9. Cache Memory

Chapter 4: sections 4.1, 4.2

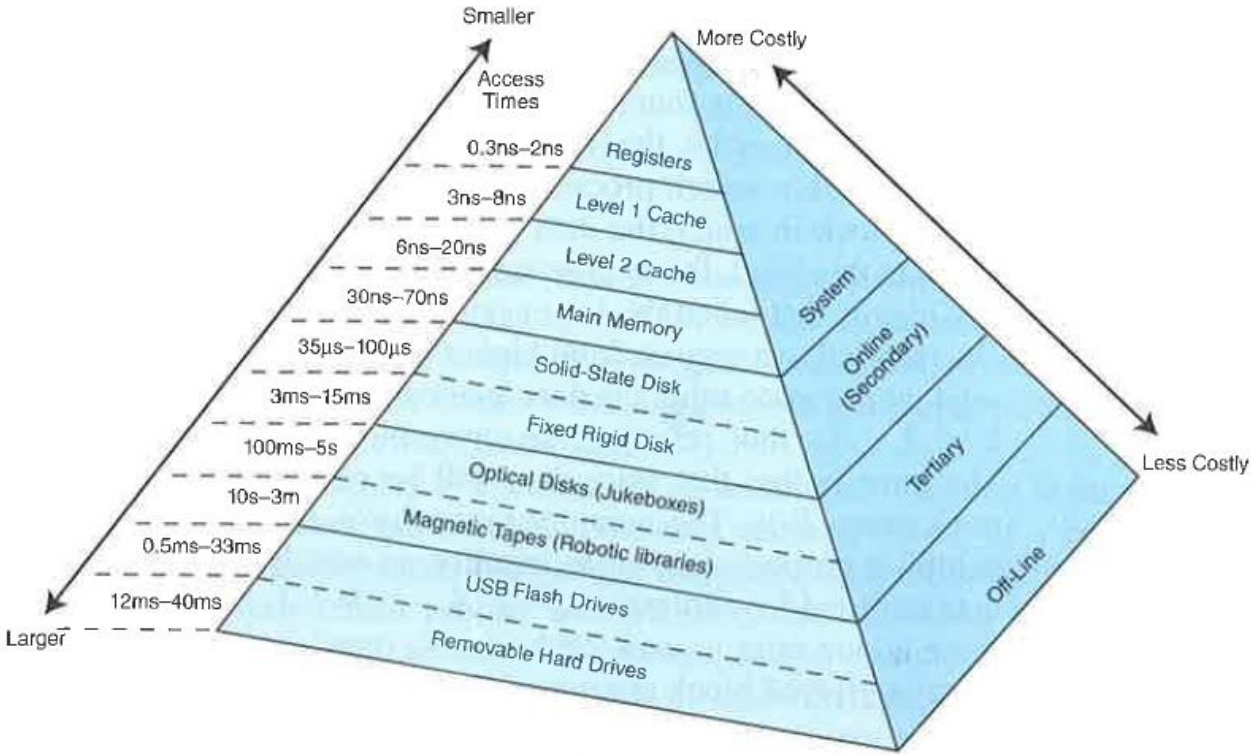
Sections 4.1, 4.2

- Reading:
  - Section 4.1 (Computer Memory System Overview)
  - Section 4.2 (Cache Memory Principles)

Chapter 4
Cache Memory

Location Internal (e.g. processor registers, cache, main memory) External (e.g. optical disks, magnetic disks, tapes) **Capacity** Number of words Number of bytes **Unit of Transfer** Word Block **Access Method** Sequential Direct Random Associative	**Performance** Access time Cycle time Transfer rate **Physical Type** Semiconductor Magnetic Optical Magneto-optical **Physical Characteristics** Volatile/nonvolatile Erasable/nonerasable **Organization** Memory modules

Memory Hierarchy



The Essentials of Computer Organization and Architecture 4th

Memory Design

- Goal
  - Provide for large-capacity memory in computer systems because the capability is needed
    - The capacity is needed
    - The cost per bit is low
  - Use lower-capacity memories
    - Short access times are needed for better performance
  - Decrease the frequency of access to slower memory

