

Lecture Six

Computer Memory System



MSc. Mohammad N.Ahmmad
2024-2025

Outline:

- Computer Memory System
- *Characteristics of Memory Systems:*
- *Location, Capacity, Unit of Transfer:*
- *RAM – Random Access Memory*
- *Dynamic random-access memory (DRAM)*
- *SRAM*
- *ROM*
- *BIOS and CMOS*
- *CMOS battery*
- *Cache Memory*

Computer Memory System

Characteristics of Memory Systems:

The complex subject of computer memory is made more manageable if we classify memory systems according to their key characteristics. The most important characteristics are listed in Table Below

Key characteristics of memory

- **Location**

- CPU
- Internal (main)
- External (secondary)

- **Capacity**

- Word size
- Number of words

- **Unit of transfer**

- Word
- Block

- **Physical Type**

- Semiconductor
- Magnetic surface
- Optical

k)

- **Access methods**

- Sequential access
- Direct access
- Random access
- Associative access

- **Performance**


- Access time
- Cycle time
- Transfer rate

- **Physical Characteristics**

- Volatile / Non-Volatile
- Erasable / Non-erasable

- **Organization**

Location:

 The term **location** in Table 3.1 refers to whether memory is **internal** and **external** to the computer. Internal memory is often equated with ⁽¹⁾ **main memory**. But there are other forms of internal memory. The processor requires its own local memory, in the form of ⁽²⁾ **registers** (see CPU Chapter). Further, ⁽³⁾ **Cache** is another form of internal memory.

External memory consists of peripheral storage devices, such as **disk** and **tape**, which are accessible to the processor via I/O controllers.

Capacity:

For internal memory, capacity is typically expressed in terms of **bytes** (1 byte 8 bits) or **words**. Common word lengths are 8, 16, and 32 bits.

Access time (latency):

For random-access memory, this is the time it takes to perform a read or write operation,

Unit of Transfer:

- For internal memory, the unit of transfer is equal to the number of electrical lines into and out of the memory module.
- *For main memory, this is the number of bits read out of or written into memory at a time*

Three performance parameters are used:

➤ **Access time (latency):**

For random-access memory, this is the time it takes to perform a read or write operation, that is, the time from the instant that an address is presented to the memory to the instant that data have been stored or made available for use.

➤ **Memory cycle time:** This concept is primarily applied to random-access memory and consists of the access time plus any additional time required before a second access can commence.

➤ **Transfer rate:** This is the rate at which data can be transferred into or out of a memory unit. For random-access memory, it is equal to **1/ (cycle time)**.

For non-random-access memory, the following relationship holds:

$$T_N = T_A + \frac{n}{R}$$

where

T_N = Average time to read or write N bits

T_A = Average access time

n = Number of bits

R = Transfer rate, in bits per second (bps).

RAM and ROM

Memory Type	Category	Erase	Write Mechanism	Volatility
Random-access memory (RAM)	Read-write memory	Electrically, byte-level	Electrically	Volatile
Read-only memory (ROM)	Read-only memory	Not possible	Masks	Nonvolatile
Programmable ROM (PROM)			Electrically	
Erasable PROM (EPROM)	Read-mostly memory	UV light, chip-level		
Electrically Erasable PROM (EEPROM)		Electrically, byte-level		
Flash memory		Electrically, block-level		

Table: Semiconductor Memory Types

RAM – Random Access Memory

The term “Random” means that any memory location can be accessed in the same amount of time, regardless of its position in the memory.

Properties of Ram:

- Read/Write
- Volatile
- Temporary storage

Static or dynamic

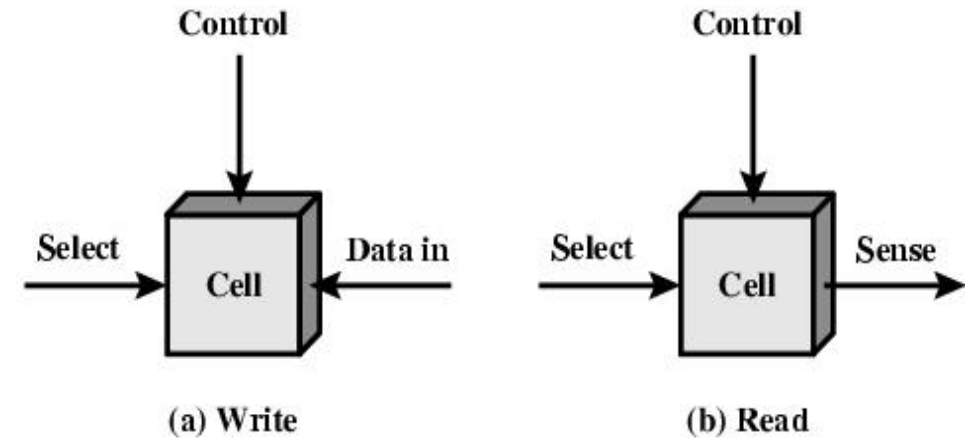
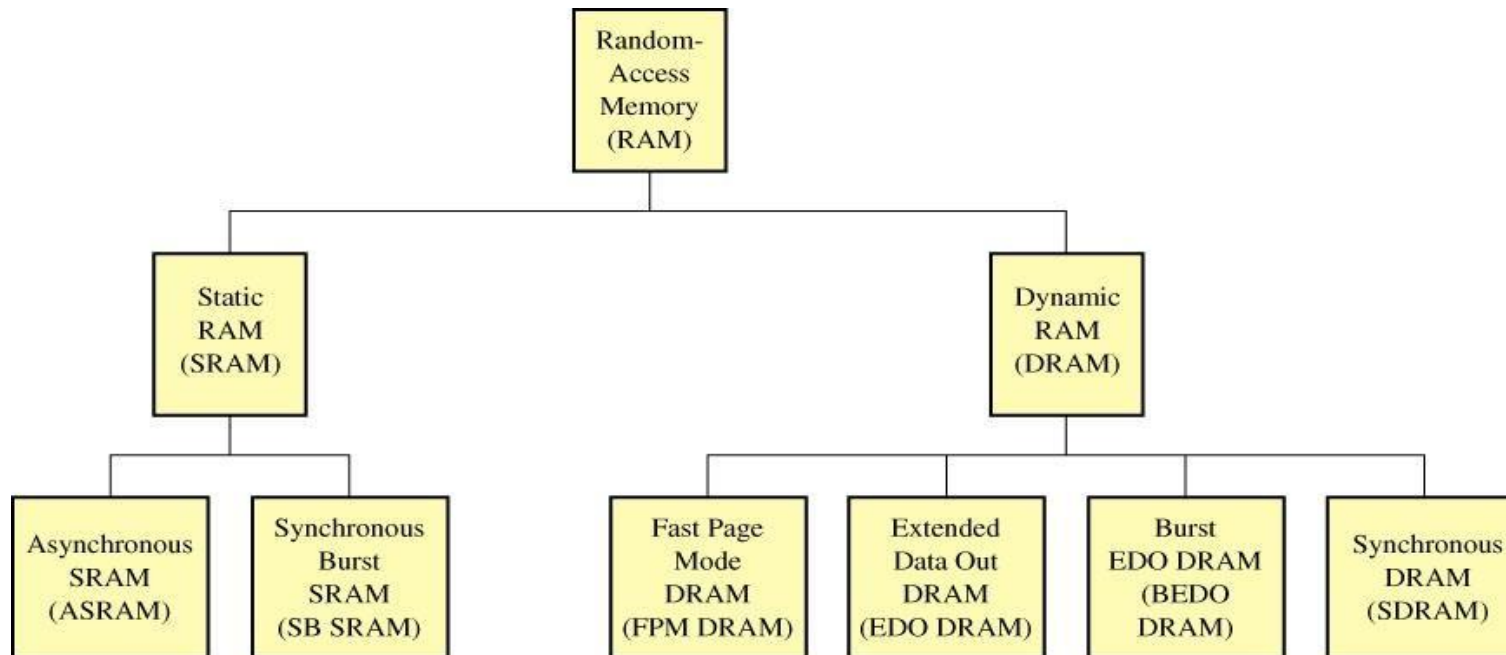
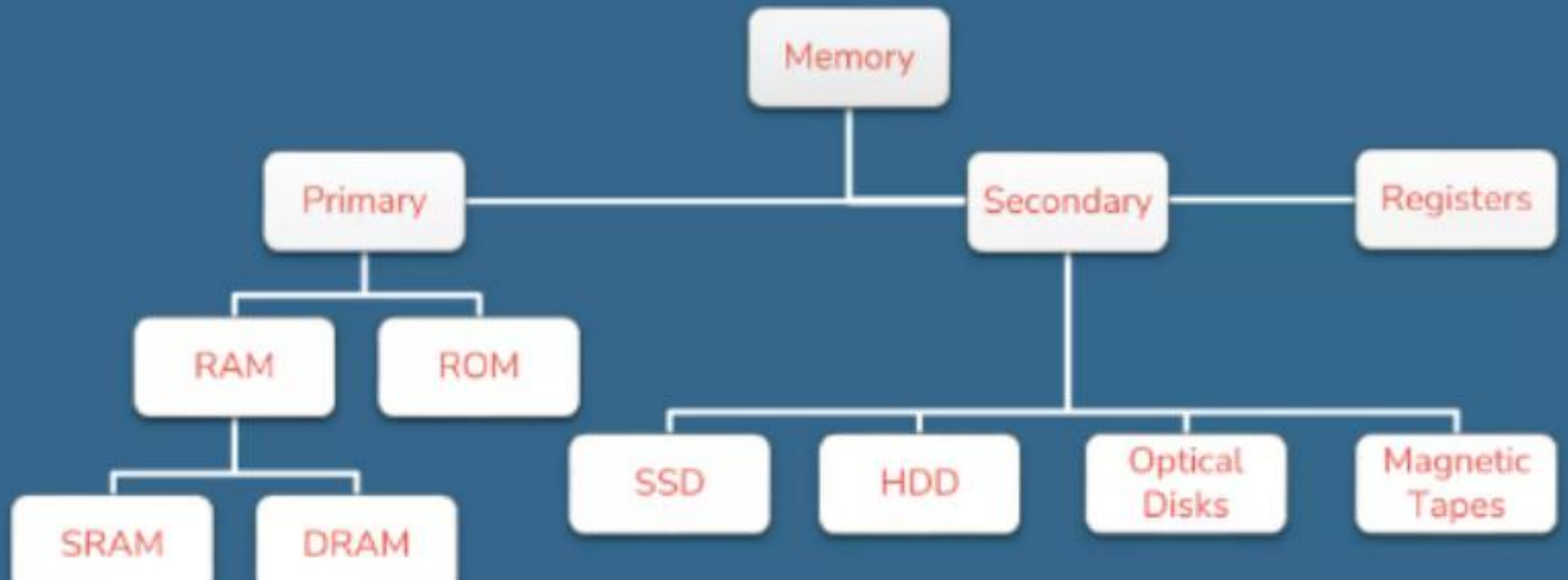


