University Of Diyala
College Of Engineering
Computer Engineering Department

COMPUTER ARCHITECTURE II

PART 2: MEMORY HIERARCHY

Asst. Prof. Ahmed Salah Hameed

Second stage

2022-2023

1.Memory hierarchy

- ❑ **Basic concepts**
- ❑ **Design techniques**

2.Caches

- ❑ **Types of caches: Fully associative, Direct mapped, Set associative**
- ❑ **Ten optimization techniques**

3.Main memory

- ❑ **Memory technology**
- ❑ **Memory optimization**
- ❑ **Power consumption**

4.Virtual memory

MEMORY TYPES

Memory is mainly of three types:

- **Secondary Memory**
- **Main Memory**
- **Cache Memory**

MEMORY TYPES

- **Secondary Memory**

CPU ⟷ HDD, SSD, and SSD M2

MEMORY TYPES

- **Main Memory**

MEMORY TYPES

- **Cache Memory**

MEMORY

What is needed?

Unlimited amounts of memory with low access time is wanted.

Consideration:

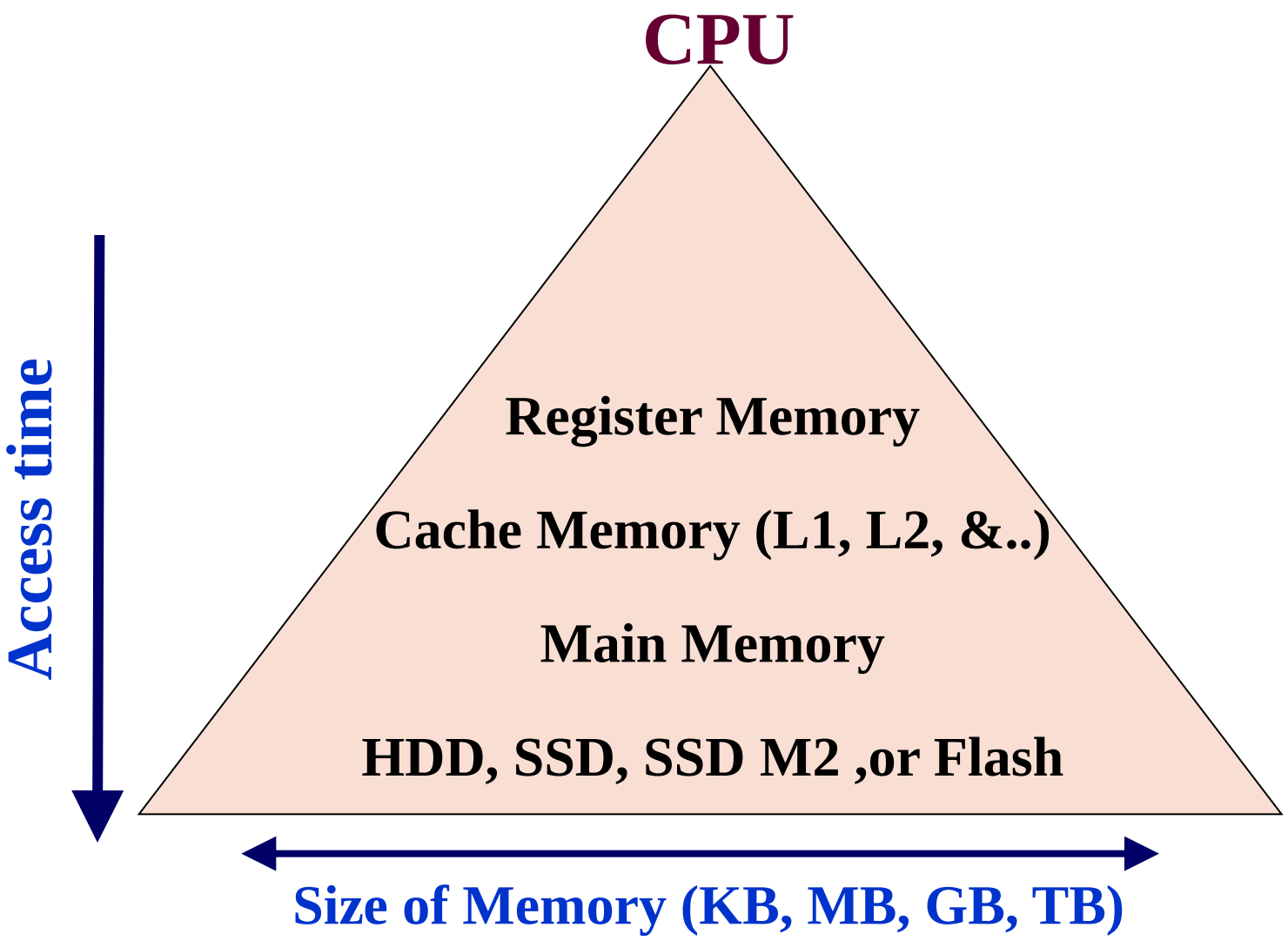
Fast memory technology is more expensive per bit than slower memory

