# Computer Architecture

Lecture-06

**Md. Biplob Hosen**

Lecturer, IIT-JU.

Email: biplob.hosen@juniv.edu

# Reference

- “Computer Organization and Architecture” by William Stallings; 8th Edition (Chapter-04).
  - Any later edition is fine.

# Content

- Computer Memory System Overview
  - Characteristics of Memory System
  - Memory Hierarchy
- Cache Memory Principles
- Elements of Cache Design
- Cache Organization

# Computer Memory System Overview

- A memory unit is an essential component in any digital computer since it is needed for storing programs and data.
- The complex concept of computer memory is made more manageable if we classify memory systems according to their key characteristics.
- The most important characteristics are as following:

# Characteristics of Memory System- **Details from Book**

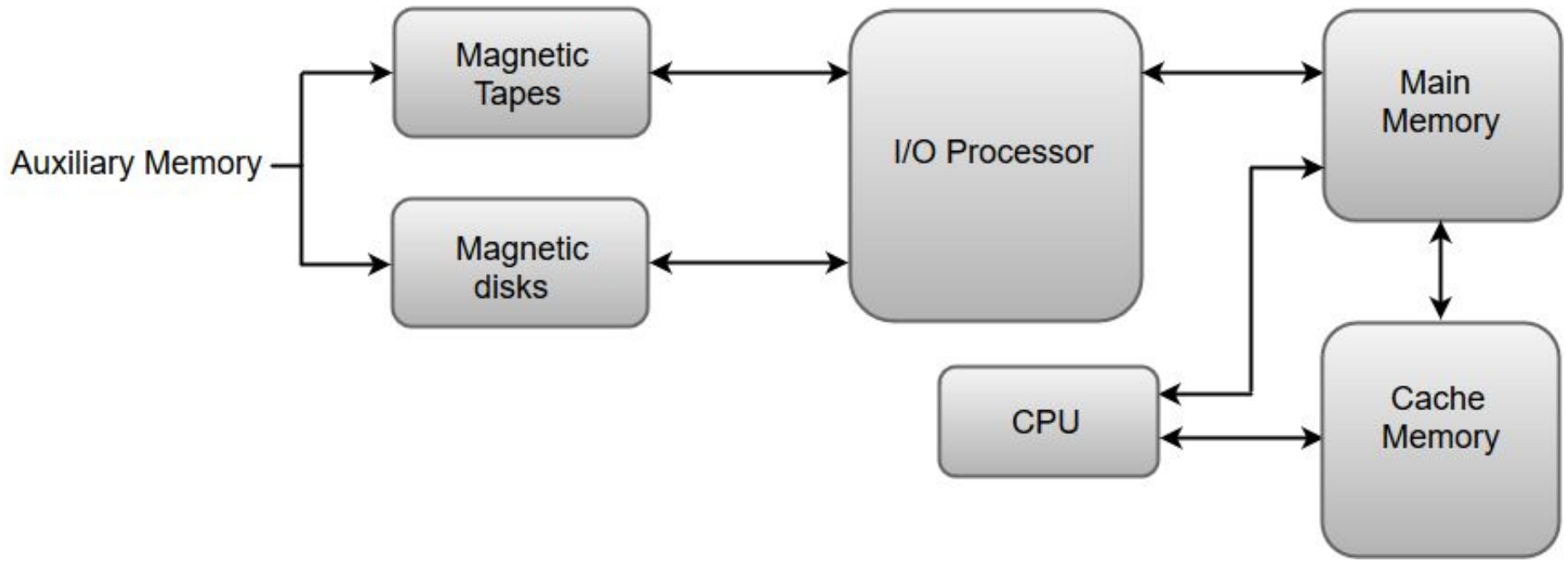
| | | | |
|---|---|---|---|
| **Location** | 1. Internal (e.g. processor registers, main memory, cache)<br>2. External (e.g. optical disks, magnetic disks) | **Capacity** | 1. Number of words<br>2. Number of bytes |
| **Unit of Transfer** | 1. Word<br>2. Block | **Performance** | 1. Access Time<br>2. Cycle Time<br>3. Transfer Time |
| **Access Method** | 1. Sequential<br>2. Direct<br>3. Random<br>4. Associative | **Physical Type** | 1. Semiconductor<br>2. Magnetic<br>3. Optical<br>4. Magneto-Optical |
| **Physical Characteristics** | 1. Volatile vs. Non-Volatile<br>2. Erasable vs. Non-Erasable | **Organization** | 1. Memory Modules |

