UNIT-IV: Memory

Table of Content

1) Memory: Basic concept and hierarchy
2) Semiconductor memories (Primary memories)
3) 2D and 2.5D memory organization
4) Cache memories: concept and design issues and performance
5) Address Mapping
6) Auxiliary memories: magnetic disk, magnetic tape and optical disk
7) Virtual memory

→ Memory

A memory is just like a human brain. It is used to store data and instructions.

Computer memory is the storage space in the computer, where data is to be processed and instructions required for processing are stored. A memory unit is the collection of storage units or devices together. The memory unit stores the binary information in the form of bits.

The memory is divided into large number of small parts called **cells**. Each location or cell has a unique address, which varies from zero to memory size minus one. **For example**, if the computer has 64k words, then this memory unit has 64 * 1024 = 65536 memory locations. The address of these locations varies from 0 to 65535.

Memory is the functional unit which stores and receives the instructions and the data being executed

MAR holds the address of the next instruction to be executed, or a piece of data that will be used. IR holds the instruction itself (What to do?); and MDR holds the data (with what?).

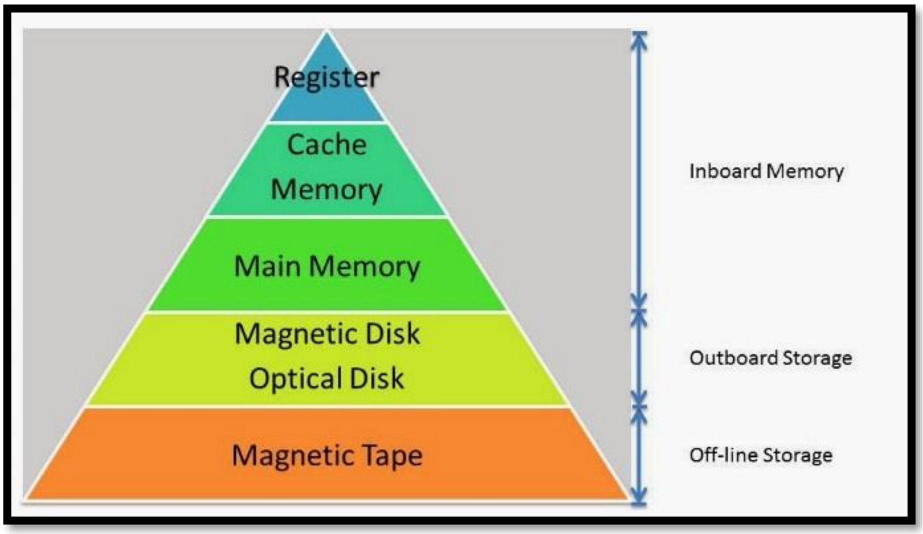
→ Memory Hierarchy

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

Typically, a memory unit can be classified into two categories:

1. 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).
2. 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.

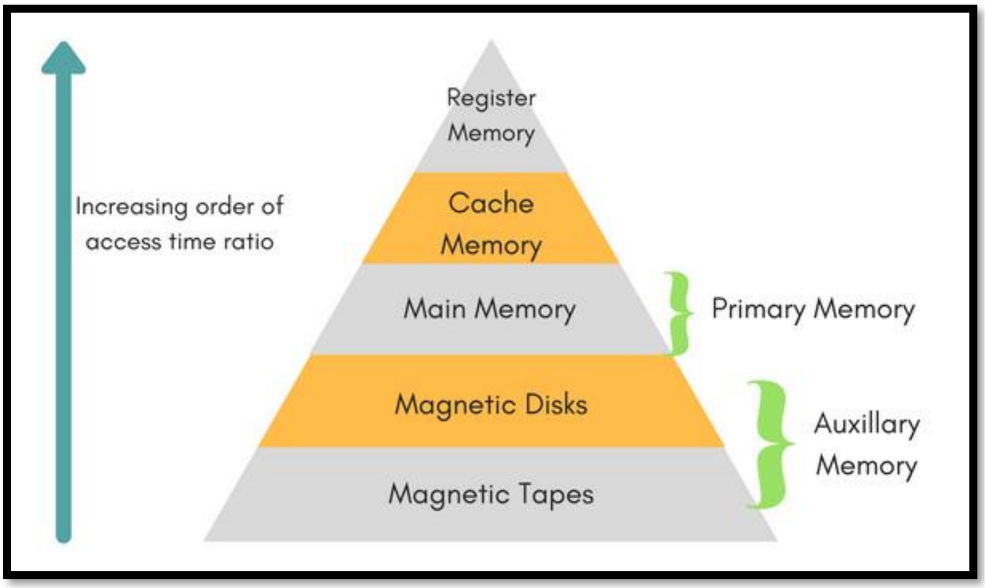
Apart from the basic classifications of a memory unit, the memory hierarchy consists all of the storage devices available in a computer system ranging from the slow but high-capacity auxiliary memory to relatively faster main memory.



