UNIT-IV Memory Management 8 hours

- Memory Hierarchy,
- Main Memory (RAM and ROM chips),
- Auxiliary Memory, and Associative memory,
- Cache Memory,
- Memory Mapping:
  - Associative mapping,
  - Direct mapping,
  - Set associative mapping.
- 2D and 2.5D memory organization

OBJECTIVE

- Study about the various types of memory according to size, cost and speed.

Introduction to Memory

In computers, **memory** is the most essential component of the normal functioning of any system, without it computer can't perform simple tasks

- **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**, known as **cells**
- Each cell has a **unique address** to store the data in memory, **ranging** from zero to memory size - one

Example:

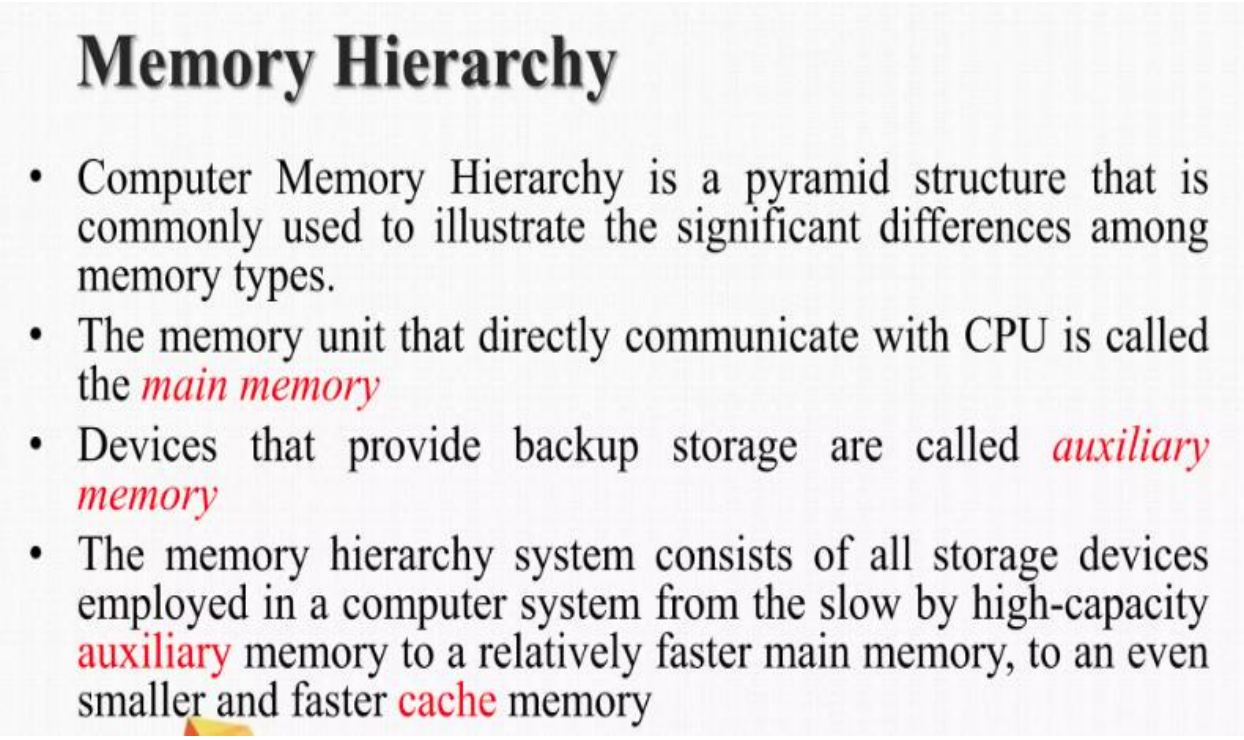
- If the size of **computer memory** is **64k words**, the memory units have 64 * 1024 = 65536 locations or cells
- The address of the memory's cells varies from 0 to 65535

In the computer system, we need computer memory to store various types of data like text, images, video, audio, documents, etc. and we can retrieve it when the data is required

Memory Hierarchy

A memory unit is an essential component in any digital computer since it is needed for storing programs and data.

The memory hierarchy is Pyramid Structure the arrangement of various types of storage on a computing system **based on access speed**. It organizes computer **storage according** to **response time**. **Since response time**, **complexity, and capacity are all connected**, the levels can also be distinguished by **their performance** and controlling technologies.



Memory Hierarchy Design is divided into 2 types:

Primary or internal memory

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).

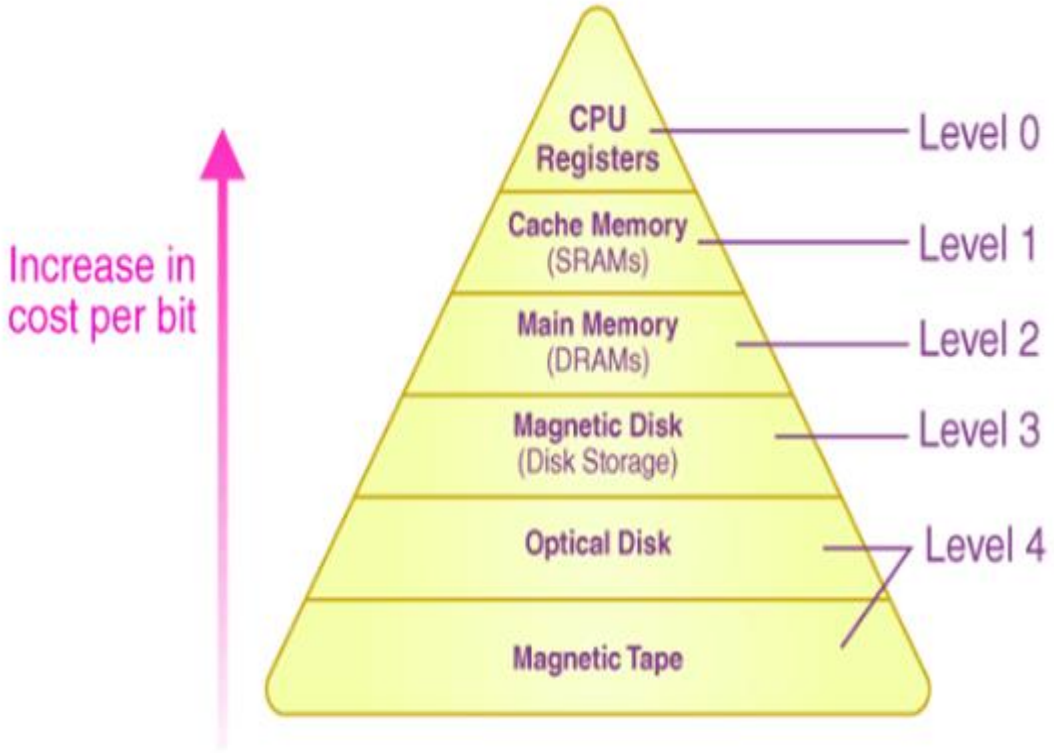
Secondary or external memory

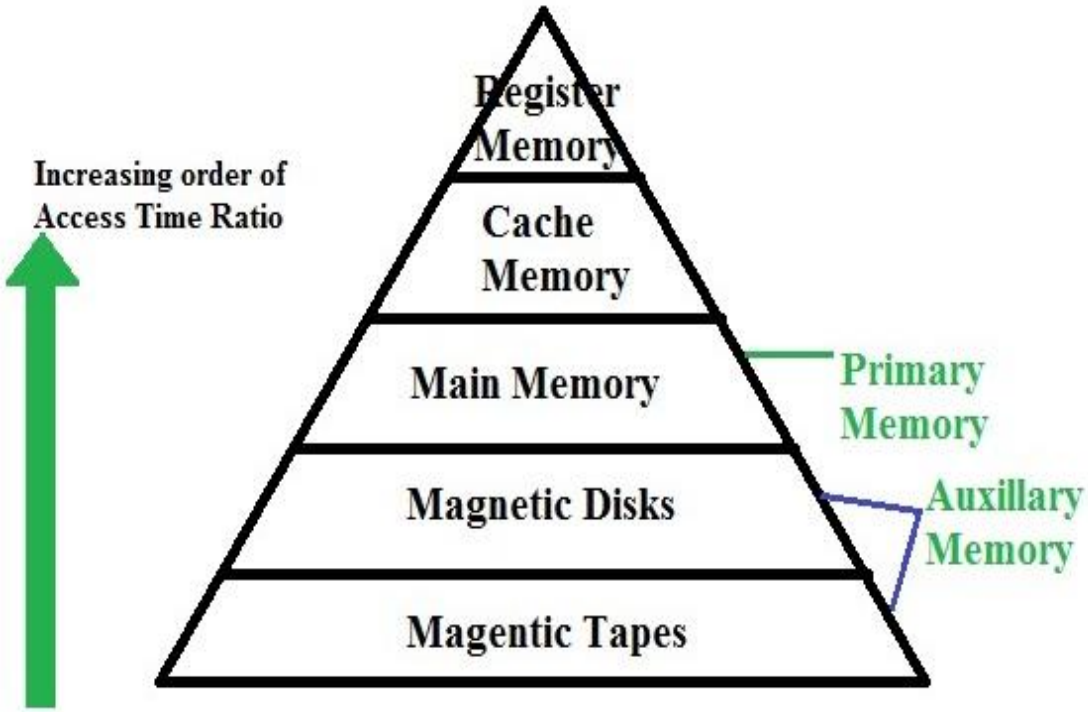
The memory units that provide backup storage are called **Auxiliary Memory**. For instance, magnetic disks and magnetic tapes are the most commonly used auxiliary memories.

1. **Volatile Memory**: This loses its data, when power is switched off.
2. **Non-Volatile Memory**: This is a permanent storage and does not lose any data when power is switched off

Memory hierarchy

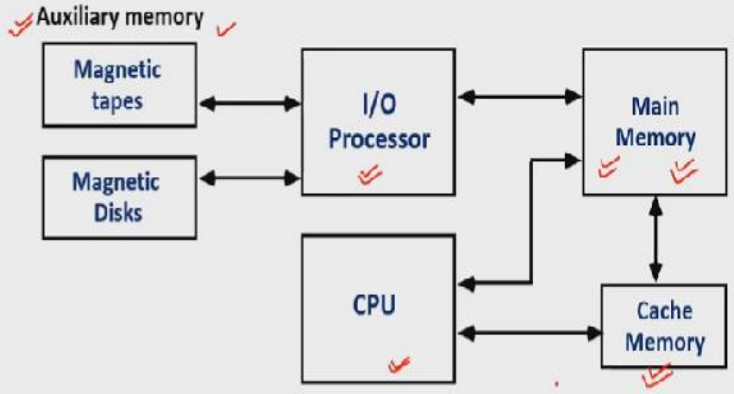
3. Level 0: CPU registers
4. Level 1: Cache memory
5. Level 2: Main memory or primary memory
6. Level 3: Magnetic disks or secondary memory
7. Level 4: Optical disks or magnetic types or tertiary Memory





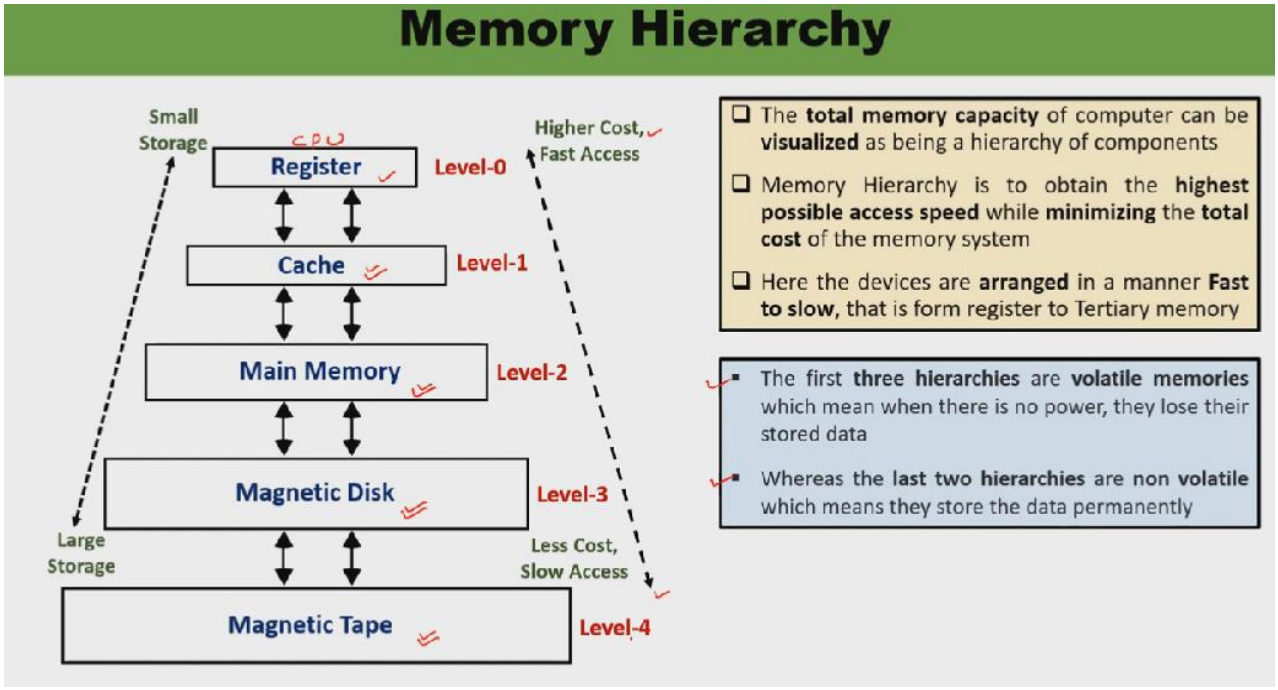
Introduction to Memory Hierarchy

- The **memory unit** is an **essential component** in any **digital computer** since it is needed for storing programs and data
- One **memory unit** does not have **enough space** to accommodate all the programs used in a typical computer
- Therefore, it is **more economical** to use **low cost storage device** to serve as **backup** for storing that information which is not currently used by CPU



Components in a typical memory hierarchy

- **Main memory** directly communicates with the CPU
- Devices which provide backup storage, known as **auxiliary memory**
- **Cache** is a special very high speed memory, sometimes which is used to increase the speed of processing



1. Registers

The register is usually an SRAM or static RAM in the computer processor that is used to hold the data word that is typically 64 bits or 128 bits. A majority of the processors make use of a status word register and an accumulator. The accumulator is primarily used to store the data in the form of mathematical operations, and the status word register is primarily used for decision making.

2. Cache Memory

The cache basically holds a chunk of information that is used frequently from the main memory. We can also find cache memory in the processor. In case the processor has a single-core, it will rarely have multiple cache levels

3. Main Memory

In a computer, the main memory is nothing but the CPU's memory unit that communicates directly. It's the primary storage unit of a computer system. The main memory is very

fast and a very large memory that is used for storing the information throughout the computer's operations. This type of memory is made up of **ROM** as well as **RAM.**

4 Magnetic Disks

A magnetic Disk is a type of secondary memory that is a flat disc covered with a magnetic coating to hold information. It is used to store various programs and files. The polarized information in one direction is represented by 1, and vice versa. The direction is indicated by 0.

