Unit IV: Memory Organization

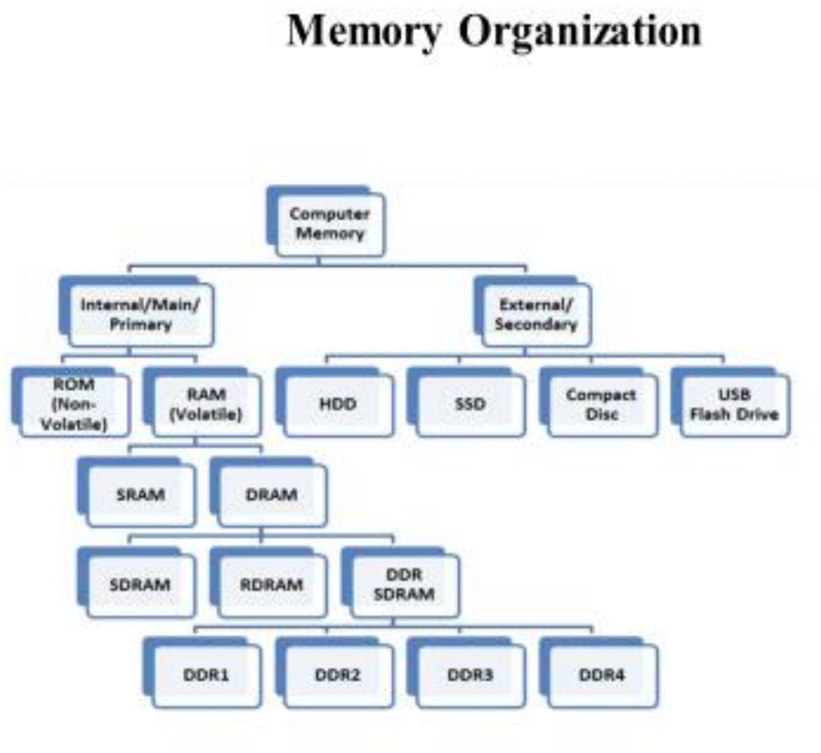


Fig 1: Organization of Memory

- RAM (Random Access Memory) is the internal memory of the CPU for storing data, program, and program result. It is a read/write memory which stores data until the machine is working. As soon as the machine is switched off, data is erased.
- Access time in RAM is independent of the address, that is, each storage location inside the memory is as easy to reach as other locations and takes the same amount of time. Data in the RAM can be accessed randomly but it is very expensive.
- RAM is similar in concept to a set of boxes in which each box can hold a 0 or a 1. Each box has a unique address that is found by counting across the columns and down the rows. A set of RAM boxes is called an array, and each box is known as a cell.
- To find a specific cell, the RAM controller sends the column and row address down a thin electrical line etched into the chip. Each row and column in a RAM array has its own address line. Any data that's read flows back on a separate data line.
- RAM is physically small and stored in microchips. It's also small in terms of the amount of data it can hold. A typical laptop computer may come with 8 gigabytes of RAM, while a hard disk can hold 10 terabytes.

- RAM is volatile, i.e. data stored in it is lost when we switch off the computer or if there is a power failure.
- RAM access time is in nanosecond, while storage memory access time is in milliseconds.
- RAM microchips are gathered together into memory modules. These plug into slots in a computer's motherboard. A bus, or a set of electrical paths, is used to connect the motherboard slots to the processor.
- When you switch on the computer the data and instructions from the hard disk are stored in the RAM, e.g., when the computer is rebooted, and when you open a program, the operating system (OS), and the program are loaded into RAM, generally from an HDD or SSD. CPU utilizes this data to perform the required tasks. As soon as you shut down the computer, the RAM loses the data.So, the data remains in the RAM as long as the computer is on and lost when the computer is turned off. The benefit of loading data into RAM is that reading data from the RAM is much faster than reading from the hard drive..

Types of RAM

Two main types of RAM are:

- Static RAM
- Dynamic RAM

Static RAM

- Static RAM is the full form of SRAM. In this type of RAM, data is stored using the state of a six transistor memory cell.
- Static RAM is mostly used as a cache memory for the processor (CPU).

