



Caching

Load balancing helps you scale horizontally across an ever-increasing number of servers, but **caching** will enable you to make vastly better use of the resources that you already have, as well as making otherwise unattainable product requirements feasible.

Caches take advantage of the locality of reference principle: recently requested data is likely to be requested again.

They are used in almost every computing layer: hardware, OS, web browsers, web applications, and more.

A **cache** is like short-term memory: it has a limited amount of space, but is typically faster than the original data source and contains the most recently accessed items.

Caches can exist at all levels in architecture, but are often found at the level nearest to the front end, where they are implemented to return data quickly without taxing downstream levels.

Application Server Cache

Placing a cache directly on a request layer node enables the local storage of response data.

Each time a request is made to the service, the node will quickly return locally cached data if it exists in the cache — if there is a cache hit.

If it is not in the cache — there is a cache miss — the requesting node will fetch the data from the disk.

The cache on one request layer node could also be located in memory (which is very fast) and on the node's local disk (faster than going to network storage).

What happens when you expand this to many nodes?

If the request layer is expanded to multiple nodes, it is still quite possible to have each node host its own cache.

However, if your load balancer randomly distributes requests across the nodes, the same request will go to different nodes, thus increasing cache misses.

Two choices for overcoming this hurdle are global caches and distributed caches.

Content Distribution (or Delivery) Networks (CDNs)

CDNs are a kind of cache that come into play for sites serving a large amount of static media.

In a typical CDN setup, a request will first ask the CDN for a piece of static media. Then, the CDN will serve that content if it has it locally available. Otherwise, the CDN will query the backend servers for the file, cache it locally, and then serve it to the requesting user.

Cache Invalidation

While caching is fantastic, it requires some maintenance to keep the cache coherent with the source of truth (e.g. database).

If the data is modified in the database, it should be invalidated in the cache. If this does not occur, it can cause inconsistent application behaviour.

Solving this problem is known as cache invalidation.

There are three main schemes that are used.

Write-Through Cache

Under this scheme, data is written into the cache and the corresponding database simultaneously.

The cached data allows for fast retrieval and, since the same data gets written in the permanent storage, we will have complete data consistency between the cache and the storage.

Also, this scheme ensures that nothing will get lost in the case of a crash, power failure, or other system disruptions.

Although write-through minimises the risk of data loss, since every write operation must be done twice before running success to the client, this scheme has the disadvantage of higher latency for write operations.

Write-Around Cache

This technique is similar to the write-through cache, but data is written directly to permanent storage, bypassing the cache.

This can reduce the cache from being flooded with write operations that will not subsequently be re-read, but has the disadvantage that a read request for recently written data will create a “cache miss” and must be read from slower backend storage and experience higher latency.

Write-Back Cache

Under this scheme, data is written to the cache alone, and completion is immediately confirmed to the client.

The write to permanent storage is done after specified intervals or under certain conditions (like when it is evicted from the cache — think dirty bit).

This results in low latency and high throughput for write-intensive applications.

However, this speed comes with the risk of data loss in the case of a crash or other adverse event because the only copy of the written data is in the cache.

Cache Eviction Policies

The following are some of the most common cache eviction policies:

- First In First Out (FIFO) — the cache evicts the first block accessed first without any regard to how often or how many times it was accessed before
- Last In First Out (LIFO) — the cache evicts the block accessed most recently first without any regard to how often or how many times it was accessed before
- Least Recently Used (LRU) — discards the least recently used items first
- Most Recently Used (MRU) — discards, in contrast to LRU, the most recently used items first
- Least Frequently Used (LFU) — counts how often an item is needed, and those that are used least often are discarded first
- Random Replacement (RR) — randomly selects a candidate item and discards it to make space when necessary

