- Memory Hierarchy
- Main Memory
- Cache Memory
- Virtual Memory
- Auxiliary Memory
- Associative Memory

Memory Hierarchy :-

Memory :- Memory is a hardware component of computer which stores the information temporarily or permanently. The size of information that can be stored depends on the number of bytes present in the memory.

- A memory unit is the collection of storage units or devices together.
- The memory unit stores the binary information in the form of bits.

- Memory / Storage is classified into two categories:

(i) volatile Memory

(ii) NON - volatile Memory.

(i) volatile Memory :- This loses it data, when power is switched off.

Ex:- RAM (Random Access Memory)

(ii) NON - volatile Memory :. This is a permanent storage and does not lose any data when power is switched off.

Memory Access Methods :-

(i) Random Access : In which each memory location has a unique address.

- Using this unique address any memory location can be reached in the same amount of time in any order.

Ex: Main Memories are random access memories.

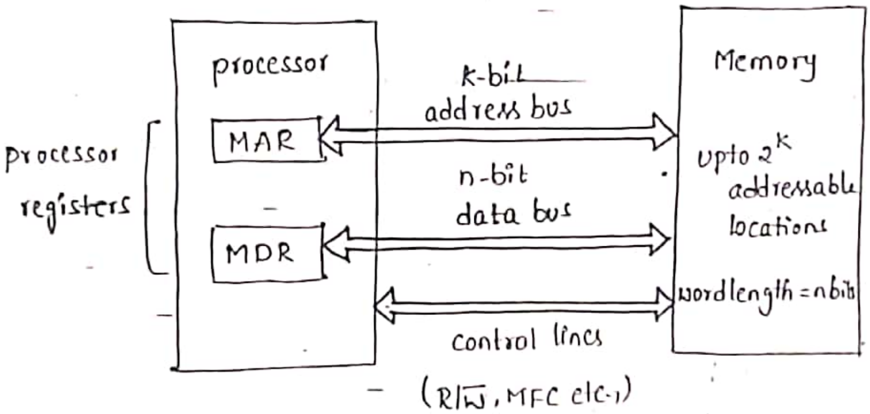
(ii) Sequential Access :-

This method allows memory access in a sequence or in order.

Ex:- Magnetic tape

(iii) Direct Access :- In this mode, information is stored in tracks, with each track having a separate read/write head.

Connection between the Memory and the processor



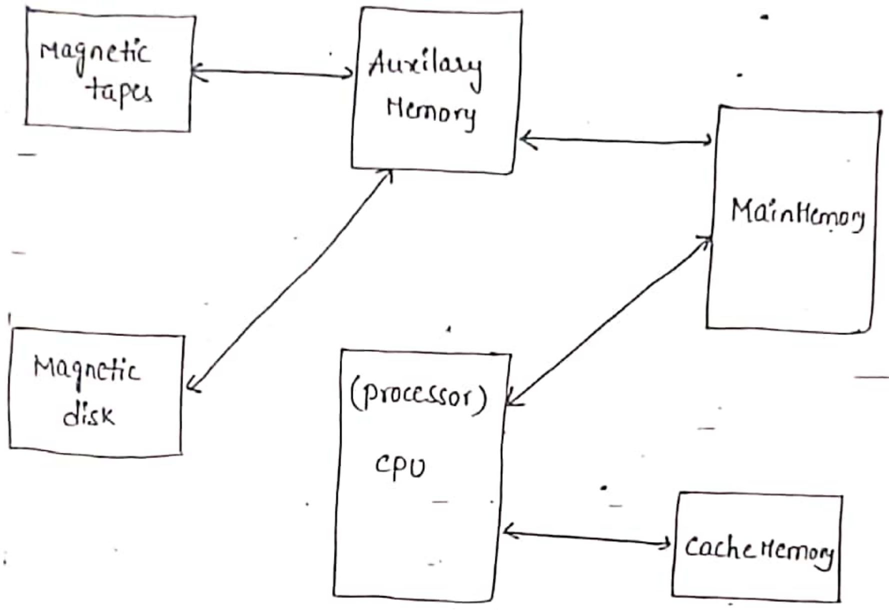
Writing Data to a Memory location :- To carry out the write operation, the processor sets the

- Sets the R/$\overline{W}$ line to '0'.
- places (load) the address of the memory location, where the data bus has to be written into MAR register.
- places (loads) the given data into MDR register.

Reading data from a Memory location :-

- Sets the R/$\overline{W}$ line to '1'.
- places the address of the memory location, where the required data is stored into MAR register.
- placing the required data onto the data bus.
- setting MFC signal to '1', to inform the processor that required data has been placed onto the bus.

Memory Hierarchy :-



- The Memory unit that directly communicates with the cpu is called the main memory.

- Devices that provide backup Storage are called auxilary memory. The most common auxiliary memory devices used in computer systems are magnetic disks and tapes.

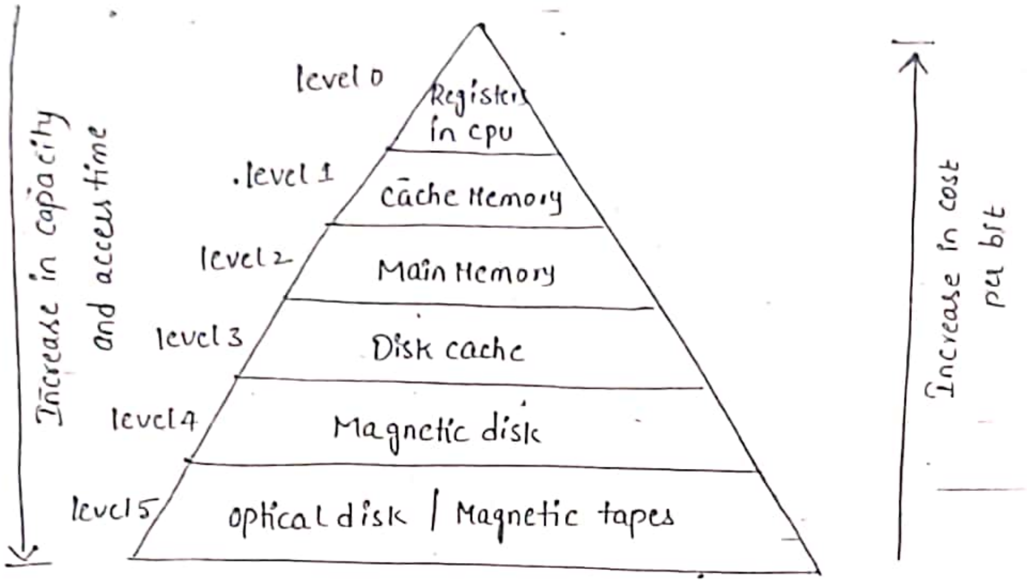
  - They are used for storing System programs, large data files, and other back up information.

- A special high-speed memory called a cache is Sometimes used to increase the speed of processing.

*- The typical access time ratio between cache and main memory is about 1 to 7.

Ex :- Typical cache memory may have an access time of 100ns, while main memory access time may be 700 ns. Auxiliary memory average access time is usually 1000 times that of main memory. Block Size in auxiliary memory typically ranges from 256 to 2048 words, while cache block size is typically from 1 to 16 words.

The comparisions in the Memory Hierarchy is



Main Memory :-

- It stores data and programs during computer operations.
- It uses semiconductor technology known as Semiconductor memory.
- It is a central storage memory unit.
- It communicates directly with the processor, Auxiliary memory, and cache memory.
- It is a fast and large memory used to store during operations.
- Main memory is divided into RAM and ROM.

