Definition

- 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
- Caching will enable you to make better use of the resources you already have
- Caches take advantage of the locality of reference principle:
    - recently requested data is likely to be requested again
- 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 Cache

- Application caching requires explicit integration in the application code itself.
- Usually, it will check if a value is in the cache; if not, retrieve the value from the database; then write that value into the 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.
- If it is not in the cache, the requesting node will fetch the data from the disk.
- The cache on one request layer node could also be located both 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's 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.
- How can we overcome this hurdle?
    - By using global caches and distributed caches.

Database Cache

- When you flip your database on, you're going to get some level of default configuration which will provide some degree of caching and performance.
- Those initial settings will be optimized for a generic use case, and by tweaking them to your system's access patterns you can generally squeeze a great deal of performance improvement.
- The beauty of database caching is that your application code gets faster "for free"
- A talented DBA or operational engineer can uncover quite a bit of performance without your code changing

Content Delivery (or Distribution) Network (CDN)

- Definition:
    - CDNs are a kind of cache that comes into play for sites serving large amounts of static media.
- How it works:
    - a request will first ask the CDN for a piece of static media.
    - the CDN will serve that content if it has it locally available.
    - If it isn't available, the CDN will query the back-end servers for the file, cache it locally, and serve it to the requesting user.
- Note:
    - If the system we are building is not large enough to have its own CDN, we can ease a future transition by serving the static media off a separate subdomain (e.g., static.yourservice.com) using a

lightweight HTTP server like Nginx, and cutting over the DNS from your servers to a CDN later.

Cache Invalidation

- Issues:
    - While caching is fantastic, it requires some maintenance to keep the cache coherent with the source of truth
    - If the data is modified in the database, it should be invalidated in the cache; if not, this can cause inconsistent application behavior.
- Solution:
    - use cache invalidation
    - there are three main schemes that are used:
        - Write-through cache:
            - 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.
            - this scheme ensures that nothing will get lost in case of a crash, power failure, or other system disruptions.
            - this scheme has the disadvantage of higher latency for write operations since every write operation must be done twice
        - Write-around cache:
            - This technique is similar to write-through cache, but data is written directly to permanent storage, bypassing the cache.
            - This can reduce the cache 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 back-end storage and experience higher latency.
    - Write-back cache:
        - Under this scheme, data is written to cache alone, and completion is immediately confirmed to the client.
        - The write to the permanent storage is done after specified intervals or under certain conditions.
        - This results in low latency and high throughput for write-intensive applications;
        - However, this speed comes with the risk of data loss in case of a crash or other adverse event because the only copy of the written data is in the cache.

In-memory Caches

- The most potent–in terms of raw performance–caches you’ll encounter are those that store their entire set of data in memory.
- Memcached and Redis are both examples of in-memory caches
- On the other hand, you’ll generally have far less RAM available than disk space, so you’ll need a strategy for only keeping the hot subset of your data in your memory cache.

Cache eviction policies

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 the most recently used items first.
- Least Frequently Used (LFU):
    - Counts how often an item is needed. 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.