# Memory Hierarchy

- Typically, a memory unit can be classified into two categories:
- The memory unit that establishes direct communication with the CPU is called **Main Memory**. The main memory is often referred to as RAM (Random Access Memory).
- The memory units that provide backup storage are called **Auxiliary Memory**. For instance, magnetic disks and magnetic tapes are the most commonly used auxiliary memories.
- Apart from the basic classifications of a memory unit, the memory hierarchy consists all of the storage devices available in a computer system ranging from the slow but high-capacity auxiliary memory to relatively faster main memory.
- The following image illustrates the components in a typical memory hierarchy.

# Continue…

**Memory Hierarchy in a Computer System:**



# Continue…

**Auxiliary Memory**

- Auxiliary memory is known as the lowest-cost, highest-capacity and slowest-access storage in a computer system.
- Auxiliary memory provides storage for programs and data that are kept for long-term storage or when not in immediate use.
- The most common examples of auxiliary memories are magnetic tapes and magnetic disks.
- A magnetic disk is a digital computer memory that uses a magnetization process to write, rewrite and access data.
- For example, hard drives, zip disks, and floppy disks.
- Magnetic tape is a storage medium that allows for data archiving, collection, and backup for different kinds of data.

**I/O Processor**

- The primary function of an I/O Processor is to manage the data transfers between auxiliary memories and the main memory.

# Continue…

**Main Memory**

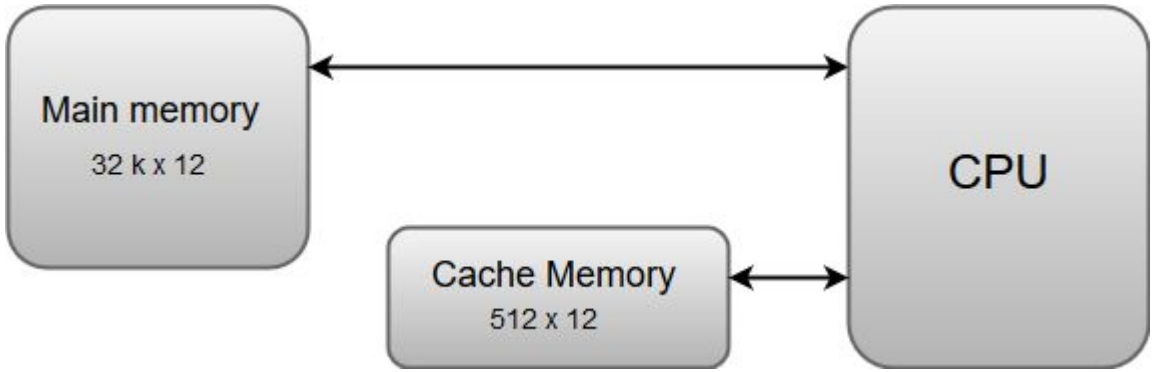
- The main memory in a computer system is often referred to as **Random Access Memory (RAM)**.
- This memory unit communicates directly with the CPU and with auxiliary memory devices through an I/O processor.
- The programs that are not currently required in the main memory are transferred into auxiliary memory to provide space for currently used programs and data.

**Cache Memory**

- The data or contents of the main memory that are used frequently by CPU are stored in the cache memory so that the processor can easily access that data in a shorter time.
- Whenever the CPU requires accessing memory, it first checks the required data into the cache memory.
- If the data is found in the cache memory, it is read from the fast memory. Otherwise, the CPU moves onto the main memory for the required data.

# Cache Memory

- Cache memory is placed between the CPU and the main memory.
- The block diagram for a cache memory can be represented as:




- The cache is the fastest component in the memory hierarchy and approaches the speed of CPU components.

# Continue…

The basic operation of a cache memory is as follows:

- When the CPU needs to access memory, the cache is examined. If the word is found in the cache, it is read from the fast memory.
- If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word.
- A block of words one just accessed is then transferred from main memory to cache memory. The block size may vary from one word (the one just accessed) to about 16 words adjacent to the one just accessed.
- The performance of the cache memory is frequently measured in terms of a quantity called **hit ratio**.
- When the CPU refers to memory and finds the word in cache, it is said to produce a **hit**.
- If the word is not found in the cache, it is in main memory and it counts as a **miss**.
- The ratio of the number of hits divided by the total CPU references to memory (hits plus misses) is the hit ratio.