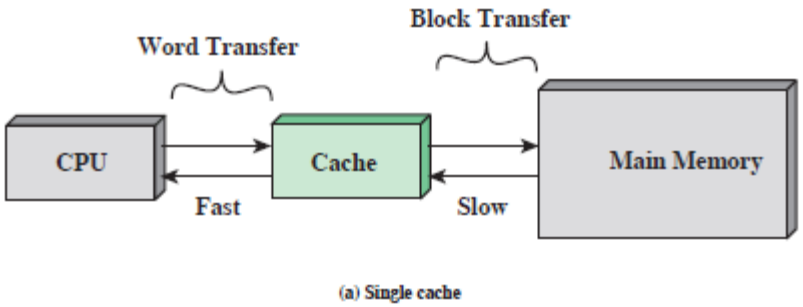
Locality of Reference

- During program execution, memory and data references tend to cluster due to the fact that programs contain
  - functions
  - Loops
- Locality of reference allows the designer to take advantage of small high speed memory

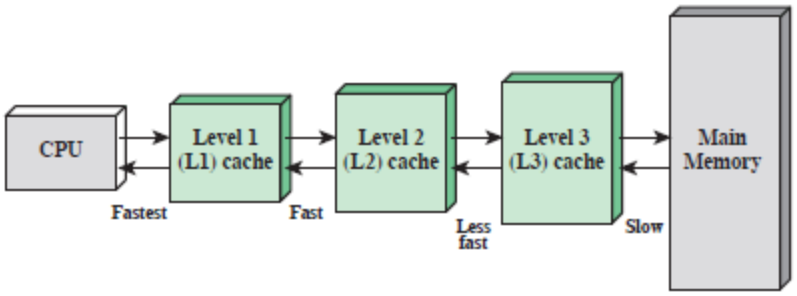
CACHE MEMORY PRINCIPLES

Cache Memory

- The goal of cache memory is to yield a memory speed of the fastest memories available while at the same time providing a much larger memory at the cheapest possible price.