The total memory capacity of a computer can be visualized by hierarchy of components. The memory hierarchy system consists of all storage devices contained in a computer system from the slow Auxiliary Memory to fast Main Memory and to smaller Cache memory.

Auxiliary memory access time is generally **1000 times** that of the main memory, hence it is at the bottom of the hierarchy.

The **main memory** occupies the central position because it is equipped to communicate directly with the CPU and with auxiliary memory devices through Input/output processor (I/O).



When the program not residing in main memory is needed by the CPU, they are brought in from auxiliary memory. Programs not currently needed in main memory are transferred into auxiliary memory to provide space in main memory for other programs that are currently in use.

The **cache memory** is used to store program data which is currently being executed in the CPU. Approximate access time ratio between cache memory and main memory is about **1 to 7~10**

→Memory Access Methods

Each memory type, is a collection of numerous memory locations. To access data from any memory, first it must be located and then the data is read from the memory location. Following are the methods to access information from memory locations:

1. **Random Access**: Main memories are random access memories, 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.

2. **Sequential Access**: This methods allows memory access in a sequence or in order.

3. **Direct Access**: In this mode, information is stored in tracks, with each track having a separate read/write head.

→Primary Memory (Main Memory)

Primary memory holds only those data and instructions on which the computer is currently working. It has a limited capacity and data is lost when power is switched off. It is generally made up of semiconductor device. These memories are not as fast as registers. The data and instruction required to be processed resides in the main memory. It is divided into two subcategories RAM and ROM. These are known as **semiconductor memories.**

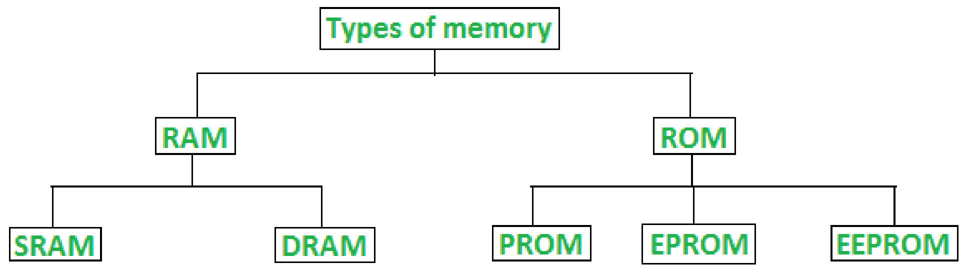
Characteristics of Main Memory

- These are semiconductor memories.
- It is known as the main memory.
- Usually volatile memory.
- Data is lost in case power is switched off.
- It is the working memory of the computer.
- Faster than secondary memories.
- A computer cannot run without the primary memory

RAM

- All information, both instruction and data, stored in memory are represented internally as binary numbers, i.e., the two bits part.
- A RAM(Random Access Technique)based computer memory has the following three characteristics:-
- It is divided into fixed-size units called cells, each of which is uniquely identified by using an address. The size of a cell used to be 8 bits, and has been going up to the current 64.

- The cell is the minimum unit of access, i.e., you cannot get a smaller amount out of the memory, in other words, data is atomic.
- The time it takes to access information is the same for all the cells. It is this feature that gives RAM its name.



Classification of computer memory

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.

Difference between SRAM and DRAM

ROM

ROM: Read Only Memory, is non-volatile and is more like a permanent storage for information. It also stores the **bootstrap loader** program, to load and start the operating system when computer is turned on. **PROM**(Programmable ROM), **EPROM**(Erasable PROM) and **EEPROM**(Electrically Erasable PROM) are some commonly used ROMs.

Read-Only Memory (ROM)

- Stores crucial information essential to operate the system, like the program essential to boot the computer.
- It is non-volatile.
- Always retains its data.
- Used in embedded systems or where the programming needs no change.
- Used in calculators and peripheral devices.
- ROM is further classified into four types- MROM, PROM, EPROM, and EEPROM.

