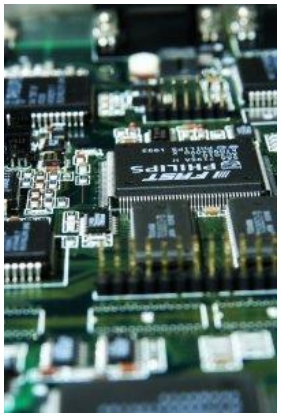
CSE 321a

Computer Organization (1)

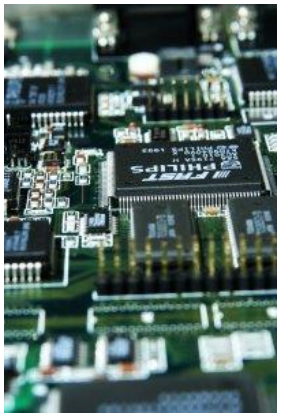
تنظيم الحاسبات (1)



3rd year, Computer Engineering

Fall 2017

Lecture #4



Dr. Hazem Ibrahim Shehata

Dept. of Computer & Systems Engineering

Credits to Dr. Ahmed Abdul-Monem Ahmed for the slides

Administrivia

- Assignment #1:
  - Released: last Sunday.
  - Due: Sunday, Oct. 15, 2017.

Website: http://hshehata.github.io/courses/zu/cse321a
Office hours: Sunday 1:00pm-2:00pm

Chapter 4. Cache Memory

Characteristics of Memory Systems

1. Location
2. Capacity
3. Unit of transfer
4. Access method
5. Performance
6. Physical type
7. Physical characteristics
8. Organization

1,2. Location and Capacity

1. Location
   - —Internal (to computer)
     - Directly accessible by CPU.
     - e.g., CPU registers, cache, MM.
   - —External (to computer)
     - Accessible by CPU via an I/O module (controller).
     - e.g., Secondary storage disks and tapes.

2. Capacity
   - —Internal memory
     - # of bytes (or words)
     - Word length (8, 16, 32, …bits).
   - —External memory
     - # of bytes.

Concepts for Internal Memory

- Word
  - —Natural unit of organization of memory.
  - —Usually holds an integer or an instruction.
  - —Not always the case!
  - —X86: word → 16 bits, instruction → 1+ words!!
- Addressable unit (i.e., location)
  - —Smallest location that can be uniquely addressed.
  - —Word, byte, or both.
  - —An A-bit address is needed for a 2^A addressable units.
  - —X86: location → 8 bits → byte-addressable memory.

3. Unit of Transfer

- Internal memory
  - —Not necessarily the addressable unit or the word!!!
  - —Number of bits read from or written to memory at a time.
  - —Governed by data bus width (# of data lines, MM).
- External memory
  - —Usually a block, which is much larger than a word.

4. Access Methods (1)

- Sequential
  - —e.g. tape (demo: https://www.youtube.com/watch?v=y-RpCzxPqac)
  - —Memory is organized into units of data, called records.
  - —Start at the beginning and read through in order.
  - —Stored addressing information is needed.
  - —Shared read/write mechanism.
  - —Access time depends on data location & previous location.