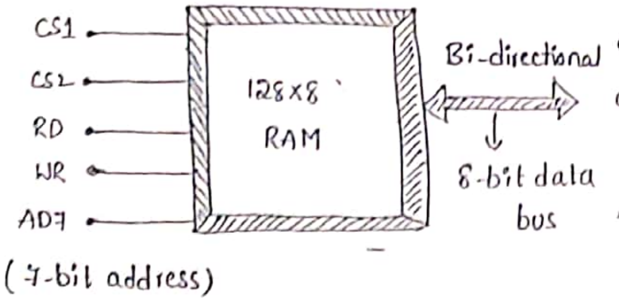
RAM

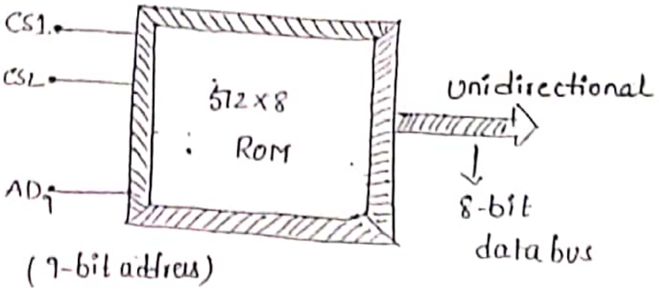
1. RAM stands for Random Access Memory.
2. It allows both read and write operations.
3. It is volatile in nature i.e, data are lost when power supply is switched off

ROM

1. ROM Stands for Read only Memory.
2. It allows only read operation.
3. It is non-volatile in nature and used for permanent storage.

4. It is used when I/o operation is performed, known as buffering.	4. It is not used for buffering purpose..
5. The second operation after booting, the computer is performed in RAM.	5. The first operation in computer system is performed in ROM (during booting process)
6. It is used to store data/Instructions while they are being processed, waiting to be processed and after being processed before it is provided to output components.	6. It is used to store program that are required for the operations of electronic devices.
7. It is usually expensive on per unit basis but while comparing on the basis of storage capacity RAM is cheaper.	7. It is usually cheaper in terms of per unit basis but while comparing on the basis of storage capacity expensive.
8. Types of RAM are (i) SRAM (ii) DRAM	8. Types of ROM are - PROM - EPROM - EE PROM - Flash memory.
9. RAM chip	9. ROM chip





Function table for RAM chip

CS1	CS2	RD	WR	Type of function	Status (databus)
0	0	X	X	undefined	High Impedence
0	1	X	X	Undefined	"
1	0	0	0	undefine d	"
1	0	0	1	WR	Write data into RAM
1	0	1	X	RD	Extract data from RAM
1	1	X	X	undefined	High Impedence

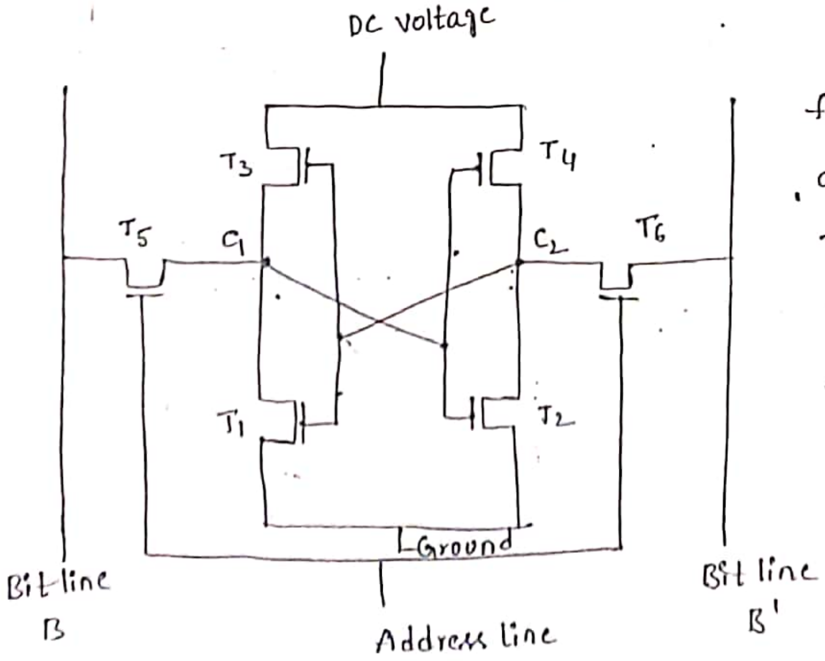
RAM types :-

Random Access Memory is divided into

(i) StaticRandom Access Memory (SRAM)

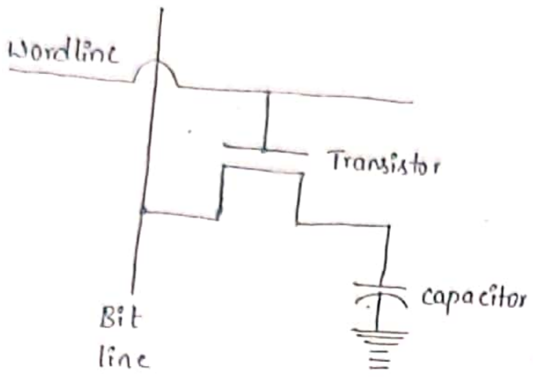
(ii) Dynamic Random Access Memory (DRAM)

SRAM :- SRAM is made up of CMOS technology and uses Six transistors. It's construction is comprised of two cross-coupled inverters to stored data (binary) Similar to flip-flops and extra two transistors for access control. SRAM Can hold the data as long as power is Supplied to it.



SRAM address line is operated for opening and closing the switch and to control the T_5 and T_6 Transistors, permitting to read and write. For read operation the Signal is applied to these address line then T_5 and T_6 gets on, and the bit value is read from . For the write operation, the signa is employed to B bit line, and it's complement is applied to B.

DRAM :-



→ To store the data in the cell, the charge must be stored in the ca This is done by turning the tran ON and applying an appropriate vo to the bit line.

→ When the data stored in the selec cell is to be read, it's transistor be turned ON, Then, the sense am which is connected to bit line, it provide full voltage to the bit lin

SRAM

1. SRAM is an on-chip memory whose access time is small.
2. SRAM is faster than DRAM.
3. SRAM is of Smaller Size.
4. SRAM is expensive
5. The cache memory is an application of SRAM.
6. SRAM is rarer
7. The construction of SRAM is complex due to the usage of a large no.of transistors.
8. In SRAM a single block of memory requires six transistors.

DRAM

1. while DRAM is an off-chip memory which has a large access time.
2. DRAM is little bit slower than SRAM
3. DRAM is available in larger storage capacity.
4. DRAM is cheap.
5. DRAM is used in Main Memory.
6. DRAM is highly dense.
   (It allows greater volumes of data to be stored in the same physical space
7. DRAM is simple to design and implment.
8. Where as DRAM needs just one transistor for a single block of memory.

Scanned with CamScanner

9. There is no issue of charge leakage in the SRAM.

9. DRAM is named as dynamic because it uses capacitor which produces leakage current due to the dielectric used inside the capacitor to separate the conductive plates is not a perfect insulator hence require power refresh circuitry.

10. Lower power consumption

10. High power consumption.

Comparision of SRAM & DRAM

Function	SRAM	DRAM
Speed	Faster	Slower
Size	Small	large
Cost	Expensive	cheap
Used in	cache Memory	Main Memory
Density	less dense	Highly Dense
construction	complex and uses transistors and latches	Simple and uses capacitors and very few transistors.
Single block of memory requires	6 Transistors	only one Transistor.