Types of Read-Only Memory (ROM)

- **PROM (Programmable read-only memory)** – It can be programmed by the user. Once programmed, the data and instructions in it cannot be changed.
- **EPROM (Erasable Programmable read-only memory)** – It can be reprogrammed. To erase data from it, expose it to ultraviolet light. To reprogram it, erase all the previous data.
- **EEPROM (Electrically erasable programmable read-only memory)** – The data can be erased by applying an electric field, with no need for ultraviolet light. We can erase only portions of the chip.
- **MROM(Mask ROM)** – Mask ROM is a kind of read-only memory, that is masked off at the time of production. Like other types of ROM, mask ROM cannot enable the user to change the data stored in it. If it can, the process would be difficult or slow.

RAM	ROM
1. Temporary Storage.	1. Permanent storage.
2. Store data in MBs.	2. Store data in GBs.
3. Volatile.	3. Non-volatile.
4.Used in normal operations.	4. Used for startup process of computer.
5. Writing data is faster.	5. Writing data is slower.

Difference between RAM and ROM

Memory unit	Description
Kilo Byte	1 KB = 1024 Bytes
Mega Byte	1 MB = 1024 KB
Giga Byte	1 GB = 1024 MB
Tera Byte	1 TB = 1024 GB
Peta Byte	1 PB = 1024 TB
Hexa Byte	1 EB = 1024 PB
Zetta Byte	1 ZB = 1024 EB
Yotta Byte	1 YB =1024 ZB
Bronto Byte	1 Bronto Byte = 1024 YB
Geop Byte	1 Geo Byte = 1024 Bronto Bytes

Q- How many 32 k x 1 RAM chips needed to provide a memory capacity of 256 k-bytes?

RAM chip's capacity = 32 K × 1 = 32 ×1024 × 1 bits

Memory Capacity = 256 K byres = 256 ×1024 × 8 bits

Formula:

$$\text{No of chips needed} = \frac{Memory\ Capacity}{RAM\ Chip's\ Capacity}$$

Calculation:

$$\text{No of chips needed} = \frac{256\times1024\times8}{32\times1024\times1} = 64\ chips$$

→2D and 2.5D memory organization

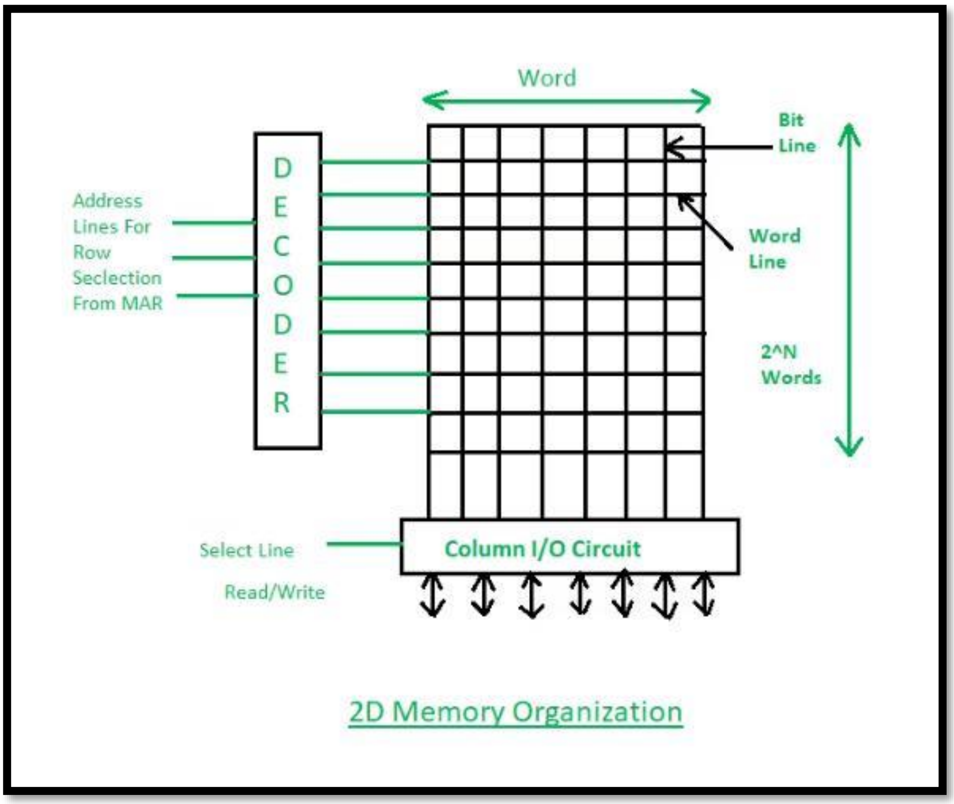
The internal structure of Memory either RAM or ROM is made up of memory cells that contain a memory bit. A group of 8 bits makes a byte. The memory is in the form of a multidimensional array of rows and columns. In which, each cell stores a bit and a complete row contains a word. A memory simply can 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.

