Reading 18 (No Reading 17)

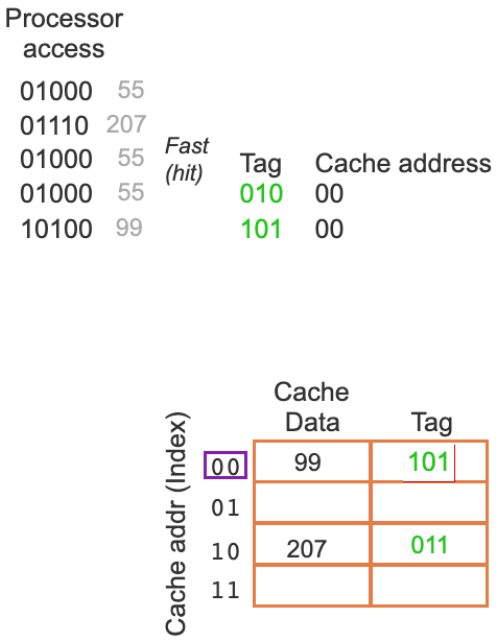
19.1 Cache basics: Part 1

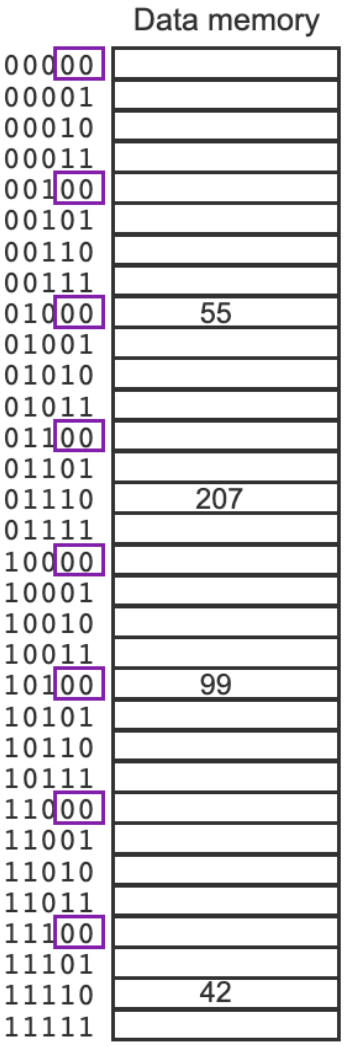
- A processor's chip has limited space, a processor's memory commonly exists on a seperate chip (or chips).
  - ↳ Access between chips is slower than within a chip.
  - ↳ So for faster access, a processor may have a small on-chip memory, called a cache.
    - ↳ The cache stores data that will be used frequently. (think of it as a favorites contact list)
    - ↳ cache is usually SRAM.

- Initially, a cache is empty.
  - ↳ For later reads, the system checks the cache first; if an item's copy exists in the cache (a hit), the system quickly reads that copy; else (a miss) the system first copies the item from memory to cache.

- A hit may require only 1 clock cycle, while a miss may require more.

- A direct-mapped cache directly maps memory addresses to cache addresses using a subset of address bits (called an index), storing the remaining the bits (called the tag) in the cache entry.





- Multiple addresses can map to the same cache address, but only one can be copied into cache at a time.