MEMORY UNIT: INTRODUCTION

A memory unit is an essential component of a computer system. It stores both data and instructions needed during processing. The memory serves as a place where information can be read from or written to at high speed.

Definition:

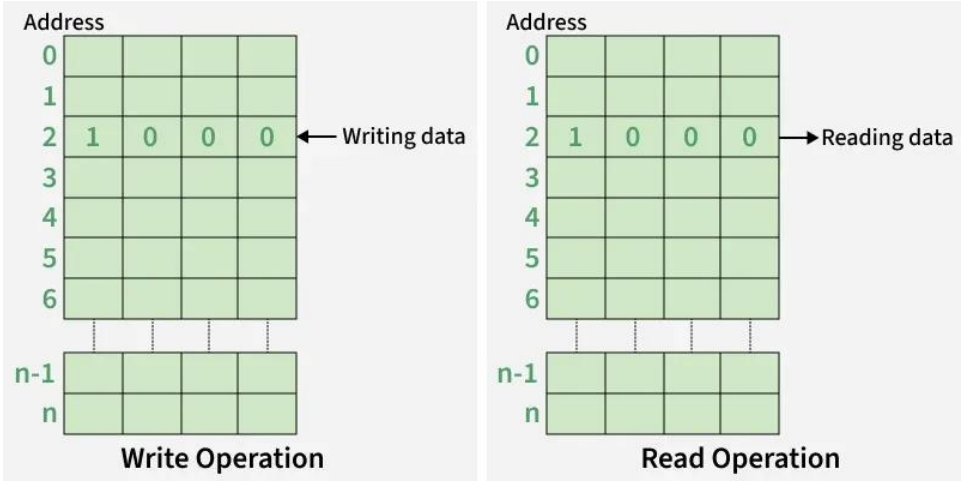
Memory is the storage space in a computer where data and instructions are kept. It is divided into small units called cells, and each cell has a unique address.

- **Storage element (Cell):** Stores 1 bit of data.
- **Register:** A memory location made of cells.
- **Capacity:** Total number of bits a memory can store.
- **Reading:** Retrieving data.
- **Writing:** Storing data.

There are two basic operations performed in memory: **'Read' (taking data out)** and **'Write' (putting data in).**

- ✓ **Read:** The computer looks for the required address (drawer number), opens the drawer, and takes out (reads) what is inside.
- ✓ **Write:** The computer chooses the drawer (by address), and puts something new inside, replacing the old data with the new one.

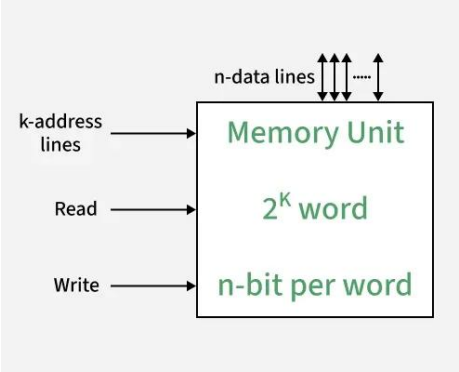
- Every memory cell is accessed through a unique address.
- The memory unit uses registers—like the address register (specifies location) and data register (holds read/write value)—to perform these operations.



- ✓ A word is a group of bits where a memory unit stores binary information. A word with a group of 8 bits is called a byte.

- ✓ A memory unit consists of data lines, address selection lines, and control lines that specify the direction of transfer.