5 Magnetic Tape

Magnetic tape refers to a normal magnetic recording designed with a slender magnetizable overlay that covers an extended, thin strip of plastic film. It is used mainly to back up huge chunks of data. When a computer needs to access a strip, it will first mount it to access the information

Characteristics of Memory

Following are the different features of the memory system that includes:

1. Location

- CPU
- Internal
- External

- It represents the **internal** or **external** location of the memory in a computer
- **Internal memory** is inbuilt in computer memory, it is also known as primary memory. Example: registers, cache and main memory
- **External memory** is the separate storage device from the computer, like disk, tape, USB pen drive

2. Capacity

- Word size
  - The natural unit of organisation
- Number of words
  - or Bytes

- Storage capacity can **vary** in **external** and **internal** memory
- **External devices'** storage capacity is measured in terms of bytes
- **Internal memory** is measured with bytes or words
- The storage word length can vary in bits, like 8, 16 or 32 bits

3. Access Methods

Sequential

- Start at the beginning and read through in order
- Access time depends on location of data and previous location
- Ex: Tape

Direct

- Individual blocks have unique address
- Access is by jumping to vicinity plus sequential search
- Access time depends on location and previous location
- Ex: Disk

Random

- Individual addresses identify locations exactly
- Access time is independent of location or previous access
- Ex: RAM

Associative

- Data is located by a comparison with contents of a portion of the store
- Access time is independent of location or previous access
- Ex: Cache

4. Unit of Transfer

A **unit of transfer** measures the transfer rate of bits that can be read or write in or out of the memory devices

- The transfer rate of data can be different in external and internal memory
- **Internal memory:** The transfer rate of bits is mostly equal to the word size
- **External memory:** The transfer rate of bit or unit is not equal to the word length, it is always greater than a word or may be referred to as **blocks**

5. Organization

- Physical arrangement of bits into words
- Not always obvious
- e.g. interleaved

6. Performance

Access Time

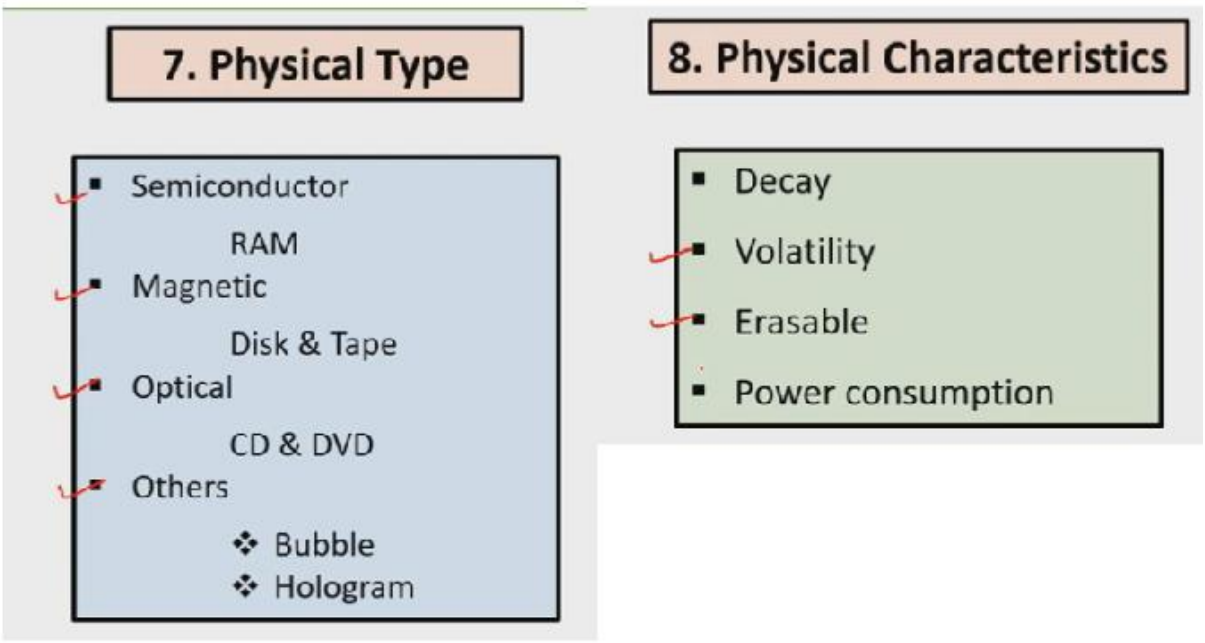
- Time between presenting the address and getting the valid data

Memory Cycle Time

- Time may be required for the memory to "recover" before next access
- Cycle time is access + recovery

Transfer Rate

- Rate at which data can be moved



1. **Capacity:**
   It is the global volume of information the memory can store. As we move from top to bottom in the Hierarchy, the capacity increases.
2. **Access Time:**
   It is the time interval between the read/write request and the availability of the data. As we move from top to bottom in the Hierarchy, the access time increases.
3. **Performance:**
   Earlier when the computer system was designed without Memory Hierarchy design, the speed gap increases between the CPU registers and Main Memory due to large difference in access time
4. **Cost per bit:**
   As we move from bottom to top in the Hierarchy, the cost per bit increases i.e. Internal Memory is costlier than External Memory.

Advantages of Memory Hierarchy

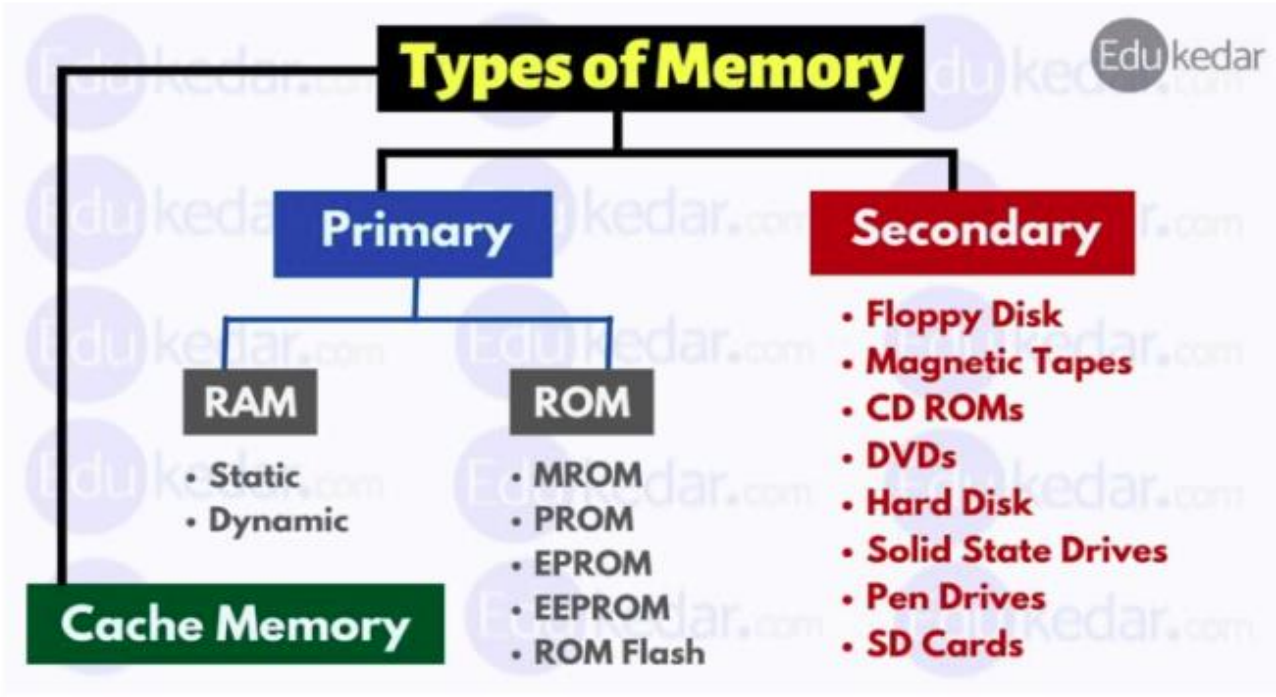
The need for a memory hierarchy includes the following.

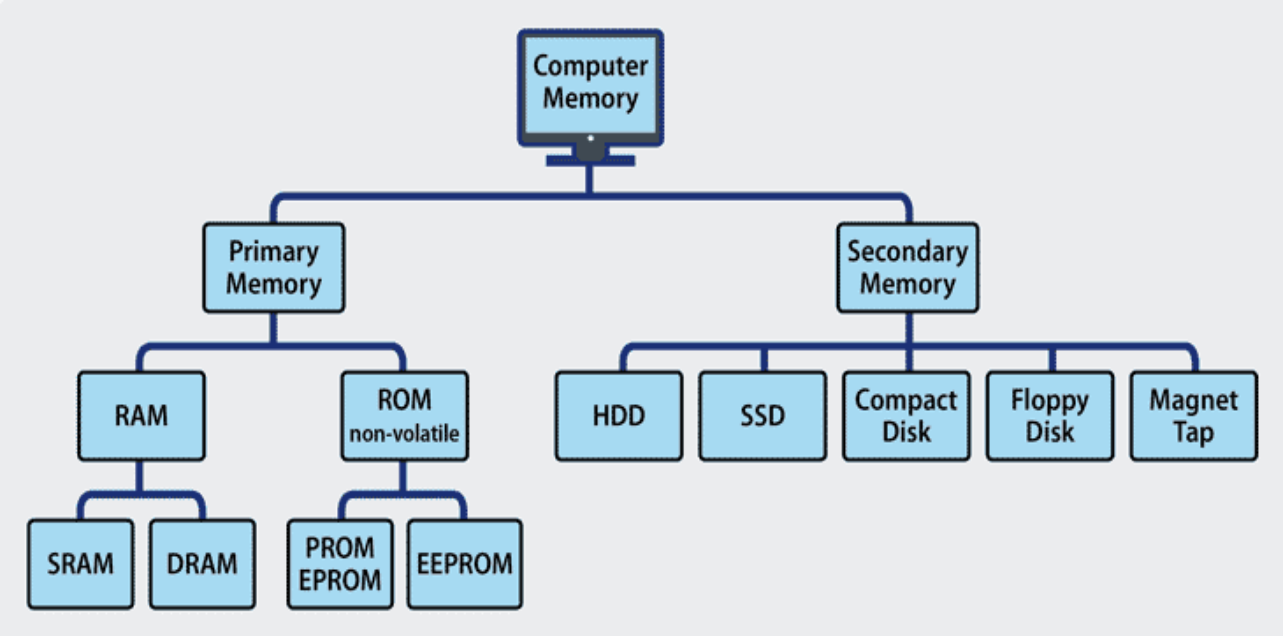
1. Memory distributing is simple and economical
2. Removes external destruction
3. Data can be spread all over
4. Permits demand paging & pre-paging
5. Swapping will be more proficient

✪ Types Of Memory In Computer:

- **Primary Memory** or Internal Memory (RAM, ROM, Cache)
- **Secondary Memory** or External Memory (SSD, CD, Floppy-disk, Magnet-tape)
- **Cache Memory** (It is part of Primary or Internal memory)

The classification of memory is depicted in the diagram below:

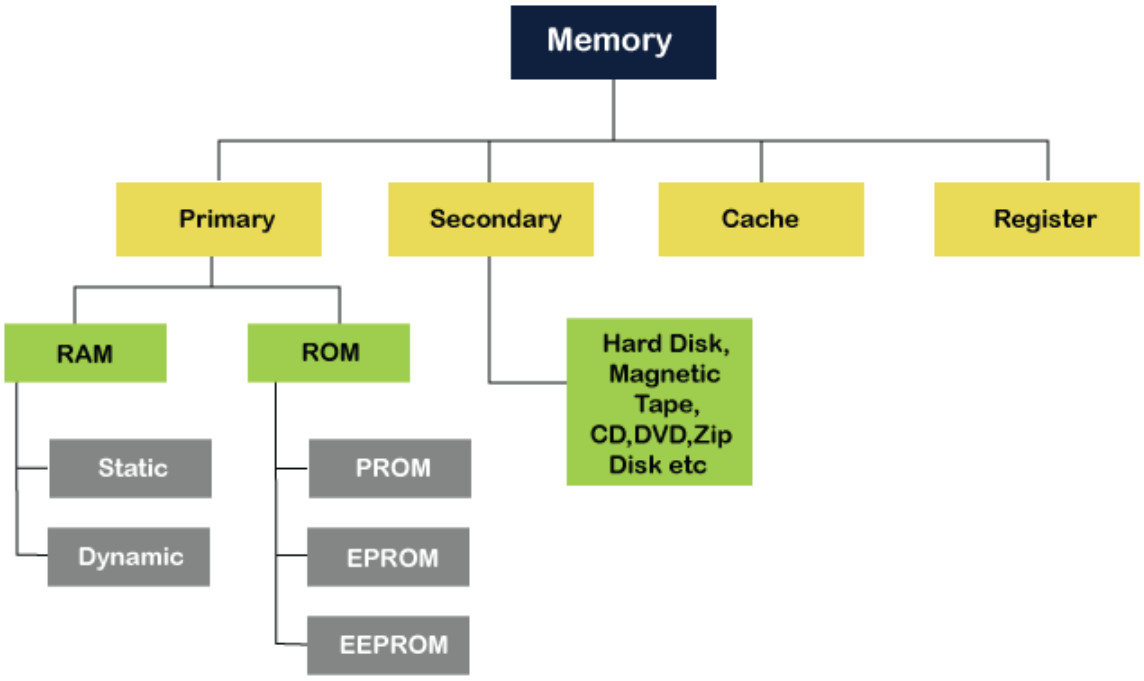




Memory

memory of the computer system. It is used to store data and programs or instructions during computer operations

Classification of Memory



Main Memory (primary memory) internal memory

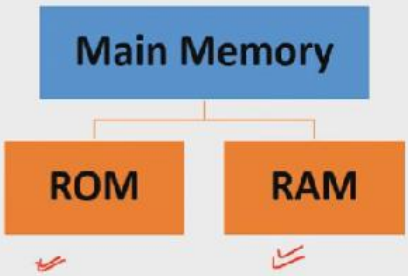
- The memory unit which directly communicates with the CPU is known as Primary Memory.
- The main memory acts as the central storage unit in a computer system.
- Primary memory is computer memory that a processor or computer accesses first or directly.

main memory of the computer system. It is used to store data and programs or instructions during computer operations. It uses semiconductor technology and hence is commonly called semiconductor memory.

