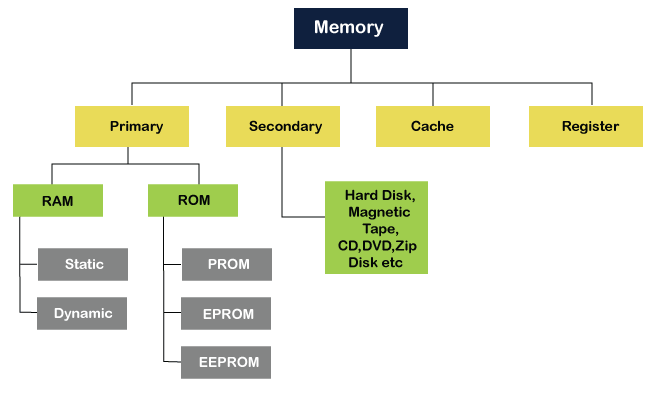
**BASIC CONCEPTS OF MEMORY SYSTEM**

## computer memory

[Computer memory](https://www.javatpoint.com/computer-memory) is any physical device, used to store data, information or instruction temporarily or permanently.

It is the collection of storage units that stores binary information in the form of bits. The memory block is split into a small number of components, called cells. Each cell has a unique address to store the data in memory, ranging from zero to memory size minus one. For example, if the size of computer memory is 64k words, the memory units have 64 \* 1024 = 65536 locations or cells. The address of the memory's cells varies from 0 to 65535.

## Classification of Memory



### **Primary or Main Memory**

Main memory is used to kept programs or data when the processor is active to use them. When a program or data is activated to execute, the processor first loads instructions or programs from secondary memory into main memory, and then the processor starts execution.

 The primary memory is volatile, which means the data in memory can be lost if it is not saved when a power failure occurs.

The primary memory is further divided into two parts:

1. RAM (Random Access Memory)
2. ROM (Read Only Memory)

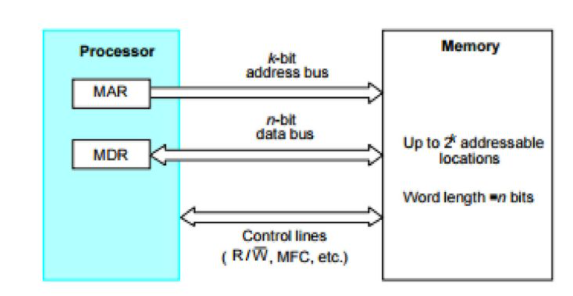
### **Random Access Memory (RAM)**

[Random Access Memory (RAM)](https://www.javatpoint.com/ram) is one of the faster types of main memory accessed directly by the CPU. It is the hardware in a computer device to temporarily store data, programs or program results. It is used to read/write data in memory until the machine is working. It is volatile, which means if a power failure occurs or the computer is turned off, the information stored in [RAM](https://www.javatpoint.com/ram-full-form) will be lost. All data stored in computer memory can be read or accessed randomly at any time

There are two types of RAM:

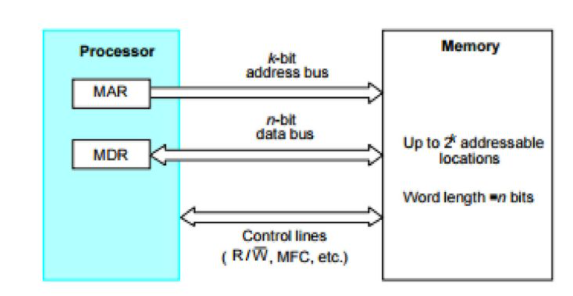
* SRAM
* DRAM

**Basic Concepts:**



From the system standpoint, the Main Memory (MM) unit can be viewed as a “black box”. Data transfer between CPU and MM takes place through the use of two CPU registers, usually called MAR (Memory Address Register) and MDR (Memory Data Register**). If MAR is K bits long and MDR is ‘n’ bits long, then the MM unit may contain upto 2k addressable locations and each location will be ‘n’ bits wide, while the word length is equal to ‘n’ bits.** During a “memory cycle”, n bits of data may be transferred between the MM and CPU. This transfer takes place over the processor bus, which has k address lines (address bus), n data lines (data bus) and control lines like Read/Write’, Memory Function completed (MFC), Bytes specifiers etc (control bus).