Block diagram of a memory unit

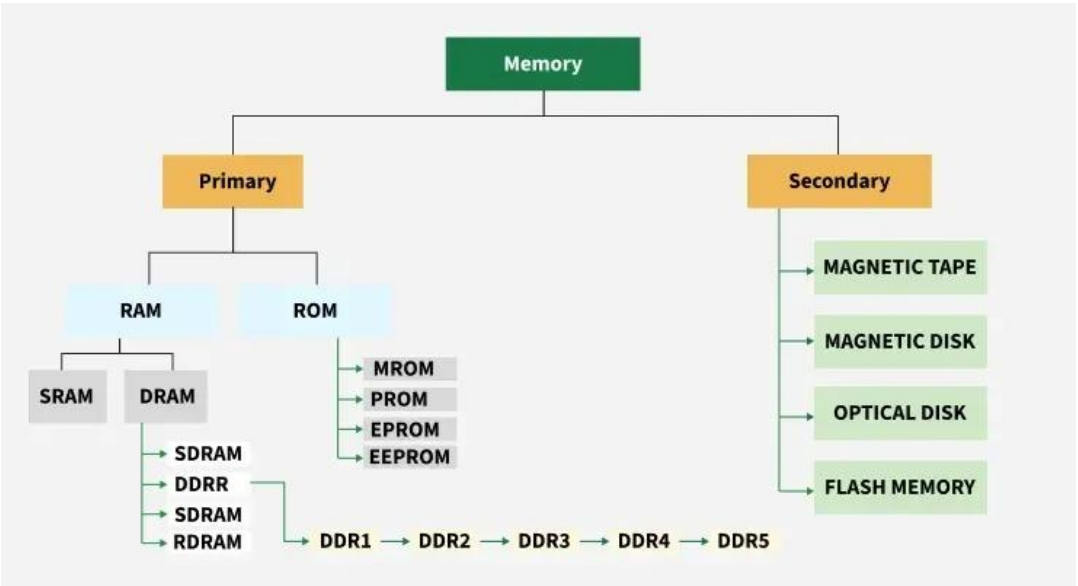


- ✓ Data lines provide the information to be stored in memory. The control inputs specify the direct transfer. The k-address lines specify the word chosen.
- ✓ When there are k address lines, 2k memory words can be accessed.

How does the computer do this?

The computer uses special lines (buses) inside it:

- ✓ Address bus: Tells which drawer (cell address) to use.
- ✓ Data bus: Carries the information that needs to be put in or taken out.
- ✓ Control bus: Tells whether to read or write, and gives the signal when the operation should happen.



1. **RAM (Random Access Memory):** Volatile, read/write, temporary data storage on motherboard.
   - **DRAM:** Uses capacitors, slower, needs refreshing.
   - **SRAM:** Uses flip-flops, faster, retains data as long as power is on.

2. **ROM (Read Only Memory):** Non-volatile, stores permanent system instructions, read-only.
   - PROM: User programmable once.
   - EPROM: Can be erased using UV light and reprogrammed.
   - EEPROM: Electrically erasable and reprogrammable, faster, reusable up to ~10,000 times.
3. **Cache Memory:** Very fast storage, close to CPU, stores frequently used data.
4. **Virtual Memory:** Uses secondary storage as an extension of RAM.
5. **Flash Memory:** Non-volatile, fast, used in SSDs, USBs, memory cards.
6. **Hybrid Memory:** Combines RAM and Flash for efficiency and reduced power use.

PRIMARY MEMORY

It is also referred to as main memory or internal memory. It is a computer system's temporary storage component which is directly accessible by the central processing unit (CPU). It houses data for immediate processing. It holds the data and instructions that the processor is currently working on.

Types of Primary Memory

- **RAM (Random Access Memory):** Temporary storage for actively used data and instructions. It is volatile and lost when power is turned off.
  There are several types of RAM, each with different features:

  (a). Static RAM (SRAM)

  - Stores data using flip-flops made of transistors.
  - Does not need to be refreshed regularly, so it is faster.
  - It is more expensive and consumes more power.
  - Commonly used as cache memory, which helps the CPU access data faster.
  - Stores data temporarily and loses it when power is off.

  (b). Dynamic RAM (DRAM)

  - Stores data using a transistor and a capacitor.
  - Needs to be refreshed thousands of times per second to maintain data.
  - It is slower than SRAM but cheaper and has a higher storage capacity.