4. Access Methods (2)

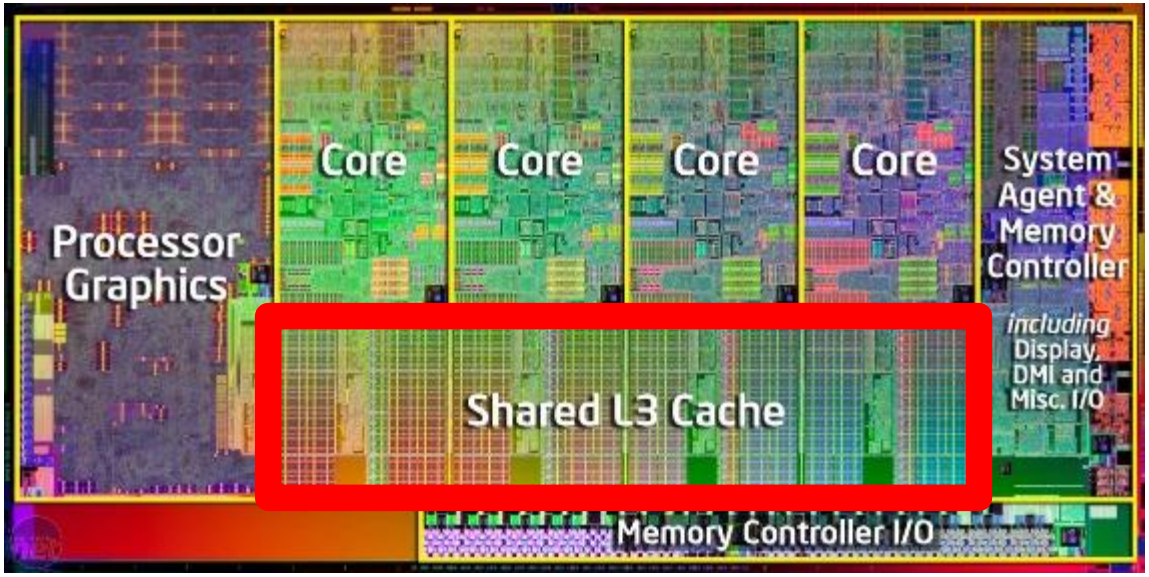
- Direct
  - —e.g. disk (demo: https://www.youtube.com/watch?v=9eMWG3fwiEU)
  - —Individual records/blocks have unique address based on physical location.
  - —Access is by jumping to vicinity plus sequential search.
  - —Shared read/write mechanism.
  - —Access time depends on data location & previous location.

4. Access Methods (3)

- Random
  - —e.g. RAM
  - —Individual addresses identify locations exactly.
  - —Access time is independent of location of previous access.

4. Access Methods (4)

- Associative
  - —e.g. cache
  - —Data is located by a comparison with contents of a portion of the store.
  - —Access time is independent of location or previous access



5. Performance

- Access time
  - Random: time between presenting add. and getting data.
  - Non-Random: time to position rd/wr mechanism at desired location
- Memory cycle time
  - Time may be required for memory to "recover" before next access.
  - Cycle time = access + recovery
- Transfer rate
  - Rate at which data can be moved (e.g. X bps)
  - Random: R = 1 / Cycle Time (in "data units per second")
  - Non-random: $R = N / (T_N - T_A)$ (in "bps")
    - T_N: Av. time to read/write N bits.
    - N: # of bits.
    - T_A: Av. access time.
    - R: transfer rate.

6,7. Physical Types & Physical Characteristics

6. Physical Types

- —Semiconductor
  - RAM & ROM
- —Magnetic
  - Disk & Tape
- —Optical
  - CD & DVD

7. Physical Characteristics

- —Volatility
  - Volatile: Information decays and lost when the power is off.
  - Non-volatile: No power is needed to retain info. (e.g., magnetic surface memory).
  - Semiconductor memory could be volatile or nonvolatile.
- —Erasablilty