Scanned with CamScanner

ROM types :-

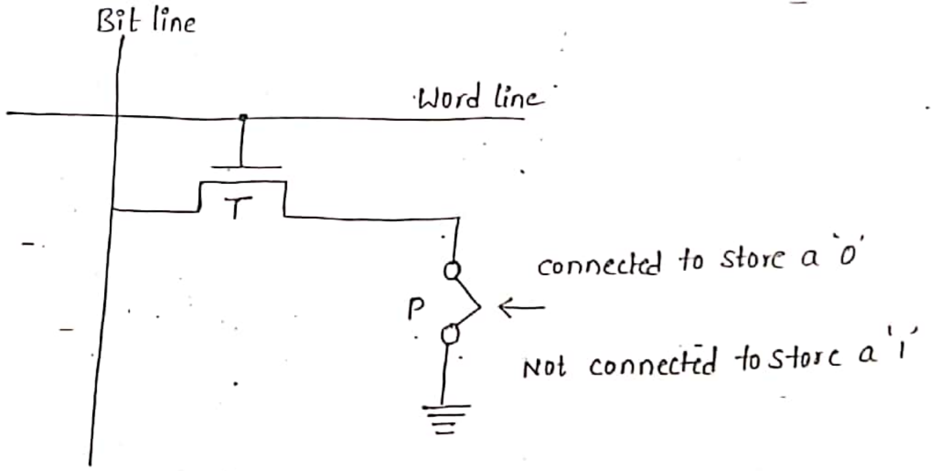
ROM :- • It is a non-volatile and is more like a permanent storage for information.

• It also stores the bootstrap loader program, to load and start operating system when computer is turned on.

The various types of ROM are

- PROM
- EPROM
- EEPROM
- Flash memory

ROM :- A ROM cell is depicted as,



- A logic value '0' is stored in the cell if the transistor is connected to ground at point P, otherwise a '1' is stored.
- The bit line is connected through a Transistor to the power supply.
- To read the state of the cell, the word line is activated.
- Thus the transistor switch is closed and the voltage on the bit line drops to near zero if there is a connection between the transistor and ground.

- If there is no connection to ground, the bit line remains at the high voltage, indicating a 1.
- A sense circuit at the end of the bit line generates the proper output value.
- Data are written into ROM when it is manufactured.

PROM:-

- It stands for programmable Read only memory.
- It was first developed in 70's by Texas Instuments.
- It is made as a blank memory.
- A PROM programmer or PROM burner is required in order to write data onto a PROM chip.
- The data stored in it cannot be modified and there-fore it is also known as one time programmable device.

EPROM :-

- It stands for Erasable Programmable ROM.
- It is different from PROM as unlike PROM the program can be written on it more than once.
- This comes as the solution to the problem faced by PROM.
- The bits of memory come back to 1, when Ultra Violet rays of some specific wavelength falls into its chips glass panel.
- The fuses are reconstituted and thus new things can be written on the memory.

EEPROM:-

- It stands for Electrically Erasable Read only memory.
- These are also erasable like EPROM, but the same work of erasing is performed with electric circuit.

- It stores computer System's BIOS. unlike, EPROM, the entire chip does not have to be erased for changing some portion of it.

Flash memory

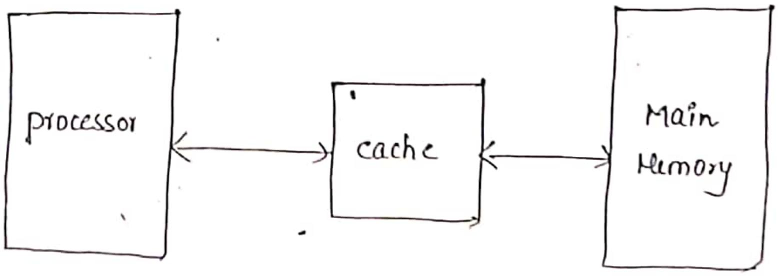
- It is an updated version of EEPROM.
- In EEPROM, it is not possible to alter many memory locations at the same time. However, Flash memory provides this advantage over the EEPROM by enabling this feature of altering many locations simultaneously.
- It was invented by Toshiba and got it's name from it capability of deleting a block of data in a flash.

Applications of flash Memory

- Digital cameras
- cell phones
- MP3 players

** CACHE MEMORY

- The cache is a small and very fast memory, interconnected between the processor and the main memory.
- It's purpose is to make the main memory appear to the processor to be much faster than it actually is.
- The effectiveness of this approach is based on a property of computer programs called locality of reference.



- When the processor issues a Read request, the contents of a block of memory words containing the location specified are transferred into the cache.

- Subsequently, when the program references any of the locations in this block, the desired contents are read directly from the cache.

- usually, the cache memory can store a reasonable number of blocks at any given time, but this number is small compared to the total number of blocks in the main memory.

- The correspondence between the main memory blocks and those in the cache is specified by a mapping function.

Hit Ratio :- The performance of cache memory is measured in terms of a quantity called hit ratio.

Hit :- when the cpu refers to memory and find the word in cache it is said to produce a hit.

Miss :- If the word is not found in cache, it is in main memory then it counts as a miss.

- The ratio of the number of ~~bits~~ hits to the total cpu references to memory is called hit ratio.

$$\boxed{\text{Hit Ratio} = \text{Hit} / (\text{Hit} + \text{Miss})}$$

MAPPING FUNCTIONS:-

- There are several possible methods for determining where memory blocks are placed in the cache.

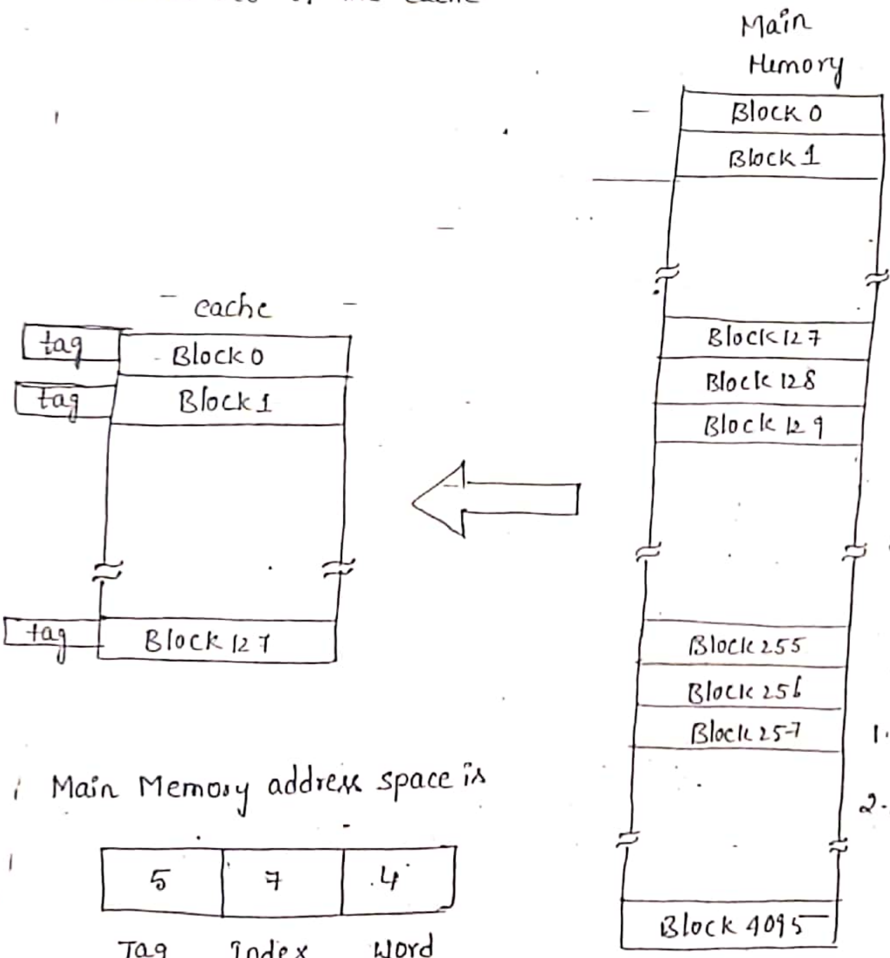
- consider a cache consisting of 128 blocks of 16 words each, for a total of 2048 (2K) => 2^{10} words, and assume that main memo is addressable by a 16-bit address.

- The main memory has 64K words, which we will view as 4K blocks of 16 words each.

(i) Direct Mapping :-

- The Simplest way to determine cache locations in which to store memory blocks is the Direct-Mapping technique.