# Continue…

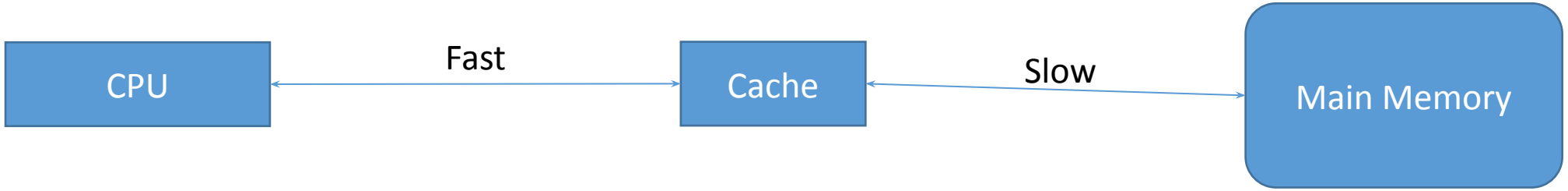



Fig-1: Single Cache

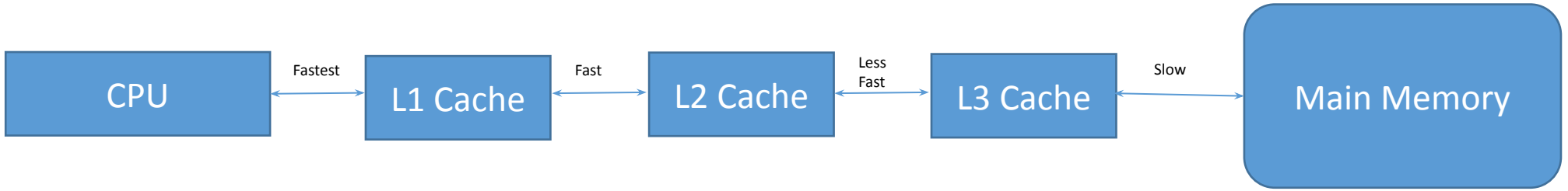