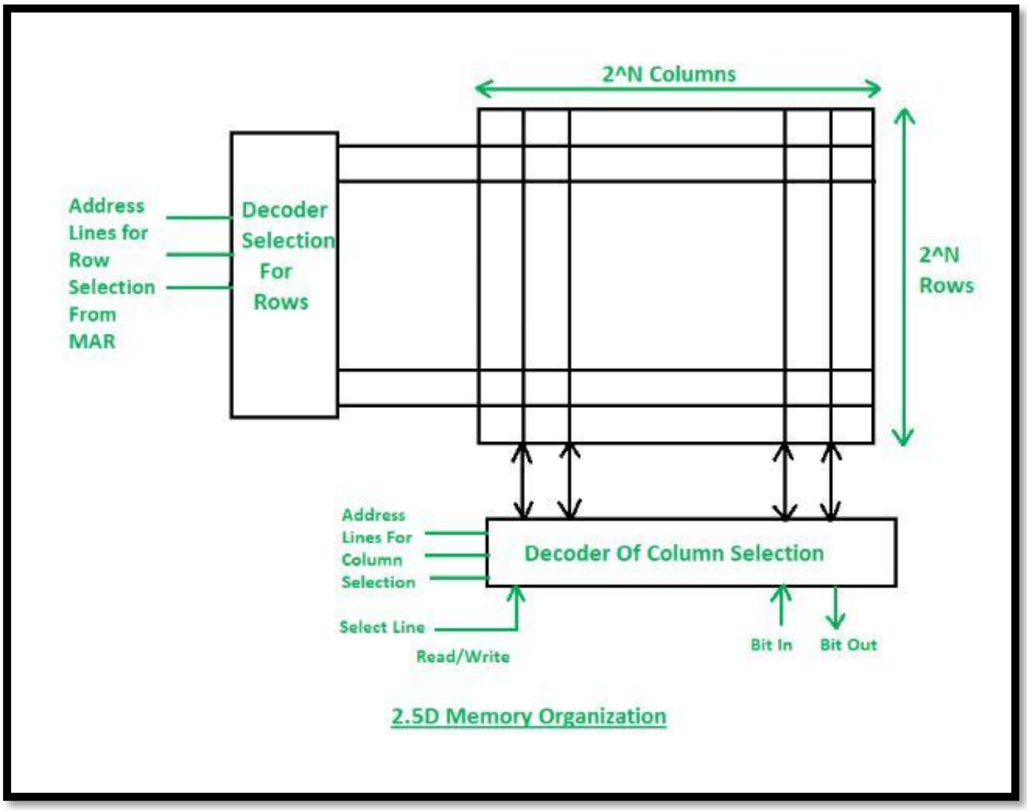
2D Memory organization –

In 2D organization, memory is divided in the form of rows an columns(Matrix). Each row contains a word, now in this memory organization, there is a decoder. A decoder is a combinational circuit that contains n input lines and 2n output lines. One of the output lines selects the row by the address contained in the MAR and the word which is represented by that row gets selected and is either read or written through the data lines.



2.5D Memory organization –

In 2.5D Organization the scenario is the same but we have two different decoders one is a column decoder and another is a row decoder. Column decoder is used to select the column and a row decoder is used to select the row. The address from the MAR goes as the decoders' input. Decoders will select the respective cell through the bit outline, then the data from that location will be read or through the bit, inline data will be written at that memory location.



Read and Write Operations –

- 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.
- 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).
- 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 –

1. In 2D organization hardware is fixed but in 2.5D hardware changes.
2. 2D Organization requires more no. of Gates while 2.5D requires less no. of Gates.
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 error correction is easy.
5. 2D is more difficult to fabricate in comparison to the 2.5D organization.

→Cache Memory

Cache memory is a very high speed semiconductor memory which can speed up the CPU. It acts as a buffer between the CPU and the main memory. It is used to hold those parts of data and program which are most frequently used by the CPU. The parts of data and programs are transferred from the disk to cache memory by the operating system, from where the CPU can access them.