- Used as the main memory in most computers to store data that the CPU actively uses.
- Also loses data when power is turned off.

(c). Synchronous DRAM (SDRAM)

- A type of DRAM synchronized with the CPU clock speed for better performance.
- Reduces waiting times for CPU to access memory.
- Widely used in modern computers.

(d). Double Data Rate SDRAM (DDR SDRAM)

- An advanced form of SDRAM that transfers data twice per clock cycle (on both rising and falling edges).
- Provides higher data transfer speeds compared to SDRAM.
- Has several generations like DDR, DDR2, DDR3, DDR4, each improving speed and efficiency.
- Commonly used as the main system memory in laptops and desktops.

- **ROM (Read-Only Memory):** Non-volatile memory that stores firmware and essential instructions for booting the computer. The data is permanent and not lost when power is off. There are several types of ROM, each with different features:

(a). Mask ROM (MROM)

This type of ROM is programmed during the manufacturing process. The data is permanently written into the chip using a mask, so it cannot be altered or erased afterward. It is low cost and used in early embedded systems and firmware.

(b). Programmable ROM (PROM)

PROM is a blank memory chip that can be programmed by the user once with special hardware called a PROM programmer. Once programmed, the data cannot be changed or erased. It is used in firmware and microcode storage.

(c). Erasable Programmable ROM (EPROM)

EPROM can be erased and reprogrammed multiple times. It is erased by exposing the chip to ultraviolet (UV) light for several minutes. After erasing, the chip can be reprogrammed with new data. EPROM was used in older computers and microcontrollers.

(d). Electrically Erasable Programmable ROM (EEPROM)

EEPROM allows data to be erased and reprogrammed electrically without removing the chip from the device. It can be rewritten multiple times and is used in microcontrollers, BIOS, and remote keyless systems. It erases data faster than EPROM and can erase data in smaller sections.

Characteristics Primary Memory

- Volatile: Data is lost upon power loss.
- High-speed access.
- Limited capacity relative to secondary storage.
- Examples: Random Access Memory (RAM), Read-Only Memory, Cache memory.

Advantages Primary Memory

- **High-speed access:** Data can be retrieved and stored very quickly.
- **Directly accessible by CPU:** No intermediate steps are required for data transfer.

Disadvantages Primary Memory

- **Volatile:** Data is lost when power is turned off.
- **Limited storage capacity**: compared to secondary storage, primary memory is relatively small.
- **Expensive**: Cost per unit is higher than secondary storage.

SECONDARY MEMORY

Secondary memory or external memory serves as long-term storage for data and programs. Unlike primary memory, it is not directly accessible by the CPU and requires input/output operations.

"The contents of the secondary memory first get transferred to the primary memory and then are accessed by the processor; this is because the processor does not directly interact with the secondary memory."

Types of Secondary Memory

- **Hard Disk Drive (HDD):** Magnetic storage device used for long-term data storage. It is slower than SSD but offers large capacity.

- **Solid-State Drive (SSD):** Faster than HDD, uses flash memory for storage. More durable and energy-efficient but typically more expensive.
- **Optical Discs (CD/DVD):** Store data using laser technology, commonly used for media storage and software distribution.
- **USB Flash Drive:** Portable storage device using flash memory, widely used for transferring and storing data.
- **External Hard Drive:** Similar to HDD, used for additional storage outside the computer, portable and used for backups.
- **Tape Drive:** Magnetic storage device used mainly for backups; offers high capacity but slower access speed.

Characteristics Secondary Memory

- Non-volatile: Data persists even when the system is powered off.
- slower access speeds compared to primary memory.
- High storage capacity.
- Examples: Hard Disk Drives (HDD), Solid-State Drives (SSD), Optical drives (CD, DVD, Blu-ray).

Advantages Secondary Memory

- **Non-volatile:** Data persists even when the power is turned off.
- **Large storage capacity:** can store vast amount of data.
- **Relatively Inexpensive:** cost-effective for storing large volumes of data.