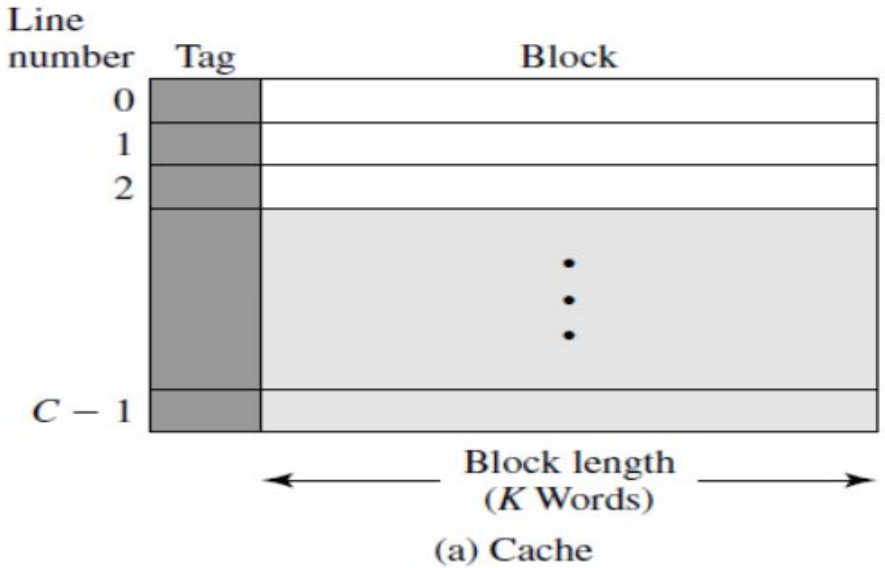
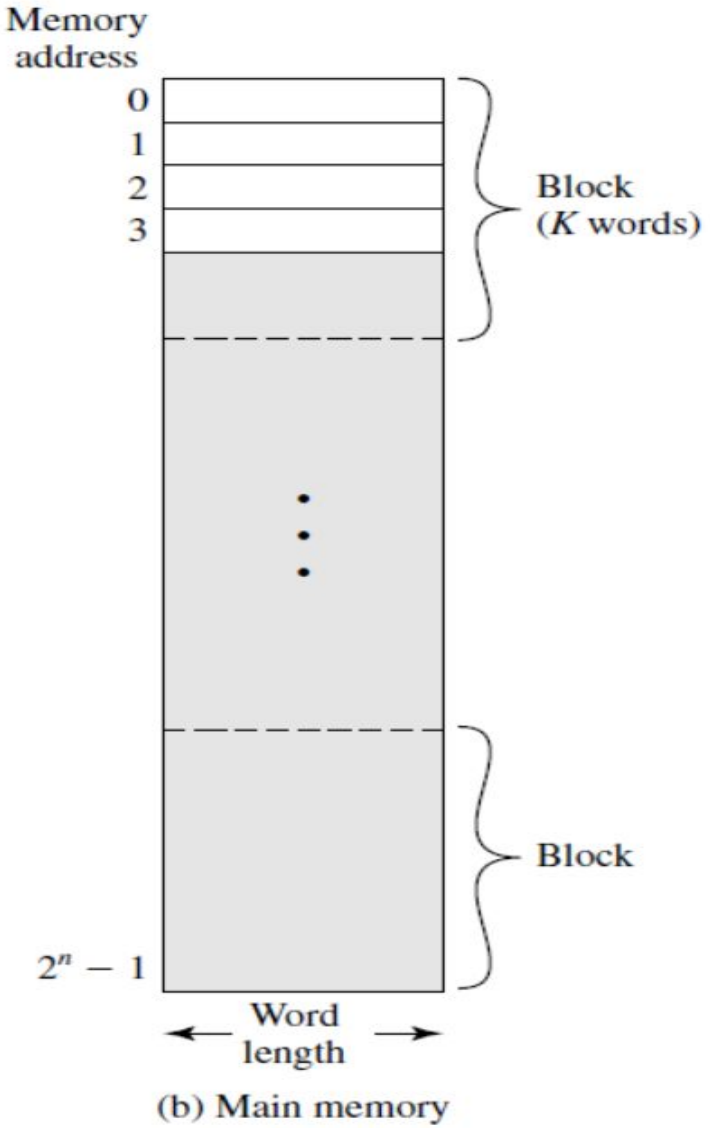