Hit Ratio

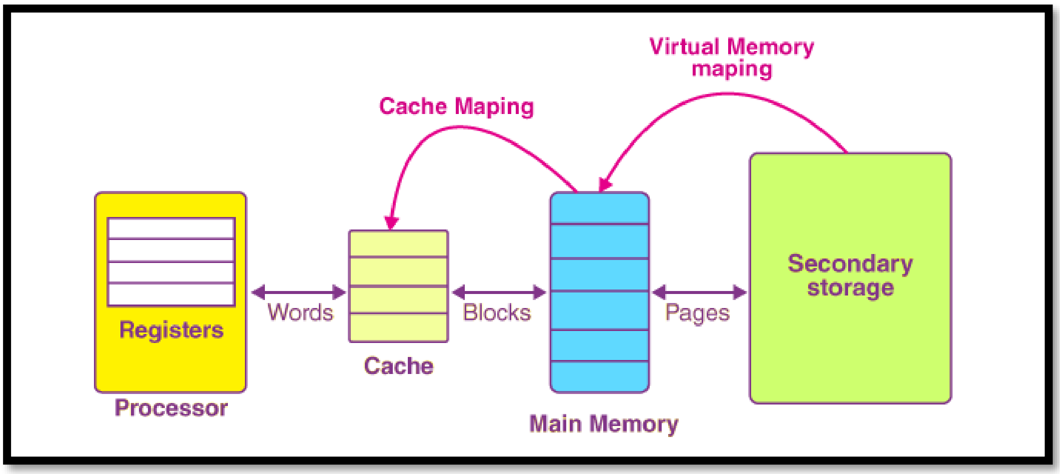
The performance of cache memory is measured in terms of a quantity called **hit ratio**. When the CPU refers to memory and finds the word in cache it is said to produce a **hit**. If the word is not found in cache, it is in main memory then it counts as a **miss**. The ratio of the number of hits to the total CPU references to memory is called hit ratio.

Hit Ratio = Hit / (Hit + Miss)

✓ Process of Cache Mapping

The process of cache mapping helps us define how a certain block that is present in the main memory gets mapped to the memory of a cache in the case of any cache miss.

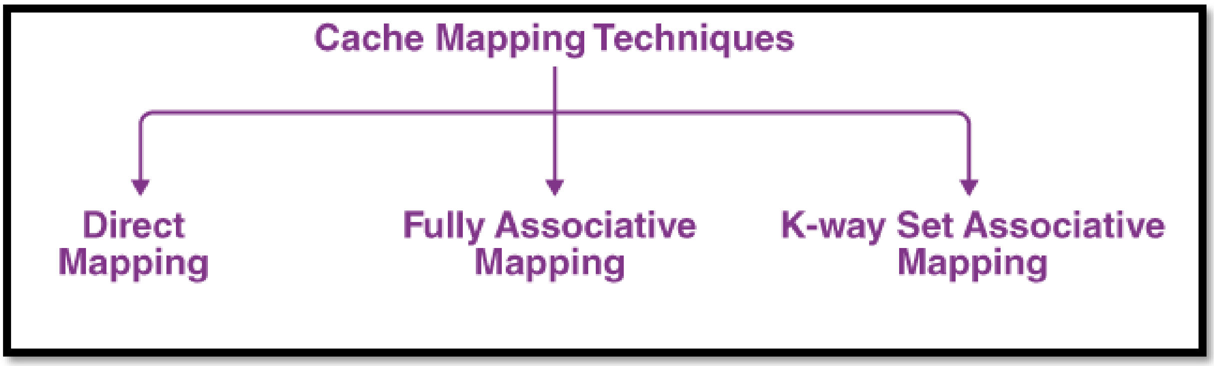
In simpler words, cache mapping refers to a technique using which we bring the main memory into the cache memory. Here is a diagram that illustrates the actual process of mapping:



- The main memory gets divided into multiple partitions of equal size, known as the **frames or blocks.**
- The cache memory is actually divided into various partitions of the same sizes as that of the blocks, known as **lines**.
- The main memory block is copied simply to the cache during the process of cache mapping, and this block isn't brought at all from the main memory.

✓ Techniques of Cache Mapping

One can perform the process of cache mapping using these three techniques given as follow