- The **main memory** acts as the **central storage unit** in a computer system
- It is a **relatively large** and **fast memory** which is **used** to store programs and data during the run time operations
- The **primary technology** used for the main memory is based on **semiconductor integrated circuits**

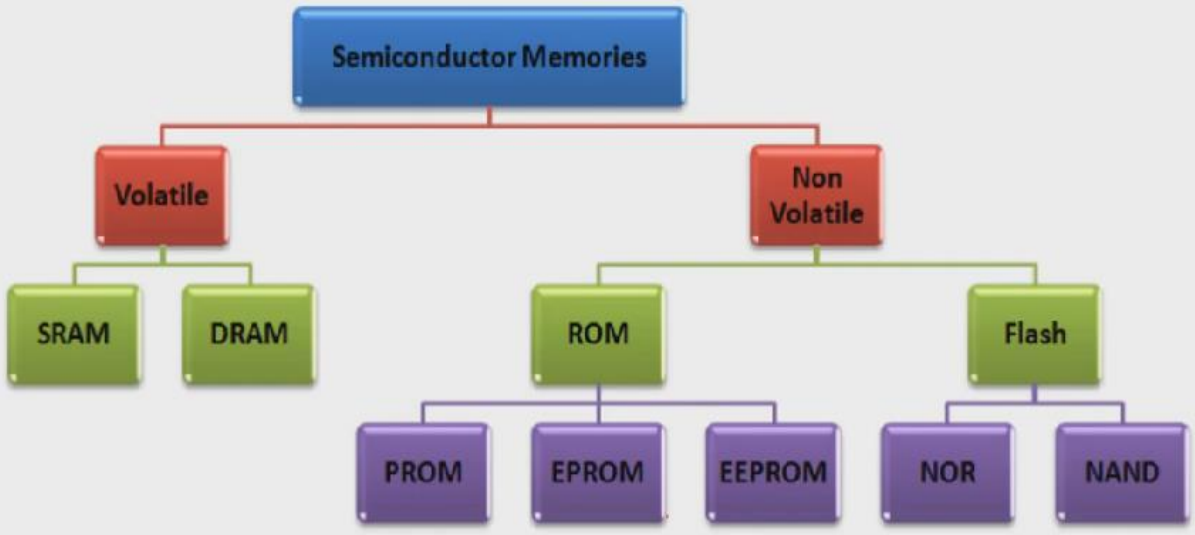
Main Memory Classification

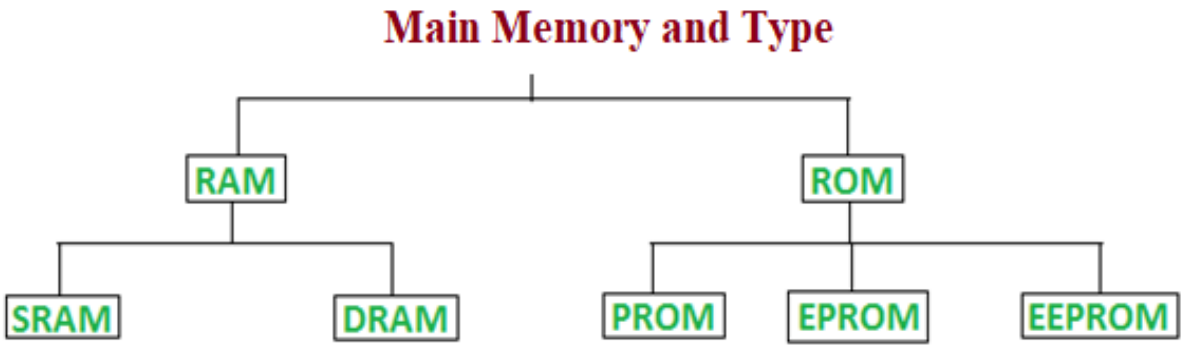
- The **main memory** is the **central storage unit** in a computer system
- It is **relatively large** and **fast memory** used to store programs and data during the computer operation
- The **principal technology** used for the main memory is based on **semiconductor integrated circuits**



- **ROM** is used for storing programs that are permanent residing in the computer
- **RAM** chips are available as **static** and **dynamic**
- **Static RAM** essentially consist of internal flip-flops which stores binary information
- **Dynamic RAM** stores the binary information in the form of electric charges that are applied to capacitors
- Most of **main memory** in a **general purpose computer** is made up of RAM integrated circuit chips, but a portion of memory may be constructed with ROM chips

Semiconductor Memory Classification





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

RAM (Random Access Memory) integrated circuit chips

- It is also called read-write memory or the main memory or the primary memory.
- The programs and data that the CPU requires during the execution of a program are stored in this memory.
- It is a volatile memory as the data is lost when the power is turned off.
- volatile memory. Volatile memory stores information based on the power supply. If the power supply fails/ interrupted/stopped,

There are two types of RAM:

- **SRAM**
- **DRAM**

- **RAM (Main Memory)**
  - Stores programs and data that the computer needs when executing a program
  - **Dynamic RAM (DRAM)**
    - Uses Tiny Capacitors
    - Needs to be recharged every few milliseconds to keep the stored data
  - **Static RAM (SRAM)**
    - Holds its data as long as the power is on
    - D Flip Flop

SRAM

The primary compositions of a static RAM are flip-flops that store the binary information This memory consists of the number of flip flops with each flip flop storing 1 bit. It has less access time and hence, it is faster. The nature of the stored information is volatile, i.e. it remains valid as long as power is applied to the system. The static RAM is easy to use and takes less time performing read and write operations as compared to dynamic RAM.

Characteristics of Static Ram

1. It does not require to refresh.
2. It is faster than DRAM
3. It is expensive.
4. High power consumption
5. Longer life
6. Large size
7. Uses as a cache memory

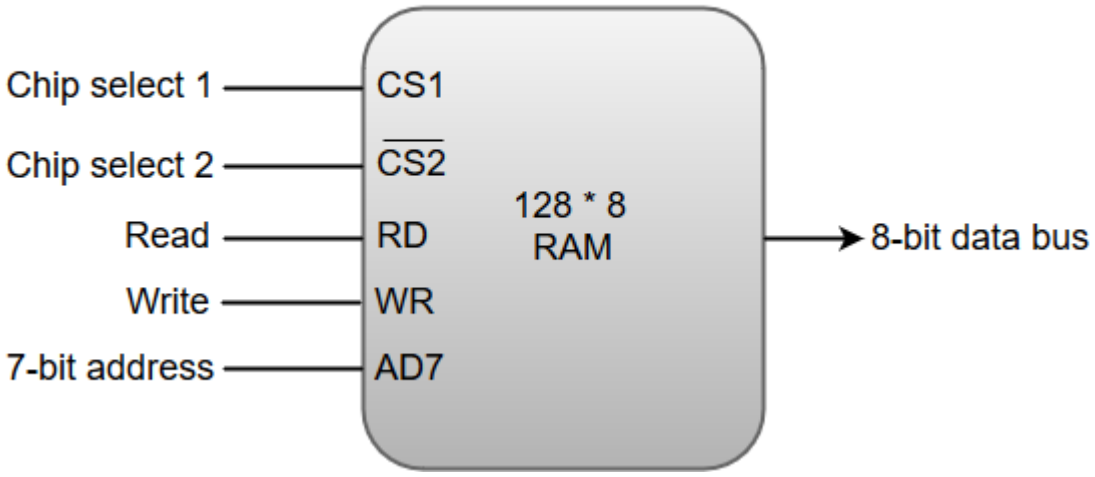
DRAM

It uses capacitors and transistors and stores the data as a charge on the capacitors. They contain thousands of memory cells. It needs refreshing of charge on capacitor after a few milliseconds. This memory is slower than S RAM.

Characteristics of DRAM

1. It requires continuously refreshed to retain the data.
2. It is slower than SRAM
3. It holds a large amount of data
4. It is the combination of capacitor and transistor
5. It is less expensive as compared to SRAM
6. Less power consumption

Typical RAM chip:



- A 128 * 8 RAM chip has a memory capacity of 128 words of eight bits (one byte) per word. This requires a 7-bit address and an 8-bit bidirectional data bus.
- The 8-bit bidirectional data bus allows the transfer of data either from memory to CPU during a **read** operation or from CPU to memory during a **write** operation.
- The **read** and **write** inputs specify the memory operation, and the two chip select (CS) control inputs are for enabling the chip only when the microprocessor selects it.
- The bidirectional data bus is constructed using **three-state buffers**.

Difference between SRAM and DRAM

DRAM	SRAM
1. Constructed of tiny capacitors that leak electricity.	1.Constructed of circuits similar to D flip-flops.
2.Requires a recharge every few milliseconds to maintain its data.	2.Holds its contents as long as power is available.
3.Inexpensive.	3.Expensive.
4. Slower than SRAM.	4. Faster than DRAM.
5. Can store many bits per chip.	5. Can not store many bits per chip.
6. Uses less power.	6.Uses more power.
7.Generates less heat.	7.Generates more heat.
8. Used for main memory.	8. Used for cache.

Read-Only Memory (ROM)

- ROM
  - Stores critical information necessary to operate the system.
  - Hardwired → can not be programmed
- Programmable Read Only Memory (PROM)
  - Can be programmed once using appropriate equipment
- Erasable PROM (EPROM)
  - Can be programmed with special tool
  - It has to be totally erased to be reprogrammed
- Electrical Erasable PROM (EEPROM)
  - No special tools required
  - Can erase a portion

ROM integrated circuit

- The primary component of the main memory is RAM integrated circuit chips, but a portion of memory may be constructed with ROM chips.
- A ROM memory is used for keeping programs and data that are permanently resident in the computer.
- Apart from the permanent storage of data, the ROM portion of main memory is needed for storing an initial program called a **bootstrap loader**. The primary function of the **bootstrap loader** program is to start the computer software operating when power is turned on

Types of ROM

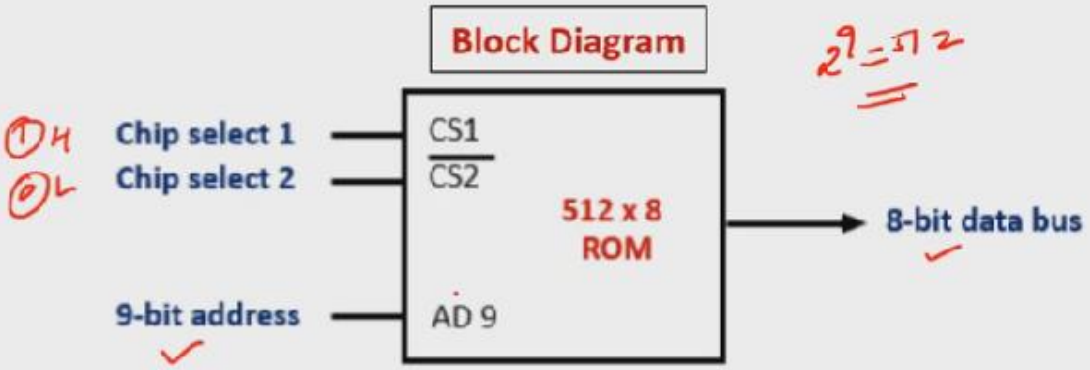
There are five types of Read Only Memory:

1. **MROM (Masked Read Only Memory):**
   MROM is the oldest type of read-only memory whose program or data is pre-configured by the integrated circuit manufacture at the time of manufacturing. Therefore, a program or instruction stored within the MROM chip cannot be changed by the user.
2. **PROM (Programmable Read Only Memory):**
   This read-only memory is modifiable once by the user. The user purchases a blank PROM and uses a PROM program to put the required contents into the PROM. Its content can't be erased once written.
3. **EPROM (Erasable and Programmable Read Only Memory):**
   It is an extension to PROM where you can erase the content of ROM by exposing it to Ultraviolet rays for nearly 40 minutes.
4. **EEPROM (Electrically Erasable and Programmable Read Only Memory):**
   written contents can be erased electrically. You can delete and reprogrammed EEPROM up to 10,000 times. Erasing and programming take very little time, i.e., nearly 4 -10

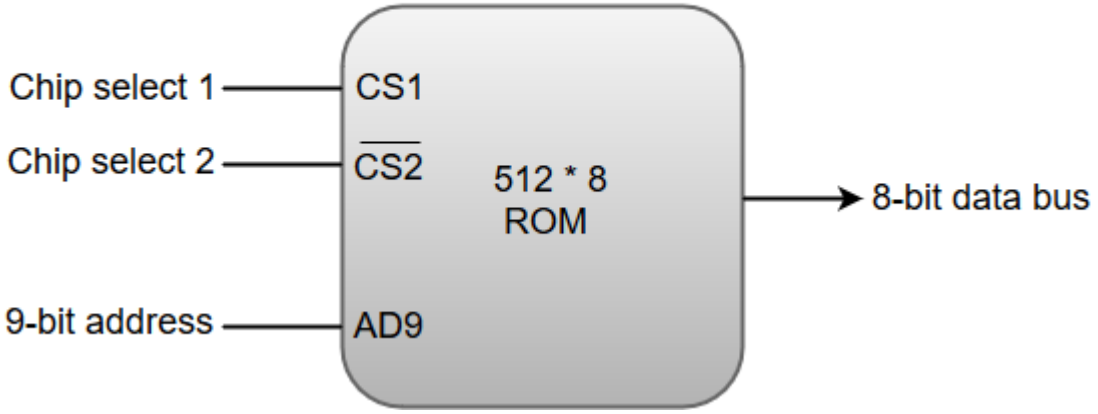
ms(milliseconds). Any area in an EEPROM can be wiped and programmed selectively.

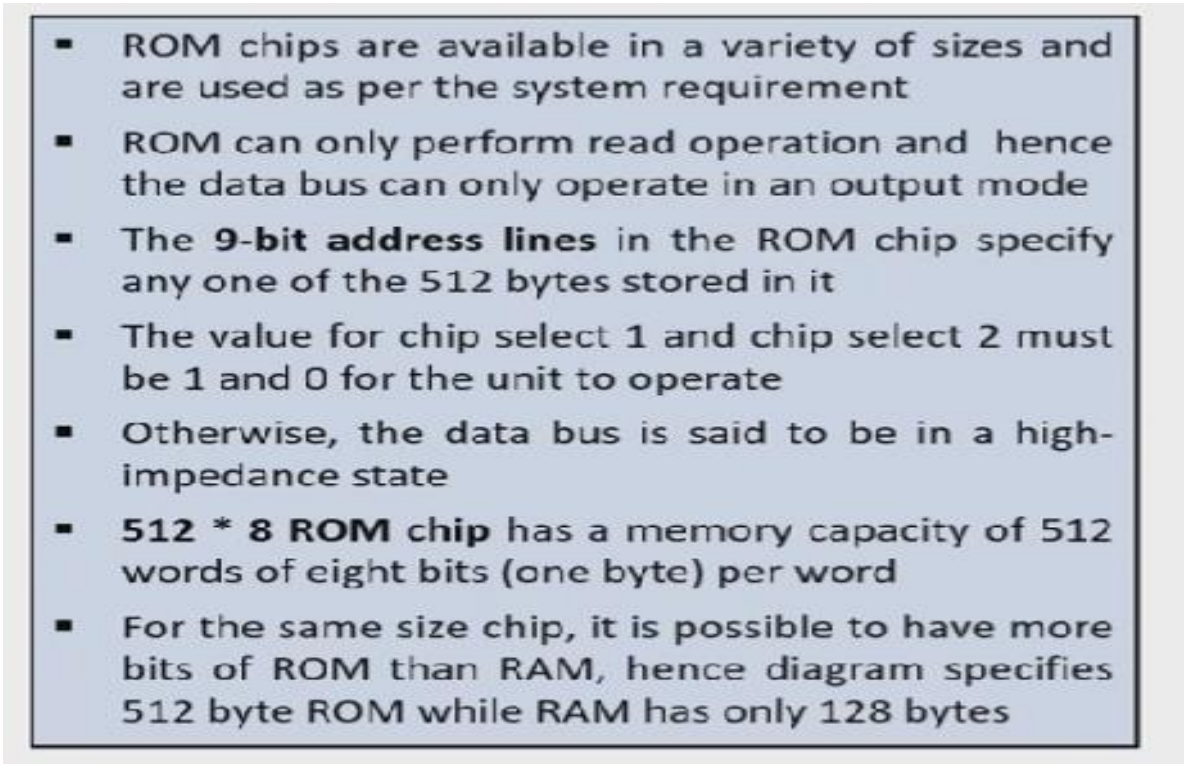
5.

- The primary component of the main memory is RAM integrated circuit chips, but a **portion of memory** may be **constructed** with ROM chips
- A ROM memory is **used** for keeping **programs and data** that are permanently resident in the computer
- Apart from the permanent storage of data, the ROM portion of main memory is required for storing an initial program known as **bootstrap loader**
- The key function of the **bootstrap loader** program is to start the computer software operating when power is turned on



Typical ROM chip:



- A ROM chip has a similar organization as a RAM chip. However, a ROM can only perform read operation; the data bus can only operate in an output mode.
- The 9-bit address lines in the ROM chip specify any one of the 512 bytes stored in it.
- The value for chip select 1 and chip select 2 must be 1 and 0 for the unit to operate. Otherwise, the data bus is said to be in a high-impedance state.
- 

Advantages of ROM

1. It is a non-volatile memory in which stored information can be lost even power is turned off.
2. It is static, so it does not require refreshing the content every time.
3. Data can be stored permanently.
4. It is easy to test and store large data as compared to RAM.
5. These cannot be changed accidently
6. It is cheaper than RAM.
7. It is simple and reliable as compared to RAM.
8. It helps to start the computer and loads the OS.