For a read operation, the processor reads the data from the memory by loading the address of the required memory location into MAR and setting the R/W’ line to 1. The memory responds by placing the data from the addressed location onto the data lines and confirms this action by asserting MFC signal. Upon receipt of MFC signal, the processor loads the data onto the data lines into MDR register. The processor writes the data into the memory location by loading the address of this location into MAR and loading the data into MDR sets the R/W’ line to 0. This organization is shown in the following block schematic.



**Memory Access Times**: - It is a useful measure of the speed of the memory unit. It is the time that elapses between the initiation of an operation and the completion of that operation (for example, the time between READ and MFC).

**Memory Cycle Time**: - It is an important measure of the memory system. It is the minimum time delay required between the initiations of two successive memory operations (for example, the time between two successive READ operations). The cycle time is usually slightly longer than the access time.

.

**SEMICONDUCTOR RAM MEMORIES:**

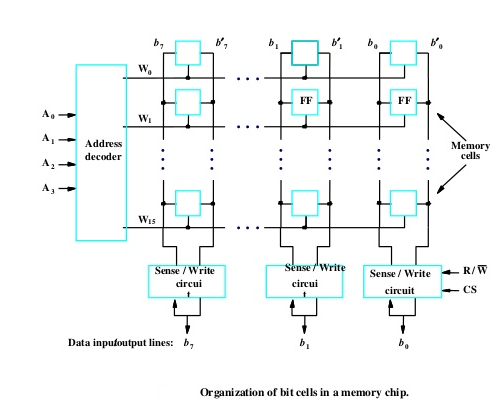
**INTERNAL ORGANIZATION OF MEMORY CHIPS:((RAM organization)**

Memory chips are usually organized in the form of an array of cells, in which each cell is capable of storing one bit of information. A row of cells constitutes a memory word, and the cells of a row are connected to a common line referred to as the word line, and this line is driven by the address decoder on the chip.

The cells in each column are connected to a sense/write circuit by two lines known as bit lines. The sense/write circuits are connected to the data input/output lines of the chip.

During a READ operation, the Sense/Write circuits sense, or read, the information stored in the cells selected by a word line and transmit this information to the output lines. During a write operation, sense/write circuits receive input information and store it in the cells of the selected word.

The following figure shows such an organization of a memory chip consisting of 16 words of 8 bits each, which is usually referred to as a 16 x 8 organization.

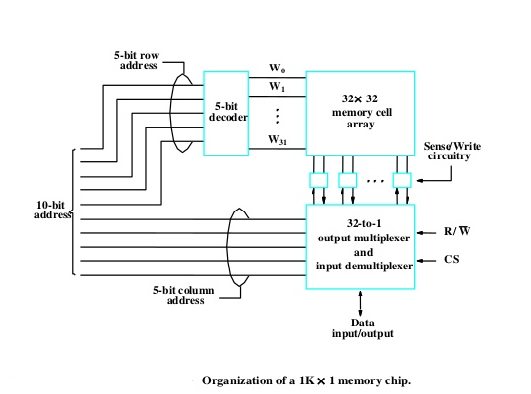


The data input and the data output of each Sense/Write circuit are connected to a single bi-directional data line that can be connected to the data bus of a computer. One control line, the R/W’ (Read/Write’) input is used a specify the required operation and another control line, the CS (Chip Select) input is used to select a given chip in a multichip memory system. This circuit requires 14 external connections (4 address line + 8 data line +R/W line +CS line), and allowing 2 pins for power supply and ground connections, can be manufactured in the form of a 16-pin chip. **It can store 16 x 8 = 128 bits**.

**STRUCTURE OF LARGER MEMORIES**

Another type of organization for 1k x 1 format is shown below(32\*32=1024 memory cell)

The 10-bit address is divided into two groups of 5 bits each to form the row and column addresses for the cell array. In this case, a 10-bit address is needed, but there is only one data line, resulting in 15 external connections. A row address selects a row of 32 cells, all of which are accessed in parallel. According to the column address only one of these cells is connected to the external data line by the output multiplexer and the input demultiplexer.