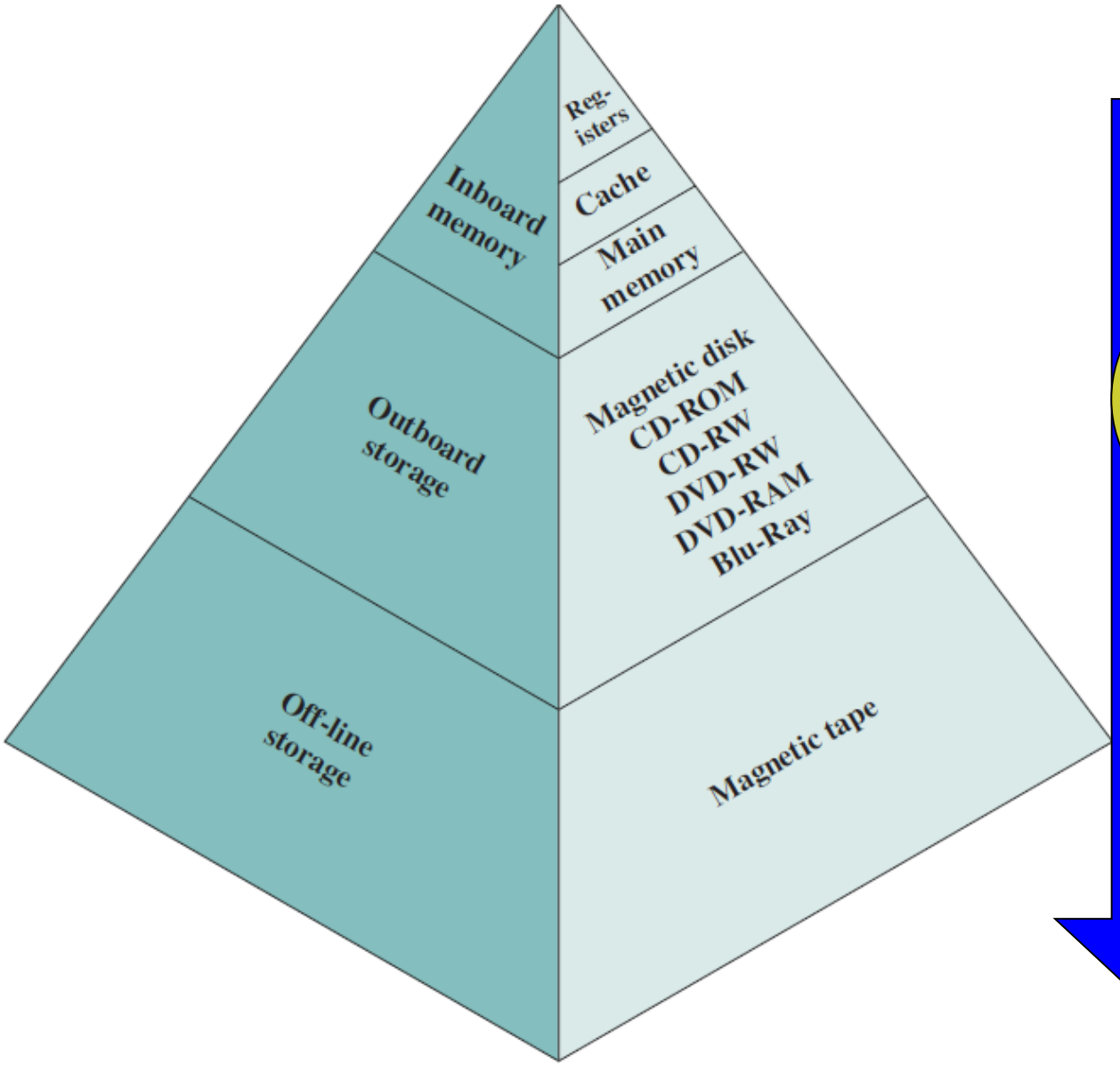
8. Organization

- Key for random-access memory
- Physical arrangement of bits to form words.
- Obvious arrangement is not always used.
  - —i.e., Cell rows may not correspond to words!!
- To be explained more later in internal memory.