RAM Vs. ROM

RAM	ROM
It is a Random-Access Memory.	It is a Read Only Memory.
Read and write operations can be performed.	Only Read operation can be performed.
Data can be lost in volatile memory when the power supply is turned off.	Data cannot be lost in non-volatile memory when the power supply is turned off.
It is a faster and expensive memory.	It is a slower and less expensive memory.
Storage data requires to be refreshed in RAM.	Storage data does not need to be refreshed in ROM.
The size of the chip is bigger than the ROM chip to store the data.	The size of the chip is smaller than the RAM chip to store the same amount of data.
Types of RAM: DRAM and SRAM	Types of ROM: MROM, PROM, EPROM, EEPROM

Secondary memory/ Auxiliary Memory: -Also known as secondary **storage or external memory.**

- It is a **non-volatile memory** (does not lose stored data when the device is powered down).
- It is not directly accessed by the CPU, because it is not accessed via input/output channels.

- These are installed for holding data/information permanently. CPU does not access these memories directly, rather they are accessed through input output routines.
- The data hold in secondary memories are transferred to the main memory, and then the **CPU can access that data.**

disk, CD-ROM, DVD, Hard Disk, Floppy Disk, CD (Compact Disc) DVD Drive/Disc, Pen Drive et

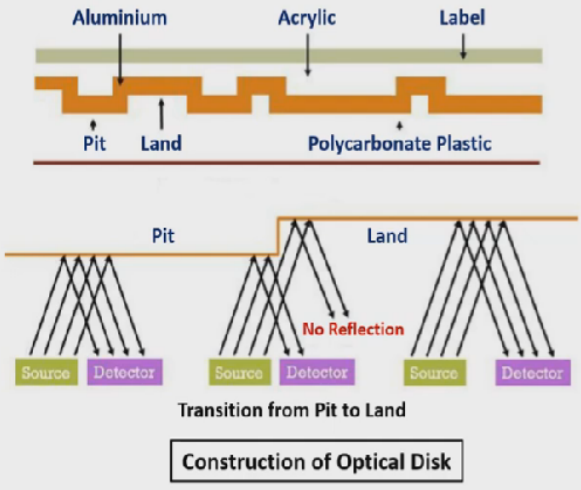
Auxiliary Memory

- **Auxiliary memory** may also refer to as **auxiliary storage, secondary storage, secondary memory, external storage** or **external memory**
- It is the lowest-cost, highest-space, and slowest-approach storage in a computer system
- In auxiliary memory, programs and information are preserved for long-term storage or when not in direct use
- Auxiliary memory holds data for future use, and that retains information even the power fails

- **Early forms** of auxiliary storage included punched paper tape, punched cards, and magnetic drums
- Since the **1980s**, the most common forms of auxiliary storage are magnetic disks, magnetic tapes and optical disks

Optical Disk

- An optical disk is **random access storage medium** and It is made from glass
- Optical disks uses **light technology** where **laser beam** is projected and reflected light is observed
- As compared to magnetic tape and disk, optical disk is **new secondary storage** medium
- It **uses** laser beam technology for READ/WRITE data, so it is also known as **laser disk** or **optical laser disks**
- It has **high storage capacity**, up to 20GB and has **long life** approximately 20 years



Optical Disk

Advantages

- The cost-per-bit of storage for optical disks is very low, because of their **low cost**
- The use of a single spiral track makes optical disks an **ideal storage medium** for reading large blocks of sequential data such as music
- Optical disks have **no mechanical read/write**, which makes optical disks a more reliable storage medium than magnetic tapes or disks
- Compact Size and Light Weight
- Easy to handle, Store and Port from one place to another

Limitations

- **Data access speed** for optical disks is **slower** than magnetic disks
- Optical disk requires a **more complicated drive mechanism** than magnetic disks

Types of Optical Disks

- **CD-ROM** (Compact Disk - Read Only Memory)
- **CD-R** (Compact Disk - Recordable)
- **CD-R or CD-RAM or Erasable Optical Disk** (Compact Disk - Rewriteable)
- **DVD-ROM, DVR and DVD-RAM**

DVD – digital versatile disk
earlier known as digital video disk

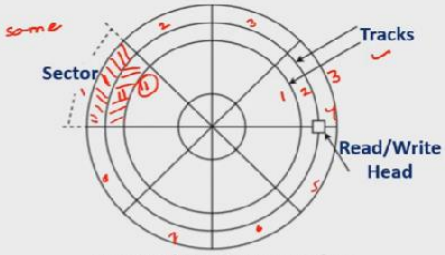
4.7 GB, 8.5 GB, 20 GB etc.
* single layer
* Double layer

CD DVD BluRay Disc

Optical Memory

Magnetic Disk

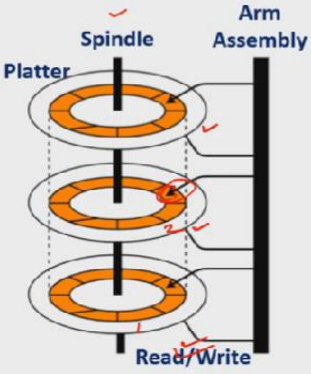
- Magnetic disks are most **popular** for direct access storage device, it **uses** a **magnetization process** to read, write, rewrite and access data
- It stores data in the form of tracks, spots, and sectors
- **Large disk storage** is created by **stacking** together multiple disks
- A **set of same tracks** on all disks forms a **cylinder**, each disk has its own read/write head which work in coordination
- Hard disks, zip disks and floppy disks are common **examples** of magnetic disks



Structural representation for a Magnetic Disk

Features:

- Cheap storage device
- Store a large amount of data
- Easy to carry
- Suitable for frequently read/write data
- Fast access device
- More reliable
- Must be prevented from dust, as read/write head flies over the disk so any dust particle in between can corrupt the disk



Multiple Platters

The diameter of each platter ranges from 1.8 to 5.25 inches

Hard Disk Drive (HDD)

Floppy Disk

Magnetic Memory

Magnetic Disk

Accessing data on the disk requires the following:

- **Seek Time:** The time taken to move the read/write head to the desired track is known as seek time
- **Latency Time or Search Time:** It is the time required to bring the starting position of the address sector of the track to come under read/write head
- **Data Transfer Rate:** Once the read/write head is positioned at the right track and sector, the data has to be written to disk or read from disk. The rate at which data is written to disk or read from disk is known as data transfer rate.
- **Access Time:** It is the time taken to move the read/write head to the address sector

Access Time = Seek Time + Latency Time

Advantages

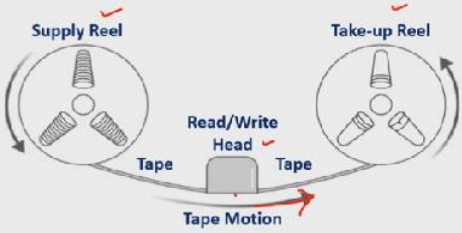
- Economical memory
- Easy and direct access to data is possible
- Store large amounts of data
- Data can be updated when required
- Better data transfer rate than magnetic tapes
- Less prone to corruption of data as compared to tapes
- Data is accessed randomly, so data access time is reduced

Magnetic Tape

- Magnetic Tape is a **Serial access type storage** device i.e. tape needs to rewind or move forward to the location where the requested data is positioned in the magnetic tape
- Due to their sequential nature, magnetic tapes are **not suitable** for data files that need to be revised or updated frequently
- They are generally **used** to store back-up data that is not frequently used or to transfer data from one system to other

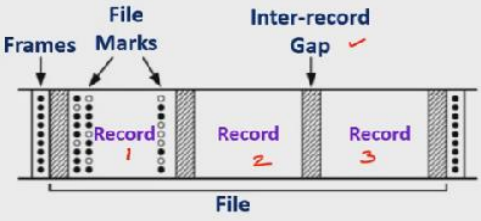
Magnetic Tape

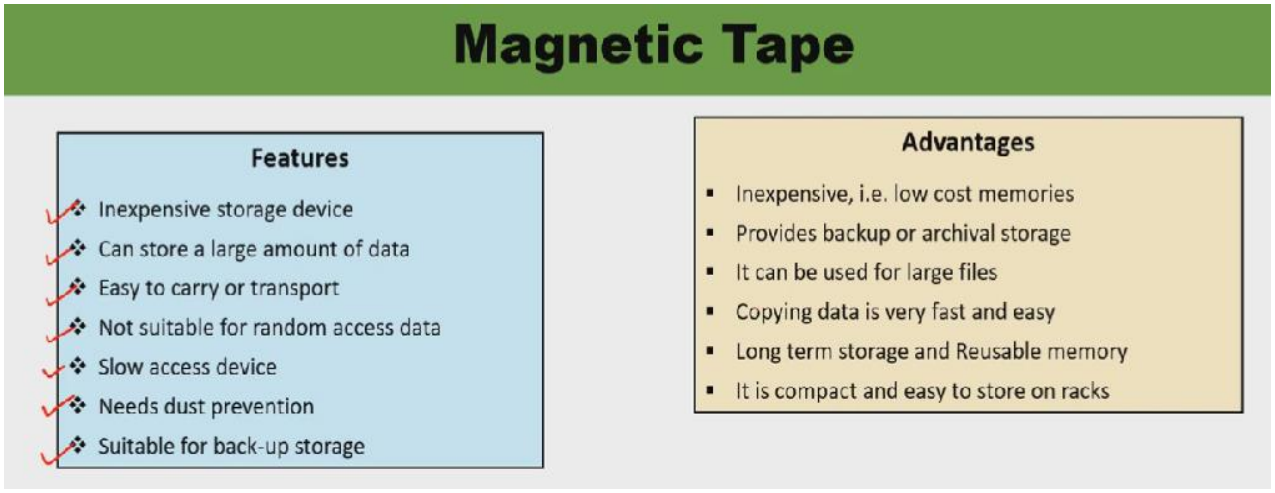
Magnetic tapes are available in ½ inch, ¼ inch, 8 mm & 3 mm sizes



Earlier tapes use 9 tracks to store a byte with parity bit

New tapes use 18 or 36 tracks to store a word or double word with parity bit





Primary Vs. Secondary Memory

Primary Memory	Secondary Memory
It is also known as temporary memory.	It is also known as a permanent memory.
Data can be access directly by the processor or CPU.	Data cannot be accessed directly by the I/O processor or CPU.
Stored data can be a volatile or non-volatile memory.	The nature of secondary memory is always non-volatile.
It is more costly than secondary memory.	It is less costly than primary memory.
It is a faster memory.	It is a slower memory.
It has limited storage capacity.	It has a large storage capacity.
It required the power to retain the data in primary memory.	It does not require power to retain the data in secondary memory.

Examples of primary memory are RAM, ROM, Registers, EPROM, PROM and cache memory.	Examples of secondary memory are CD, DVD, HDD, magnetic tapes, flash disks, pen drive, etc.

Associative Memory

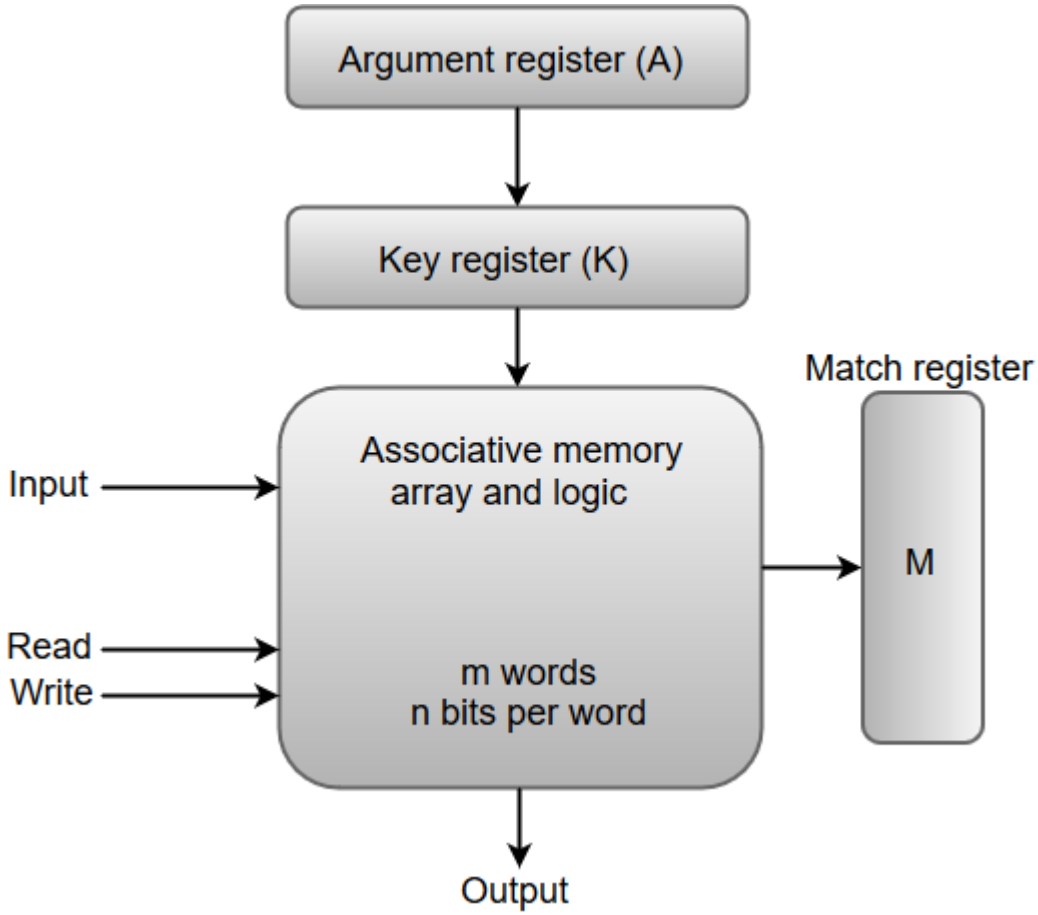
An associative memory can be considered as a memory unit whose stored data can be identified for access by the content of the data itself rather than by an address or memory location.

Associative memory is often referred to as **Content Addressable Memory (CAM)**.

When a write operation is performed on associative memory, no address or memory location is given to the word. The memory itself is capable of finding an empty unused location to store the word.

On the other hand, when the word is to be read from an associative memory, the content of the word, or part of the word, is specified. The words which match the specified content are located by the memory and are marked for reading.

The following diagram shows the block representation of an Associative memory.



From the block diagram, we can say that an associative memory consists of a memory array and logic for 'm' words with 'n' bits per word.