Figure: Read and Write memory cell



Memory Unit Types



Memory Unit Types

Dynamic random-access memory (DRAM)

DRAM is a type of random-access memory that stores each bit of data in a **separate capacitor within an integrated circuit**. The capacitor can be *either charged or discharged*; these two states are taken to represent the two values of a bit, conventionally called 0 and 1. Since even "non-conducting" transistors always *leak* a small amount, the capacitors will slowly discharge, and the information eventually fades unless the capacitor charge is refreshed periodically. Because of this refresh requirement, it is a *dynamic* memory.

Properties:

- Bits stored as charge in capacitors
- Charges leak
- Need refreshing even when powered
- Simpler construction
- Smaller per bit
- Less expensive
- Need refresh circuits
- Slower

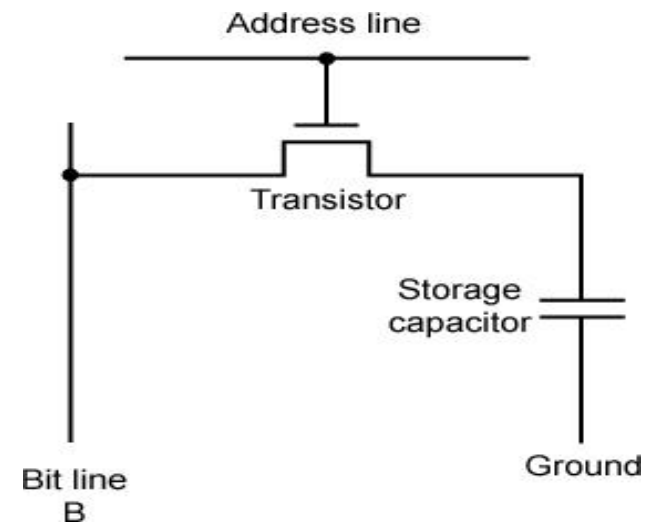


Figure: DRAM-Architecture

SRAM

- Stands for "Static Random Access Memory." SRAM is a type of RAM that **stores data using a static method**, in which the data remains constant as long as electric *power* is supplied to the memory chip.
- This is different than DRAM (dynamic RAM), which stores data dynamically and constantly needs to refresh the data stored in the memory. Because SRAM stores data statically, it is faster and requires less power than DRAM.
- DRAM is most often used as the main memory for personal computers. However, SRAM is commonly used in smaller applications, such as CPU cache memory and hard drive buffers. It is also used in other consumer electronics, from large appliances to small children's toys.