- In this technique, block j of the main memory maps onto block j modulo 128 of the cache.



Advantages

1. It is very simple to implement

2. It can access any block address directly

Disadvantages

1. It is not effective

2. It causes contention of data.

Main Memory address space is

5	7	4
Tag	Index	Word

- Whenever one of the main memory blocks 0, 128, 256 --- is loaded into the cache, it is stored in cache block '0'.

- Blocks 1, 129, 257 ---- are stored in cache block 1 and so on.

- The memory address can be divided into three fields.

- The low-order 4 bits select one of 16 words in a block.

- When a new block enters the cache, the 7-bit cache block field determines the cache position in which this block must be stored.

- The highorder 5 bits of memory address ~~of~~ the block are stored in 5 tag bits associated with it's location in the cache.

- Tha tag bits identify which of the 32 main memory blocks mapped into this cache position is currently resident in the cache.

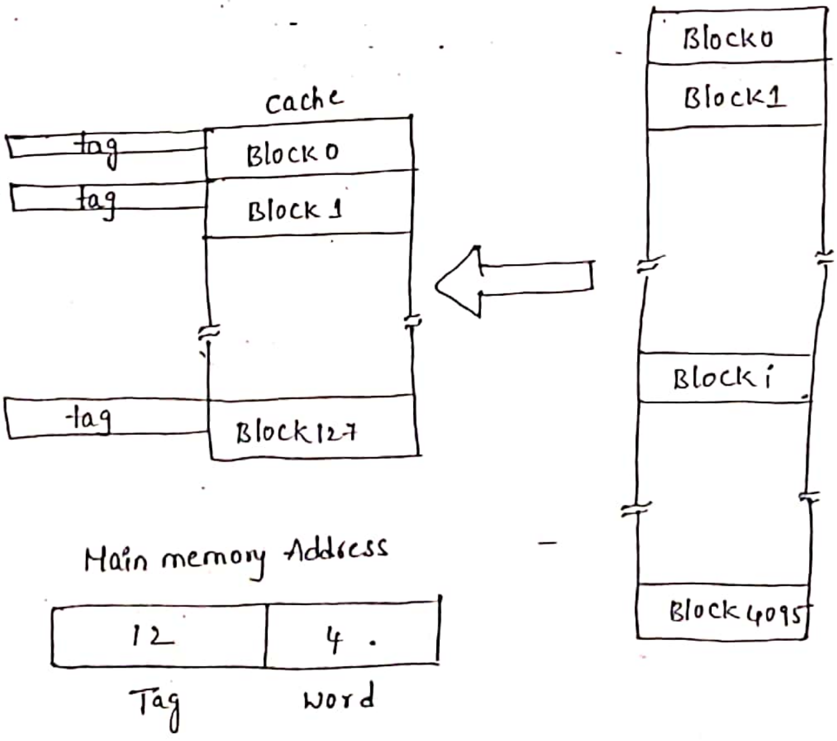
- As execution proceeds, the 7-bit cache block field of each address generated by the processor points to a particular block location in the cache.

- The highorder 5 bits of the address are compared with the tag bits associated with that cache location. If they match, then the desired word is in that block of the cache.

- If there is no match, then the block containing the required word must first be read from the mainmemory and loaded into the cache.

(ii) Associative Mapping :-

- The most flexible mapping method, in which a main memory block can be placed into any cache block position.



As

1. This method is introduced to overcome the disadvantage of direct mapping.

2. In this method cache allows any line mapped to Main memory word line.

Disadvantage :- find out, whether a particular block is in cache, all cache lines would have to be examined.

* Set-associative :-

1. This technique is introduced to overcome the disadvantage of Associative and direct mapping.

2. In this, cache is divided into a no.of sets, each set containing equal no.of cache lines.

3. Each block in Main memory maps into one set in cache memory is Similar to that of direct mapping.

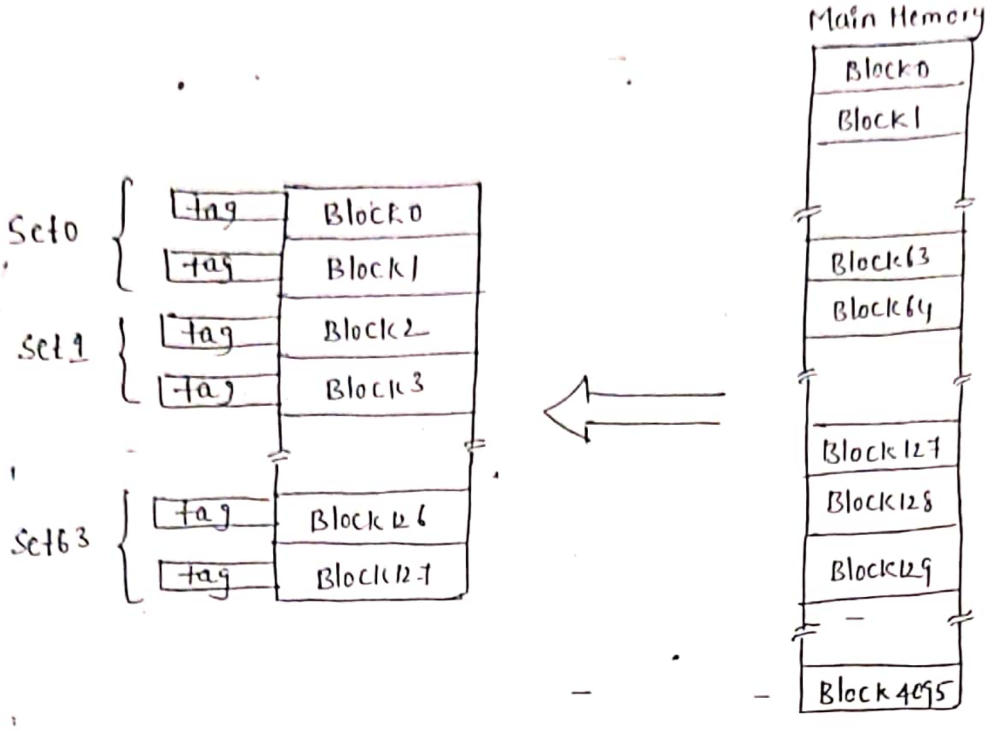
4. With in the set, the main memory block can occupy any line is Similar to Associative mapping.

In this technique Main Memory Address space will be divided into

← 16 bits →

Tag	set	Word
6	6	4

Main Memory Address space



2-way set-associative Mapping

Advantages of cache Memory

1. It is faster than the memory

2. It takes less time to Execute

3. It stores the data less period of time till the Execution complete.

4. It does not require System bus to transfer the data.

Locality of Reference:- in any computer program, Some instructions are of primary focus, which gets repeatedly executed all the time and rest of the instructions are Executed rarely. This behaviour of computer programs are called as "Locality of Reference".

Write through:- In write-through cache, whenever a write is made to cache, immediately the write is also made to main memory.

Write Back:- In write-back cache, when the written block is about to be replaced, then the cache block is written entirely to main memory

*** Virtual Memory

→ Virtual memory is a separation of user logical Memory from physical memory.

→ Virtual Memory which appear to be present actually it is not.

→ It is just gives illusion to user.

→ A programmer can write a program Which requires more memory than the capacity of Main memory, Such programs is Executed by virtual Memory technique.

→ logical address space is always greater than physical address space.

→ It is a non-contiguous memory allocation.

→ It does not require more I/O to load and store.

Virtual memory can be implemented via

(i) paging technique

(ii) Demand paging technique.