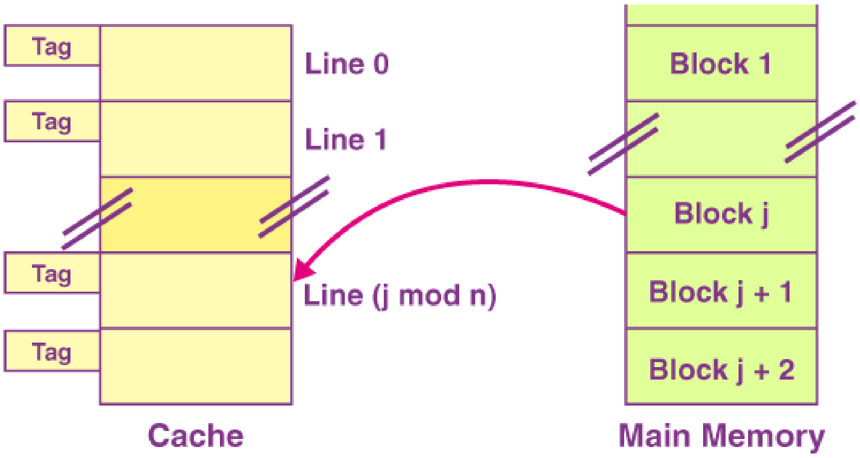
1. Direct Mapping

In the case of *direct mapping*, a certain block of the main memory would be able to map a cache only up to a certain line of the cache. The total line numbers of cache to which any distinct block can map are given by the following:

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

For example,

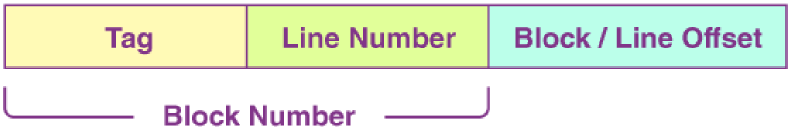
- Let us consider that particular cache memory is divided into a total of 'n' number of lines.
- Then, the block 'j' of the main memory would be able to map to line number only of the cache (j mod n).



Direct Mapping

Division of Physical Address

In the case of direct mapping, the division of the physical address occurs as follows



Division of Physical Address in Direct Mapping

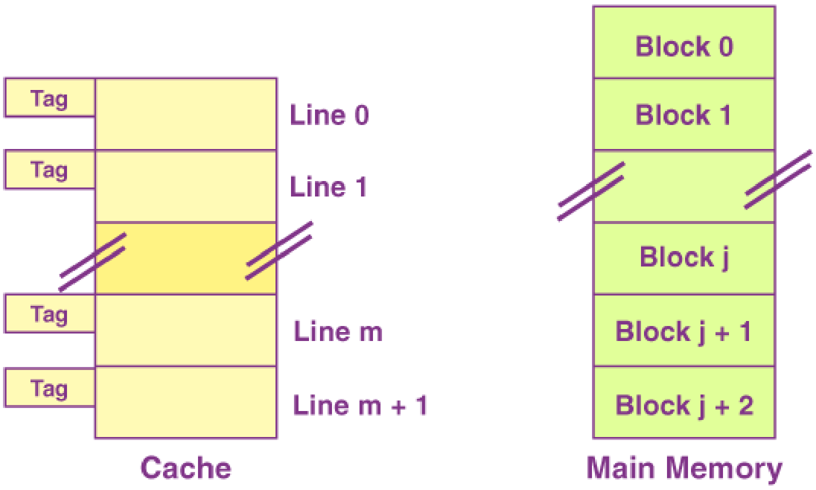
2. Fully Associative Mapping

In the case of fully associative mapping,

- The main memory block is capable of mapping to any given line of the cache that's available freely at that particular moment.
- It helps us make a fully associative mapping comparatively more flexible than direct mapping.

For Example

Let us consider the scenario given as follows:



Fully Associative Mapping

Here, we can see that,

- Every single line of cache is available freely.
- Thus, any main memory block can map to a line of the cache.
- In case all the cache lines are occupied, one of the blocks that exists already needs to be replaced.

Division of Physical Address

In the case of fully associative mapping, the division of the physical address occurs as follows:

Division of Physical Address in Fully Associative Mapping

3. K-way Set Associative Mapping

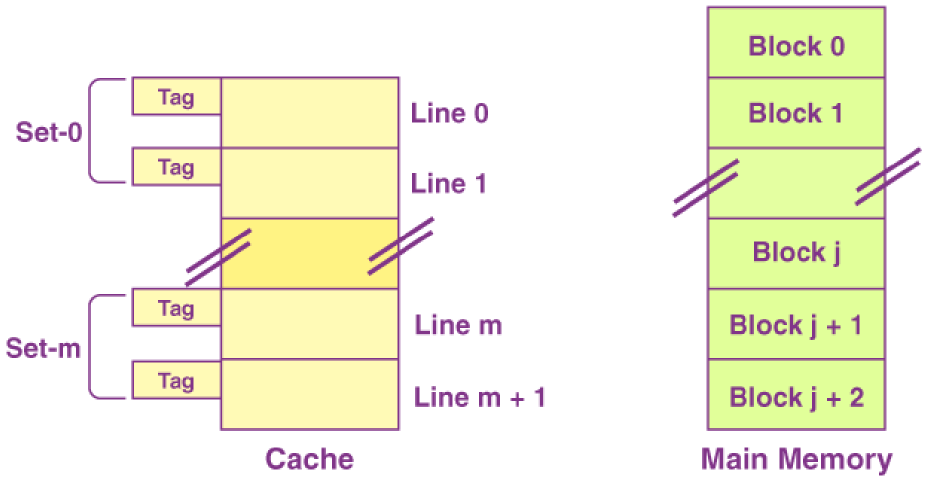
In the case of k-way set associative mapping,