- **Bits stored as on/off switches**
- **No charges to leak**
- **No refreshing needed when powered**
- **More complex construction**
- **Larger per bit**
- **More expensive**
- **Does not need refresh circuits**
- **Faster**
- **Cache**
- **Digital**
- **Uses flip-flops**

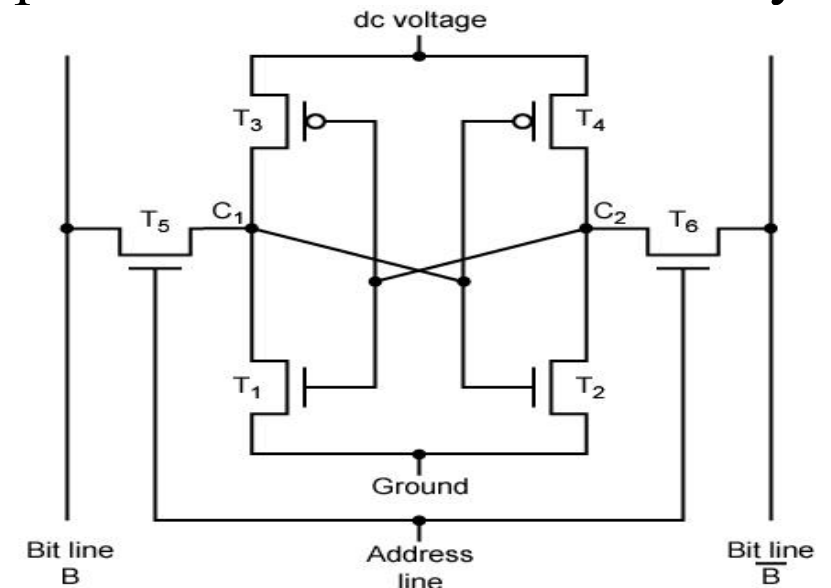


Figure: SRAM Architecture

ROM

Read-only memory (ROM) is a type of non-volatile memory used in computers and other electronic devices.

- Data stored in ROM can only be modified slowly, with difficulty, or not at all,
- It mainly used to store firmware (software that is closely tied to specific hardware and unlikely to need frequent updates).
- Every stored-program computer may use a form of non-volatile storage (that is, storage that retains its data when power is removed) to store the initial program that runs when the computer is powered on or otherwise begins execution (a process known as bootstrapping).
- Since ROM (at least in hard-wired mask form) cannot be modified, it is really only suitable for storing data which is not expected to need modification for the life of the device.

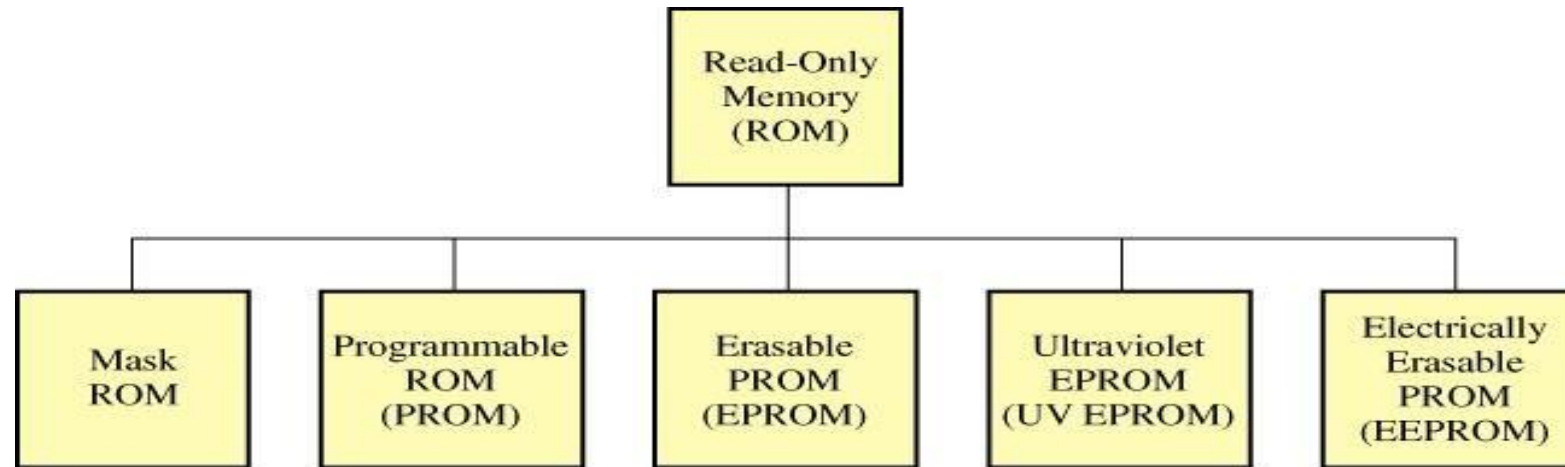


Figure: Rom Family

BIOS and CMOS

The terms **BIOS** and **CMOS** both **refer to essential parts of your computer's motherboard**. They work together and they're both important, but they are not the same thing.