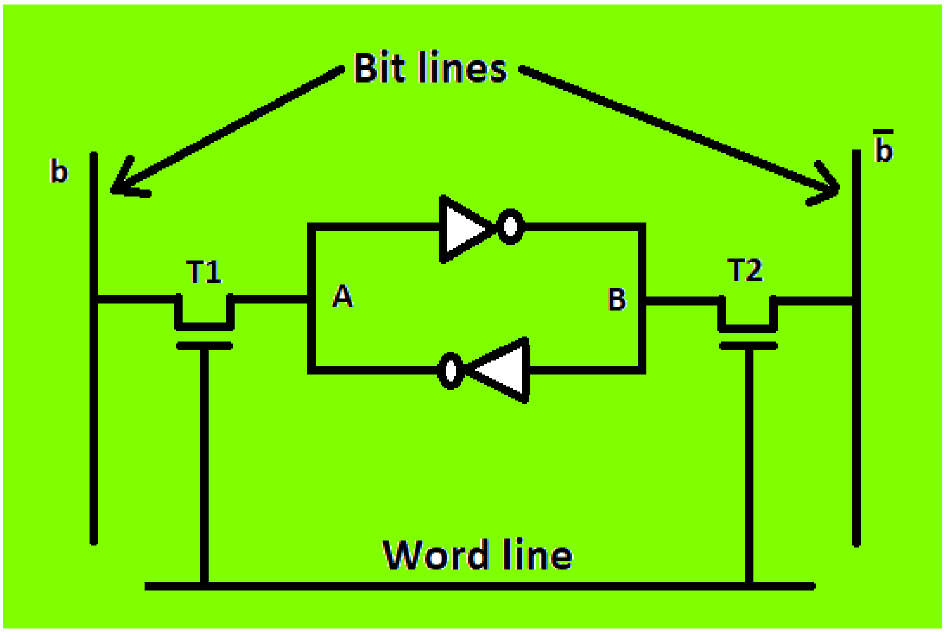


Fig.2 : Static RAM cell circuit

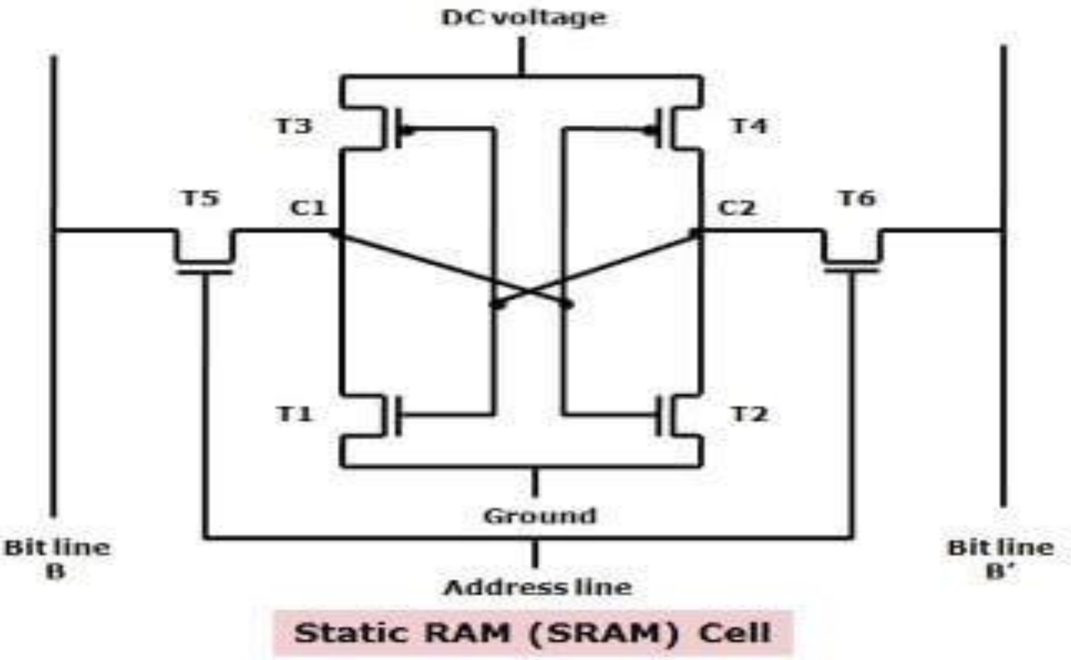


Fig.3 : Detailed (With all Transistors)Static RAM(SRAM) cell circuit

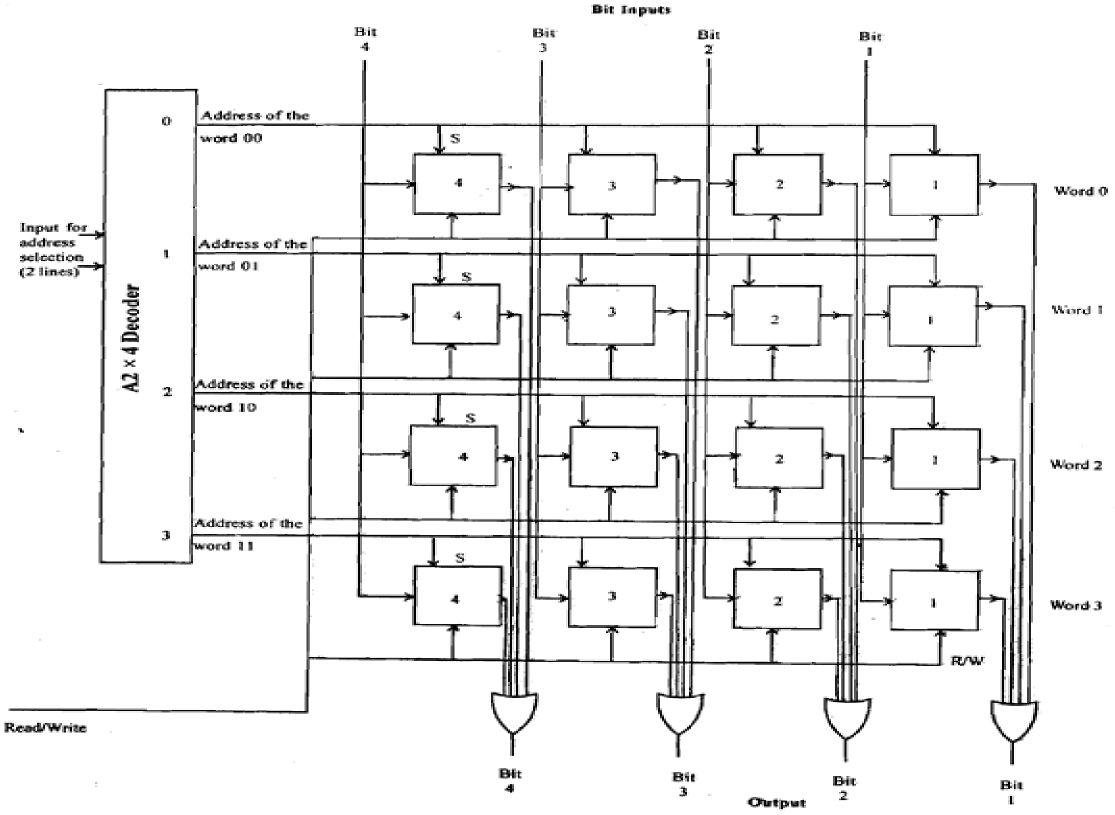


Fig.4 Memory Structure

- Static RAM (SRAM) is a type of random access memory that retains its state for data bits or holds data as long as it receives the power.
- It is made up of memory cells and is called a static RAM as it does not need to be refreshed on a regular basis because it does not need the power to prevent leakage, unlike dynamic RAM. So, it is faster than DRAM.
- It has a special arrangement of transistors that makes a flip-flop, a type of memory cell.
- One memory cell stores one bit of data. Most of the modern SRAM memory cells are made of six CMOS transistors, but lack capacitors.
- Furthermore, its cycle time is much shorter than that of DRAM as it does not pause between accesses. Due to these advantages associated with the use of SRAM, It is primarily used for system cache memory, and high-speed registers, and small memory banks such as a frame buffer on graphics cards.
- The Static RAM is fast because the six-transistor configuration of its circuit maintains the flow of current in one direction or the other (0 or 1).
- The 0 or 1 state can be written and read instantly without waiting for the capacitor to fill up or drain.

DRAM:

- Dynamic RAM is made from one transistor and one capacitor. each cell represents or stores a single bit of data in its capacitor within an integrated circuit.
- Many of these tiny cells combine to form a large memory chunk. Since a capacitor is used, it needs to be refreshed from time to time to maintain the charge. Capacitors leak, hence they need to be recharged as soon as they are read, they need to be written back.
- The transistor, which is also present in the cell, acts as a switch that allows the electric circuit on the memory chip to read the capacitor and change its state.
- The capacitor needs to be refreshed after regular intervals to maintain the charge in the capacitor. This is the reason it is called dynamic RAM as it needs to be refreshed continuously to maintain its data or it would forget what it is holding.
- This is achieved by placing the memory on a refresh circuit that rewrites the data several hundred times per second.
- The access time in DRAM is around 60 nanoseconds.

Features of Dynamic RAM

DRAM computer RAM is useful to have as a cheaper memory option. It usually serves as the main memory.

- DRAM has a much higher access time of around 50 nanoseconds.
- It is slower than SRAM because memory cells need to be continuously refreshed.
- It consumes less power because the information is stored in one capacitor.
- DRAM is less expensive than SRAM.
- One memory cell is made up of one transistor and one capacitor so it occupies less space on the same-sized chip, providing you with more memory than an SRAM of similar size.