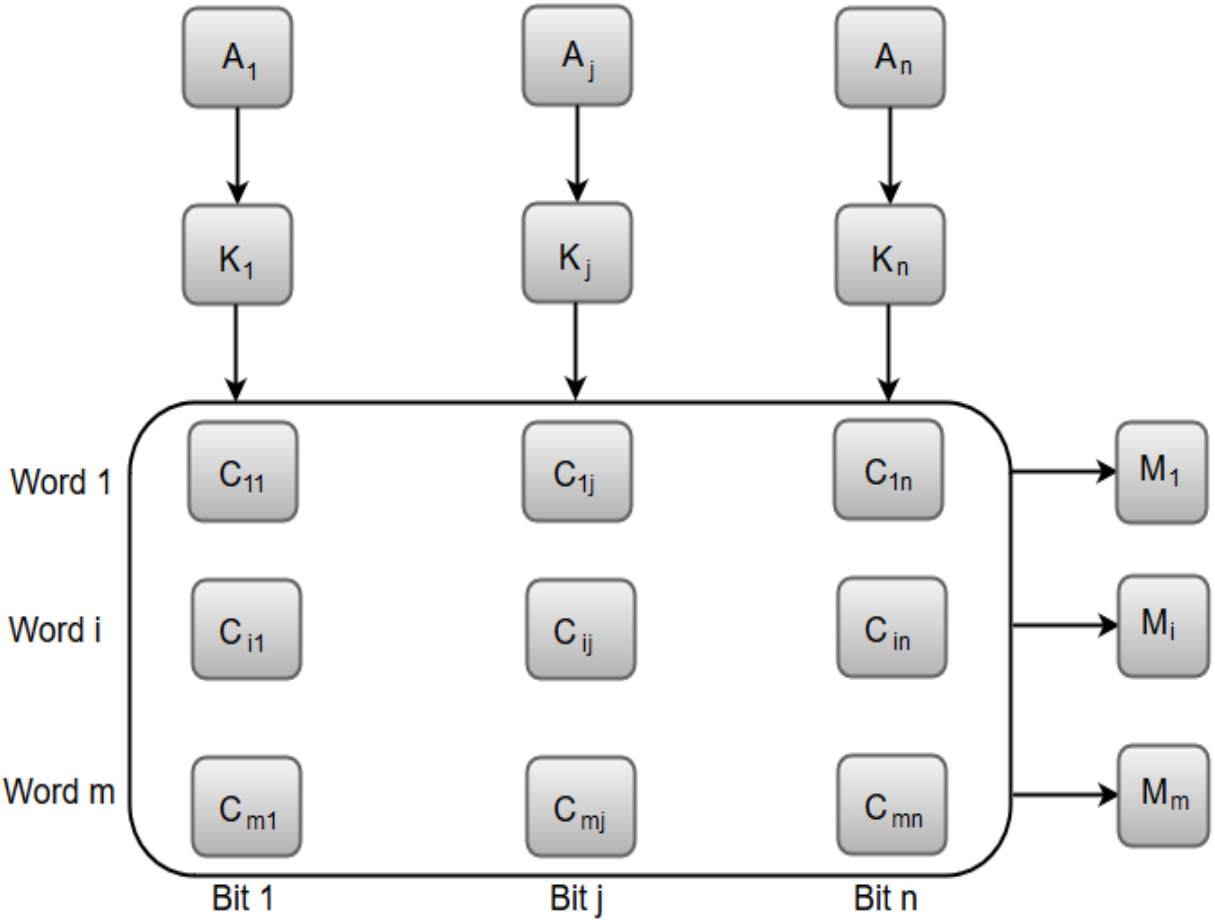
The functional registers like the argument register **A** and key register **K** each have **n** bits, one for each bit of a word. The match register **M** consists of **m** bits, one for each memory word.

The words which are kept in the memory are compared in parallel with the content of the argument register.

The key register (K) provides a mask for choosing a particular field or key in the argument word. If the key register contains a binary value of all 1's, then the entire argument is compared with each memory word. Otherwise, only those bits in the argument that have 1's in their corresponding position of the key register are compared. Thus, the key provides a mask for identifying a piece of information which specifies how the reference to memory is made.

The following diagram can represent the relation between the memory array and the external registers in an associative memory.

Associative memory of m word, n cells per word:



The cells present inside the memory array are marked by the letter C with two subscripts. The first subscript gives the word number and the second specifies the bit position in the word. For instance, the cell C_{ij} is the cell for bit **j** in word **i**.

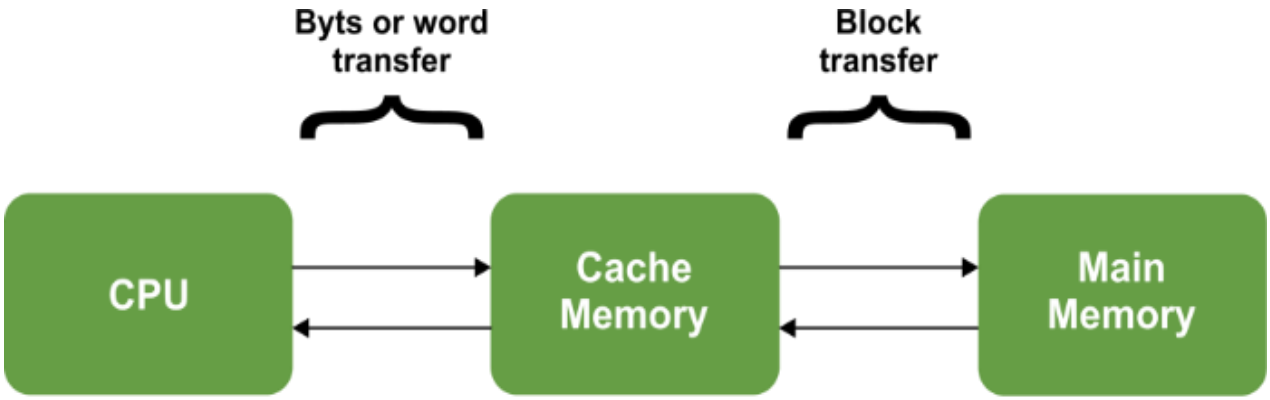
A bit A_j in the argument register is compared with all the bits in column **j** of the array provided that $K_j = 1$. This process is done for all columns **j** = 1, 2, 3....., n.

If a match occurs between all the unmasked bits of the argument and the bits in word **i**, the corresponding bit M_i in the match register is set to 1. If one or more unmasked bits of the argument and the word do not match, M_i is cleared to 0.

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 needs to access memory, it first checks the cache memory. If the data is not found in cache memory, then the CPU moves into the main memory.

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



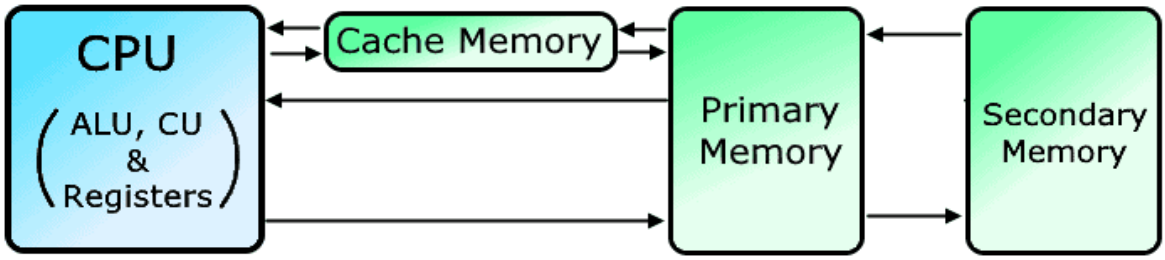
Cache Memory

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

Cache memory is organised as distinct set of blocks where each set contains a small fixed number of blocks.

Types of Computer Memories
The computer memory is classified into three main types;

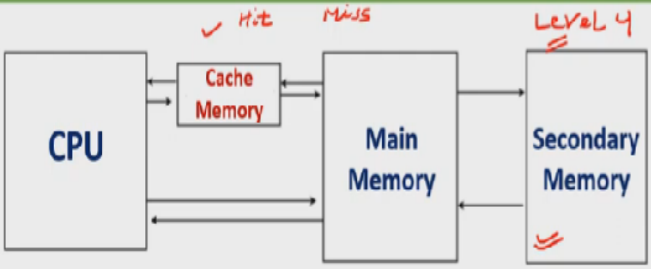
- **Cache Memory**,
- **Primary Memory**
- **Secondary Memory**

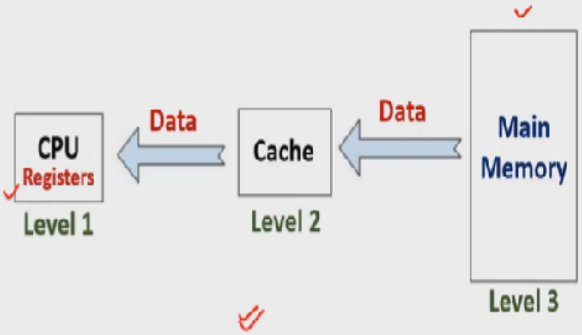


Cache Memory

Why cache is needed?

- The cache memory is required to **balance the speed mismatch** between the main memory and the CPU
- The clock of the processor is very fast, while the main memory access time is comparatively slower
- Hence, the **processing speed depends** more on the speed of the main memory





Levels of memory

- **Level 1 or Register:** Here data is stored and accepted that are immediately stored in CPU. Most commonly used register is accumulator, Program counter, address register etc.
- **Level 2 or Cache memory:** It is the fastest memory which has faster access time where data is temporarily stored for faster access.
- **Level 3 or Main Memory:** It is memory on which computer works currently. Once power is off data no longer stays in this memory.
- **Level 4 or Secondary Memory:** It is external memory which is not as fast as main memory but data stays permanently.

Levels of memory:

- **Level 1 or Register –** It is a type of memory in which data is stored and accepted that are immediately stored in CPU. Most commonly used register is accumulator, Program counter, address register etc.
- **Level 2 or Cache memory –** It is the fastest memory which has faster access time where data is temporarily stored for faster access.
- **Level 3 or Main Memory –** It is memory on which computer works currently. It is small in size and once power is off data no longer stays in this memory.
- **Level 4 or Secondary Memory –** It is external memory which is not as fast as main memory but data stays permanently in this memory.

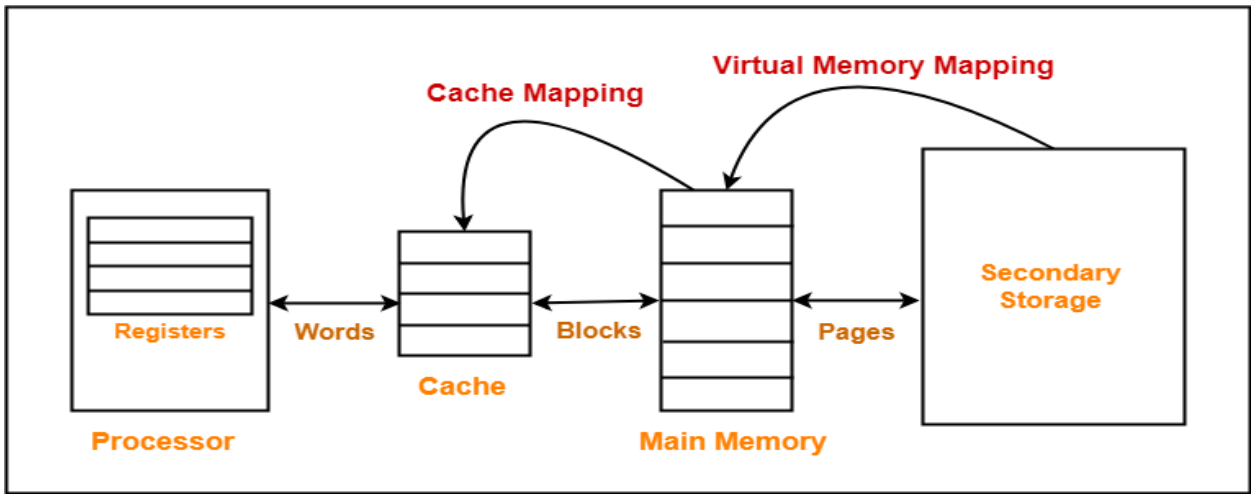
Cache Performance: When the processor needs to read or write a location in main memory, it first checks for a corresponding entry in the cache.

- If the processor finds that the memory location is in the cache, a **cache hit** has occurred and data is read from the cache.
- If the processor **does not** find the memory location in the cache, a **cache miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from main memory, then the request is fulfilled from the contents of the cache.

The performance of cache memory is frequently measured in terms of a quantity called **Hit ratio.**

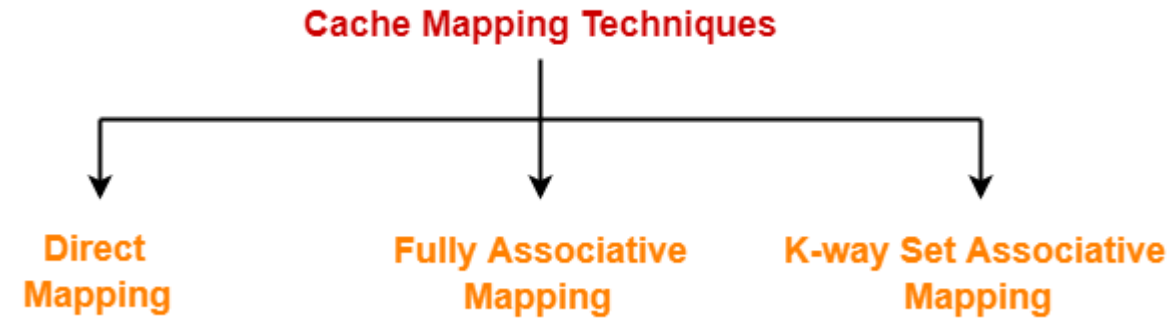
Hit ratio = hit / (hit + miss) = no. of hits/total accesses

Cache Mapping: There are three different types of mapping used for the purpose of cache memory which is as follows: Direct mapping, Associative mapping, and Set-Associative mapping. These are explained below.



- Main memory is divided into equal size partitions called as **blocks** or **frames**.
- Cache memory is divided into partitions having same size as that of blocks called as **lines**.

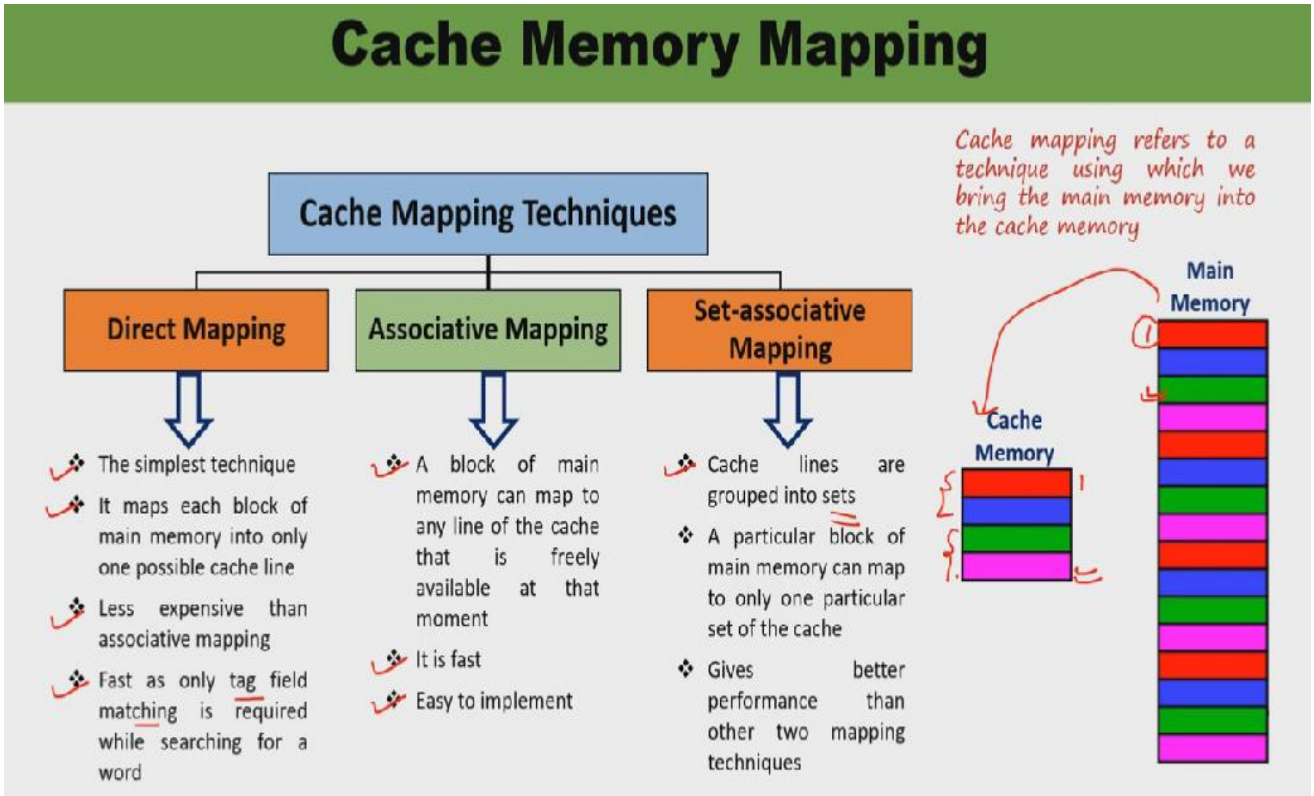
- During cache mapping, block of main memory is simply copied to the cache and the block is not actually brought from the main memory.