*32 words are there & each word is having 32 bits.32 words are represented using w1,w2…w31& to select one of these words we are using 5 bit decoder. Each word is having 32 bits. So, we should be able to select one of these 32 bits also. In this case we use 32 to 1 output multiplexer.*

*If we want to perform a read operation, R/W’ will become 1 and 10 bit address will be provided. Using 5 bits one of the word will be selected. So, all the 32 bits from that corresponding word will be coming in to the multiplexer. The 5 bit column address will help to take only one bit as an output. If we want to perform write operation R/W’ will become 0 & 1 bit will give as input of the demultiplexer. Where it has to go to out of these 32 bits will be decided by 5 bit column address and which word it has to go that will has to be decided by 5 bit row address.*

**STATIC MEMORIES**

Memories that consist of circuits capable of retaining their state as long as power is applied are known as static memory.  Thus, this type of memory is called volatile memory.

How a static RAM (SRAM) cell may be implemented:

**Two inverters are cross connected to form a latch.** The latch is connected to two-bit lines by transistors T1 and T2. These transistors act as switches that can be opened / closed under the control of the word line. When the word line is at ground level, the transistors are turned off and the latch retains its state.

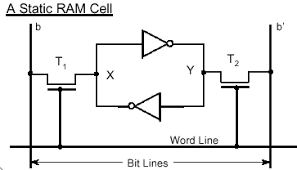
Eg:

Assume that the cell is in state1:

Logic value at point X is 1

Logic value at point Y is 0

This state is maintained as long as the signal on the word line is at ground level.



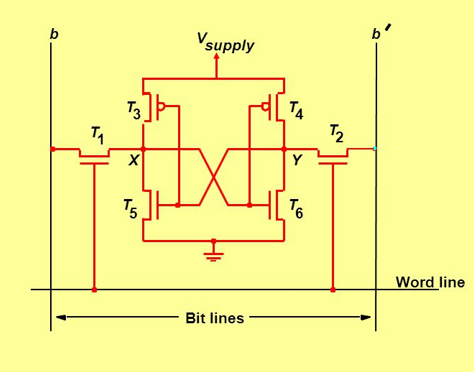
**Read operation: -**

The activated word line closes both the transistors (switches) T1 and T2. Then the bit values at points X and Y can transmit to their respective bit lines. The sense/write circuit at the end of the bit lines sends the output to the processor.

**Write Operation: -**

The state of the cell is set by placing the appropriate value on bit line b and its complement on bꞌ and then activating the word line. This forces the cell into the corresponding state. The required signals on the bit lines are generated by Sense / Write circuit.

**[[CMOS Memory Cell**: Transistor pairs (T3, T5) and (T4, T6) form the inverters in the latch. In state 1, the voltage at point X is high by having T3, T6 ON and T4, T5 are OFF. Thus T1 and T2 returned ON (Closed), bit line b and bꞌ will have high and low signals respectively. The CMOS requires 5V (in older version) or 3.3.V (in new version) of power supply voltage. The continuous power is needed for the cell to retain its state.

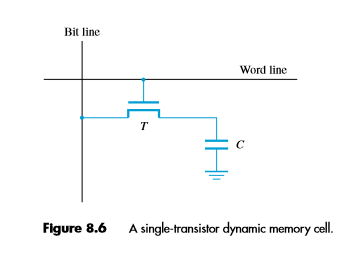


Merit: It has low power consumption because the current flows in the cell only when the cell is being accessed. Static RAMs can be accessed quickly. It access time is few nanoseconds.

Demerit: SRAMs are said to be volatile memories because their contents are lost when the power is interrupted.]]

Dynamic RAM:

* Static RAMs are fast ,but they come at a high cost because their cells require several transistors.
* Less expensive RAM’s can be implemented if simpler cells are used.
* Such cells cannot retain their state indefinitely.
* Hence they are called Dynamic RAM‘s (DRAM).
* The information stored in a dynamic memory cell in the form of a charge on a capacitor and this charge can be maintained only for a few milliseconds.
* The contents must be periodically refreshed by restoring this capacitor charge to its full value.



