Cache

In the hierarchy of computer memory, the cache occupies a critical role, acting as a high-speed buffer between the processor and memory (RAM). The principle of locality is crucial to cache. Temporal locality refers to reuse of recently used data and spatial locality refers to the use of data near recently used data.

The CPU first checks L1 cache, which is on-chip. A 'hit' happens if data is found. If not, then it is a 'miss', and the next lower level of memory needs to be checked.

The multi-layer cache approach ensures the processor spends most of its time working. Replacement policies include Least Recently Used. (LRU).

Browser cache also exists for recently visited pages, etc. Content Delivery Networks are also a variant of cache.