Solution:

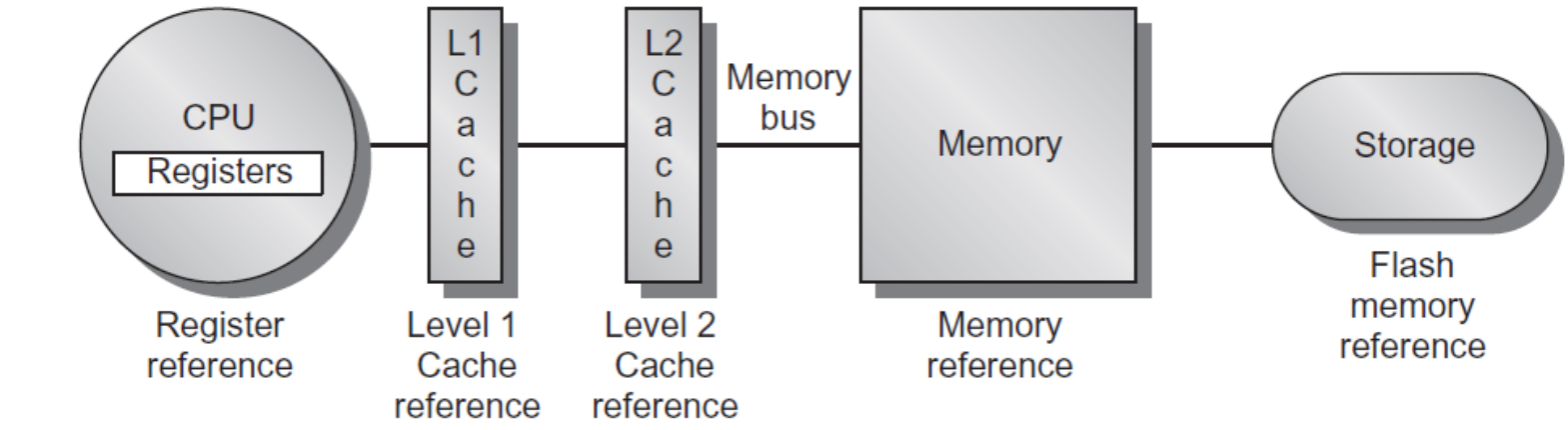
Organize memory system into a hierarchy

- Entire addressable memory space available in largest, slowest memory
- Incrementally smaller and faster memories, each containing a subset of the memory below it, proceed in steps up toward the processor

