**Cache Memory-**

Cache memory is a Random Access Memory.

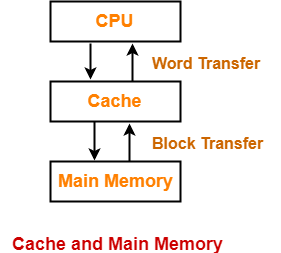
The main advantage of cache memory is its very fast speed.

It can be accessed by the CPU at much faster speed than main memory.

**Location-**

Cache memory lies on the path between the CPU and the main memory.

It facilitates the transfer of data between the processor and the main memory at the speed which matches to the speed of the processor.



Data is transferred in the form of words between the cache memory and the CPU.

Data is transferred in the form of blocks or pages between the cache memory and the main memory.

**Purpose-**

The fast speed of the cache memory makes it extremely useful.

It is used for bridging the speed mismatch between the fastest CPU and the main memory.

It does not let the CPU performance suffer due to the slower speed of the main memory.

**Execution Of Program-**

Whenever any program has to be executed, it is first loaded in the main memory.

The portion of the program that is mostly probably going to be executed in the near future is kept in the cache memory.

This allows CPU to access the most probable portion at a faster speed.

**Step-01:**

 Whenever CPU requires any word of memory, it is first searched in the CPU registers.

Case-01:

 If the required word is found in the CPU registers, it is read from there.

 Case-02:

 If the required word is not found in the CPU registers, Step-02 is followed.

**Step-02:**

When the required word is not found in the CPU registers, it is searched in the cache memory.

Tag directory of the cache memory is used to search whether the required word is present in the cache memory or not.

 Now, there are two cases possible-

Case-01:

 If the required word is found in the cache memory, the word is delivered to the CPU.

This is known as Cache hit.

Case-02:

If the required word is not found in the cache memory, Step-03 is followed.

This is known as Cache miss.

**Step-03:**

When the required word is not found in the cache memory, it is searched in the main memory.

Page Table is used to determine whether the required page is present in the main memory or not.

Now, there are two cases possible-

Case-01:

If the page containing the required word is found in the main memory, The page is mapped from the main memory to the cache memory. This mapping is performed using cache mapping techniques.

Then, the required word is delivered to the CPU.

Case-02:

If the page containing the required word is not found in the main memory,

A page fault occurs.

The page containing the required word is mapped from the secondary memory to the main memory.

Then, the page is mapped from the main memory to the cache memory.

Then, the required word is delivered to the CPU.

Multilevel Cache Organization-

A multilevel cache organization is an organization where cache memories of different sizes are organized at multiple levels to increase the processing speed to a greater extent.

The smaller the size of cache, the faster its speed.

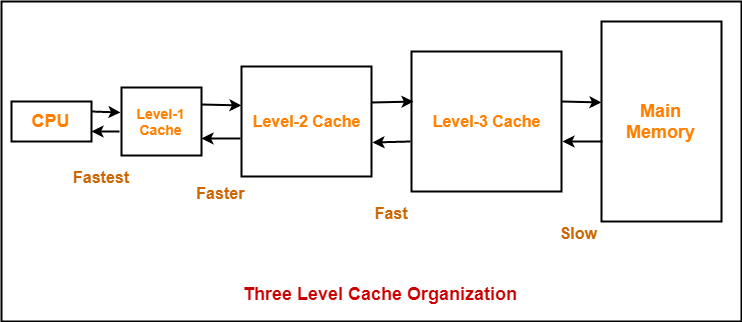
The smallest size cache memory is placed closest to the CPU.

This helps to achieve better performance in terms of speed.

Example-

Three level cache organization consists of three cache memories of different size organized at three different levels as shown below-

Size (L1 Cache) < Size (L2 Cache) < Size (L3 Cache) < Size (Main Memory)



## What are Cache Hit and Misses?

Similarly to know what caching is, it’s also essential to understand what hit and miss ratios are as well as miss penalties. That way, it’s easier to know what hit and miss ratios in caches are and why they’re important in helping determine your cache’s performance.

A cache hit refers to the situation wherein the cache is able to successfully retrieve data and content that was saved to it, and then display it on a web page.

A cache miss occurs in the opposite situation. The site requests the content from the cache, but after searching for it, the cache determines that content wasn’t saved. That’s when the cache saves the content so it’s available the next time it’s requested.

When a cache miss occurs, it takes up extra time and server resources which ends up slowing down your page speed load times.

One cache miss isn’t a big deal, but the more that happens, the worse it is for your server’s resources, and page load times.

This is where miss penalties come into the picture.

This delay that occurs in your page load times due to cache misses are known as miss penalties.

You can check out [**Cache Miss vs Cache Hit: What’s the Difference?**](https://wp-rocket.me/blog/cache-miss-vs-cache-hit/) for details.

## Hit and Miss Ratios in Caches

Hit and miss ratios in caches have a lot to do with cache hits and misses.

A hit ratio is a calculation of cache hits, and comparing them with how many total content requests were received.

A miss ratio is the flip side of this where the cache misses are calculated and compared with the total number of content requests that were received.

### The Importance of Hit and Miss Ratios in Caches

Hit and miss ratios are significant because, as mentioned earlier, they can give you a good idea of how well your cache is performing, and if its performance is optimized.

If you have a high hit ratio and low miss ratio, that means your cache is operating well. It also means that content is likely being retrieved from the cache quickly, and page load times are also as fast as possible for end users.

If the opposite is true, and you’re able to calculate that your cache’s miss ratio is high, and you hit ratio is low, then your cache isn’t running as well as it could be, and your users are seeing slower page load times than they should be seeing.

Knowing this, you can troubleshoot the issue to fix it up, and get your cache running more smoothly.

For example, if you have a high miss ratio, an option could be expanding your cache since the larger it is, the more data it can hold, and the less cache misses you should have as a result.

But, how do you calculate hit and miss ratios in caches?

Many CDNs display cache hits, misses, and the total number of [**content requests**](https://wp-rocket.me/blog/reduce-http-requests-speed-wordpress-site/). If they also show other metrics such as hit and miss ratios, then you now know what they mean, and you can start troubleshooting, if necessary.

On the other hand, if you don’t have access to these ratios, you calculate them with the number of cache hits, misses, and total cache content accesses.

### How to Calculate a Hit Ratio

To calculate a hit ratio, divide the number of cache hits with the sum of the number of cache hits, and the number of cache misses.

For example, if you have 51 cache hits and three misses over a period of time, then that would mean you would divide 51 by 54. The result would be a hit ratio of 0.944.

You can also choose to express this as a percentage by multiplying the end result by 100.

In the example above, the 0.944 result would be multiplied by 100 to get a hit ratio of 94.4%.

Alternatively, you can find out the hit ratio if you already know the miss ratio. Then, you can subtract one by the miss ratio.

You can do this because the hit and miss ratios equal to one.