COA

Chapter 04: | Cache Memory – Principles

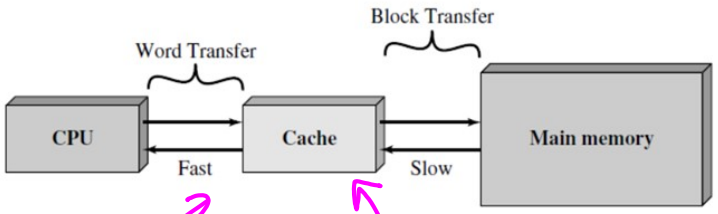
Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

AGENDA

- Cache Memory Principles
  - Cache / Main Memory Structure
  - Cache Read Operation

Cache and Main Memory



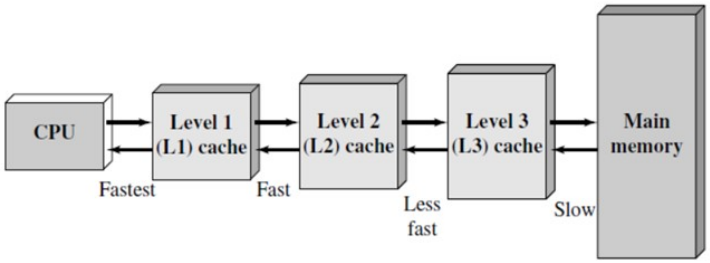
1 level of cache

Access time ↓

Word

Block

Cache and Main Memory



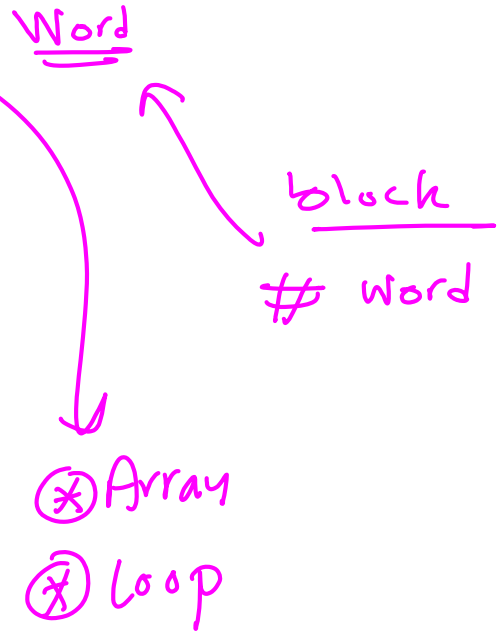
- Memory Capacity ↑
- Speed ↓

Cache and Main Memory

- The cache contains a copy of portions of main memory
  - When the processor attempts to read a word of memory, a **check is made** to determine if the word is in the cache. If so, the word is delivered to the processor.
  - **If not**, a **block of main memory**, consisting of some fixed number of words, is read into the cache and then the word is delivered to the processor.

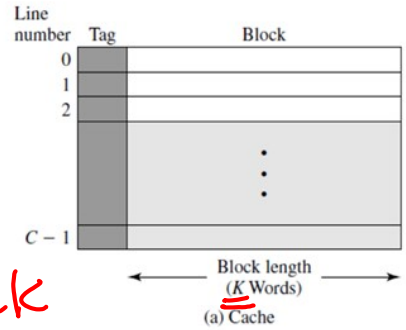
Cache and Main Memory

- Because of the phenomenon of **locality of reference**, when a block of data is fetched into the cache to satisfy a single memory reference, **it is likely that** there will be **future references** to that same memory location or to other words in the block