MEMORY HIERARCHY

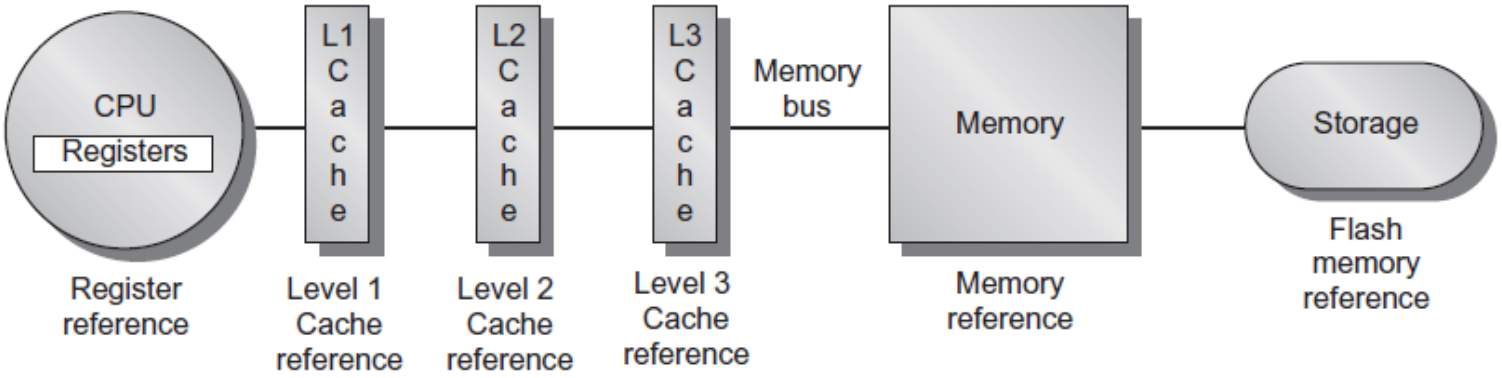


MEMORY HIERARCHY



(A) Memory hierarchy for a personal mobile device

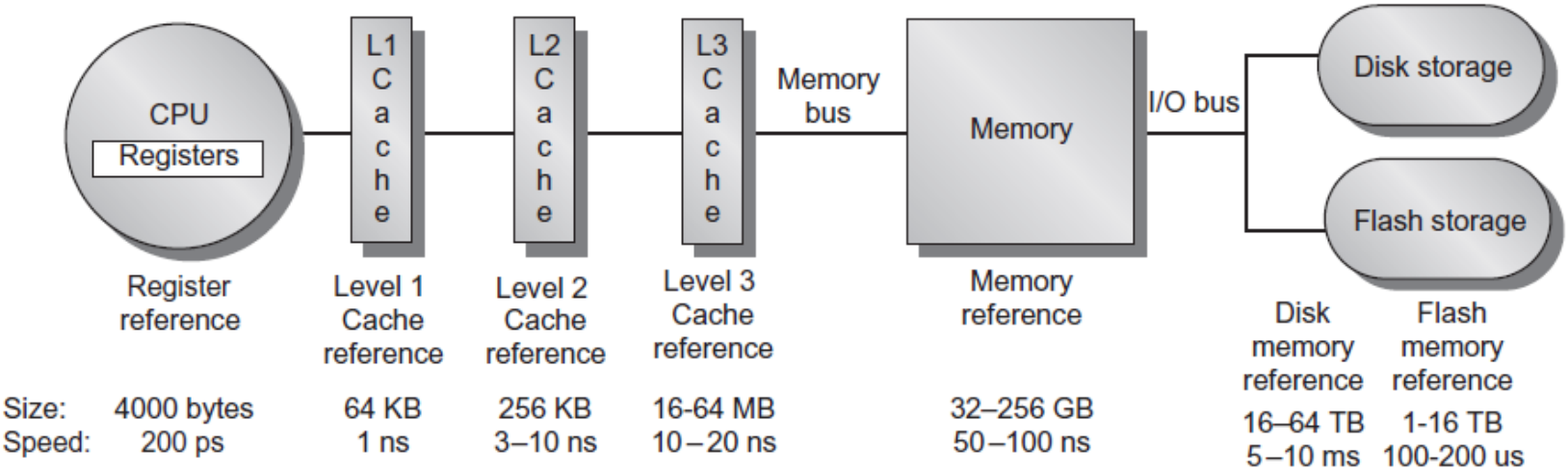
MEMORY HIERARCHY



Laptop	Size:	1000 bytes	64 KB	256 KB	4-8 MB	4–16 GB	256 GB-1 TB
	Speed:	300 ps	1 ns	3–10 ns	10–20 ns	50–100 ns	50-100 uS
Desktop	Size:	2000 bytes	64 KB	256 KB	8-32 MB	8–64 GB	256 GB-2 TB
	Speed:	300 ps	1 ns	3–10 ns	10–20 ns	50–100 ns	50-100 uS

(B) Memory hierarchy for a laptop or a desktop

MEMORY HIERARCHY



(C) Memory hierarchy for server

MEMORY HIERARCHY DESIGN

Designers of memory hierarchies focused on optimizing:

- Average memory access time (cache access time)

how long it takes for a character in memory to be transferred to or from the CPU.