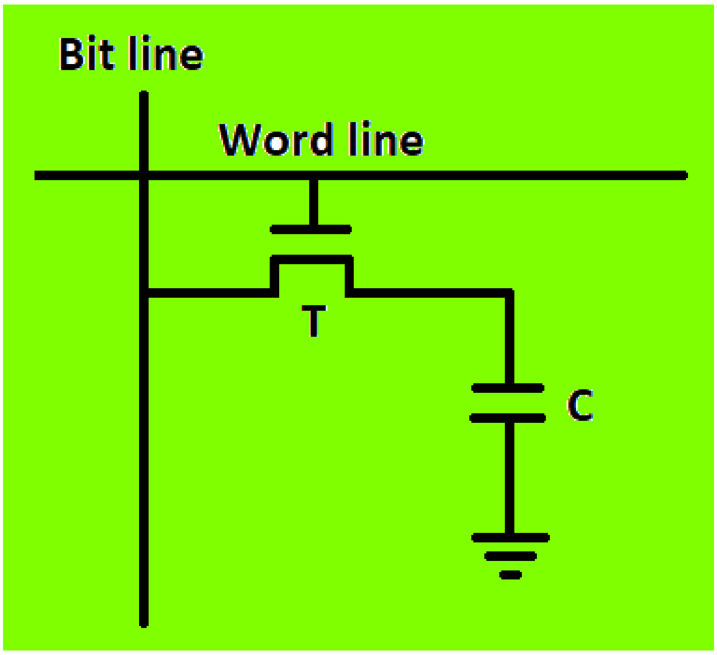


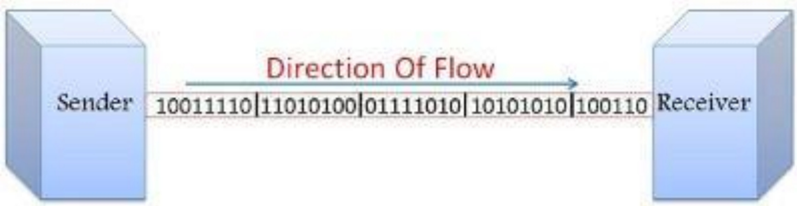
Fig 5: DRAM Structure.

Types of DRAM:

i) Asynchronous DRAM:

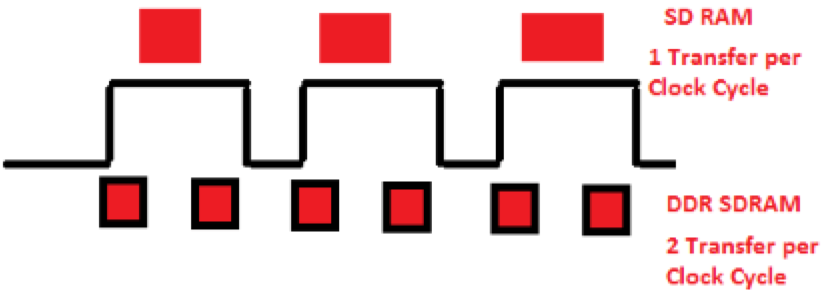
This type of DRAM is not synchronized with the CPU clock. So, the drawback with this RAM is that CPU could not know the exact timing at which the data would be available from the RAM on the input-output bus. This limitation was overcome by the next generation of RAM, which is known as the synchronous DRAM. RAM was originally asynchronous because the RAM microchips had a different clock speed than the computer's processor. This was a problem as processors became more powerful and RAM couldn't keep up with the processor's requests for data.

ii) Synchronous DRAM:



In SDRAM, the RAM was synchronized with the CPU clock. It allowed the CPU or to be precise the memory controller to know the exact clock cycle or timing or the number of cycles after which the data will be available on the bus. So, the CPU does not need for the memory accesses and thus the memory read and write speed can be increased. The SDRAM is also known as the single data rate SDRAM (SDR SDRAM) as data is transferred only at each rising edge of the clock cycle.

iii) DDR SDRAM:

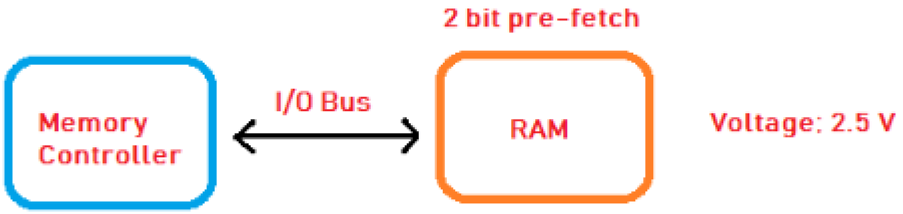


The next generation of the synchronous DRAM is known as the DDR RAM. It was developed to overcome the limitations of SDRAM and was used in PC memory at the beginning of the year 2000. In DDR SDRAM (DDR RAM), the data is transferred twice during each clock cycle; during the positive edge (rising edge) and the negative edge (falling edge) of the cycle. So, it is known as the double data rate SDRAM.

There are different generations of DDR SDRAM which include DDR1, DDR2, DDR3, and DDR4. Today, the memory that we use inside the desktop, laptop, mobile, etc., is mostly either DDR3 or DDR4 RAM.

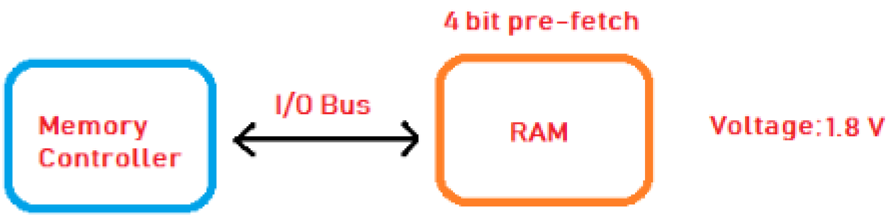
Types of DDR SDRAM:

a) DDR1 SDRAM:



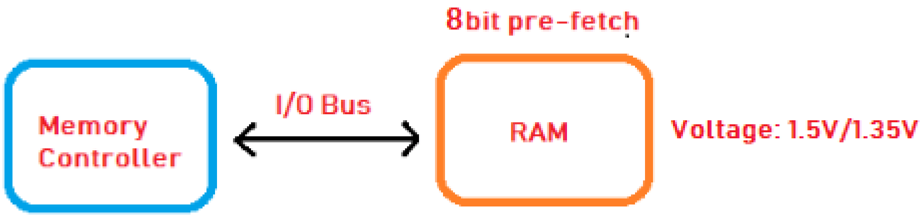
DDR1 SDRAM is the first advanced version of SDRAM. In this RAM, the voltage was reduced from 3.3 V to 2.5 V. The data is transferred during both the rising as well as the falling edge of the clock cycle. So, in each clock cycle, instead of 1 bit, 2 bits are being pre-fetched which is commonly known as the 2 bit pre-fetch. It is mostly operated in the range of 133 MHz to the 200 MHz. Furthermore, the data rate at the input-output bus is double the clock frequency because the data is transferred during both the rising as well as falling edge. So, if a DDR1 RAM is operating at 133 MHz, the data rate would be double, 266 Mega transfer per second.

ii) DDR2 SDRAM:



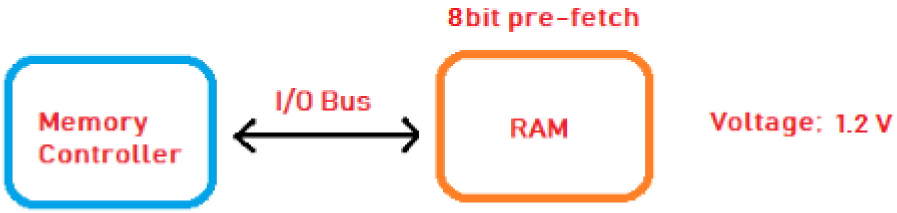
It is an advanced version of DDR1. It operates at 1.8 V instead of 2.5V. Its data rate is double the data rate of the previous generation due to the increase in the number of bits that are pre-fetched during each cycle; 4 bits are pre-fetched instead of 2 bits. The internal bus width of this RAM has been doubled. For example, if the input-output bus is 64 bits wide, the internal bus width of it will be equal to 128 bits. So, a single cycle can handle double the amount of data.

iii) DDR3 SDRAM:



In this version, the voltage is further reduced from 1.8 V to the 1.5 V. The data rate has been doubled than the previous generation RAM as the number of bits that are pre-fetched has been increased from 4 bits to the 8 bits. We can say that the internal data bus width of RAM has been increased 2 times than that of the last generation.

iv) DDR4 SDRAM:



In this version, the operating voltage is further reduced from 1.5 V to 1.2 V, but the number of bits that can be pre-fetched is same as the previous generation; 8 bits per cycle. The Internal clock frequency of the RAM is double of the previous version. If you are operating at 400 MHz the clock frequency of the input-output bus would be four times, 1600 MHz and the transfer rate would be equal to 3200 Mega transfer per second.

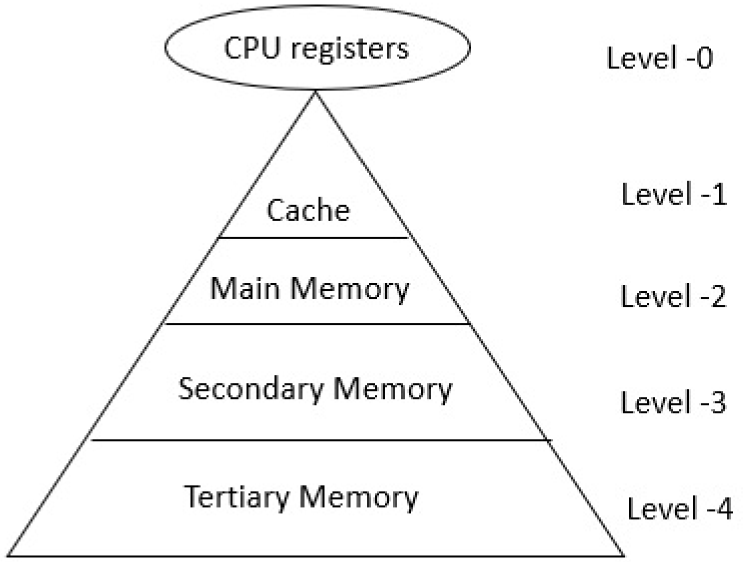
Difference between Static RAM and Dynamic RAM:

SRAM	DRAM
It is a static memory as it does not need to be refreshed repeatedly.	It is a dynamic memory as it needs to be refreshed continuously or it will lose the data.
Its memory cell is made of 6 transistors. So its cells occupy more space on a chip and offer less storage capacity (memory) than a DRAM of the same physical size.	Its memory cell is made of one transistor and one capacitor. So, its cells occupy less space on a chip and provide more memory than a SRM of the same physical size.
It is more expensive than DRAM and is located on processors or between a processor and main memory.	It is less expensive than SRAM and is mostly located on the motherboard.
It has a lower access time, e.g. 10 nanoseconds. So, it is faster than DRAM.	It has a higher access time, e.g. more than 50 nanoseconds. So, it is slower than SRAM.
It stores information in a bistable latching circuitry. It requires regular power supply so it consumes more power.	The information or each bit of data is stored in a separate capacitor within an integrated circuit so it consumes less power.
It is faster than DRAM as its memory cells don't need to be refreshed and are always available. So, it is mostly used in registers in the CPU and cache memory of various devices.	It is not as fast as SRAM, as its memory cells are refreshed continuously. But still, it is used in the motherboard because it is cheaper to manufacture and requires less space.
Its cycle time is shorter as it does not need to be paused between accesses and refreshes.	Its cycle time is more than the SRAM's cycle time.
Examples: L2 and LE cache in a CPU.	Example: DDR3, DDR4 in mobile phones, computers, etc.