(i) **Paging :-**

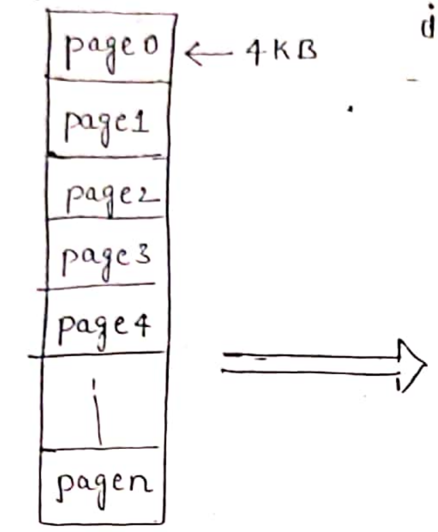
1. It is a NON-contiguous allocation.
2. operating system giving illusion to the user such that user can write a very big Program, user thinks that entire program is present in RAM, All the space allocated to the user is contiguous. But reality only small portion of program stored in RAM. which may or may not be contiguous, remain will be stored in Secondary Memory.
3. A computer can add more memory than the amount of Physically installed on system. This Extra memory is called "Virtual memory".

Paging technique can be done in 3 steps.

Step1 :- Before adding logical Address space with physical Address space, logical address space is divided into equal size of pages.

Default page size = 4KB

logical Add.space



Virtual Memory

ii) All pages are stored into page table along with page offset.

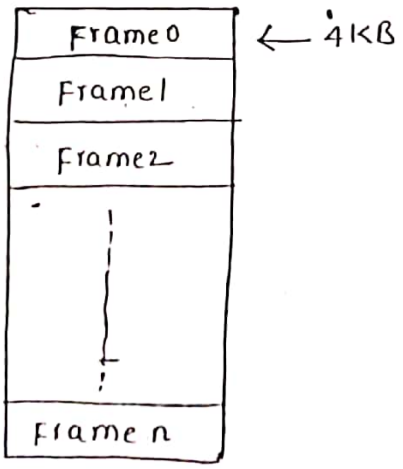
pageno	pageoffset
page0	4001
page1	4002
⋮	⋮
pagen	5040

Page map table

Step2 :- The Size of process can be measured in no. of pages Similarly Main memory divided in small fixed blocks (physically) called frames.

Size of the frame = Size of page

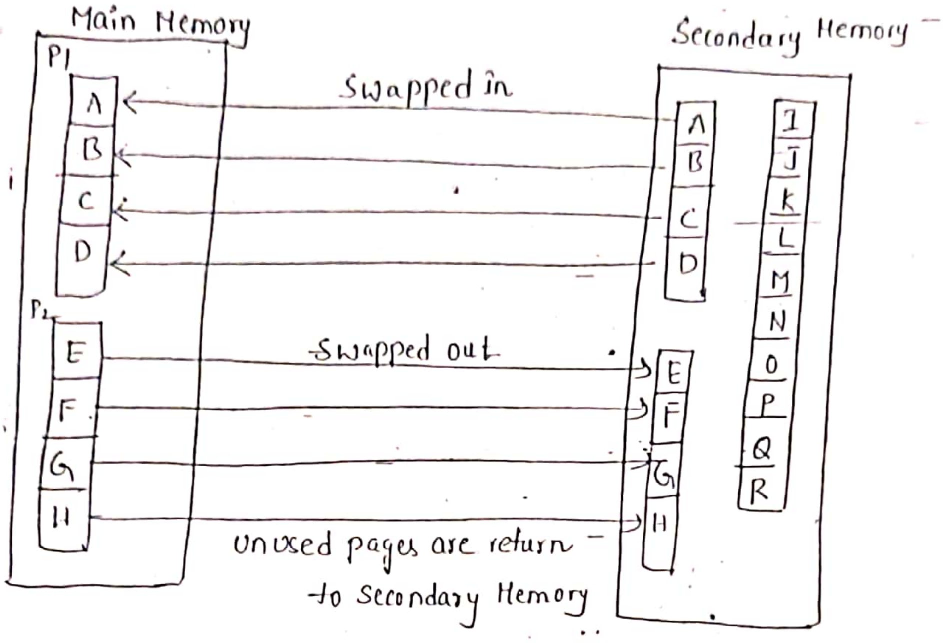
physical Add-space



Main Memory

Demand paging :-

- While Executing which pages are not used in Main Memory those are swapped into Secondary Memory and take the new data.



Suppose while Executing A, B, C, D this process demands one more page i.e., E but it is not in Main memory is called "page fault".

Page fault :- Which page is referenced by the Main Memory, and that page is not available in Main Memory is called "page faults".

To overcome the page faults cpu uses "demand paging" to reload the missing pages.

Demand paging :- The process resides in secondary memory and pages loaded only on demand not in advance is called demand paging.

Disadvantage :- The only major issue with Demand paging is after reload the new page, the process starts Execution from the begining. It is not a big issue for small programs, but for larger programs it affects performance drastically.

i.e., It causes "Thrashing problem".

Steps :- Address Translation

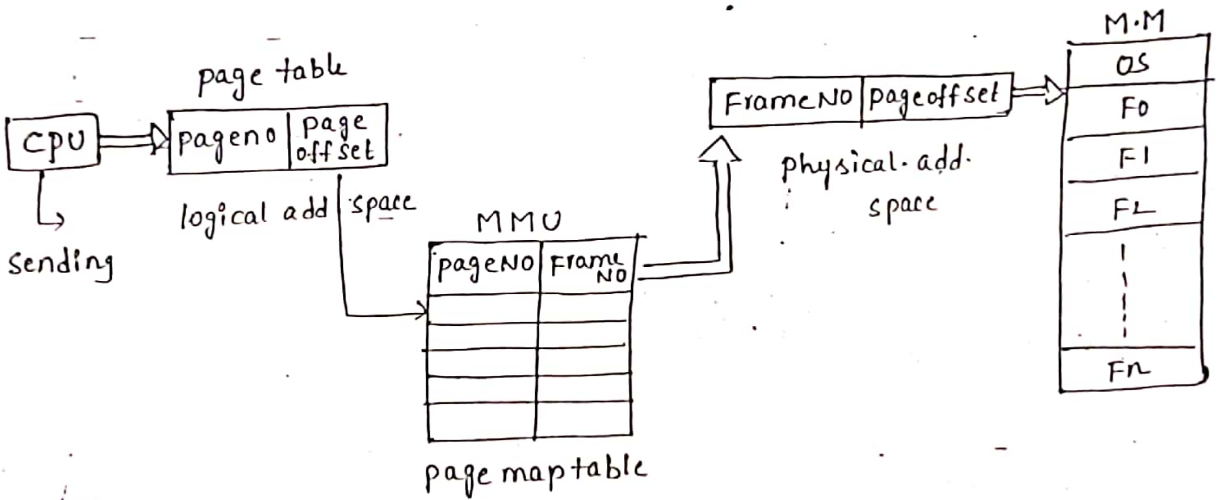
- page address is logical address represented by

  logical address space = page no + page offset.