- Miss rate

Ratio of no. of memory access leading to a cache miss to the total number of instructions

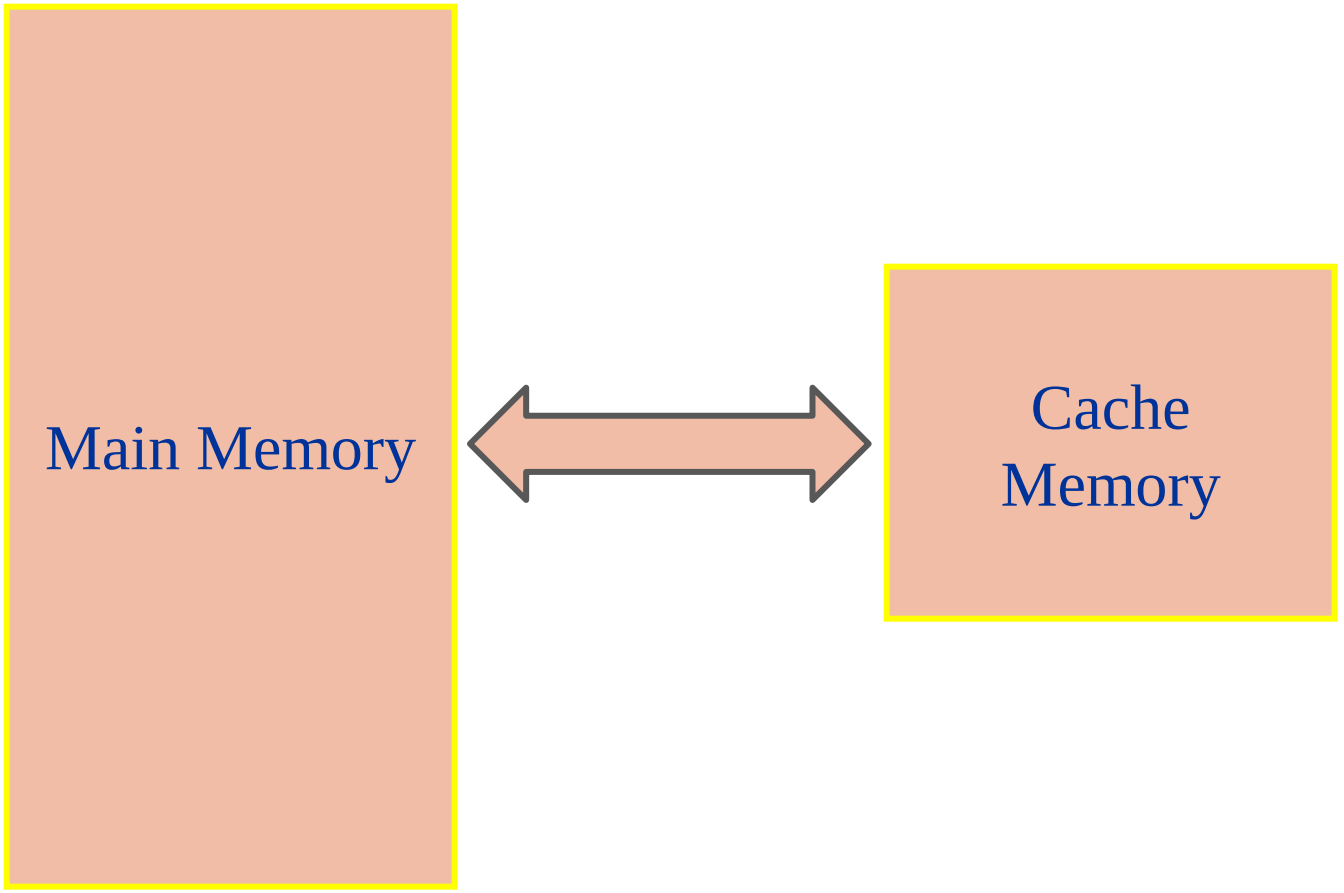
- Miss penalty

time/cycles required for making a data item in the cache

- Power (new factor)

MAIN MEMORY AND CACHE MEMORY

Block placement and **Mapping** between main and cache memories.