Cache Mapping:

There are three different types of mapping

- Direct mapping,
- Associative mapping
- Set-Associative mapping



Direct Mapping –

The simplest technique, known as direct mapping, maps each block of main memory into only one possible cache line. or In Direct mapping, assign each memory block to a specific line in the cache.

Cache line number
= (Main Memory Block Address) Modulo (Number of lines in Cache)

or

Direct Cache Mapping A[i] = A[j] mod Main memory

- Consider cache memory is divided into 'n' number of lines.
- Then, block 'j' of main memory can map to line number (j mod n) only of the cache.

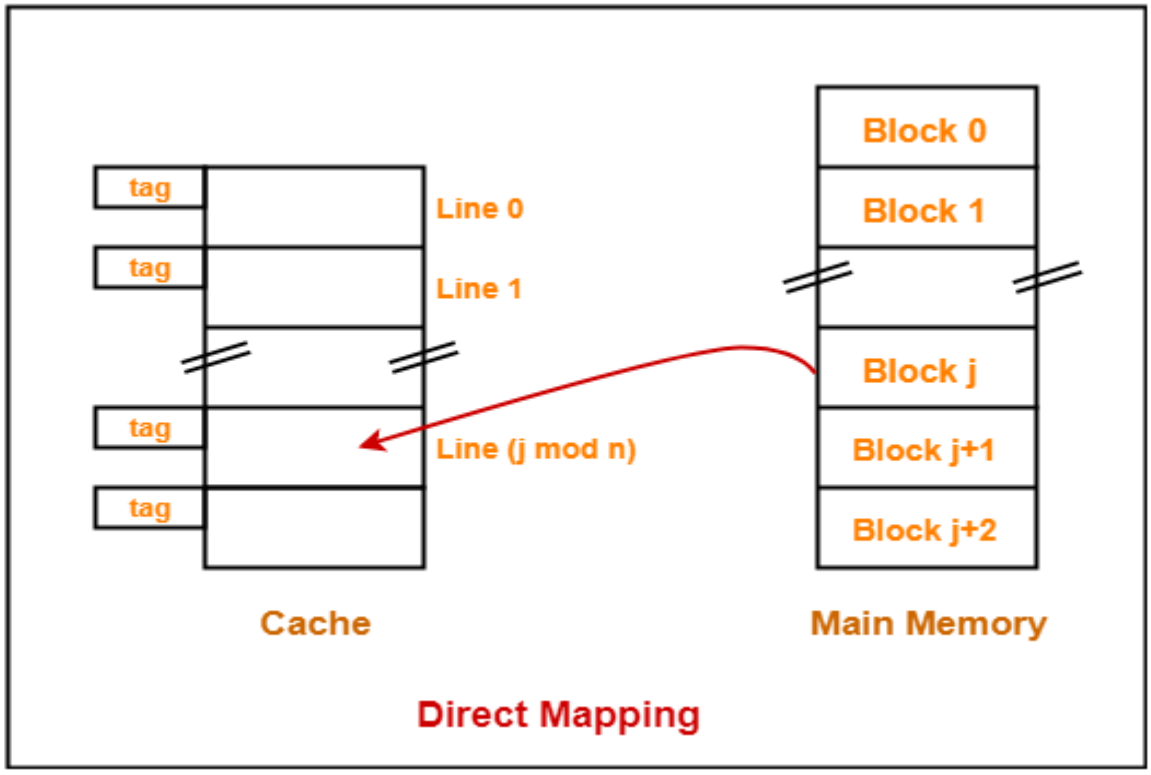
i = j modulo m

where

i=cache line number

j= main memory block number

m=number of lines in the cache



Direct Mapping Cache

A particular block of main memory can map to only one particular line of the cache

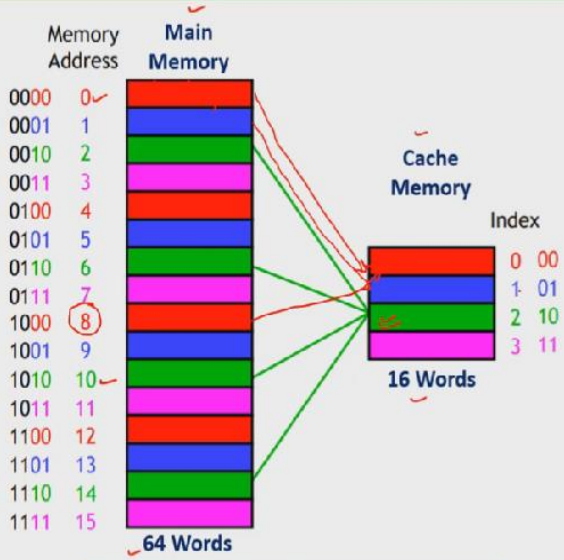
Cache line number = (Main Memory Block Address) **Modulo** (Number of lines in Cache)

mod operator i.e. (remainder)

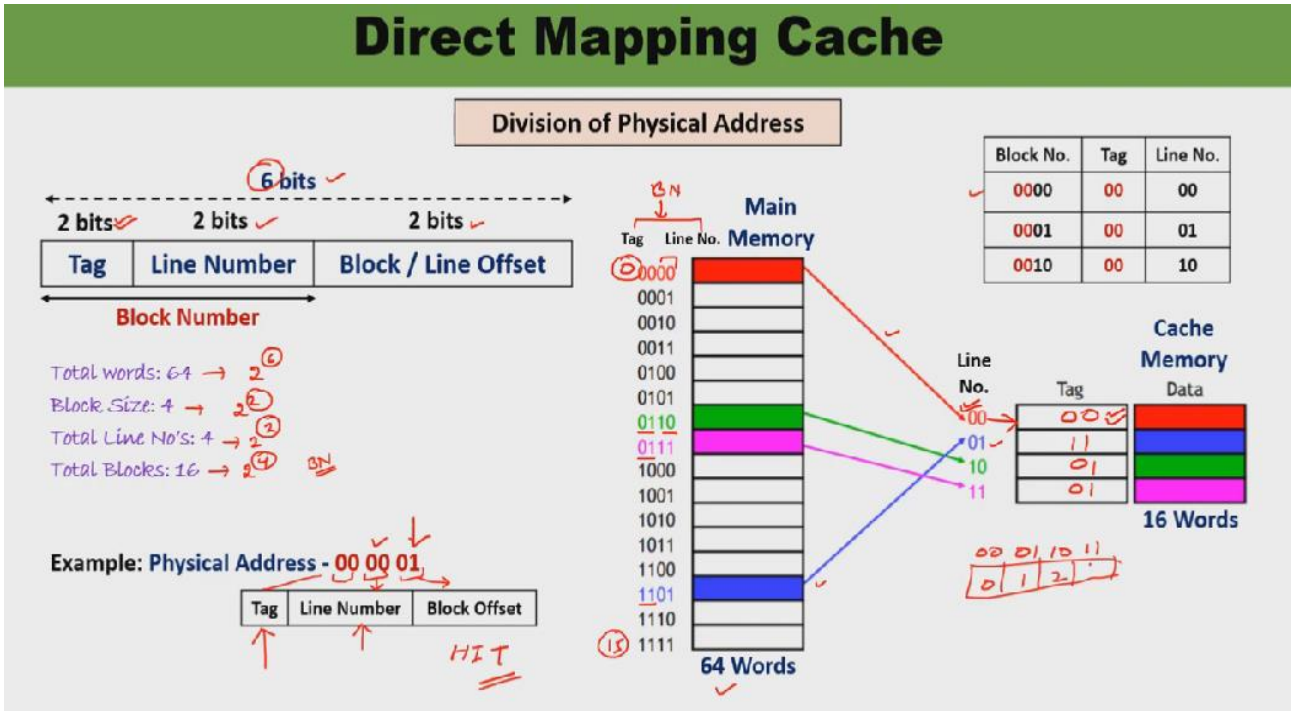
0 mod 4 = 0
1 mod 4 = 1
8 mod 4 = 0
10 mod 4 = 2

- Example: there is a 16-byte main memory and a 4-byte cache
- Memory locations 0, 4, 8 and 12 all map to cache block 0
- Addresses 1, 5, 9 and 13 map to cache block 1, etc.
- For instance, with the four-block cache here, address 14 would map to cache block 2

14 mod 4 = 2



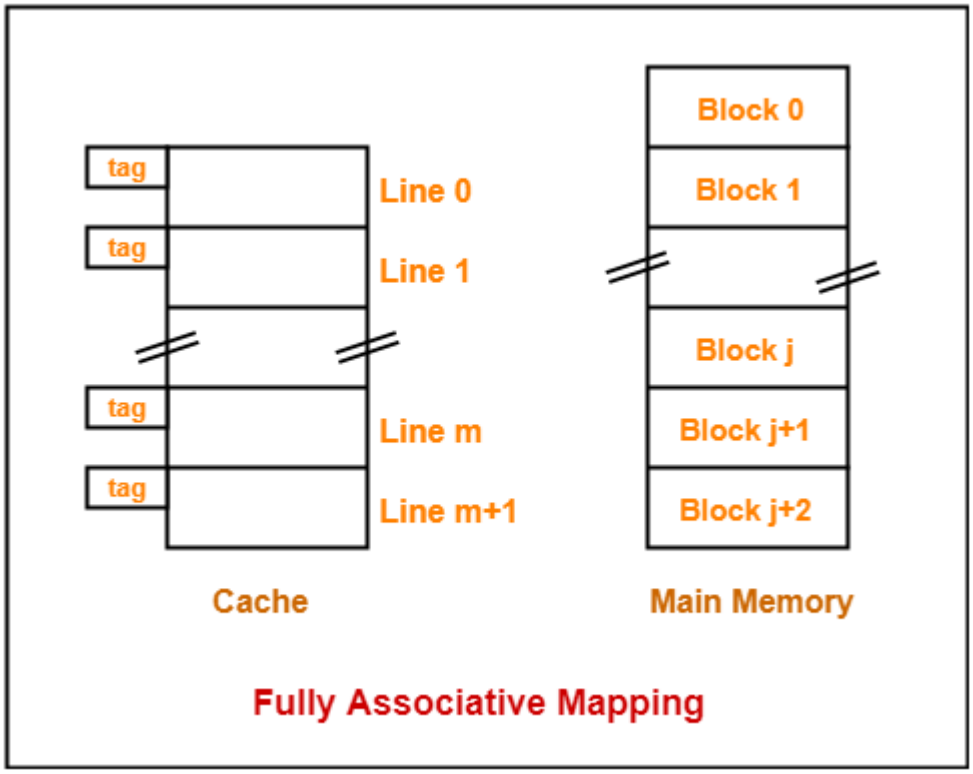
The Direct Mapping Cache Organization



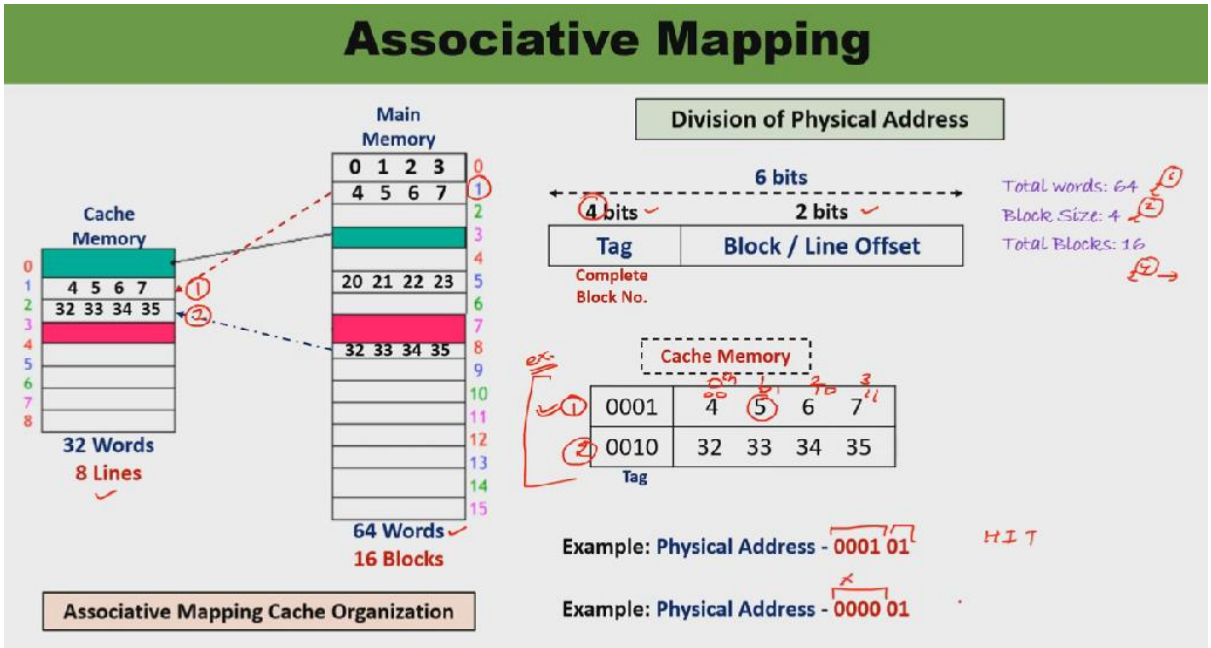
Associative Mapping-

In this type of mapping, the associative memory is used to store content and addresses of the memory word. Any block can go into any line of the cache. This means that

- A block of main memory can map to any line of the cache that is freely available at that moment.
- This makes fully associative mapping more flexible than direct mapping.