Fig-2: Three-Level Cache

# Cache Memory Structure

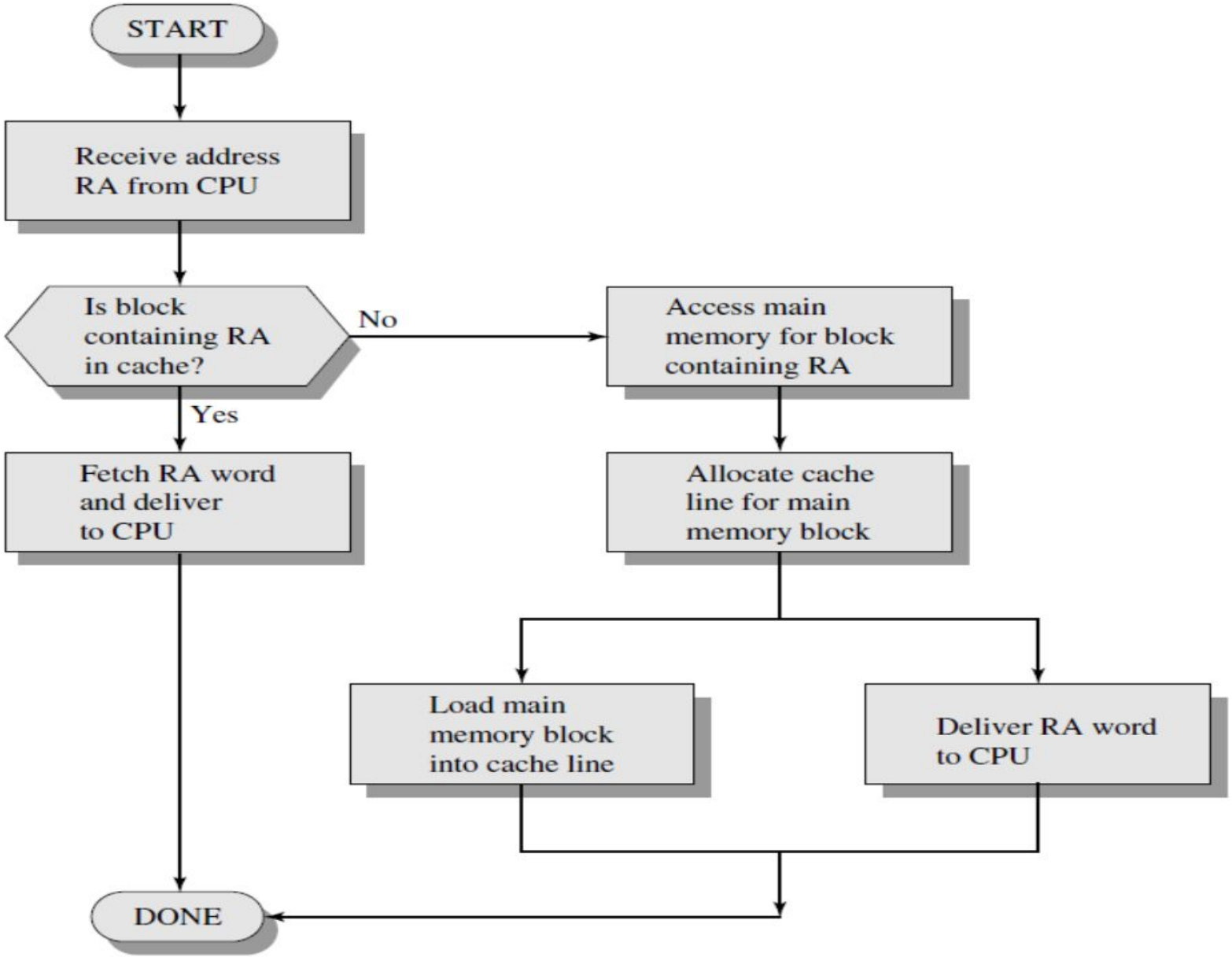



(a) Cache

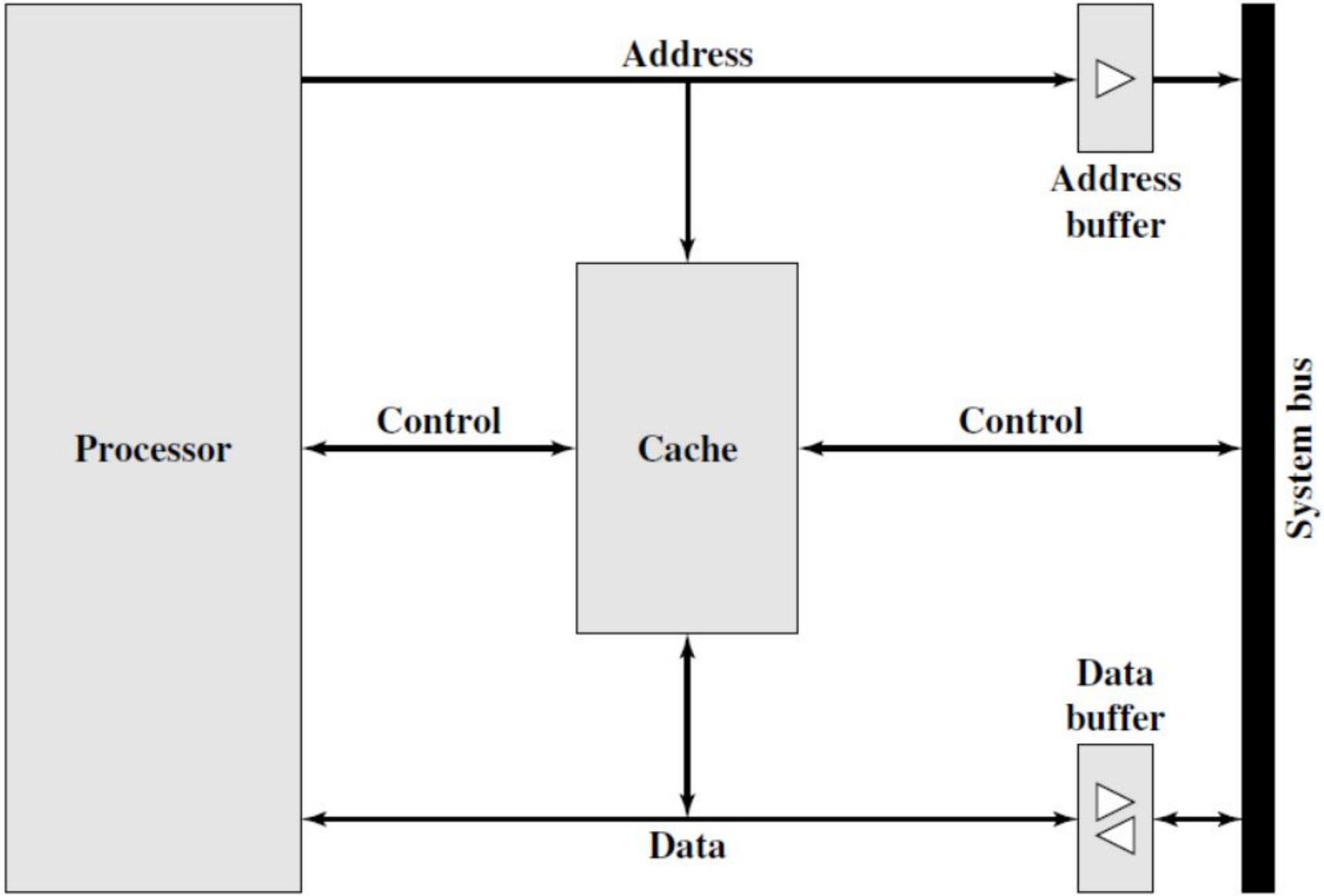



(b) Main memory

# Cache Read Operation



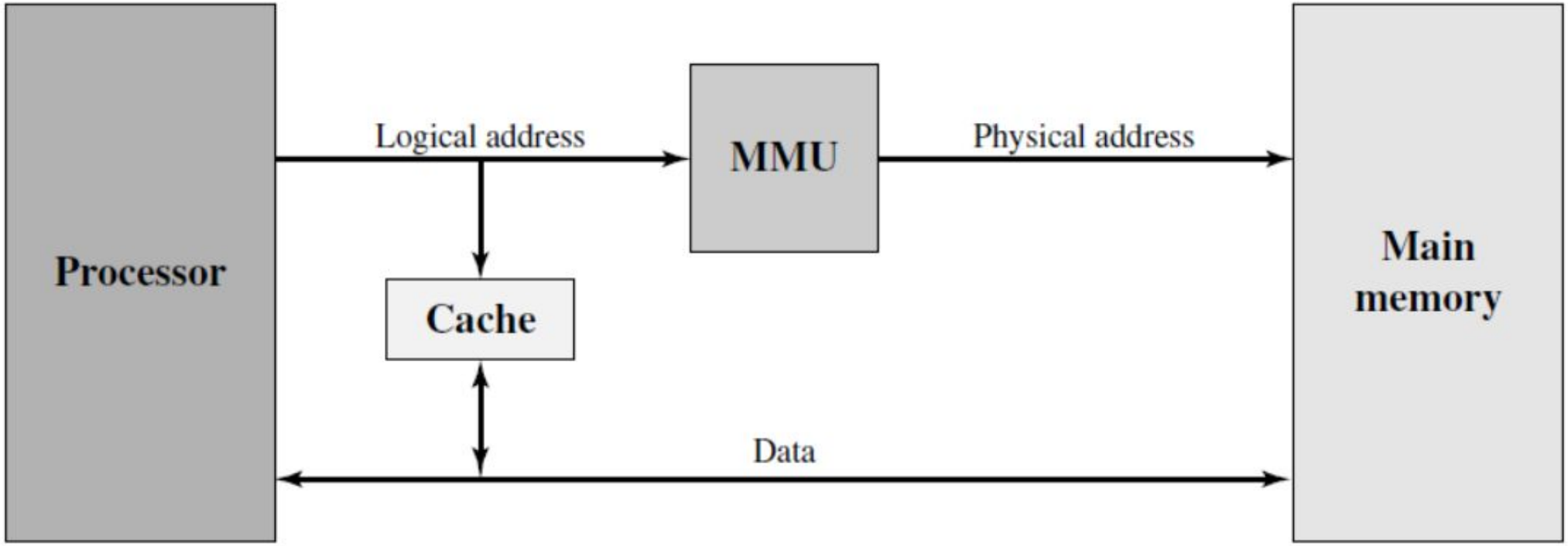