BIO

- The BIOS, or "Basic Input/Output System", is special firmware stored in a chip on your computer's motherboard. **It is the first program that runs every time you turn on your computer.**
- **The BIOS performs the POST**, which initializes and tests your computer's hardware. Then it locates and runs your boot loader, or loads your operating system directly.
- The BIOS also provides a simple interface for configuring your computer's hardware. When you start your computer, you will often see a message like "Press F2 for setup." This setup is your BIOS configuration interface.



CMOS

When you make changes your BIOS configuration, the settings are not stored on the BIOS chip itself.

- Instead, they are stored on a special memory chip, which is referred to as "the CMOS." **CMOS** stands for "[Complementary Metal-Oxide-Semiconductor](#)".
- It holds a small amount of data, usually 256 [bytes](#). This information includes what types of disk drives are installed on your computer, the current date and time of your [system clock](#), and your computer's [boot sequence](#).
- On some motherboards the CMOS is a separate chip, but on most modern motherboards it is integrated with the realtime clock ([RTC](#)) .
- Your BIOS memory is [non-volatile](#): it retains its information even when your computer has no power. This is important, because your computer needs to remember its BIOS settings even when it's turned off. That's why the CMOS has its own dedicated power source, the CMOS battery.

CMOS battery

- The CMOS battery is a [Lithium-ion battery](#) about the size of a coin.
- It can hold a charge for up to ten years before needing to be replaced.
- If your CMOS battery dies, your BIOS settings will reset to their [defaults](#) when your computer is turned off.



Cache Memory

Where is cache located?

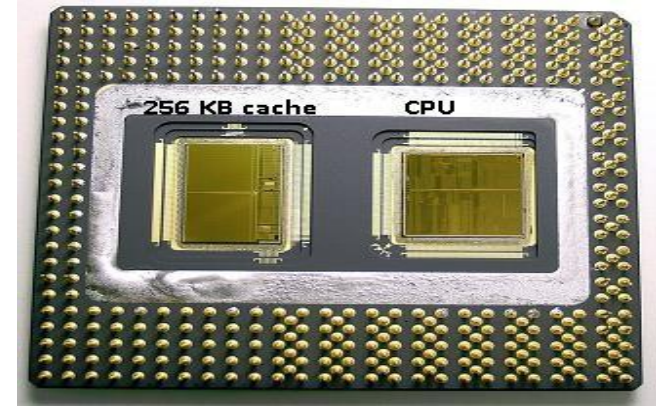
- Sits between normal main memory and CPU
- May be located on CPU chip or module

History of Cache Memory

80486: (1989)

This is the first CPU of this generation that has some cache on the CPU

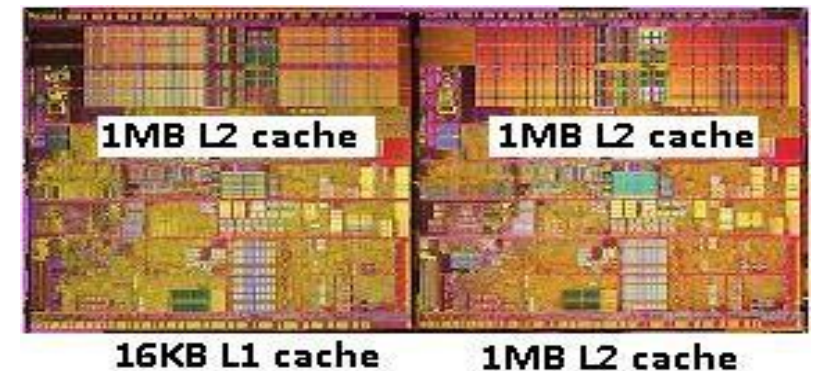
1. 80586 (1993)