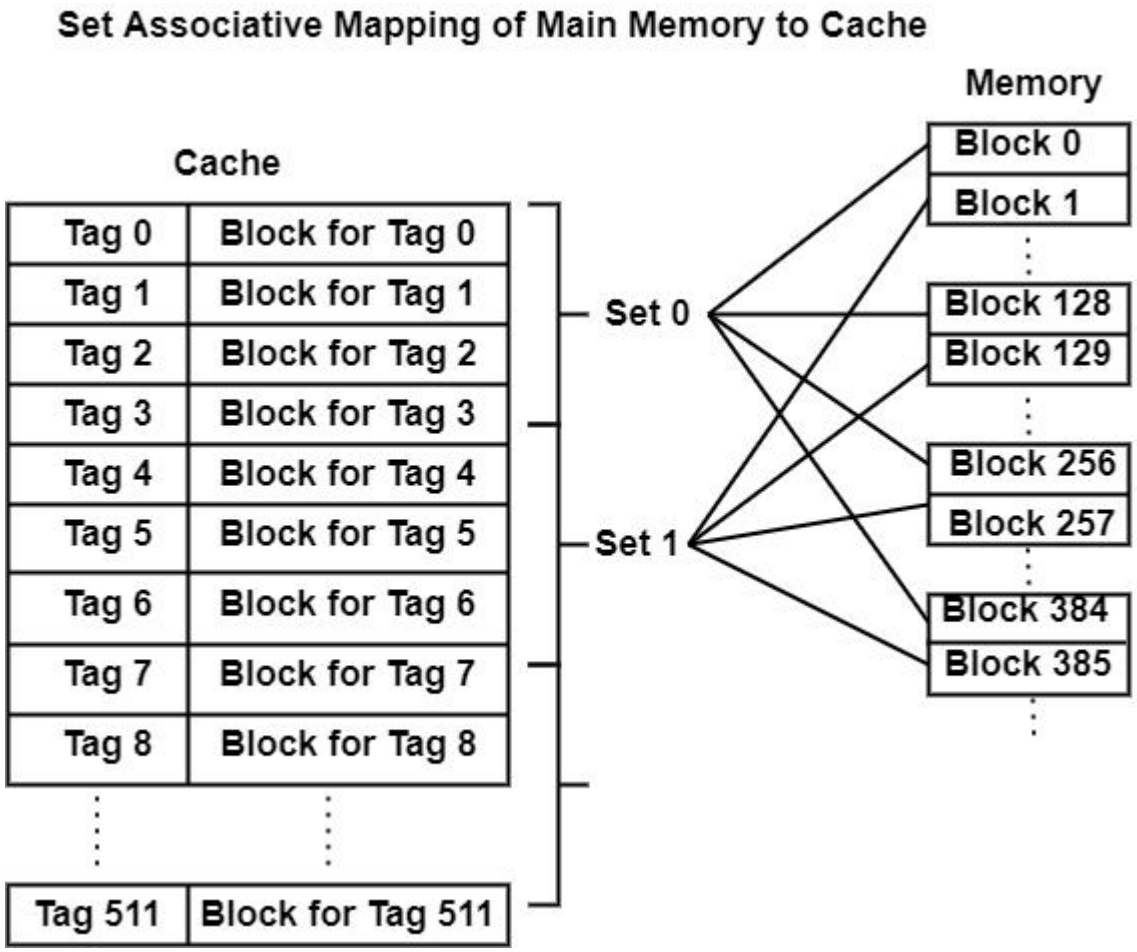
- All the lines of cache are freely available.
- Thus, any block of main memory can map to any line of the cache.
- Had all the cache lines been occupied, then one of the existing blocks will have to be replaced.



Set Associative Mapping –

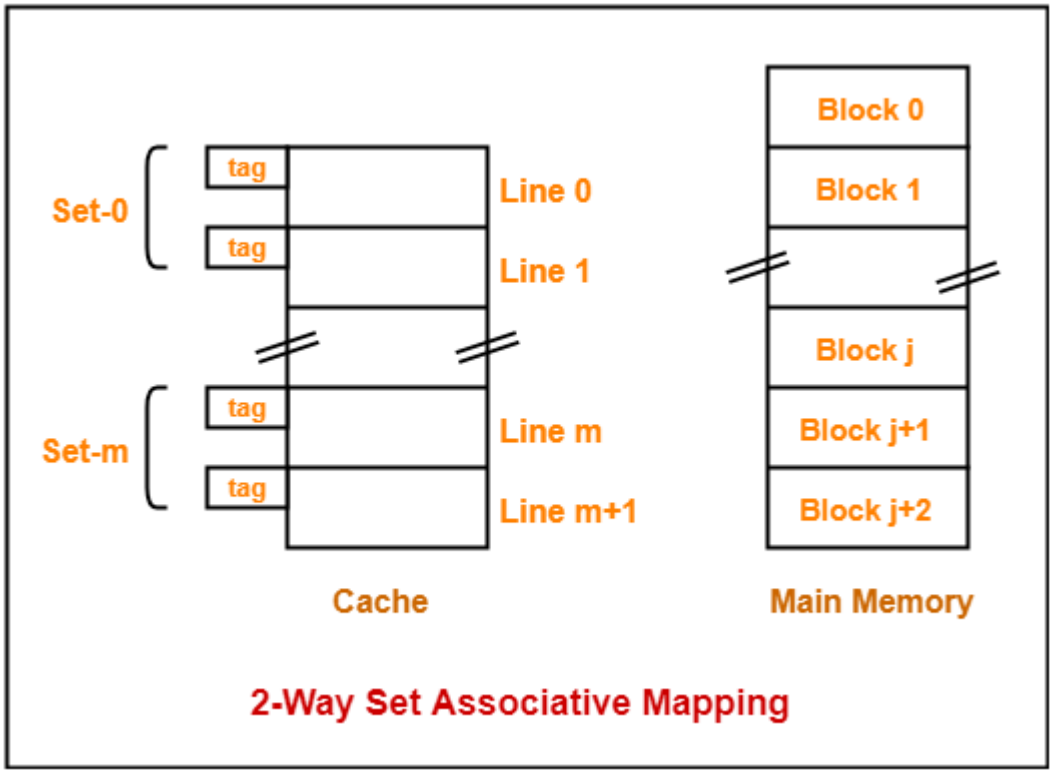
Set associative mapping combines direct mapping with fully associative mapping by arrangement lines of a cache into sets. The sets are persistent using a direct mapping scheme. It contains several groups of direct mapped blocks that operate as several direct mapped caches in parallel.

- Cache lines are grouped into sets where each set contains k number of lines.
- A particular block of main memory can map to only one particular set of the cache.
- However, within that set, the memory block can map any cache line that is freely available.
- The set of the cache to which a particular block of the main memory can map is given by-

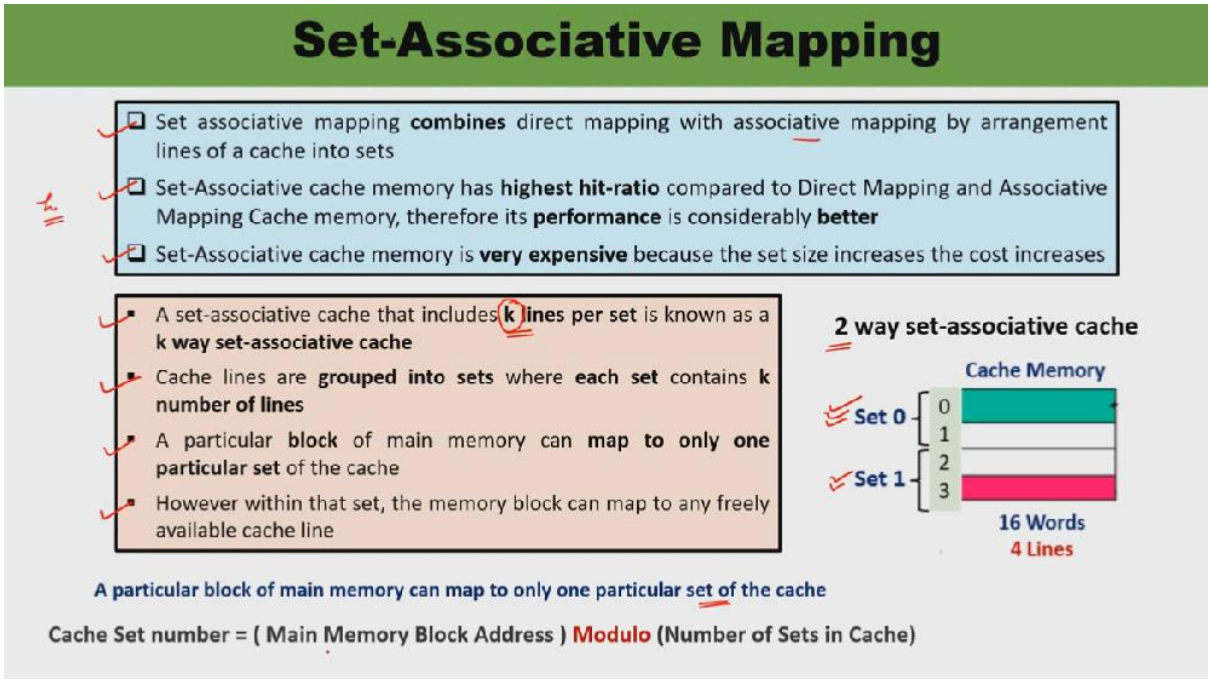


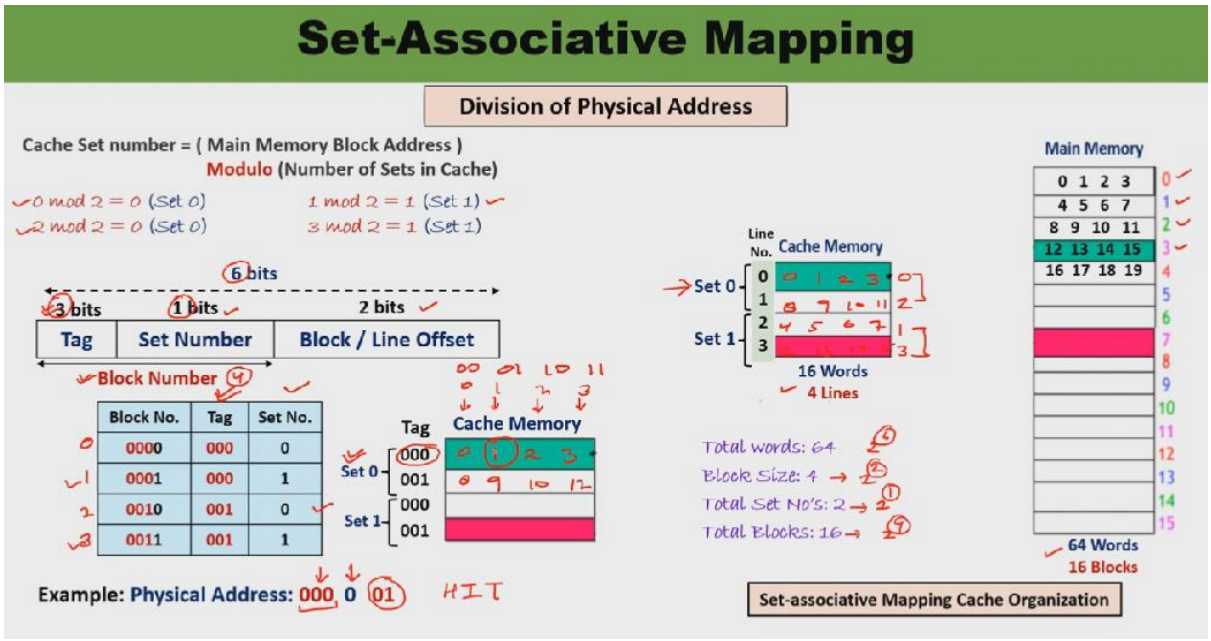
Cache set number

= (Main Memory Block Address) Modulo (Number of sets in Cache)



- k = 2 suggests that each set contains two cache lines.
- Since cache contains 6 lines, so number of sets in the cache = 6 / 2 = 3 sets.
- Block 'j' of main memory can map to set number (j mod 3) only of the cache.
- Within that set, block 'j' can map to any cache line that is freely available at that moment.
- If all the cache lines are occupied, then one of the existing blocks will have to be replaced.





In this type of cache, the following steps are used to access the data from a cache:

1. The index of the address from the processor is used to access the set.
2. Then the comparators are used to compare all tags of the selected set with the incoming tag.
3. If a match is found, the corresponding location is accessed.

4. If no match is found, an access is made to the main memory.

Problem-01:

Consider a 2-way set associative mapped cache of size 16 KB with block size 256 bytes. The size of main memory is 128 KB. Find-

1. Number of bits in tag
2. Tag directory size

Solution-

Given-

- Set size = 2
- Cache memory size = 16 KB
- Block size = Frame size = Line size = 256 bytes
- Main memory size = 128 KB

We consider that the memory is byte addressable.

Number of Bits in Physical Address-

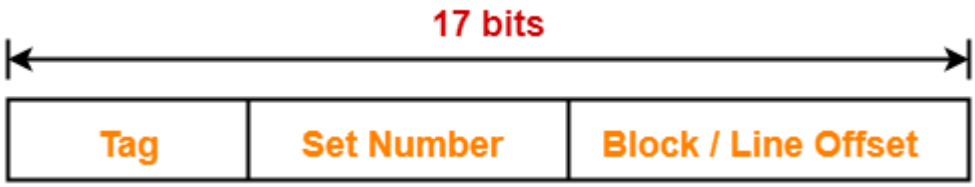
We have,

Size of main memory

= 128 KB

= 2^{17} bytes

Thus, Number of bits in physical address = 17 bits



Number of Bits in Block Offset-

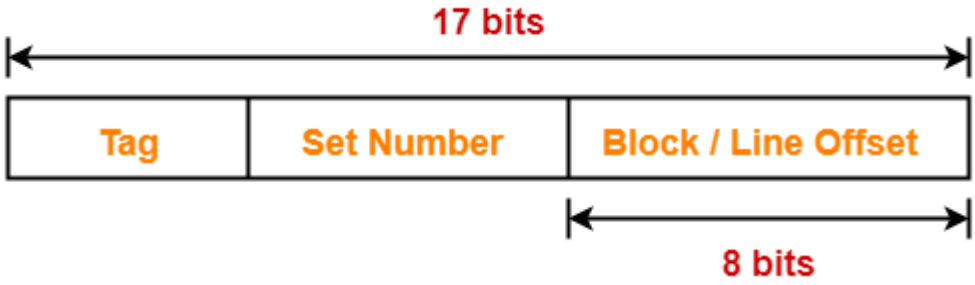
We have,

Block size

= 256 bytes

= 2^8 bytes

Thus, Number of bits in block offset = 8 bits



Number of Lines in Cache-

Total number of lines in cache

= Cache size / Line size

= 16 KB / 256 bytes

= 2^{14} bytes / 2^8 bytes

= 64 lines

Thus, Number of lines in cache = 64 lines

Number of Sets in Cache-

Total number of sets in cache

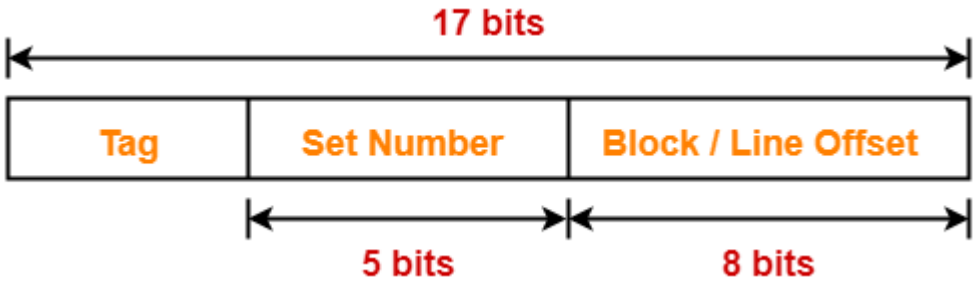
= Total number of lines in cache / Set size

= 64 / 2

= 32 sets

= 2^5 sets

Thus, Number of bits in set number = 5 bits



Number of Bits in Tag-

Number of bits in tag

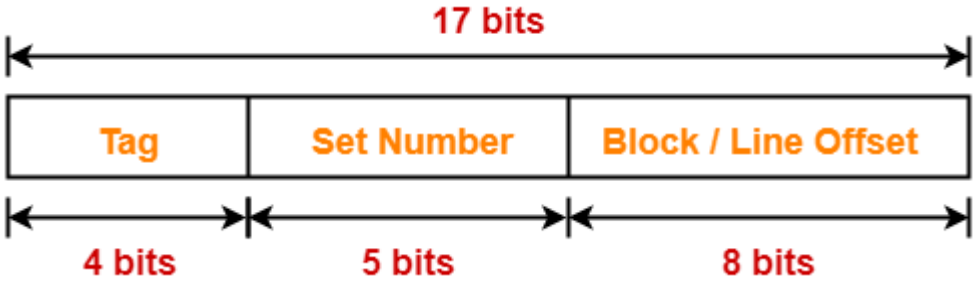
= Number of bits in physical address – (Number of bits in set number + Number of bits in block offset)

= 17 bits – (5 bits + 8 bits)

= 17 bits – 13 bits

= 4 bits

Thus, Number of bits in tag = 4 bits



Tag Directory Size-

Tag directory size

= Number of tags x Tag size

= Number of lines in cache x Number of bits in tag

= 64 x 4 bits

= 256 bits

= 32 bytes

Thus, size of tag directory = 32 bytes

Problem-02:

Consider a 8-way set associative mapped cache of size 512 KB with block size 1 KB. There are 7 bits in the tag. Find-

1. Size of main memory
2. Tag directory size

Solution-

Given-

- Set size = 8
- Cache memory size = 512 KB
- Block size = Frame size = Line size = 1 KB
- Number of bits in tag = 7 bits

We consider that the memory is byte addressable.