Design Constraints on Memory

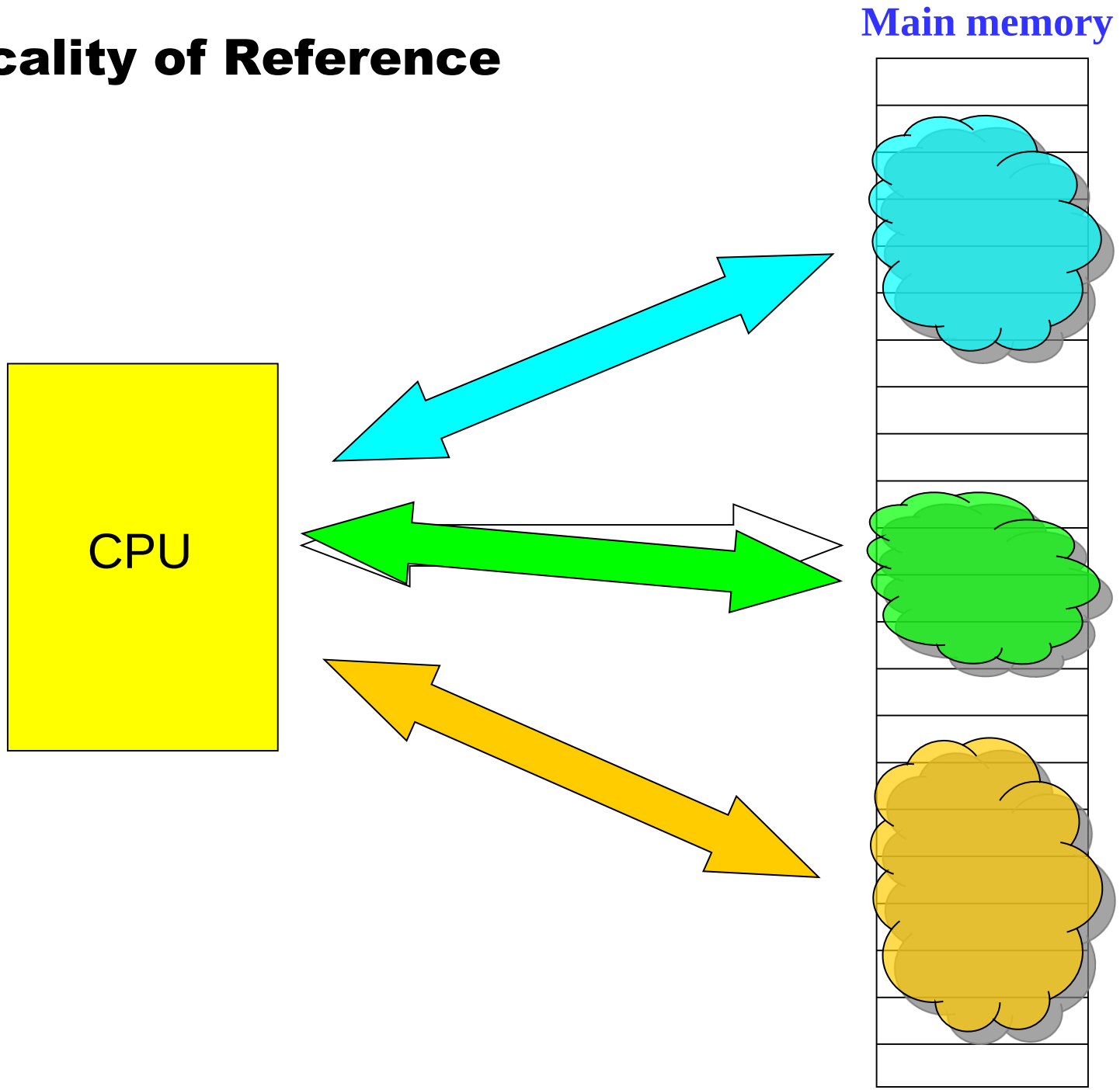
- How much?
  - —Capacity: bigger is better!
- How fast?
  - —Speed: keep up with CPU.
- How expensive?
  - —Cost: reasonable compared to other components.
- **Problem**: Trade-off among these three characteristics!!
  - —No single memory technology has it all!!
- **Solution**: memory hierarchy.