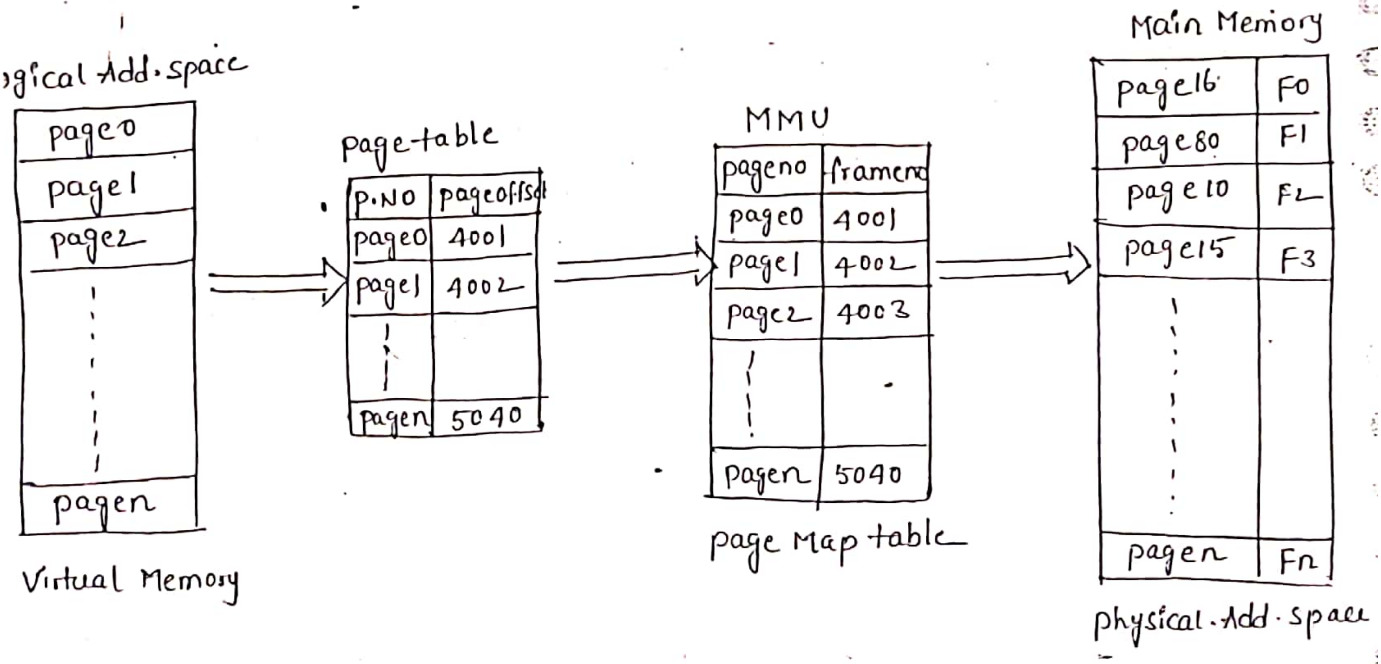
- frame address is called physical address & represented by

  physical address space = frame no + page off set.

- Translating Address (logical address space to physical Add. space) cpu used one special H/w component is called MMU (Memory Management unit is located at cpu.



paging



page0
page1
page2
⋮
pagen

P.NO	pageoffset
page0	4001
page1	4002
⋮	
pagen	5040

pageno	frameno
page0	4001
page1	4002
page2	4003
⋮	
pagen	5040

page16	F0
page80	F1
page10	F2
page15	F3
⋮	
pagen	Fn

Thrashing :- A process that is spending more time on paging than Executing is said to be "Thrashing".

To avoid this os uses one more technique is called "Anticipatory paging"

Anticipatory paging :-

This implies that os guesses the pages that can be referenced and fetches from the secondary memory to put onto the MainMemory even before they need to be referenced by the main memory.

Advantages of Virtual Memory :-

1. We can run more applications at once.
2. logical address space is much larger than physical address space.
3. Need to allow pages to be swapped in and swapped out
4. We need to maintain entire image of process in disk storage.
5. Reduce internal fragmentation and eliminates External fragmentation.

Fragmentation :- means wastage of memory

divided into two types.

(1) External fragmentation
(2) Internal fragmentation

External fragmentation :- External fragmentation is the breaking up of free memory into small chunks via partitioning, which can mean a request for a larger partition later may fail due to lack of contiguous memory, even though enough total memory is available.

Internal Fragmentation :- Internal fragmentation means allocating a larger than necessary partition (because of fixed-size partitions), leading to wasted space.

Disadvantages :-

(1) Using virtual memory that increased overhead for handling paging interrupts, Software complexity and hardware costs.

(2) Applications run slower.

(3) It takes more time to switch between applications.

(4) possibility of Thrashing due to excessive paging and page faults.

SECONDARY MEMORY

1. The Secondary Memory is also called the Secondary storage device or Auxiliary Memory.

2. Before data is stored in secondary memory must be fetched into RAM before processing done by the cpu.

3. Magnetic disk, Magnetic tapes, and optical disk are the different types of Storage devices.

4. Access type of Storage devices

The information stored in Storage devices can be in two ways.

1. Sequential Access
2. Direct Access.

Sequential access :- means that computer must run through the data in sequence, Starting from the begining inorder to locate a particular piece of data.

EX: Magnetic tape

let us suppose that Magnetic tape has 80 records to access 25th record the computer starts from first record then reaches second, third etc until it reaches 25th record. generally they are slow devices.

Direct Acces :- Data can be retrieved in a non-sequential manner (Random order) by locating It's using data's address.

Ex: Magnetic disk & optical disks are example.

Magnetic tape :-

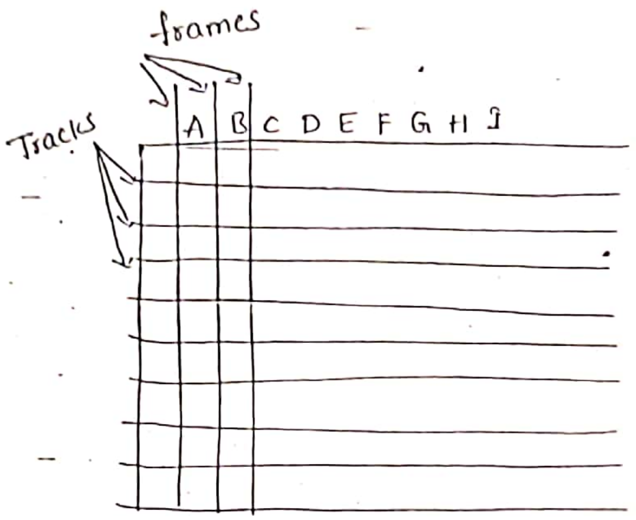
→ It is a plastic tape with magnetic coating. It is a storage medium on a large open reel or in a smaller cartridge or cassette.

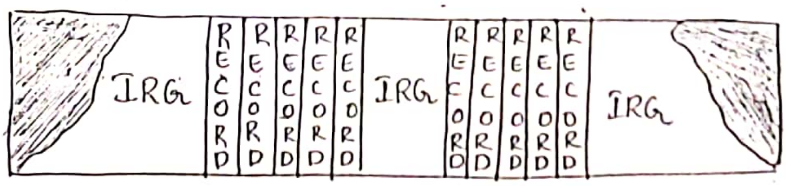
→ Magnetic tape are cheaper storage media. They are durable.

→ Magnetic tapes are Sequential access devices, Due to their Sequential nature, magnetic tapes are, not Suitable for data file that need to be revised or updated often.

→ They are generally used to store back-up data that is not frequently to transfer data from one system to other.

Working of Magnetic tape :- -





→ Magnetic tape is divided horizontally into tracks (7 or 9) and into frames. A frame stores one byte of data and a track in a frame can store 1 bit of data is stored in Successive frames as a string with one byte per frame.

→ Data is recorded on tape in the form of blocks, Where a block consists of a group of data also called as records.

→ Each block is read continuously, There is an Inter-Record Gap (IRG between two blocks that provides time for the tape to be stopped and started b/w records.

→ The basic magnetic tape mechanism consists of the supply reel, take-up reel, and read/write head assembly. The magnetic tape moves on tape drive from the supply reel to take-up reel with it's magnetic coated side passing over the read/write head.

→ The Magnetic tape width is 1/4 inch, 1/2 inch etc.,

→ Storage capacity a 10-inch diameter reel of tape which is 2400 feet long can store up to 180 million characterstics.

Features of Magnetic tape are

1. Inexpensive storage devices.
2. can store a large amount of data.
3. easy to carry or transport.
4. Not suitable for random access data.
5. Slow. access devices.
6. Needs dust prevention, a dust can harm the tape.
7. Suitable for backup storage.