Capacitors are also used to **maintain the voltage at a certain level**.

Read Operation:

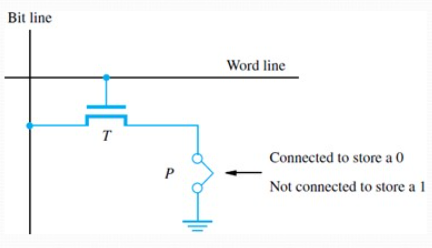
* During a read operation, the transistor is turned on and a sense amplifier connected to the bit line detects whether the charge on the capacitor is above the threshold value.
* If charge on capacitor > threshold value →Bit line will have logic value 1.
* If charge on capacitor < threshold value → Bit line will set to logic value 0.

Write Operation:

* In order to store information in the cell, the transistor T is turned on and the appropriate voltage is applied to the bit line, which charges the capacitor.
* After the transistor is turned off, the capacitor begins to discharge which is caused by the capacitor‘s own leakage resistance.
* Hence the information stored in the cell can be retrieved correctly before the threshold value of the capacitor drops down
* Therefore, for dynamic memory to work, either the CPU or the **memory controller** has to come along and recharge all of the capacitors holding a 1 before they discharge. To do this, the memory controller reads the memory and then writes it right back. This refresh operation happens automatically thousands of times per second.

**READ ONLY MEMORY (ROM):**

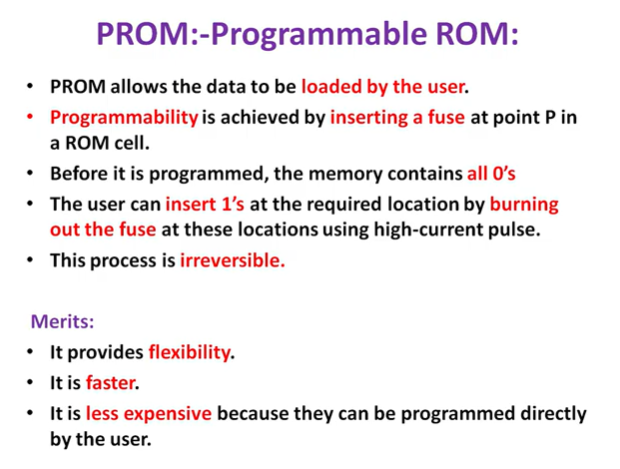
* Both SRAM and DRAM chips are volatile, which means that they lose the stored information if power is turned off.
* Many applications require Non-volatile memory (which retains the stored information if power is turned off).
* Non- volatile memory is used in embedded system. Since the normal operation involves only reading of stored data, a memory of this type is called ROM.
* Both static and dynamic RAM chips are *volatile*, which means that they retain information only while power is turned on. Different types of nonvolatile memories have been developed. A special writing process is needed to place the information into a nonvolatile memory. A memory is called a read-only memory, or ROM, when information can be written into it only once at the time of manufacture



* At Logic value 0 is stored in the cell → Transistor (T) is connected to the ground point (P); otherwise a 1 stored.
* The bit line is connected through a resistor to the power supply.
* To read the state of the cell, the word line is activated.
* A Sense circuit at the end of the bit line generates the proper output value.

**Different types of non-volatile memory (ROM):**

* PROM
* EPROM
* EEPROM
* Flash Memory



### **EPROM (Erasable and Programmable Read Only Memory)**

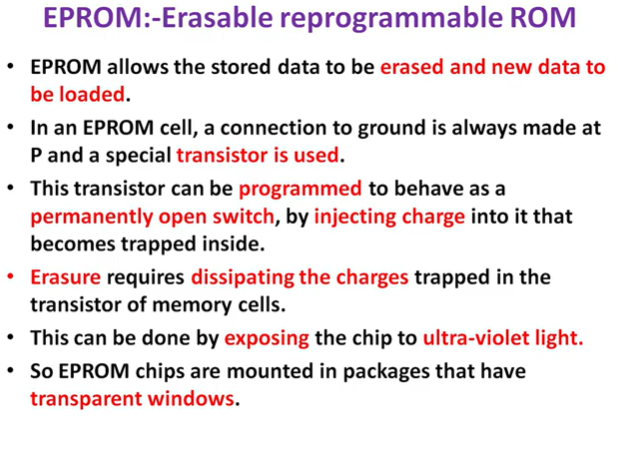
We can reprogram this memory by erasing the data.