- The grouping of the cache lines occurs into various sets where all the sets consist of k number of lines.
- Any given main memory block can map only to a particular cache set.
- However, within that very set, the block of memory can map any cache line that is freely available.
- The cache set to which a certain main memory block can map is basically given as follows:

Cache set number = (Block Address of the Main Memory) Modulo (Total Number of sets present in the Cache)

For Example

Let us consider the example given as follows of a two-way set-associative mapping:



2-Way Set Associative Mapping

In this case,

- k = 2 would suggest that every set consists of two cache lines.
- Since the cache consists of 6 lines, the total number of sets that are present in the cache = 6 / 2 = 3 sets.
- The block ‘j’ of the main memory is capable of mapping to the set number only (j mod 3) of the cache.

- Here, within this very set, the block 'j' is capable of mapping to any cache line that is freely available at that moment.
- In case all the available cache lines happen to be occupied, then one of the blocks that already exist needs to be replaced.

Division of Physical Address

In the case of fully k-way set mapping, the division of the physical address occurs as follows:

Tag	Set Number	Block / Line Offset

Division of Physical Address in K-way Set Associative Mapping

Special Cases

- In case k = 1, the k-way set associative mapping would become direct mapping. Thus,

Direct Mapping = one-way set associative mapping

- In the case of k = The total number of lines present in the cache, then the k-way set associative mapping would become fully associative mapping.

Advantages

The advantages of cache memory are as follows −

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

Disadvantages

The disadvantages of cache memory are as follows −

- Cache memory has limited capacity.
- It is very expensive.

→ Secondary Memory

This type of memory is also known as external memory or non-volatile. It is slower than the main memory. These are used for storing data/information permanently. CPU directly does not access these memories, instead they are accessed via input-output routines. The contents of secondary memories are first transferred to the main memory, and then the CPU can access it. For example, disk, CD-ROM, DVD etc.

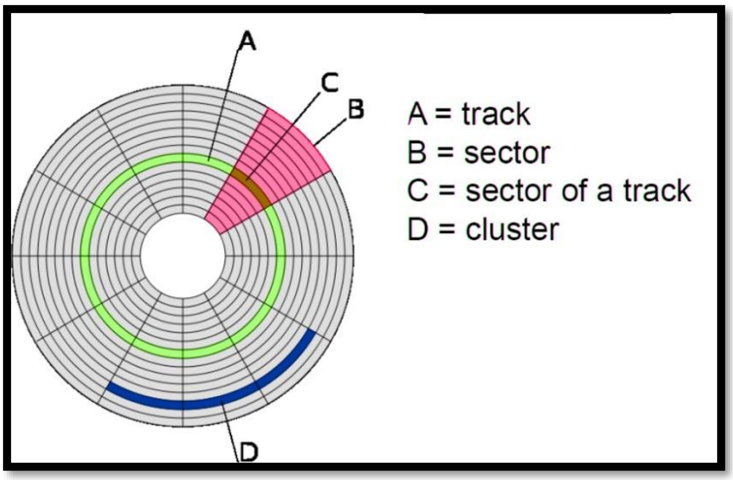
Characteristics of Secondary Memory

- These are magnetic and optical memories.
- It is known as the backup memory.
- It is a non-volatile memory.
- Data is permanently stored even if power is switched off.
- It is used for storage of data in a computer.
- Computer may run without the secondary memory.
- Slower than primary memories.