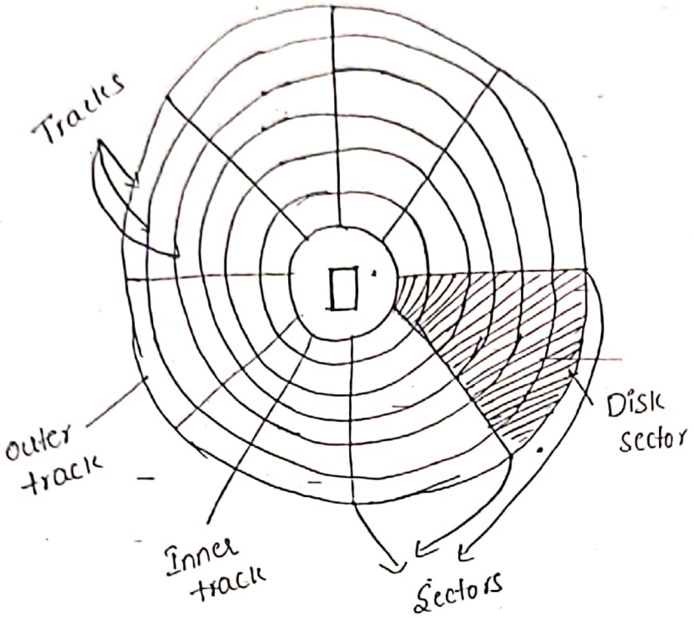
Magnetic disk :-

→ Magnetic disk is a direct access secondary storage device. It is a thin plastic or metallic circular plate coated with magnetic oxide

→ Data is stored on Magnetic disk as magnetized spots. The presence of magnetic spot represents the bit 1 and it's absence the bit 0.

Working:- The surface of disk is divided into concentric circles known as tracks. The outermost track is numbered '0' and innermost track is the last frame. Tracks are further divided into Sectors. A sector is a pie slice that cuts accross all tracks.

→ The data on disk is stored in sector. Sector is the smallest unit that can be read or written on disk. A disk has 8 or more sectors per track.



Seek time :- The time taken to move the read/write head to the desired track is called the seek time.

Latency time :- The time taken for desired sector of the track to come under read/write head is called latency time.

Data Transfer Rate :- The rate at which data is written to disk or read from disk is called data Transfer rate.

The sum of seek time, latency time and Data transfer time is the access time of disk.

Features :-

1. cheap storage device
2. can store large amount of data.
3. Easy to carry or transport.
4. Suitable for frequently read/write data
5. fast access device.
6. More reliable storage device.
7. To be prevented from dust, as the read/write head moves properly.

8. Floppy disk, hard disk and zip disk are the diffaent types of magnetic disk's.

Floppy disk :-

→ Floppy disk is a flat, round, single disk made of Mylar plastic and enclosed in square plastic Jacket.

→ Floppy disk has a write- protect slide tab that prevents a user from writing to it.

→ They are portable. They can be removed from the disk drive, carried or stored separately.

→ They are small and inexpensive.

→ They are slower to access than Hard disk, less storage capacity.

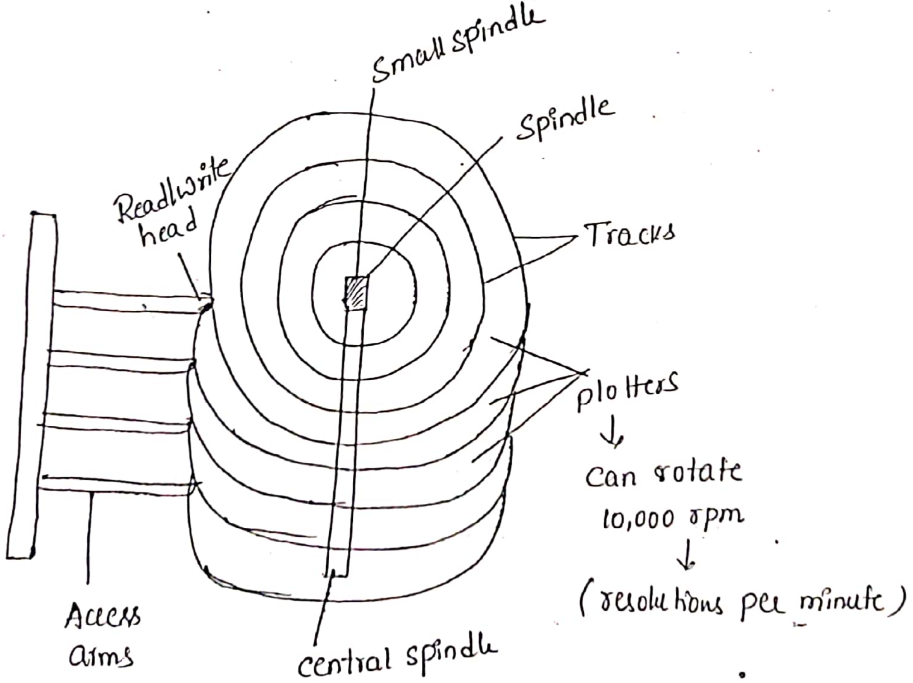
→ their basic size is 5¼ inch and 3½ inch.

→ 5½ inch disk came arround 1987. It can store 360 KB to 1.2MB.

→ 3½ inch disk " " " 400 KB to 1.44 MB.

→ It usually contains 40 tracks and 18 sectors per track it can store 512 bytes per sector.

Hard disk :-



1. It contains one or more plotters divided into concentric tracks and sectors.

2. It is mounted on a central spindle, like a stack.

3. central spindle allows platter to rotate at high speed.

4. Small spindle allows read-write arm to across plotter.

5. Magnetic plotter stores information in binary form.

6. It can be read by read/write head that pivots across the rotating disk.

7. The data is stored on the plotters covered with magnetic coating.

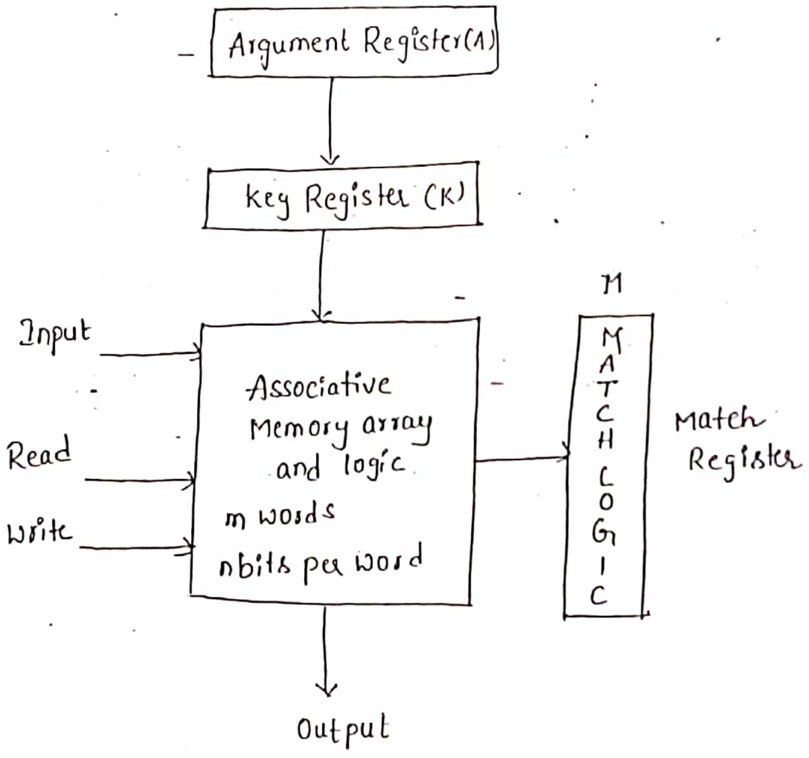
zip disk :-

1. They are high-capacity removable disk and drive.
2. They have the speed and capacity of harddisk and portability of floppy disk.
3. zip disks are the Same Size as floppy disk. i.e., 3½ inch but have a much higher capacity than the floppy disk.
4. They can be used to store large files, audio and video data.
5. It is easy to destroy it by bare hand.
6. It is expensive

Associative Memory :-

- A Memory unit accessed by content is called an associative memory or content Addressable Memory (CAM).
- It is also called as parallel search memory or multi-access memory.
- It is a search procedure and algorithm.
- Stored data can be identified for access by content of data rather than address.
- used in very critical and short search time applications.