# Typical Cache Organization



# Elements of Cache Design - Details from Book

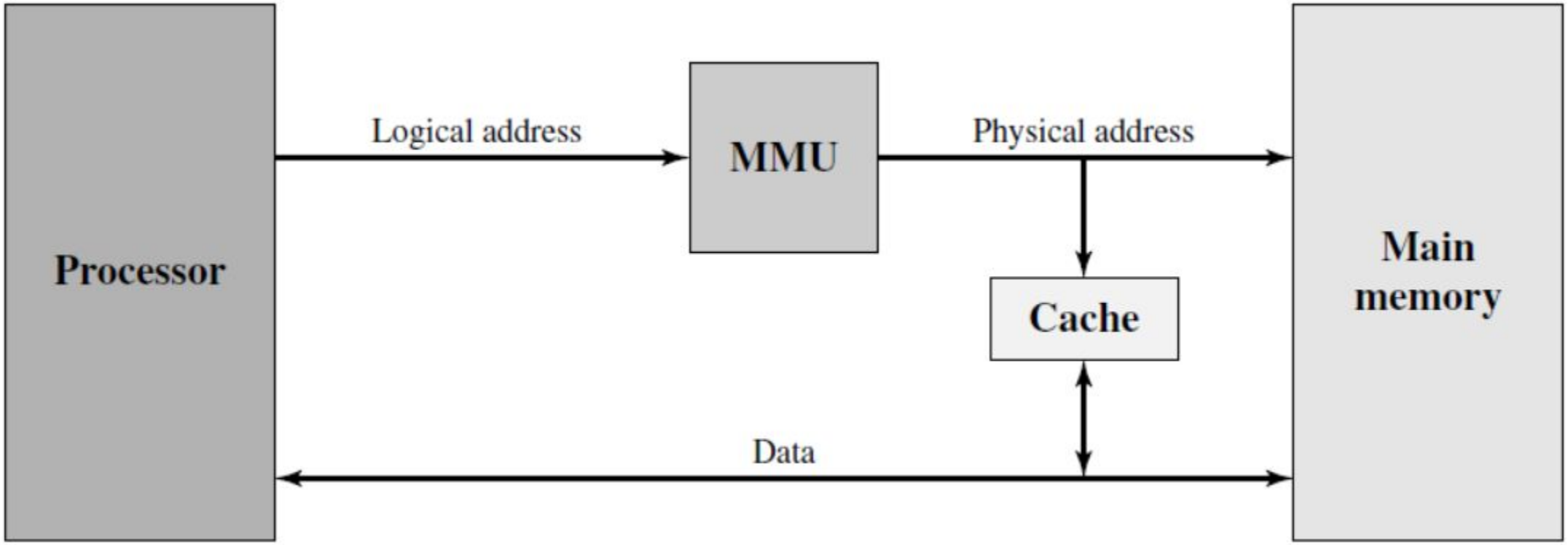
| | |
|---|---|
| **Cache Addresses** | **Write Policy** |
| Logical | Write through |
| Physical | Write back |
| **Cache Size** | Write once |
| **Mapping Function** | **Line Size** |
| Direct | **Number of caches** |
| Associative | Single or two level |
| Set Associative | Unified or split |
| **Replacement Algorithm** | |
| Least recently used (LRU) | |
| First in first out (FIFO) | |
| Least frequently used (LFU) | |
| Random | |