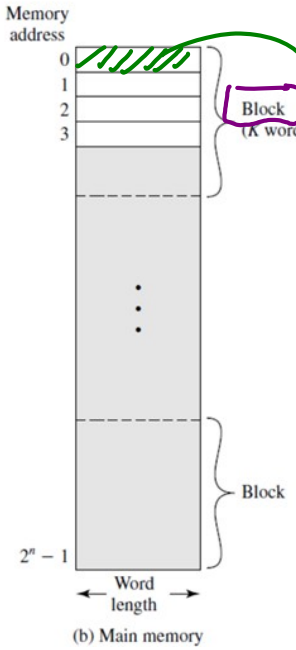


Cache / Main Memory Structure

cache



RAM



line = Block

tag → control bit

lines = m << M

Address = 2^n

address bw = n-bit

3-bit ⇒ 2^3 ⇒ 8

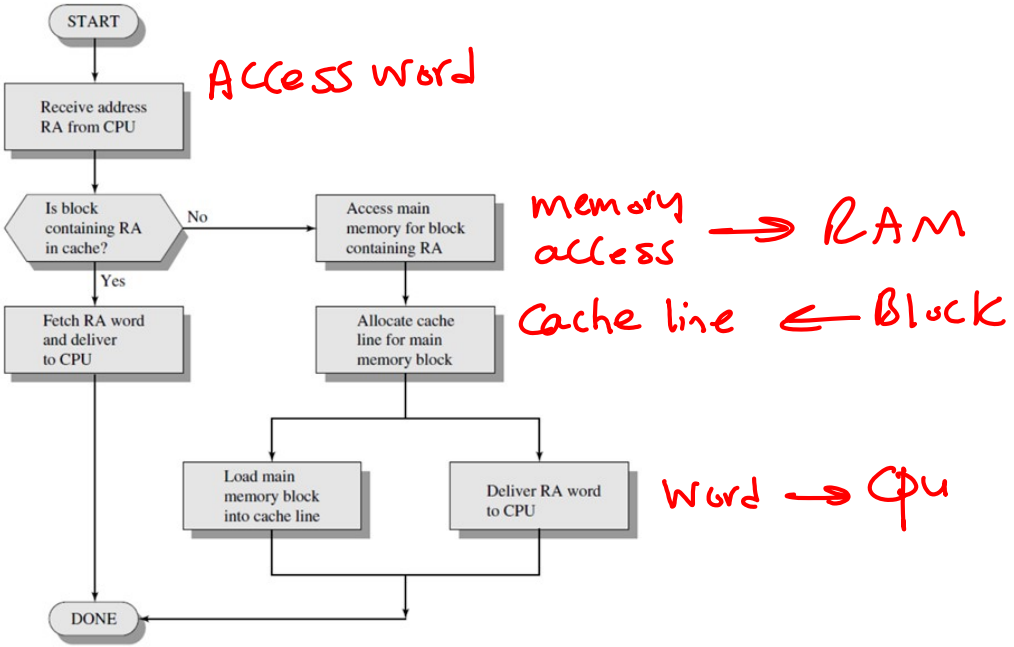
8-bit ⇒ 2^8 ⇒ 256

block ⇒ # word

K word

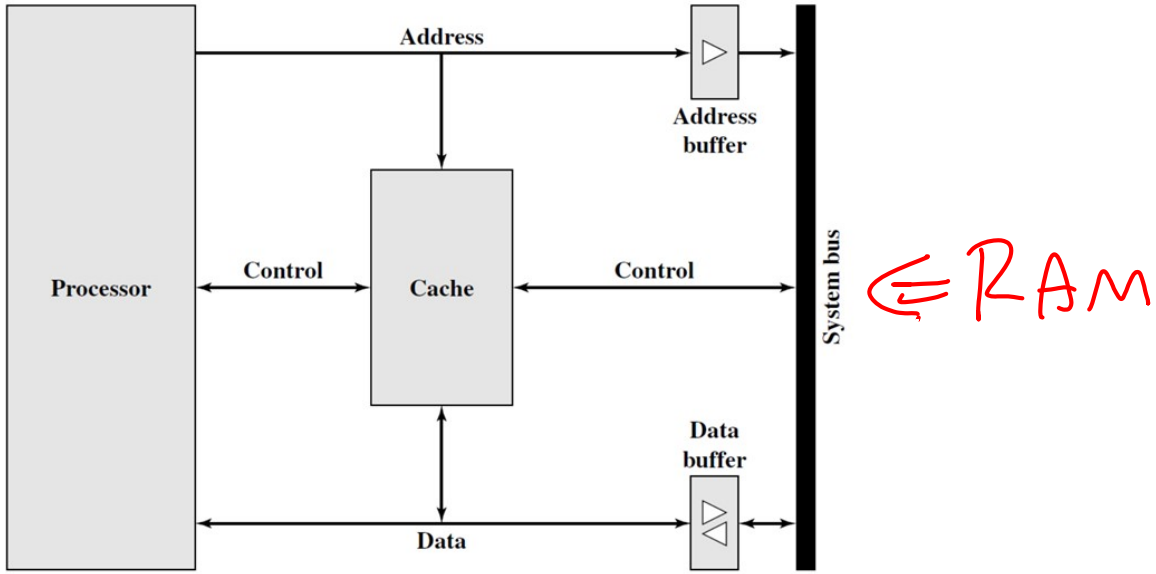
block = $2^n / K$ ⇒ M

Cache Read Operation



Typical Cache Organization

disable ⇒ in cache



⇐ RAM