Disadvantages Secondary Memory

- **Slower access time:** Data retrieval is slower compared to primary memory.
- **Requires input/output operations**: Data transfer involves additional steps

FUNCTIONS OF MEMORY UNIT

The memory unit of a computer has several functions:

1. **Data Storage:** Store temporary data (RAM) and permanent data (ROM).
2. **Quick Access:** CPU retrieves stored data and instructions for fast processing.
3. **Data Transfer:** Move data between CPU, RAM, and storage devices.
4. **Program Execution:** Provide instructions and space for active processes.
5. **Reliability:** Non-volatile memory (ROM/EEPROM) keeps data safe even without power.

The size of the memory unit affects its speed, power, and capabilities. without a memory unit, the processor would have to wait longer for data retrieval.

MEMORY OPERATIONS: STEP-BY-STEP

For example, when the CPU wants to read something from memory:

Step # 1: The CPU places the address of the required cell on the address bus.
Step # 2: It sends a read signal on the control bus.
Step # 3: The memory finds the data at that address and puts it on the data bus.
Step # 4: The CPU reads the data from the data bus and uses it for processing.

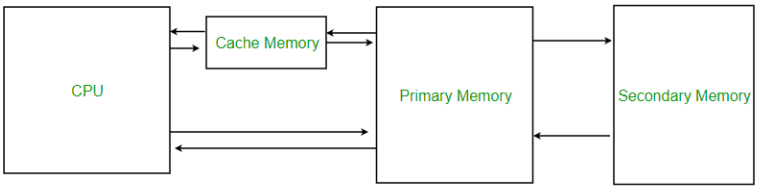
For writing, the CPU does almost the same, but puts data on the data bus and sends a write signal instead

DIFFERENCE BETWEEN PRIMARY AND SECONDARY MEMORY

Feature	Primary Memory	Secondary Memory
Definition	Main memory used to store data temporarily for CPU	Additional storage used to store data permanently
Nature	Temporary (volatile)	Permanent (non-volatile)
Speed	Faster, directly accessible by CPU	Slower, not directly accessible by CPU
Volatility	Volatile (loses data when power is off)	Non-volatile (retains data without power)
Accessibility	Directly accessible by processor/CPU	Accessed through I/O channels after loading into primary memory
Cost	More expensive per byte	Less expensive per byte
Type of Memory Devices	Semiconductor memories (e.g., RAM, ROM)	Magnetic and optical memories (e.g., HDD, CD, tapes)
Data Storage	Holds data and instructions currently used by CPU	Holds data not currently used by CPU
Capacity	Relatively small (16 to 32 GB typical)	Larger capacity (200 GB to multiple terabytes)
Also Known As	Main memory or Internal memory	External memory or Auxiliary memory
Examples	RAM, ROM, PROM, EPROM	Hard Disk, Floppy Disk, Magnetic Tape, USB Flash Drive

CACHE MEMORY

Cache memory is a small, very fast type of memory in a computer that stores copies of data and instructions that the CPU uses frequently. It acts as a buffer between the CPU and the main memory (RAM) to reduce the time it takes the CPU to access data.



- The key function of cache memory is to reduce the average time needed to retrieve data from the main memory.
- Cache memory works on the principle of locality of reference, meaning the CPU is likely to reuse data or instructions it accessed recently or data near it. By keeping this data close, cache speeds up processing significantly, making computers much faster.
- Cache memory stores data close to the CPU, which helps speed up processing. It's much faster than the main memory (RAM). When the CPU needs data, it checks the cache first. If the data is there, it's quickly accessed. If not, the CPU gets it from the slower main memory.

Key Features of Cache Memory

1. **Speed:** Faster than the main memory (RAM), which helps the CPU retrieve data more quickly.
2. **Proximity:** Located very close to the CPU, often on the CPU chip itself, reducing data access time.
3. **Function:** Temporarily holds data and instructions that the CPU is likely to use again soon, minimizing the need to access the slower main memory.