Memory Hierarchy - Diagram



- Less cost
- Higher capacity
- Greater access time
- Less access frequency

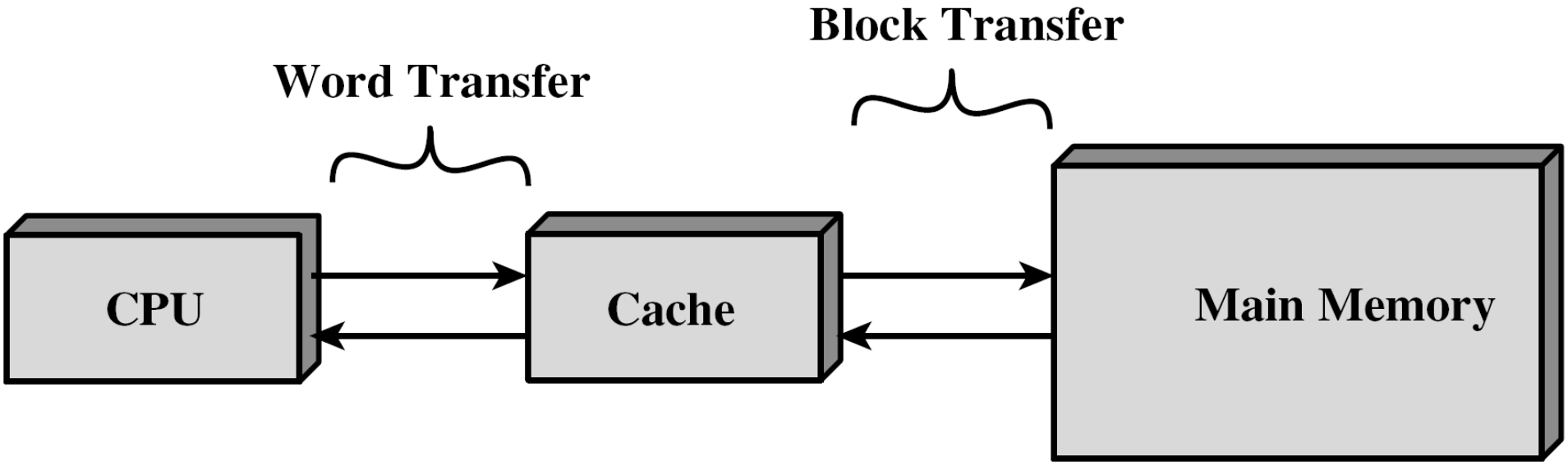
Locality of Reference



Locality of Reference

- During the course of execution of a program, memory references tend to cluster (for both instructions and data).
  - —e.g., loops, subroutines.
  - —e.g., operations on tables and arrays.
- Over a short period of time, CPU is working with fixed clusters of memory references.
- Over a long period of time, the clusters in use change.
- This principle can be applied across all levels of the memory hierarchy.

