execution speed of programs and enabling efficient processing of complex tasks by minimizing memory access latency.