Size ranges from 1 MB to 16MB.	Size ranges from 1 GB to 3 GB in smartphones and 4GB to 16GB in laptops.

Memory Hierarchy

Memory hierarchy is arranging different kinds of storage present on a computing device based on speed of access.



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

Level-0 − Registers

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

Level-1 − Cache

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

Level-2 − Primary or Main Memory

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

Ex:RAM

Level-3 − Secondary storage

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

Ex: Pen drives, Hard Disc, CD,DVD etc.

Level-4 − Tertiary storage

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

Ex :An optical jukebox is a robotic data storage device that can automatically load and unload optical discs.

Cache Memory:

- **cache memory**, also called **cache**, supplementary memory system that temporarily stores frequently used instructions and data for quicker processing by the central processing unit (CPU) of a computer.
- The cache augments, and is an extension of, a computer’s main memory.
- Both main memory and cache are internal random-access memories (RAMs) that use semiconductor-based transistor circuits.
- Cache holds a copy of only the most frequently used information or program codes stored in the main memory. The smaller capacity of the cache reduces the time required to locate data within it and provide it to the CPU for processing.
- When a computer’s CPU accesses its internal memory, it first checks to see if the information it needs is stored in the cache.
- If it is, the cache returns the data to the CPU. If the information is not in the cache, the CPU retrieves it from the main memory.

Memory Cache

Memory cache is a type of cache that uses CPU memory to speed up data access from the main memory. It is known as L1, L2, L3, and so on, and it is considerably smaller than RAM memory but much quicker.

Disk Cache

Disk cache creates a duplicate of anything you're working on in RAM memory. The whole folder is usually copied into the cache since the computer expects you may require some of the information. That's why opening a folder for the first time may take substantially longer than opening a file within it.

Browser Cache (Web Cache)

Different portions of the online sites such as graphics, JavaScript, and queries, are stored on your hard drive by web browsers. You should be able to check how much storage has been utilized for cached pictures if you go to your browser settings and decide to clear your history.

App Cache

An app cache functions just like a web cache. It saves data such as code and files to the app's memory so that it can access them more quickly the next time you need them.

Cache Memory Mapping:

- Cache mapping is a technique by which the contents of main memory are brought into the cache memory.

• cache memory is a small and fast memory between CPU and main memory

• A block of words have to be brought in and out of the cache memory continuously

• Performance of the cache memory mapping function is key to the speed .

- 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.