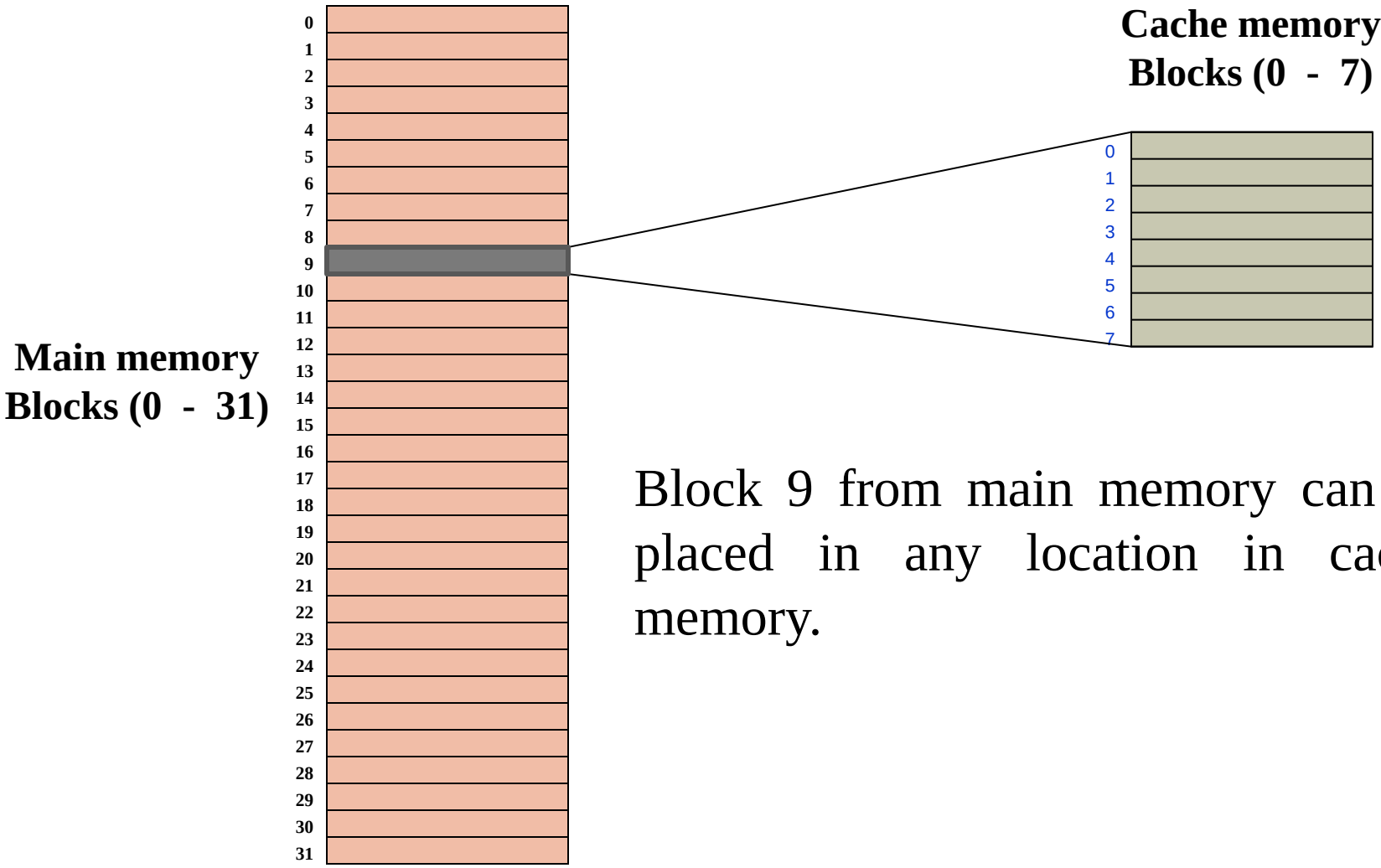
CACHE ORGANIZATION (PLACEMENT POLICIES)