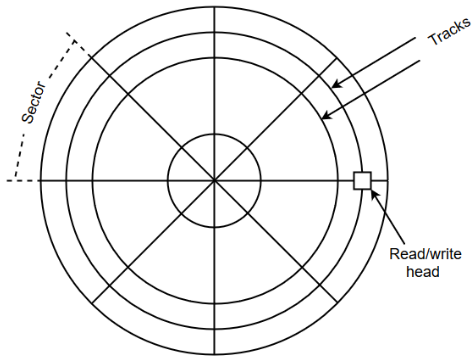
Magnetic Disks

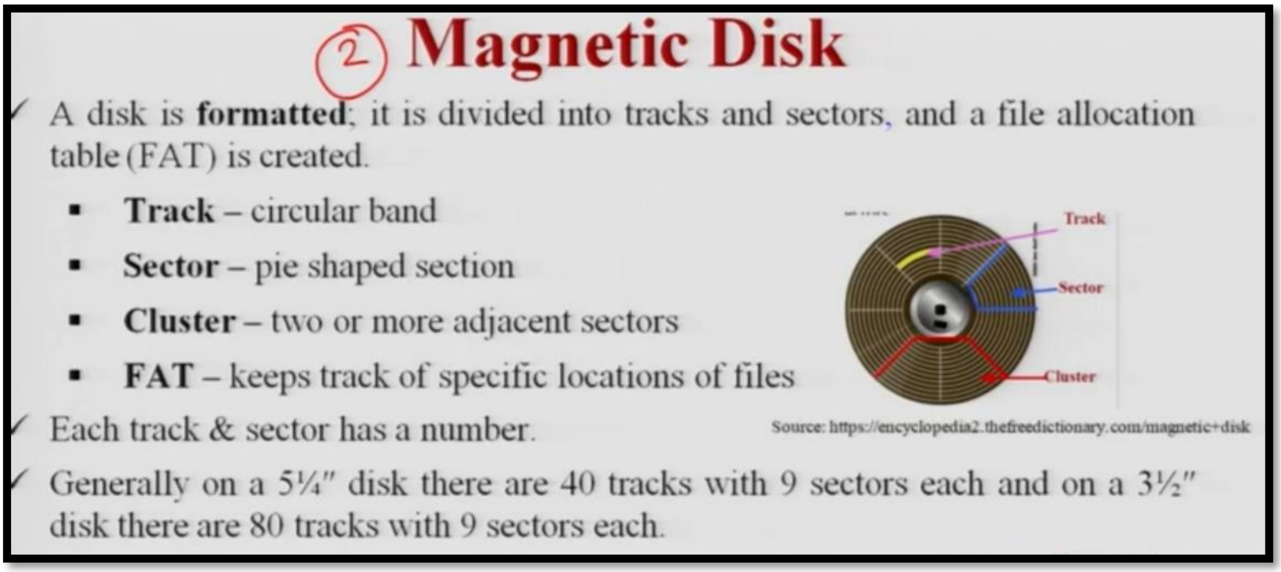
A magnetic disk is a type of memory constructed using a circular plate of metal or plastic coated with magnetized materials. Usually, both sides of the disks are used to carry out read/write operations. However, several disks may be stacked on one spindle with read/write head available on each surface.

The following image shows the structural representation for a magnetic disk.



Magnetic disks





- The memory bits are stored in the magnetized surface in spots along the concentric circles called tracks.
- The concentric circles (tracks) are commonly divided into sections called sectors.

Magnetic Tape

Magnetic tape is a storage medium that allows data archiving, collection, and backup for different kinds of data. The magnetic tape is constructed using a plastic strip coated with a magnetic recording medium.

The bits are recorded as magnetic spots on the tape along several tracks. Usually, seven or nine bits are recorded simultaneously to form a character together with a parity bit.

Magnetic tape units can be halted, started to move forward or in reverse, or can be rewound. However, they cannot be started or stopped fast enough between individual characters. For this reason, information is recorded in blocks referred to as records.

Optical Disk

- ✓ is an electronic data storage medium that supports writing and reading using a low-powered laser beam.
- ✓ can store data up to 6 gigabytes (6 billion bytes).
- ✓ made up of a resin (such as polycarbonate) coated with a highly reflective material (Aluminum and also silicon, silver, or gold in double-layered DVDs).
- ✓ Store software, data, digital photos, movies, and music.

Source: https://encyclopedia2.thefreedictionary.com/optical+disk

Optical Disk-Types

CD-ROM	WORM	EOD	DVD
Stands for Compact Disc-Read Only Memory.	Stands for write-once, read-many.	stands for Erasable Optical Disk.	A DVD can have data in two layers.
Used for storing OS, large application and multimedia programs.	With WORM , one can write data only once.	The disk contents can be erased and new data can be rewritten to it.	Data can be accessed from 1 or both sides.
capable of storing 650 MB of data.	advantages are high capacity, longer life and better reliability.	Used as secondary memory of computer.	Data can't be changed once it is created on DVDs.

Reference: Book/Websites

Computer System Architecture: Morris Mano

Computer Organization and Architecture: Sudam Pawar

www.geeksforgeeks.org

www.javatpoint.com

www.learncomputerscienceonline.com