• There are a number of mapping techniques

– Direct mapping

– Fully Associative mapping

– Set associative – mapping

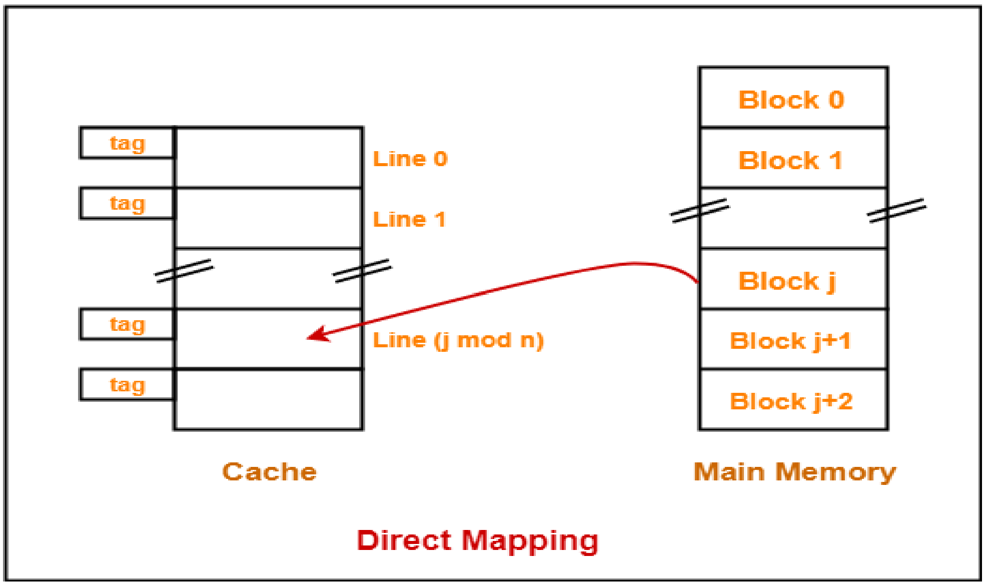
Direct mapping

In direct mapping,

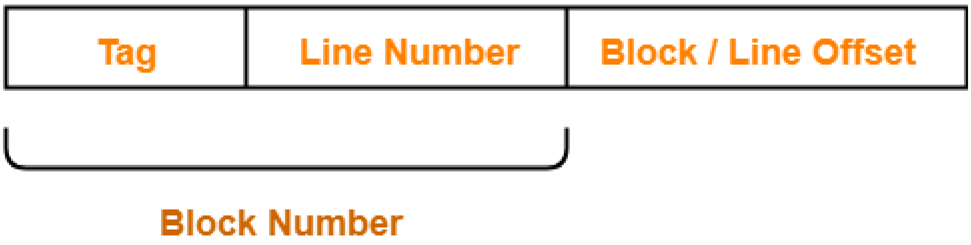
- A particular block of main memory can map only to a particular line of the cache.
- The line number of cache to which a particular block can map is given by-

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

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



In direct mapping, the physical address is divided as-

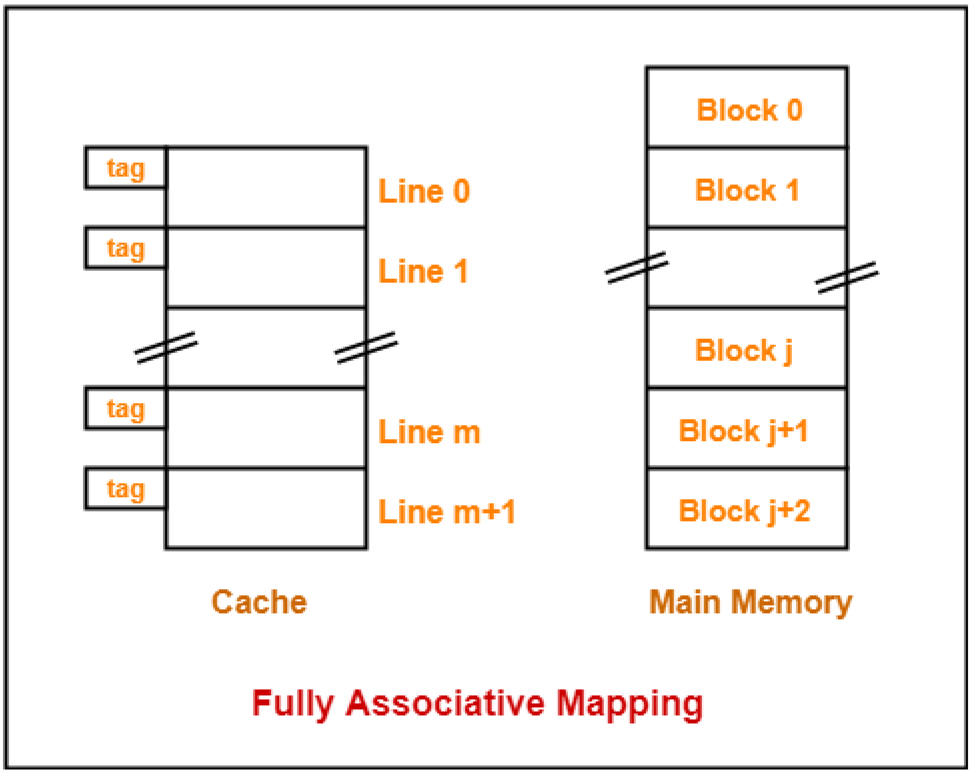


Division of Physical Address in Direct Mapping

2. Fully Associative Mapping-

In fully associative mapping,

- 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.

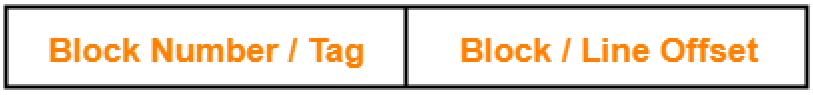


Fully Associative Mapping

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

Division of Physical Address-

In fully associative mapping, the physical address is divided as-



Division of Physical Address in Fully Associative Mapping

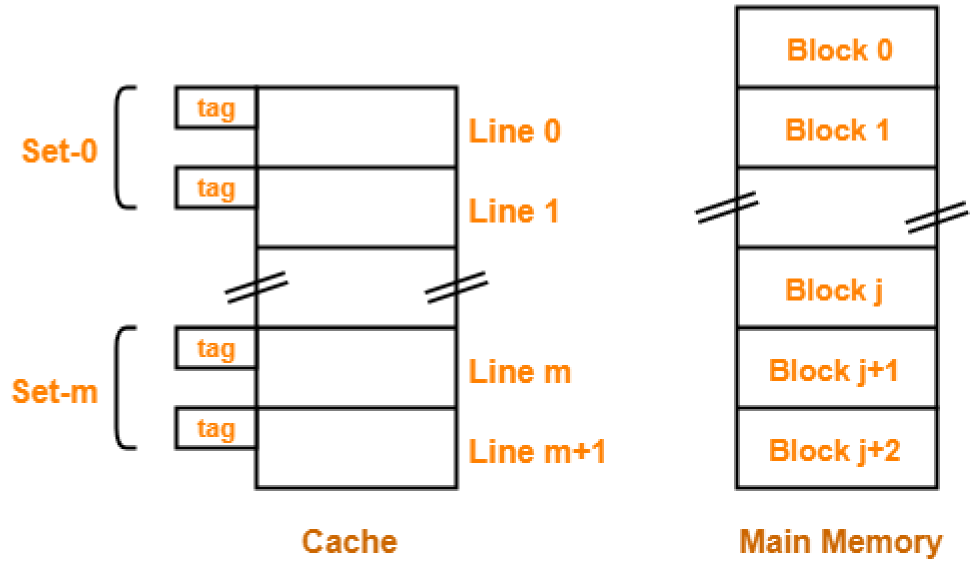
3. K-way Set Associative Mapping or Set Associative Mapping-