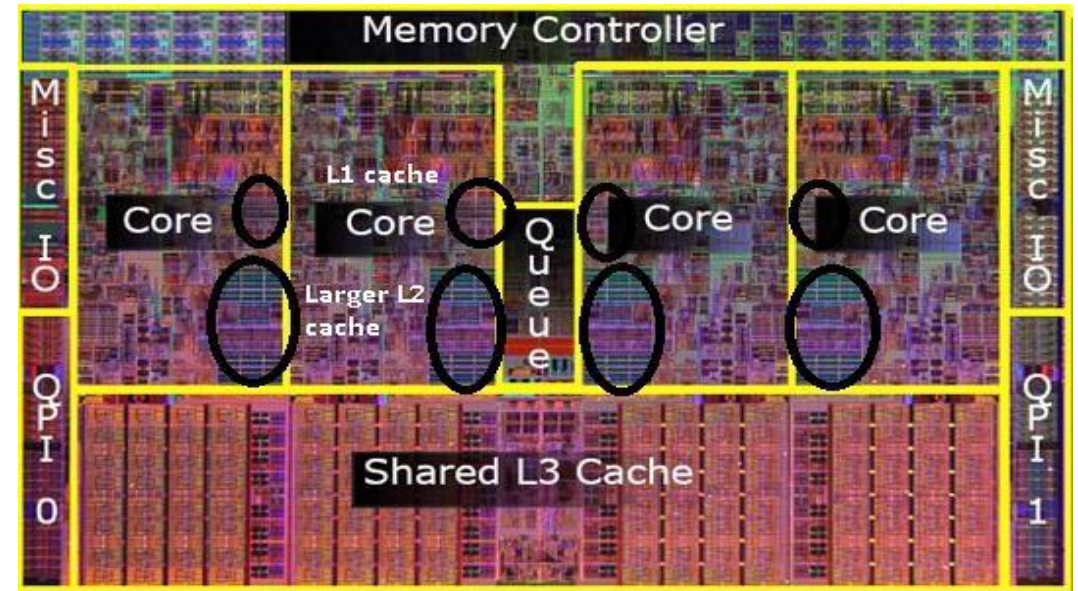
2. Pentium-2



3. Pentium-D (duo)



4. Core I7



What is cache?

A small amount of very high speed memory between the “main memory” and the CPU

How is it organized?

Organized in a number of uniform sized blocks of memory that have a high likelihood of being used.

How is kept “current”?

When a block in main memory is more likely to be needed, that block replaces a block in the cache.

Advantages of cache memory

- Cache memory is faster than main memory.
- It consumes less access time as compared to main memory.
- It stores the program that can be executed within a short period of time.
- It stores data for temporary use.
- A cache reduces the traffic to the lower-level store.

Why is cache so fast?

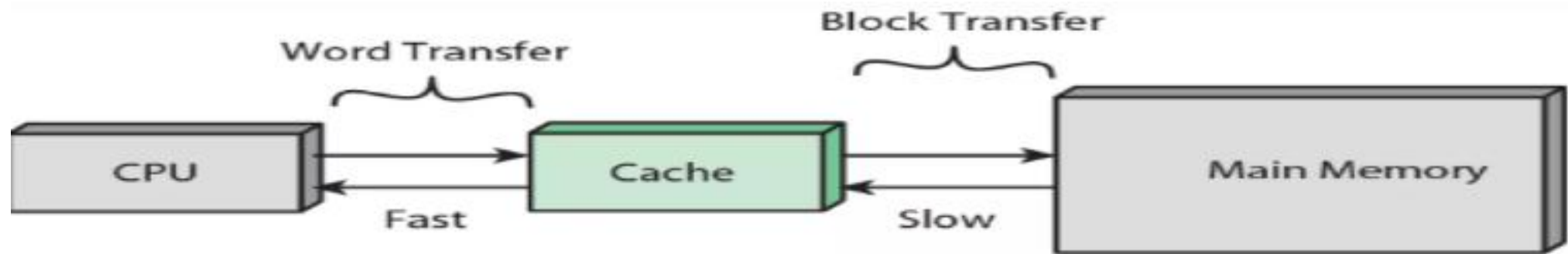
- Faster electronics in a small size cost
- Fewer locations to access means faster access time
- Physically closer to the CPU, avoids communication delays

Cache memory Levels

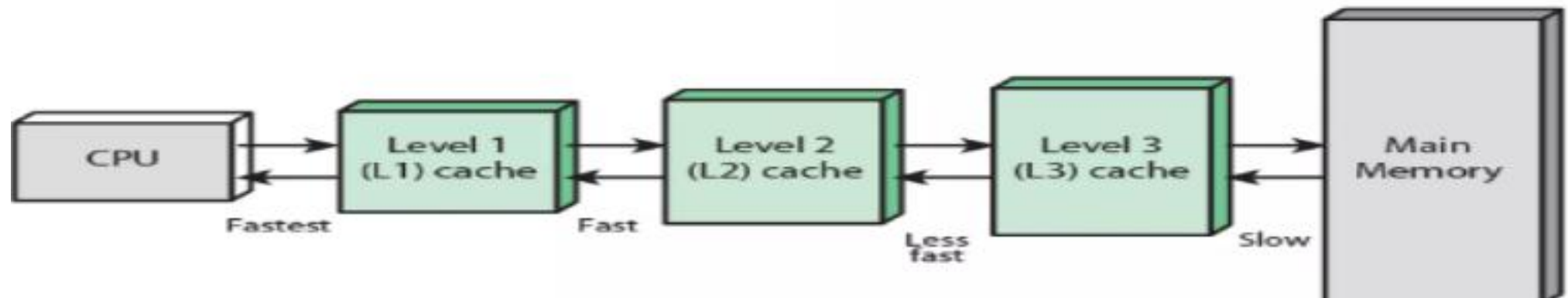
- Level 1 ([L1](#)) cache is extremely fast but relatively small, and is usually embedded in the processor chip (CPU).
- Level 2 (L2) cache is often more capacious than L1; it may be located on the CPU or on a separate chip or [coprocessor](#) with a high-speed alternative system bus interconnecting the cache to the CPU, so as not to be slowed by traffic on the main system bus.