Furthermore, to erase the data it has to be exposed to **ultraviolet light.**

During the programming, a charge is trapped in the insulated gate region.

Besides, on exposing it to the ultraviolet light for around 40 minutes this charge destroys.

Hence, in this way, the data gets erased. After erasing the data we can now reprogram the ROM.



Merits:

* It provides flexibility during the development phase of digital system.
* It is capable of retaining the stored information for a long time.

Demerits:

* The chip must be physically removed from the circuit for reprogramming and its entire contents are erased by UV light.

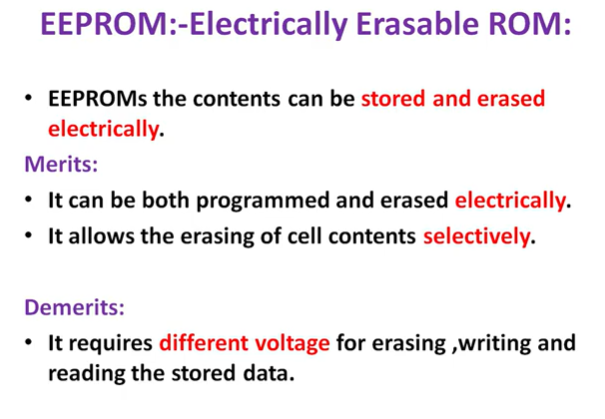
### **EEPROM (Electrically Erasable and Programmable Read Only Memory)**

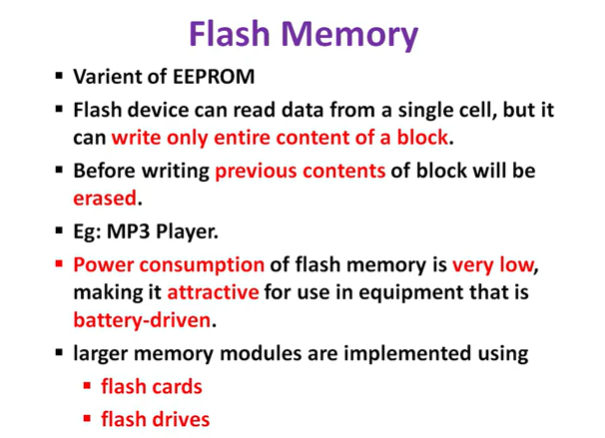
We can program and erase this memory electrically. Furthermore, we do not require any ultraviolet light to erase the data.

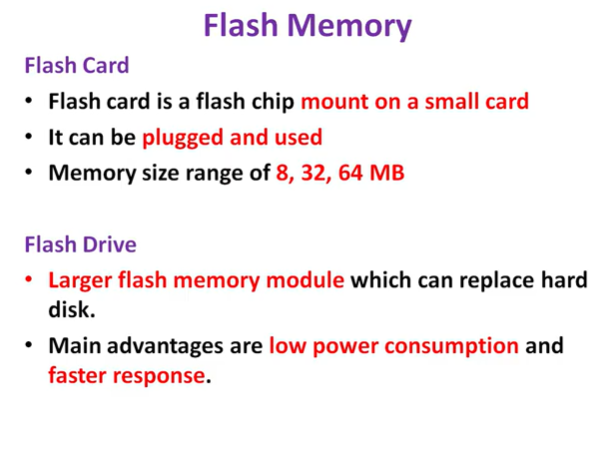
Moreover, erasing and reprogramming is possible many times.

Besides, we can erase any particular location of the memory selectively.

At the same time, we can delete only one byte from the memory at a time rather than erasing the whole chip. Therefore, the process of reprogramming is flexible and slow.