Working of Cache Memory

To understand the working of the cache, we must understand a few points:

1. **Fast but Small**: Cache is way faster than RAM but can only hold a small amount of data because it's limited in size.
2. **Checking the Cache First**: When the CPU needs data, it looks in the cache first because it's so quick. If the data is there (called a **cache hit**), the CPU grabs it and gets to work.
3. **What Happens if Data Isn't in Cache?** : If the data isn't in the cache (called a **cache miss**), the CPU looks in the slower RAM, gets the data, and copies it to the cache for next time.

4. **Speeding Things Up**: By keeping frequently used data in the cache, the CPU spends less time waiting for data from RAM. This makes your computer run faster.

Types of Cache Memory

1. **L1 or Level 1 Cache:** It is the first level of cache memory that is present inside the processor. It is present in a small amount inside every core of the processor separately. The size of this memory ranges from 2KB to 64 KB.
2. **L2 or Level 2 Cache:** It is the second level of cache memory that may present inside or outside the CPU. If not present inside the core, it can be shared between two cores depending upon the architecture and is connected to a processor with the high-speed bus. The size of memory ranges from 256 KB to 512 KB.
3. **L3 or Level 3 Cache:** It is the third level of cache memory that is present outside the CPU and is shared by all the cores of the CPU. Some high processors may have this cache. This cache is used to increase the performance of the L2 and L1 cache. The size of this memory ranges from 1 MB to 8MB.

IMPORTANT

Qns: Why Cache Memory is Important?

Ans: Cache memory acts as a bridge between the CPU and RAM, helping the CPU access data more quickly. It stores frequently used data so that the CPU doesn't have to go all the way to the slower RAM. By keeping this data close, cache memory speeds up the CPU's work and improves the overall performance of the computer.

Qns: How Cache Memory Improves CPU Performance?

Ans: Cache memory helps improve the CPU's performance by reducing the time it takes to fetch data. By keeping the most frequently accessed data closer to the CPU, cache minimizes the need to access slower main memory (RAM). This reduction in wait time results in a much faster and more efficient system.

Qns: What is a Cache Hit and a Cache Miss?

Ans: Cache Hit: When the CPU finds the required data in the cache memory, allowing for quick access. On searching in the cache if data is found, a cache hit has occurred.

Cache Miss: When the required data is not found in the cache, forcing the CPU to retrieve it from the slower main memory. On searching in the cache if data is not found, a cache miss has occurred

DIFFERENCE BETWEEN CACHE AND RAM

Although Cache and RAM both are used to increase the performance of the system but there exists a lot of differences in which they operate to increase the efficiency of the system.

Cache Memory	RAM (Random Access Memory)
Located close to the CPU.	Connected to the CPU via the memory bus.
Stores frequently accessed data and instructions.	Serves as the main working memory for the CPU.
Very fast, with access times in nanoseconds.	Fast, but slower than cache memory, with access times in tens of nanoseconds.
Smaller in size, typically measured in kilobytes (KB) to a few megabytes (MB).	Larger in size, ranging from gigabytes (GB) to terabytes (TB).
Uses SRAM (Static RAM), which is faster but more expensive.	Uses DRAM (Dynamic RAM), which is slower but more cost-effective.
Extremely fast access times due to proximity to the CPU.	Slightly slower access times compared to cache memory.
More expensive per unit of memory due to its speed and proximity to the CPU.	Less expensive per unit of memory compared to cache memory.
Typically organized into multiple levels (L1, L2, L3), with each level increasing in size and latency.	Single level, serving as the primary working memory for the CPU.
Acts as a buffer between the CPU and main memory (RAM), speeding up data access.	Used for storing data and instructions currently being processed by the CPU.
Limited capacity due to its small size and high-speed nature.	Larger capacity, providing ample storage space for running applications and processes.