Fully Associative Cache

Direct Mapped Cache

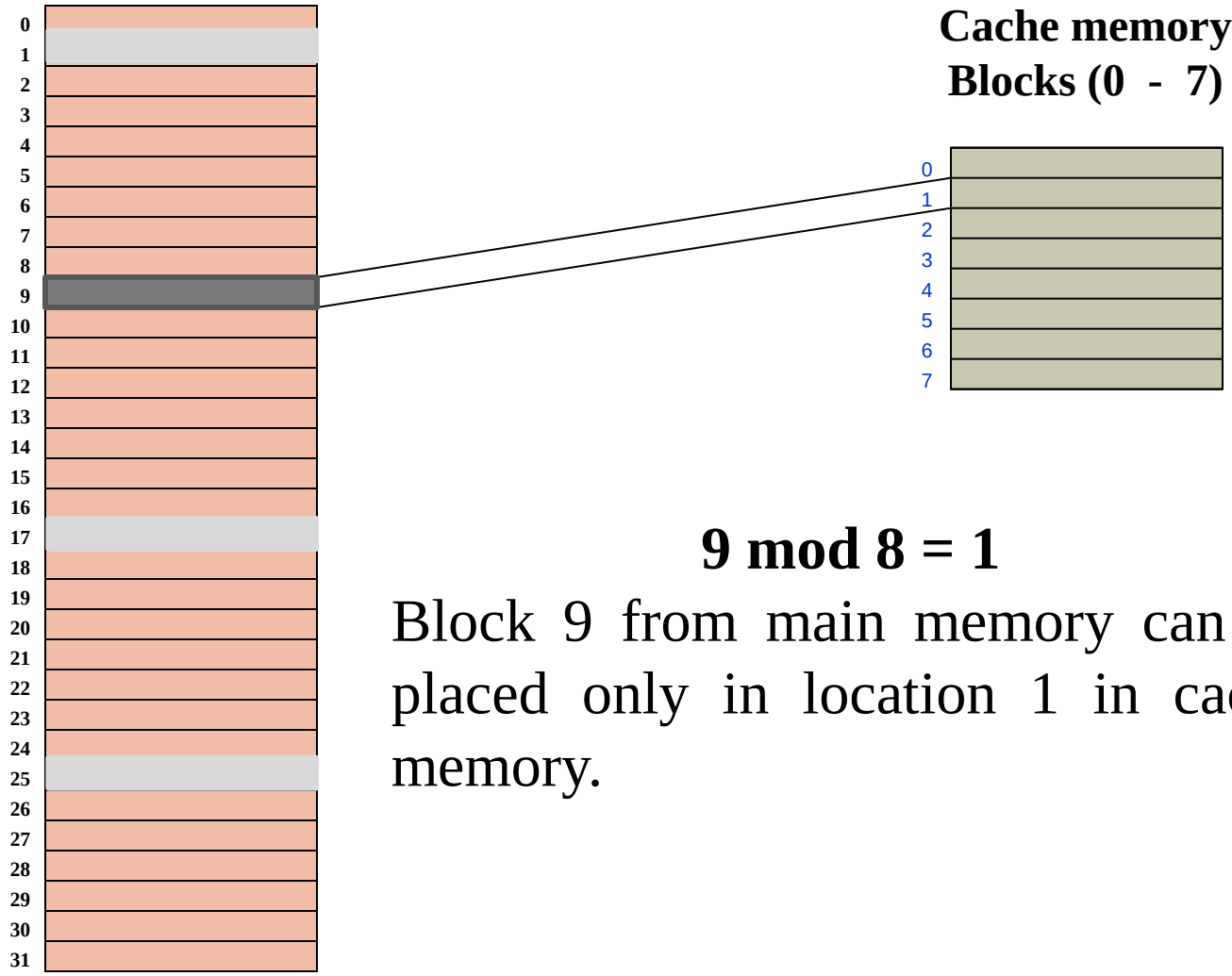
Set Associative Cache

FULLY ASSOCIATIVE CACHE ORGANIZATION



Block 9 from main memory can be placed in any location in cache memory.