## ****Applications of Flash Memory****

1. **Used in SSDs:**Flash memory is used in SSDs to increase the speed of read/write of operations.
2. **Embedded systems:** Flash memory is used in [embedded systems](https://www.geeksforgeeks.org/introduction-of-embedded-systems-set-1/). Examples: digital cameras, camcorders, MP3 players etc.
3. **Smartphones and tablets:**Flash memory is used in smartphones and tablets.
4. **USB drives:** Flash memory is commonly used in USB drives.

Flash memory has many features. It is a lot less expensive than EEPROM and does not require batteries for solid-state storage such as static RAM (SRAM).

**Speed, Size, and Cost**

The Computer memory hierarchy looks like a pyramid structure which is used to describe the differences among memory types. It separates the computer storage based on hierarchy.

Level 0: CPU registers

Level 1: Cache memory

Level 2: Main memory or primary memory

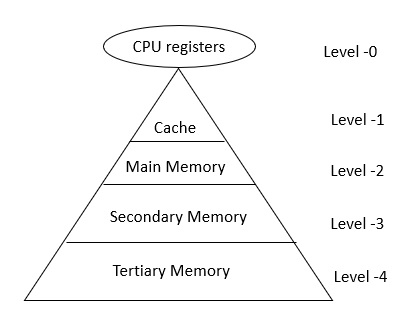
Level 3: Magnetic disks or secondary memory

Level 4: Optical disks or magnetic types or tertiary Memory

**Why memory Hierarchy is used in systems?**

Memory hierarchy is arranging different kinds of storage present on a computing device based on speed of access. At the very top, the highest performing storage is CPU registers which are the fastest to read and write to. Next is cache memory followed by conventional DRAM memory, followed by disk storage with different levels of performance including SSD, optical and magnetic disk drives.

To bridge the processor memory performance gap, hardware designers are increasingly relying on memory at the top of the memory hierarchy to close / reduce the performance gap. This is done through increasingly larger cache hierarchies (which can be accessed by processors much faster), reducing the dependency on main memory which is slower.

In Memory Hierarchy the cost of memory, capacity is inversely proportional to speed. Here the devices are arranged in a manner Fast to slow, that is form register to Tertiary memory.

Let us discuss each level in detail:

Level-0 − Registers

The registers are present inside the CPU. As they are present inside the CPU, they have least access time. Registers are most expensive and smallest in size generally in kilobytes. They are implemented by using Flip-Flops.

Level-1 − Cache

Cache memory is used to store the segments of a program that are frequently accessed by the processor. It is expensive and smaller in size generally in Megabytes and is implemented by using static RAM.

Level-2 − Primary or Main Memory

It directly communicates with the CPU and with auxiliary memory devices through an I/O processor. Main memory is less expensive than cache memory and larger in size generally in Gigabytes. This memory is implemented by using static and dynamic RAM.

Level-3 − Secondary storage

Secondary storage devices like Magnetic Disk are present at level 3. They are used as backup storage. They are cheaper than main memory and larger in size generally in a few TB.

 Secondary storage includes memory devices that are not a part of the CPU chipset or motherboard, for example, magnetic disks, optical disks (DVD, CD, etc.), hard disks, flash drives, and magnetic tapes.

Level-4 − Tertiary storage

Tertiary storage devices like magnetic tape are present at level 4. They are used to store removable files and are the cheapest and largest in size (1-20 TB).

Tertiary storage is used to store huge volumes of data. Since such storage devices are external to the computer system, they are the slowest in speed. These storage devices are mostly used to take the back up of an entire system. Optical disks and magnetic tapes are widely used as tertiary storage.

Let us see the memory levels in terms of size, access time, bandwidth.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Level** | **Register** | **Cache** | **Primary memory** | **Secondary memory** |
| Bandwidth | 4k to 32k MB/sec | 800 to 5k MB/sec | 400 to 2k MB/sec | 4 to 32 MB/sec |
| Size | Less than 1KB | Less than 4MB | Less than 2 GB | Greater than 2 GB |
| Access time | 2 to 5nsec | 3 to 10 nsec | 80 to 400 nsec | 5ms |
| Managed by | Compiler | Hardware | Operating system | OS or use |