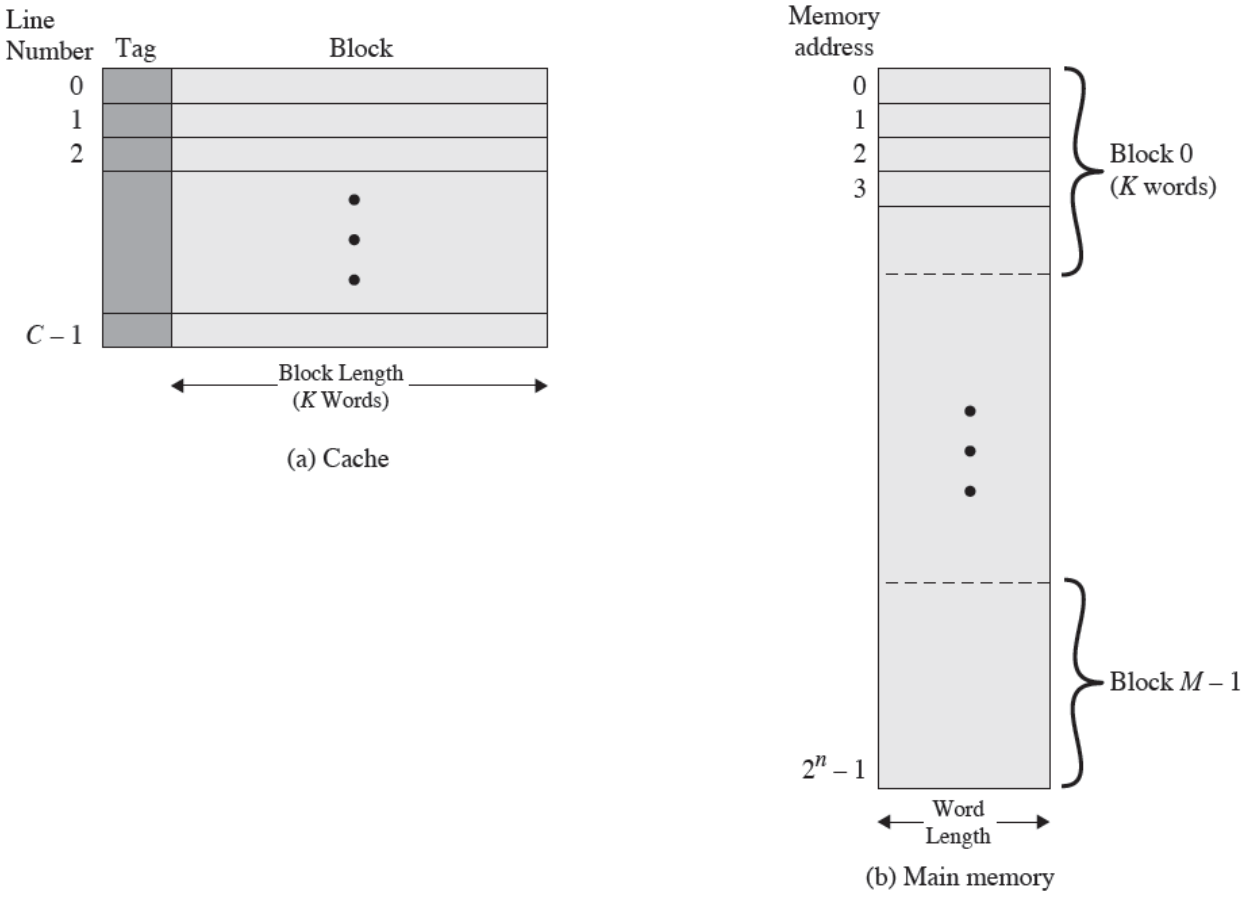


(a) Single cache



(b) Three-level cache organization

Cache/Memory Structure



Cache/Memory Structure

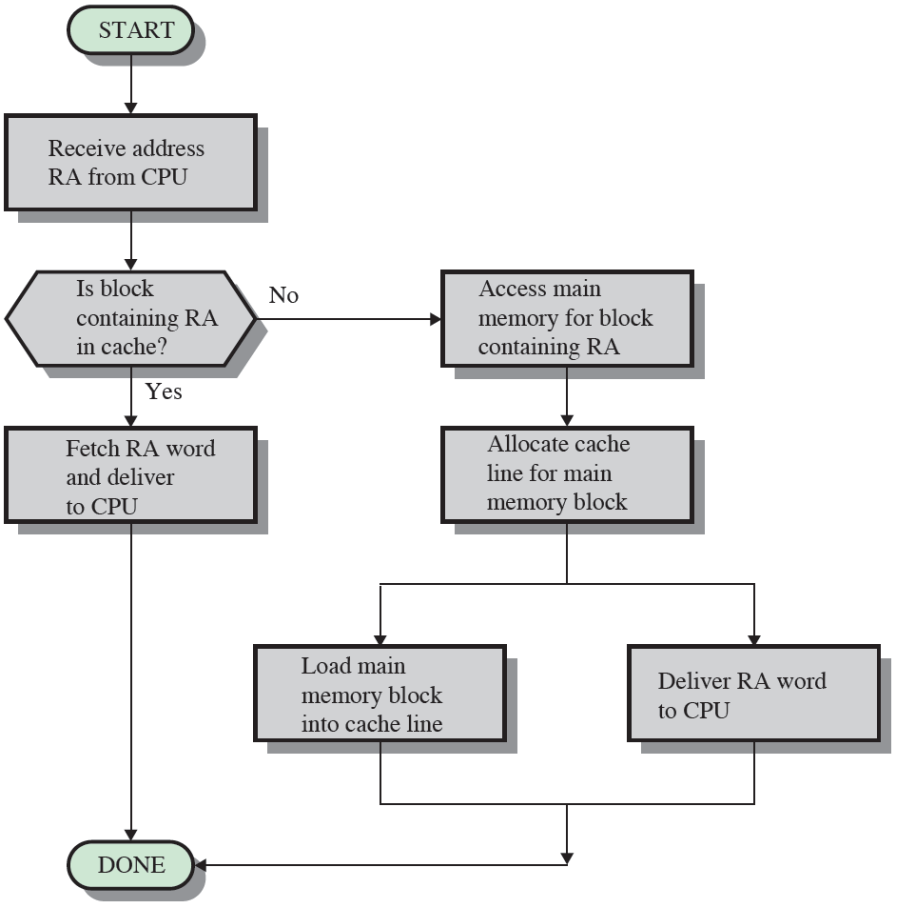
- Things to note:
  - A main memory has 2^n addressable memory locations
  - Each word has a unique $n - bit$ address
  - A block of $K - words$ is mapped to a line in the cache
  - The number of words in a cache line is called the line size

Questions

- How many blocks of information exist in main memory?
- What is the line size in the previous picture?
- If M is the number of main memory blocks and C is the number of cache lines, would you say that C is a little smaller than M or a lot smaller than M? Why?

Cache Read

- RA is read address



Cache Organization