Level 3 (L3) cache is typically specialized memory that works to improve the performance of L1 and L2. It can be significantly slower than L1 or L2, but is usually double the speed of RAM. In the case of **multicore processors**, each core may have its own dedicated L1 and L2 cache, but share a common L3 cache. When an instruction is referenced in the L3 cache, it is typically elevated to a higher tier cache

Cache Memory



(a) Single cache



Locality

Programs tend to use data and instructions with addresses near or equal to those they have used recently.

Types of Locality:

1. Temporal locality:

Recently referenced items are likely to be referenced again in the near future

2. Spatial locality:

Items with nearby addresses tend to be referenced close together in time.

Reasons:

- iterations
- commonly used subroutines

Cache operation

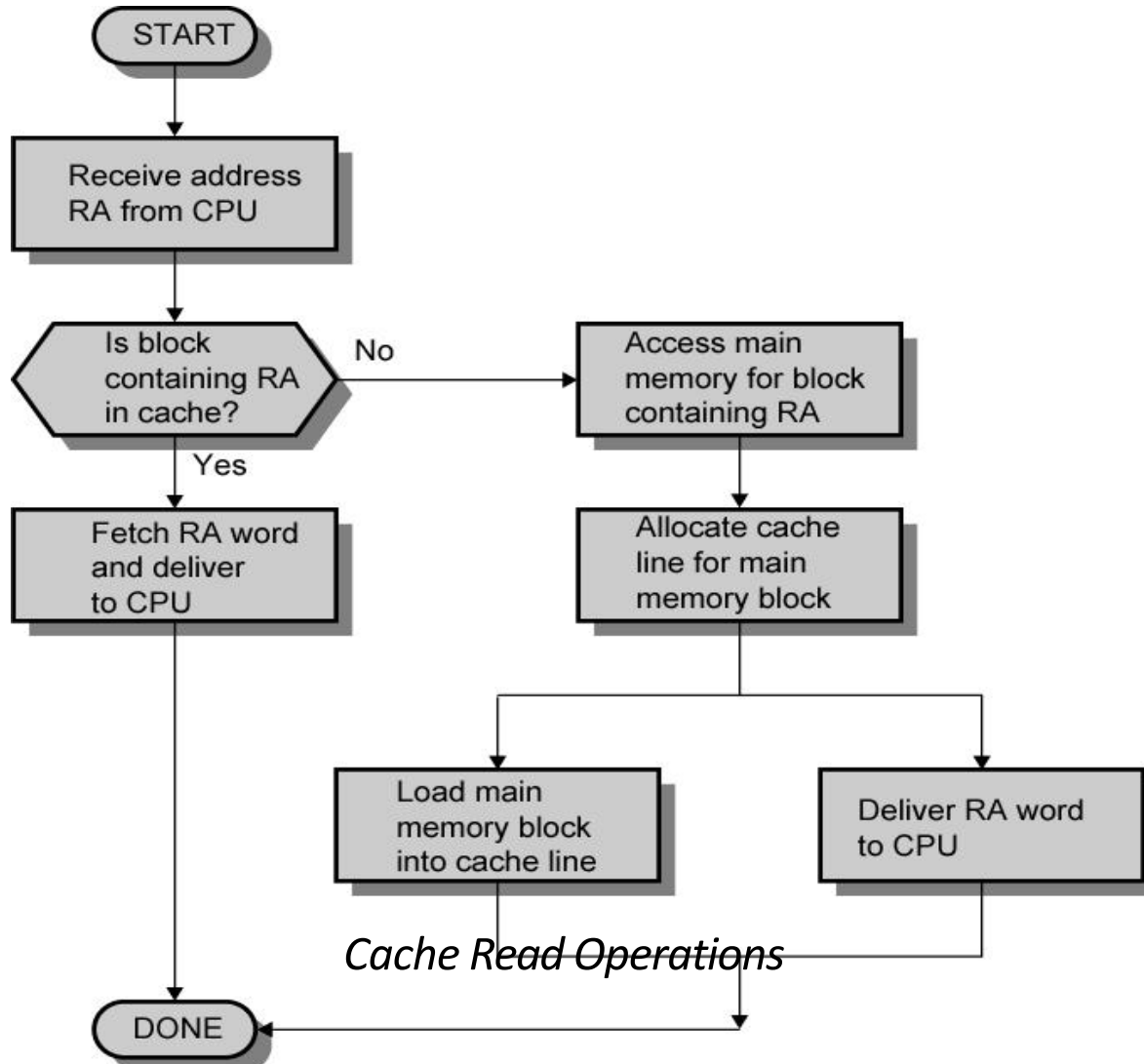
- CPU requests contents of memory location
- Check cache for this data
- If present, get from cache (fast)
- If not present, read required block from main memory to cache
- Then deliver from cache to CPU
- Cache includes tags to identify which block of main memory is in each cache slot