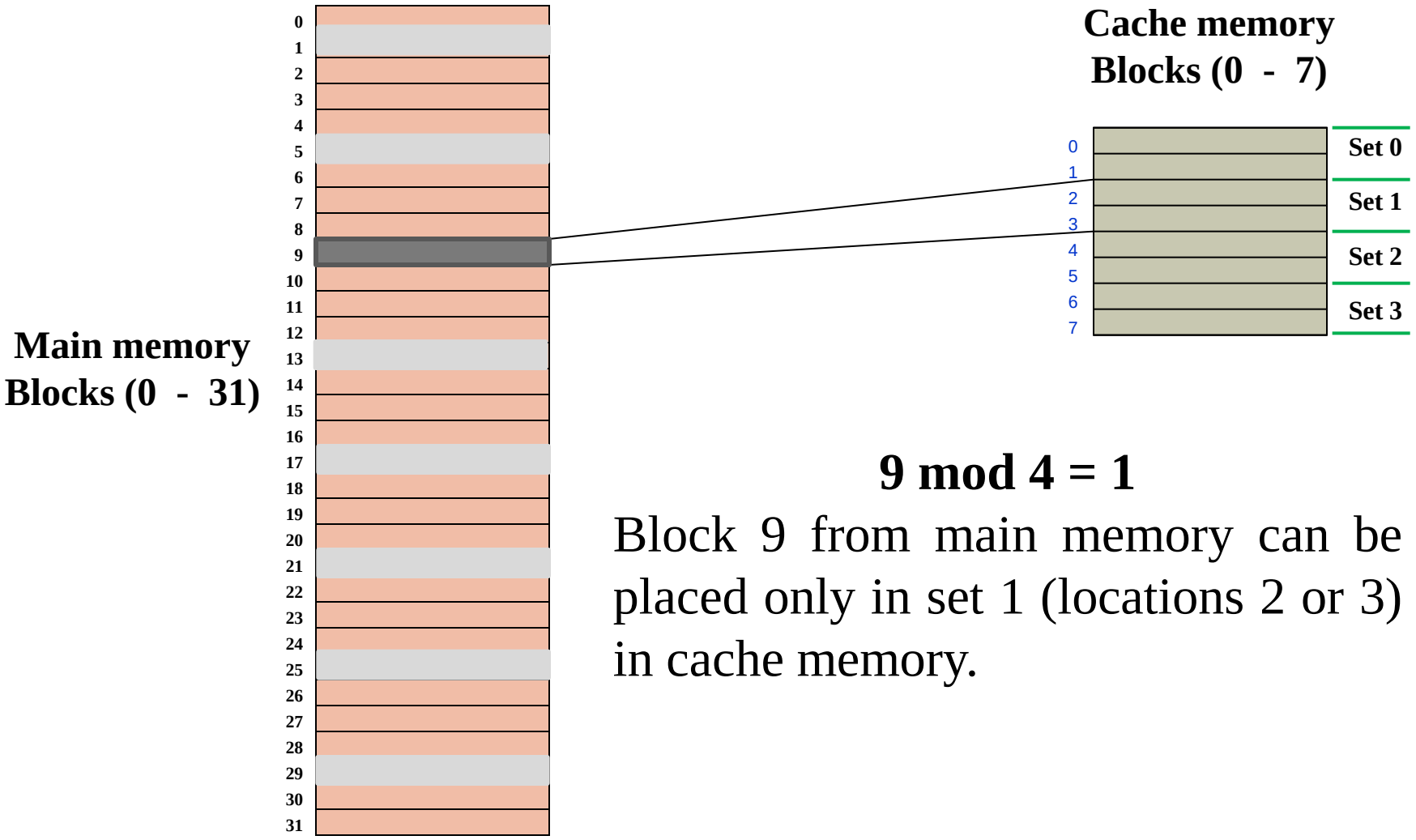
DIRECT MAPPED CACHE ORGANIZATION



9 mod 8 = 1

Block 9 from main memory can be placed only in location 1 in cache memory.

SET ASSOCIATIVE CACHE ORGANIZATION



9 mod 4 = 1

Block 9 from main memory can be placed only in set 1 (locations 2 or 3) in cache memory.

HOMEWORK 1

A computer uses a mapping procedure between main and cache memory. If the main memory has **128 blocks (0 - 127)** and cache memory has **32 blocks (0 - 31)** with **four sets (set0 - set3),** List the entire block addresses of the main memory used in each of the following situations:

(1) The entire block addresses of the main memory that can be placed in **location 15** of the cache memory. (When the computer uses ***Direct Mapped Cache***)

(2) The entire block addresses of the main memory that can be placed in **set2** of the cache memory. (When the computer uses ***Set Associative Cache***)

(3) The entire block addresses of the main memory that can be placed in **location 10** of the cache memory. (When the computer uses ***Fully Associative Cache***)