Number of Bits in Block Offset-

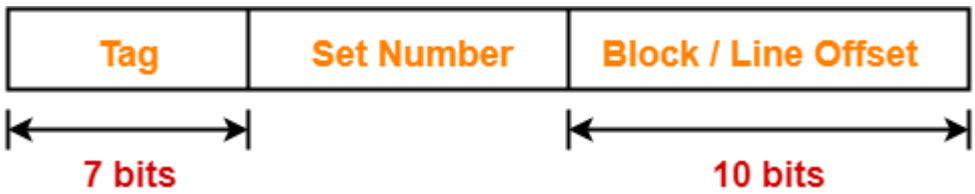
We have,

Block size

= 1 KB

= 2^{10} bytes

Thus, Number of bits in block offset = 10 bits



Number of Lines in Cache-

Total number of lines in cache

= Cache size / Line size

= 512 KB / 1 KB

= 512 lines

Thus, Number of lines in cache = 512 lines

Number of Sets in Cache-

Total number of sets in cache

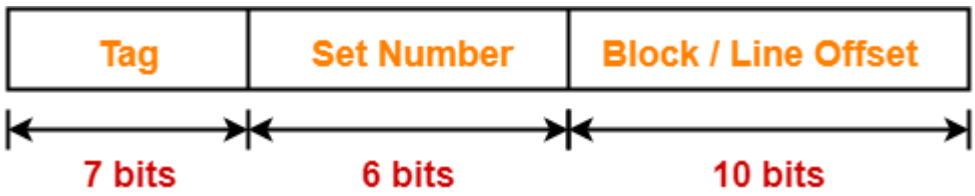
= Total number of lines in cache / Set size

= 512 / 8

= 64 sets

= 2^6 sets

Thus, Number of bits in set number = 6 bits



Number of Bits in Physical Address-

Number of bits in physical address

= Number of bits in tag + Number of bits in set number + Number of bits in block offset

= 7 bits + 6 bits + 10 bits

= 23 bits

Thus, Number of bits in physical address = 23 bits

Size of Main Memory-

We have,

Number of bits in physical address = 23 bits

Thus, Size of main memory

= 2^{23} bytes

= 8 MB

Tag Directory Size-

Tag directory size

= Number of tags x Tag size

= Number of lines in cache x Number of bits in tag

= 512 x 7 bits

= 3584 bits

= 448 bytes

Thus, size of tag directory = 448 bytes

Q1. If the cache size is 64KB and Block/line size is 8B then how many bits are required to represent lines of 4-way set associative cache memory?

Total lines = Cache size/ Block size = 64KB/8B = 213B
So 13 bits are require to represent lines in cache.
Total bits for set = Total lines / K-way = 213B/22 = 211
So 11 bits are require to represent sets in cache
Q2. If we found 10 bits for Set in 4-way set associative and block size is 16 kb. Then Cache size will be.
Cache size in K-set associative = total set * total lines per set * line size
Cache size = 210 *4* 214 B= 64 MB.

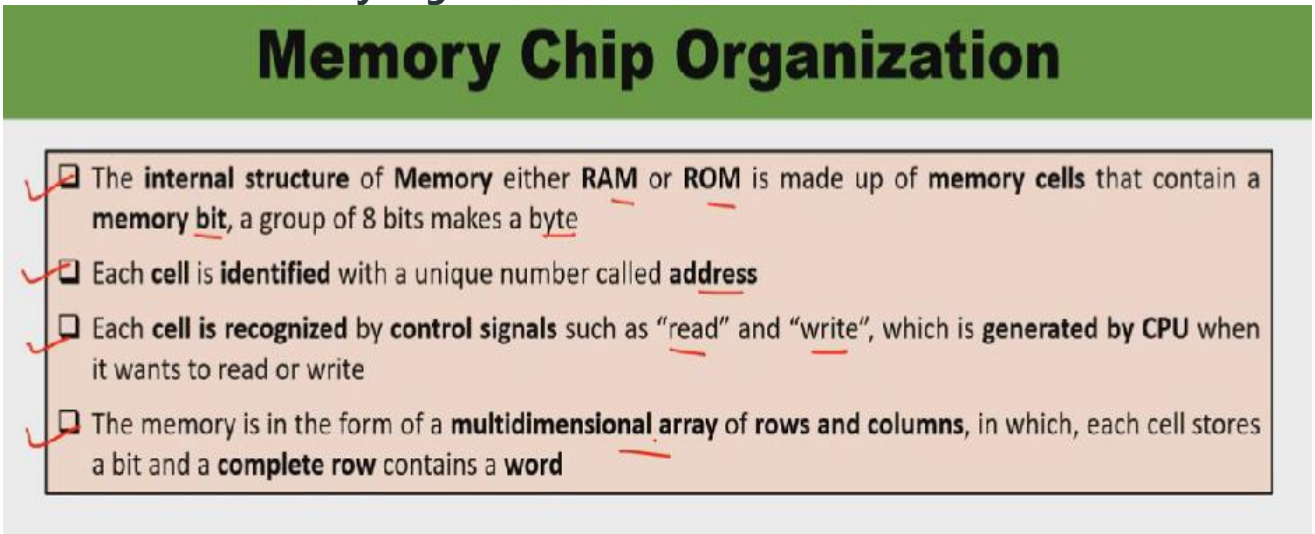
Advantages of Cache Memory

1. Cache memory is the faster memory as compared to the main memory.
2. It stores all data and instructions that are repeatedly used by the CPU for improving the performance of a computer.
3. The access time of data is less than the main memory.

Disadvantage of Cache Memory

1. It is very costly as compared to the Main memory and the Secondary memory.
2. It has limited storage capacity.

2D and 2.5D Memory organization



he **internal structure** of Memory either RAM or ROM is made up of memory cells that contain a memory bit. A group of 8 bits makes a byte.

The memory is in the form of a multidimensional array of rows and columns. In which, each cell stores a bit and a complete row contains a word. A memory simply can be divided into this below form

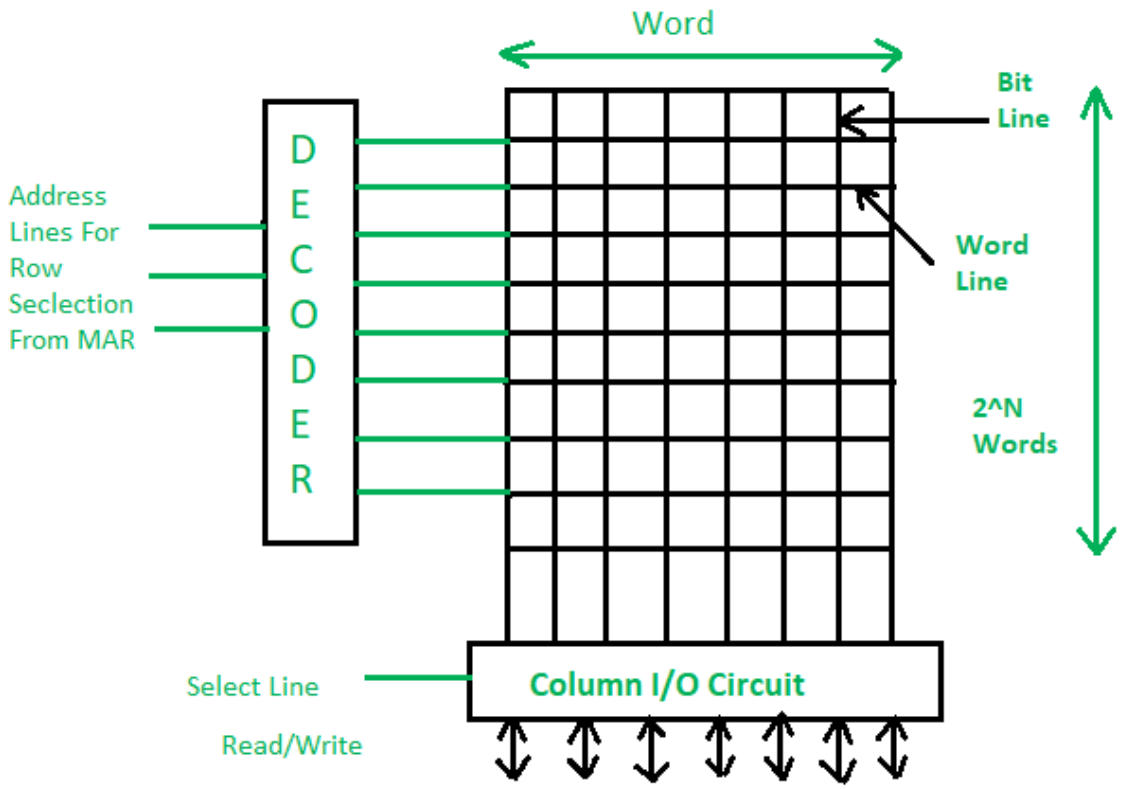
$$2^n = N$$

where **n** is the no. of address lines
and **N** is the total memory in bytes.
There will be 2^n words.

2D Memory organization –

1. Basically in 2D organization memory is divides in the form of rows and columns. (Matrix).
2. Each row contains a word now in this memory organization there is a decoder.
3. A decoder is a combinational circuit which contains n input lines and 2n output lines. One of the output line will select the row which address is contained in the MAR.

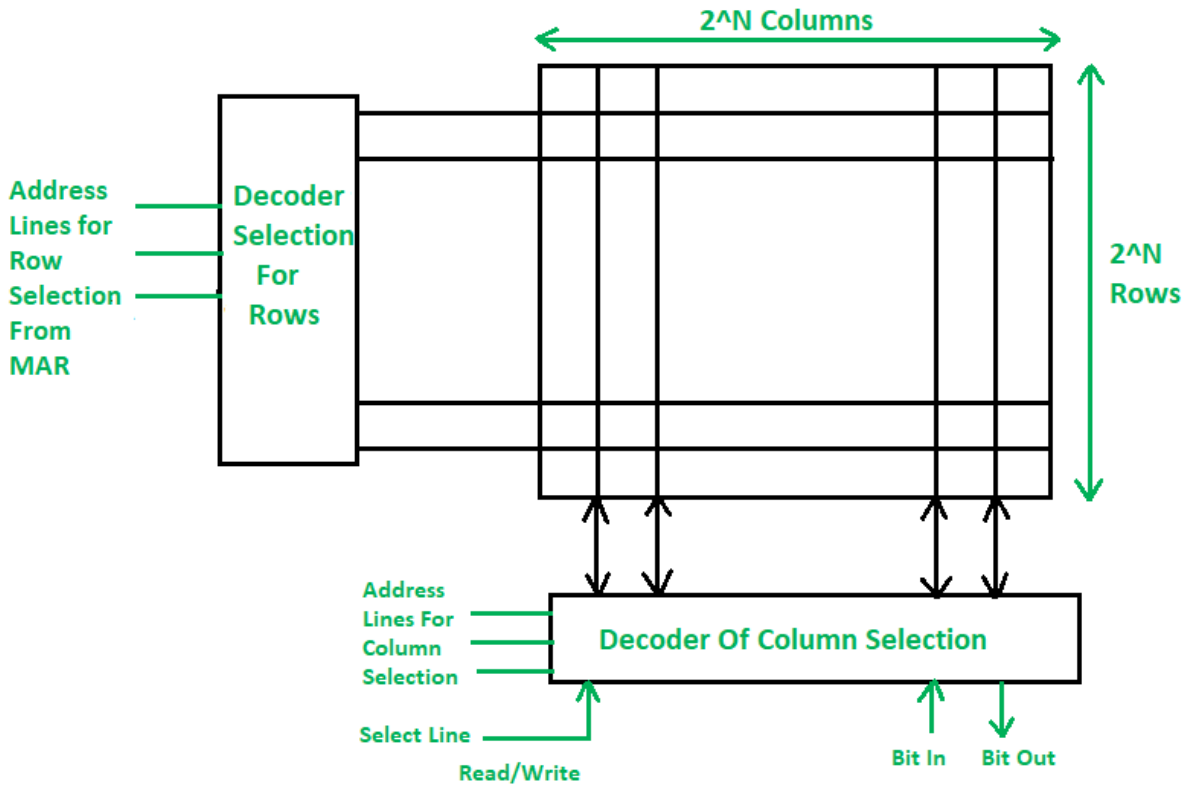
4. And the word which is represented by the row that will get selected and either read or write through the data lines.



2D Memory Organization

2.5D Memory organization –

1. In 2.5D Organization the scenario is the same but we have two different decoders one is column decoder and another is row decoder.
2. Column decoder used to select the column and row decoder is used to select the row. Address from the MAR will go in decoders' input. Decoders will select the respective cell.
3. Through the bit outline, the data from that location will be read or through the bit in line data will be written at that memory location.



2.5D Memory Organization

Read and Write Operations –

1. If the select line is in Reading mode then the Word/bit which is represented by the MAR will be available to the data lines and will get read.
2. If the select line is in write mode then the data from the memory data register (MDR) will be sent to the respective cell which is addressed by the memory address register (MAR).
3. With the help of the select line, we can select the desired data and we can perform read and write operations on it.

Comparison between 2D & 2.5D Organizations

2D Organization	2.5D Organization
Hardware is fixed	Hardware can be change
Requires more no. of gates	Requires less no. of gates
More complex	Less complex
Relatively difficult to implement	Easy to implement
Error correction is not possible	Error correction could be done easily
More difficult to fabricate	Relatively easy to fabricate
ROM Circuits mainly uses it	Currently most of RAM circuits uses it

Comparison between 2D & 2.5D Organizations –

1. In 2D organization hardware is fixed but in 2.5D hardware changes.
2. 2D Organization requires more gates while 2.5D requires less.
3. 2D is more complex in comparison to the 2.5D organization.
4. Error correction is not possible in the 2D organization but in 2.5D it could be done easily.
5. 2D is more difficult to fabricate in comparison to the 2.5D organization