Hardware Implementation :-



- Memory Array – m words with n bits per word.
- Each word in memory is compared in parallel with content of argument register. It has n bits
- Key register – provides mask for choosing a particular field or key in argument word.
- Every argument word is compared with memory word if key register contains 1's.

Write operation :-

1. When a word is written in associative memory no address is given.

2. The memory is capable of finding an unused location to store the word.

Read operation:-

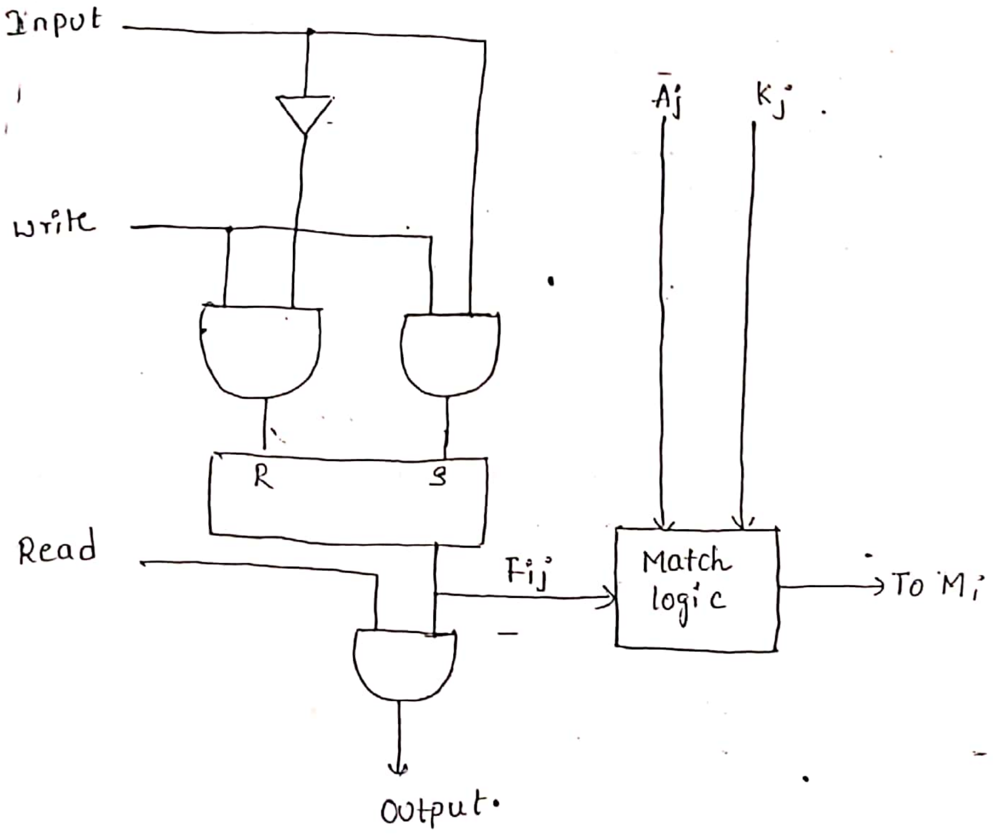
1. When a word is to be read from an associative memory, the content of the word, or a part of the word is specified.

2. The memory locates all words which match the specified content and mark them for reading.

Associative Memory array :- It contains the words which are to be compared with the argument word.

Match Register (M) :-

It has m bits, one bit corresponding to each word in the memory array. After the matching process, the bits corresponding to matching words in match register are set to 1.



One cell of associative Memory

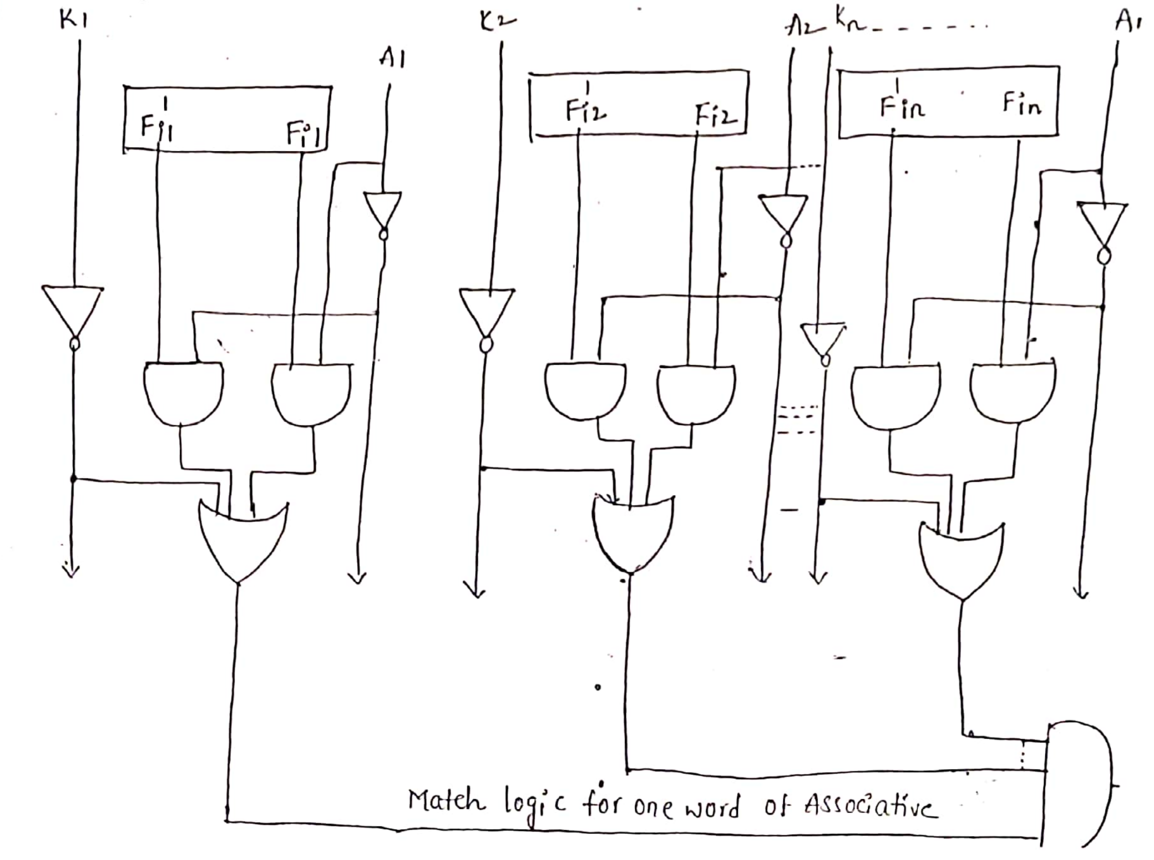
- Key register provides the mask for choosing the particular field in A register.
- The entire content of A register is compared if Key register content all 1.
- If the compared data is matched corresponding bits in the match register are set.

Ex:			
	A	101 111100	
	k	111 000000	
	Word1	100 111100	no match
	Word2	101 000001	match

Match logic :-

- let us include key register. If $K_j = 0$ then there is no need to compare A_j and F_{ij}.
- only when $K_j = 1$, comparison is needed.
- This is achieved by ORing each term with K_j.

$$M_i = (x_1 + K_1')(x_2 + K_2')(x_3 + K_3') \dots\dots (x_n + K_n')$$



Match logic for one word of Associative

Advantages of Associative Memory :-

1. This is suitable for parallel searches. It is also used where Search time needs to be short.

2. Associative Memory is often used to speed up databases, in neural networks and in the page tables used by the virtual memory of modern computers.

3. CAM-design challenge is to reduce power consumption associated with the large amount of parallel active circuitry, without sacrificing speed or memory density.

Disadvantages :-

1. An associative memory is more expensive than a random access memory because each cell must have an extra storage capability as well as logic circuits for matching.

2. usually associative memories are used in applications where the Search time is very critical and must be very Short.