Cache Memory – Concept

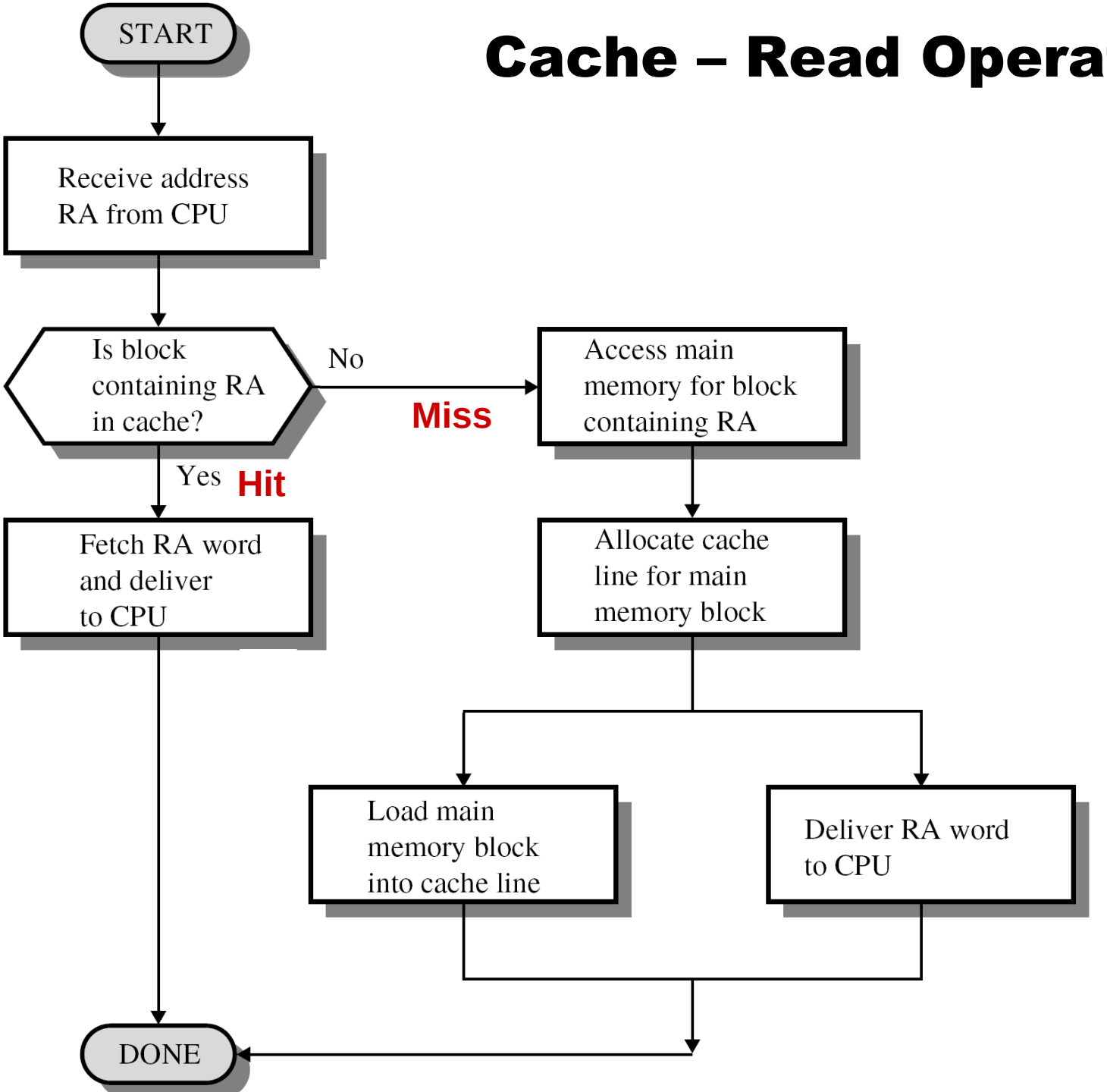


- Small amount of fast memory.
- Sits between normal main memory and CPU.
- May be located on CPU chip.
- Not usually visible to the programmer or CPU
- Volatile, uses semiconductor technology.

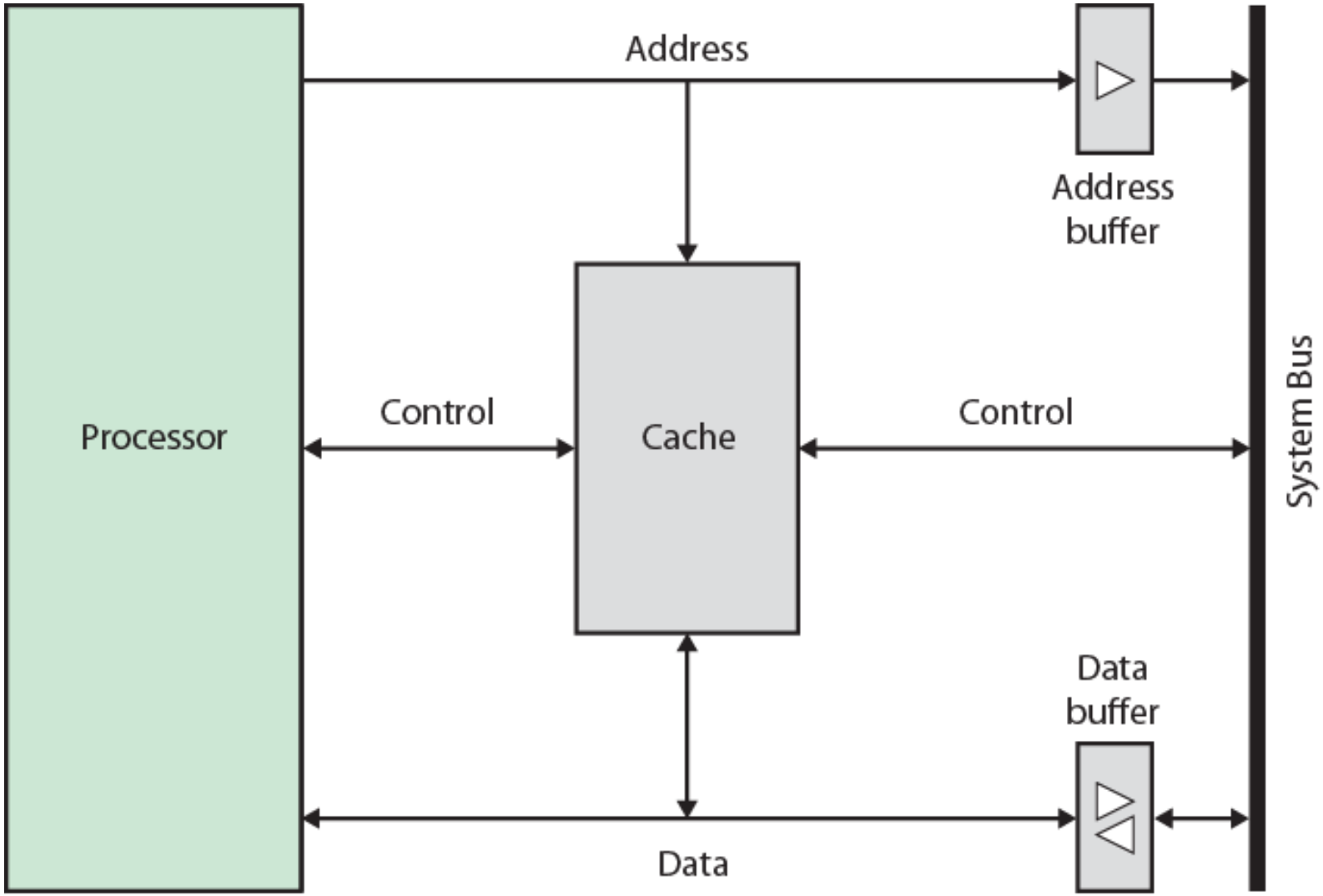
Cache Memory – Operation

- CPU requests contents of memory location.
- Check cache for this data.
- If present ➔ cache hit, get from cache (fast).
  - Because of locality of reference, this location, or a close one, is likely to be referenced soon.
- If not present ➔ cache miss, read required block from MM to cache.
- Then deliver from cache to CPU.
- Cache includes tags to identify which block of main memory is in each cache slot.