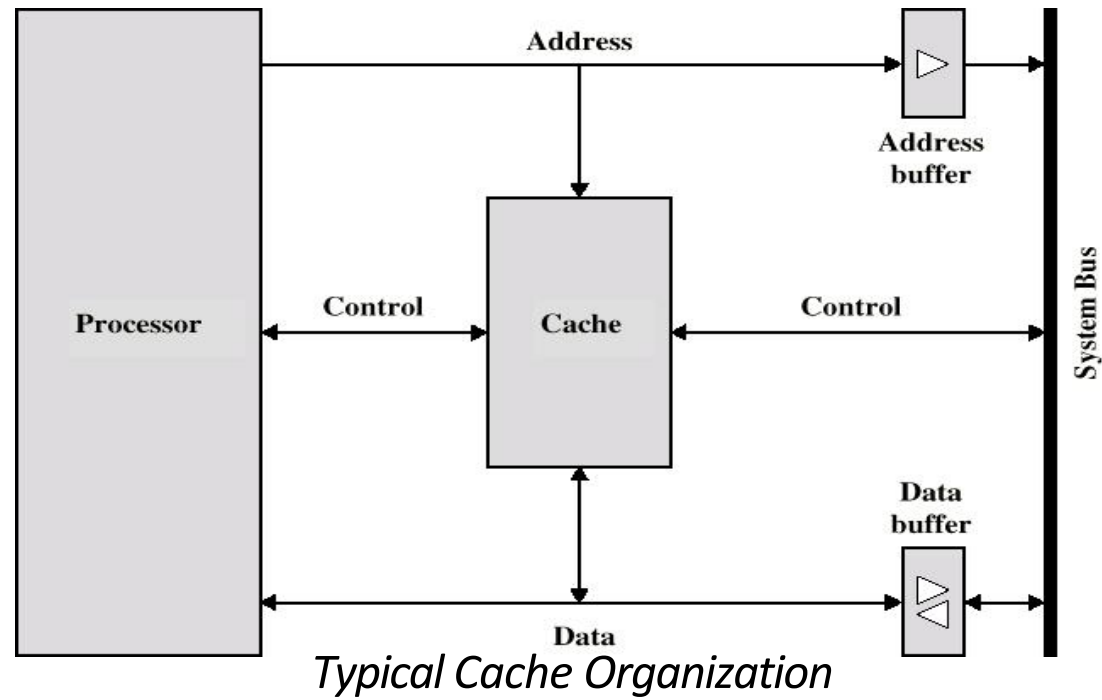
Two questions to answer:

Q1: How do we know if a data item is in the cache?

Q2: If it is, how do we find it?

Cache Read Operation





Cache hit

A *cache hit* occurs when the requested data can be found in a cache.

Cache miss

A cache miss refers to a failed attempt to read or write a piece of data in the cache, which results in a main memory access with much longer latency. There are three kinds of cache misses: instruction read miss, data read miss, and data write miss.

Cache Mapping Types

1. Direct-Mapped Cache
2. Associative Mapped Cache
3. Set-Associative Mapped Cache

Hierarchy of memory

❑ The **hierarchy of memory** in a microprocessor refers to the structured arrangement of memory types based on speed, cost, and size. Here's a brief overview from fastest (and most expensive) to slowest (and least expensive):

1. Register:

- Located inside the CPU.
- Fastest memory.
- Very limited in size.
- Used for immediate instruction execution.

2. Cache Memory.

- Small, fast memory close to or within the CPU.
- Levels: L1 (fastest), L2, and sometimes L3.
- Stores frequently accessed data and instructions.

3. Main Memory (RAM).

- Larger and slower than cache.
- Temporary storage used during program execution.
- Volatile (loses data when powered off).

Hierarchy of memory

4. Secondary Storage (e.g., SSD, HDD)
 - Much larger but much slower than RAM.
 - Non-volatile (data persists after power off).
 - Used for long-term data storage.
5. Tertiary Storage (e.g., external drives, cloud storage)
 - Slowest and often used for backup or archival.
 - Not directly accessed by the CPU during execution.

Each level in the hierarchy balances **speed**, **cost**, and **capacity**, helping optimize system performance and cost-effectiveness

Hierarchy of memory