# Logical Cache




(a) Logical cache

# Physical Cache



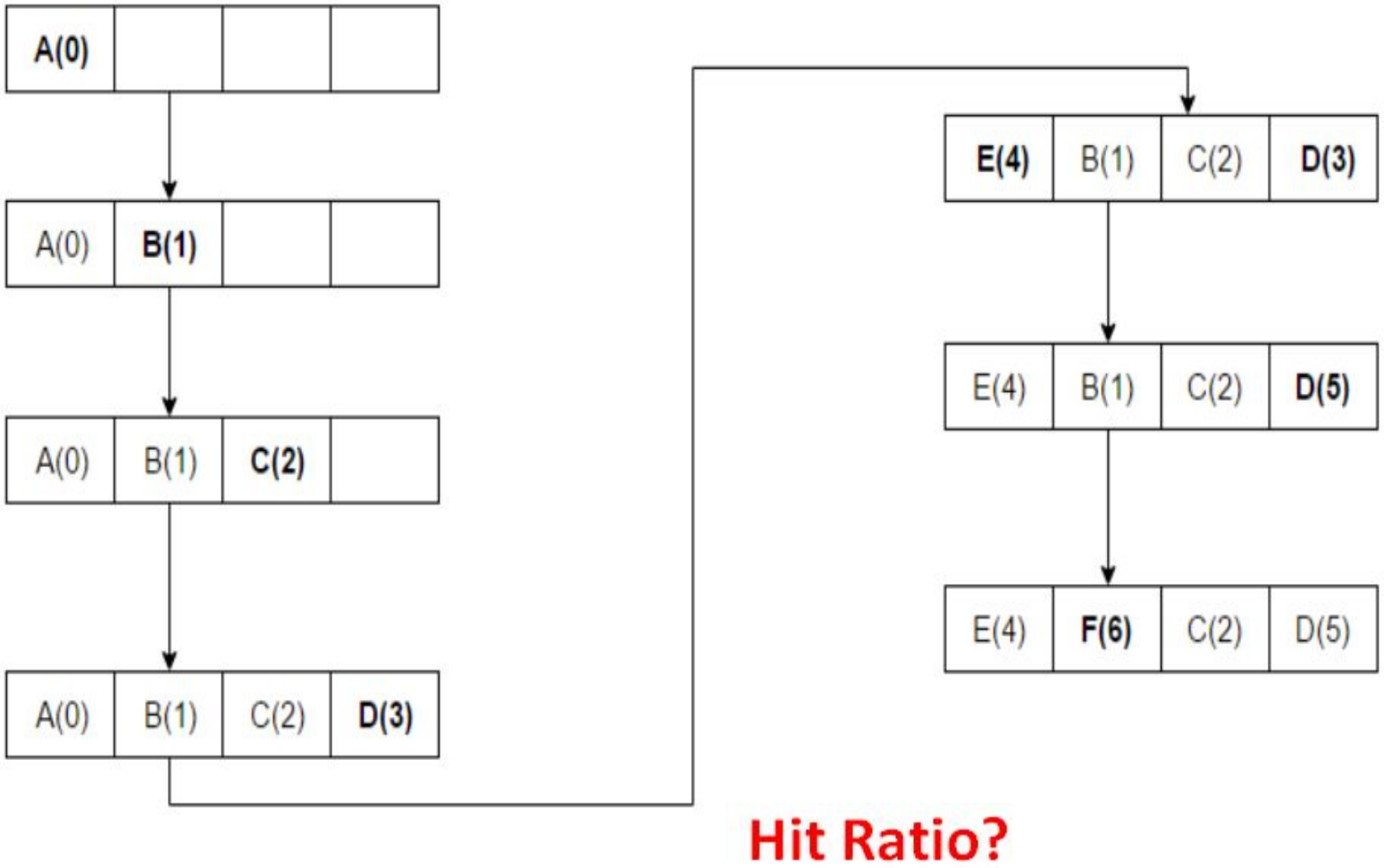

(b) Physical cache

# Replacement Algorithms

- Once the cache has been filled, when a new block is brought into the cache, one of the existing blocks must be replaced.
- For direct mapping, there is only one possible line for any particular block, and no choice is possible.
- For the associative and set-associative techniques, a replacement algorithm is needed.
  - To achieve high speed, such an algorithm must be implemented in hardware
- Least Recently Used (LRU)
  - Most effective
- First-in-First-Out (FIFO)
- Least Frequently Used (LFU)

# Least Recently Used

The access sequence for the below example is A B C D E D F.




**Hit Ratio?**

# Least Frequently Used

| 7 | 0 | 1 | 2 | 0 | 3 | 0 | 4 | 2 | 3 | 0 | 3 | 2 | 1 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 7 | 7 | 2 | 2 | 2 | 2 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | 1 | 1 | 1 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |

**Hit Ratio?**

# First-in-First-out (FIFO)

- Self Study

Thank You