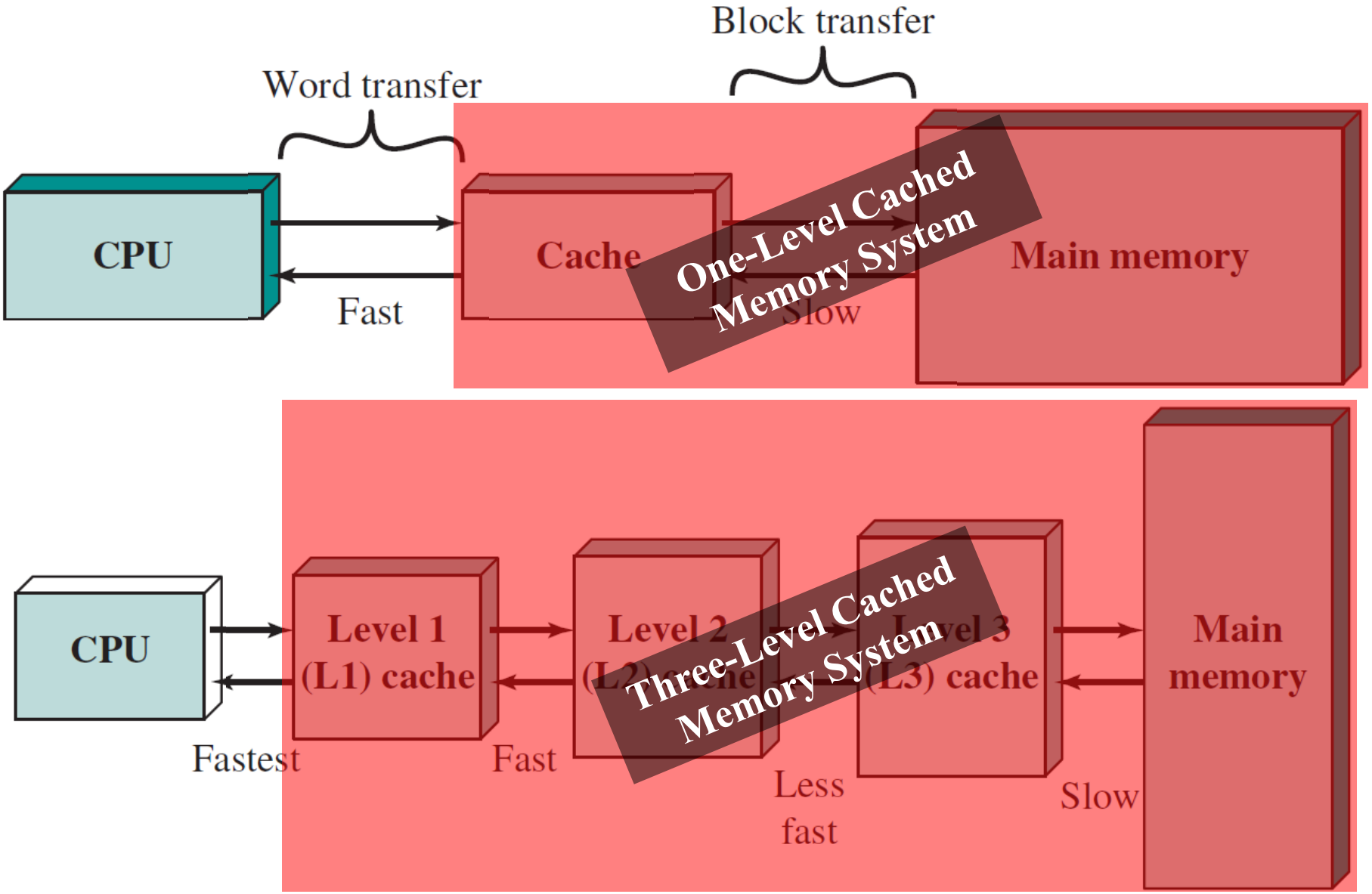
Cache – Read Operation



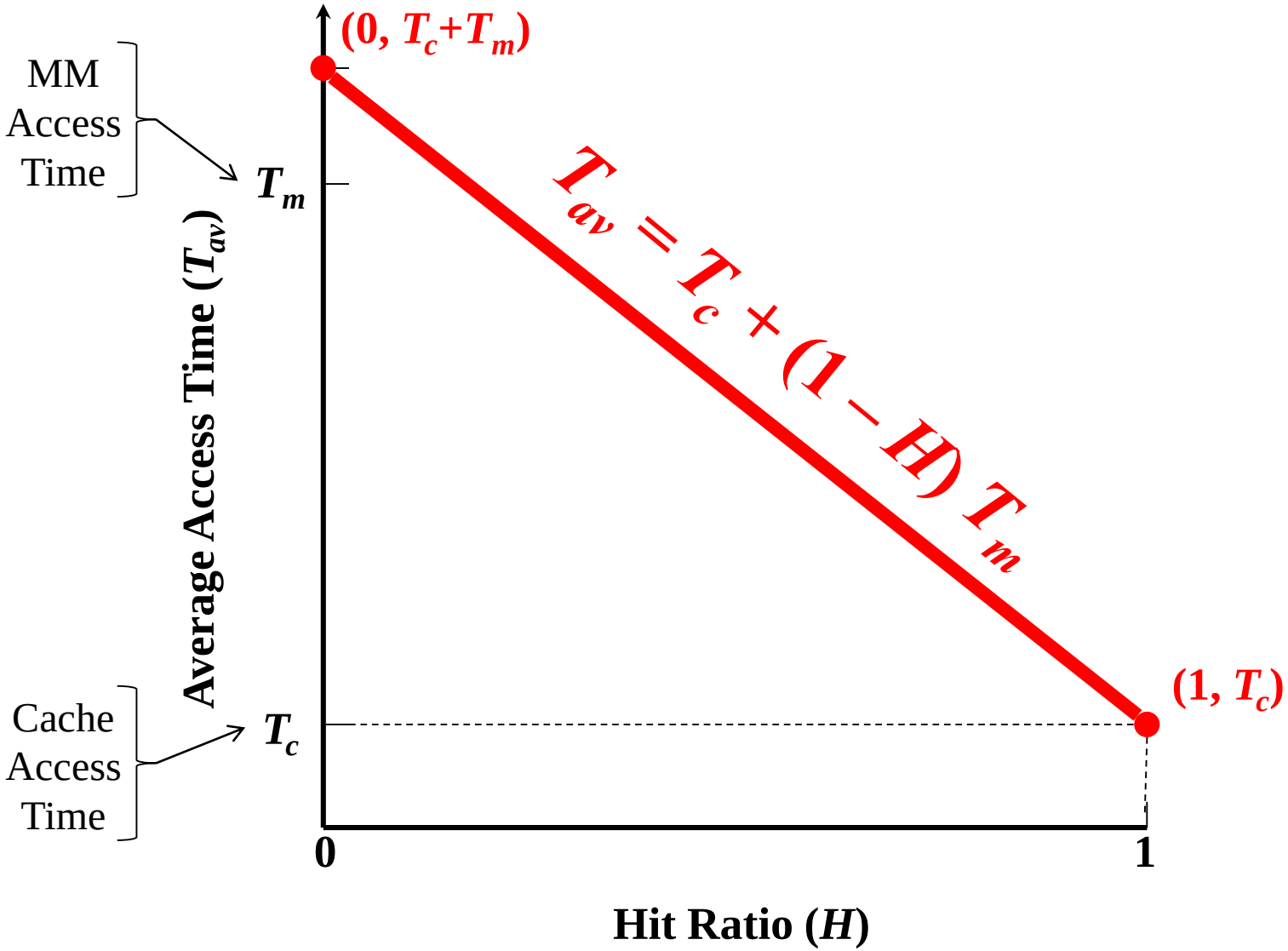
Typical Cache Organization



Cache Memory - Concept



Average Access Time of a One-Level Cached Memory System



Reading Material

- Stallings, Chapter 4:
  - Pages 113 – 123