In k-way set associative mapping,

- Cache lines(or blocks) are grouped into sets where each set contains k number of lines (or blocks).
- 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-

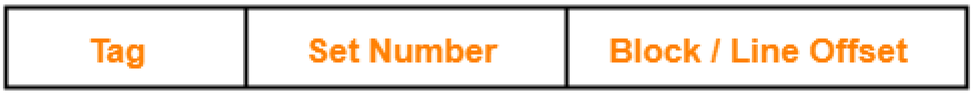
Cache set number = (Main Memory Block Address) Modulo (Number of sets in Cache)



2-Way Set Associative Mapping

The above diagram shows

- k = 2 suggests that each set contains two cache lines or blocks.
- 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 set associative mapping, the physical address is divided as-



Tag	Set Number	Block / Line Offset

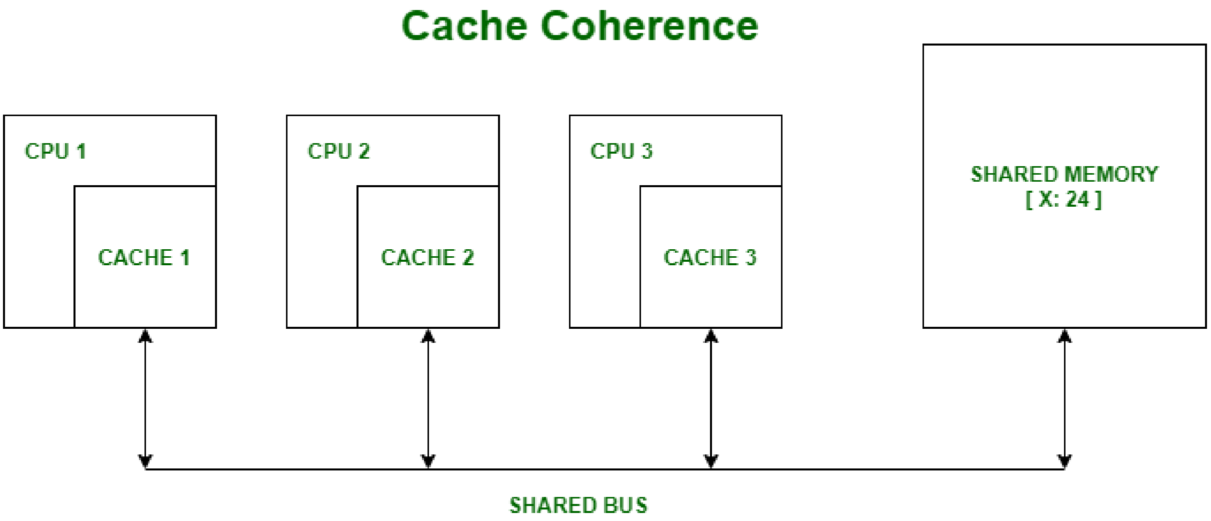
Division of Physical Address in K-way Set Associative Mapping

- If k = 1, then k-way set associative mapping becomes direct mapping i.e. **1-way Set Associative Mapping ≡ Direct Mapping**
- If k = Total number of lines in the cache, then k-way set associative mapping becomes fully associative mapping.

Cache Coherence:

In a shared memory multiprocessor with a separate cache memory for each processor, it is possible to have many copies of any one instruction operand: one copy in the main memory and one in each cache memory. When one copy of an operand is changed, the other copies of the operand must be changed also. Example : Cache and the main memory may have inconsistent copies of the same object.

Cache Coherence problem is the problem where the updated value in cache memory is not reflected in main memory.



Suppose there are three processors, each having cache. Suppose the following scenario:-

- **Processor 1 read X :** obtains 24 from the memory and caches it.
- **Processor 2 read X :** obtains 24 from memory and caches it.
- **Again, processor 1 writes as X :** 64, Its locally cached copy is updated. Now, processor 3 reads X, what value should it get?
- Memory and processor 2 thinks it is 24 and processor 1 thinks it is 64.

As multiple processors operate in parallel, and independently multiple caches may possess different copies of the same memory block, this creates a cache coherence problem.

To deal with Cache coherence problem , snoopy cache coherence protocol is used.

Snoopy Cache Coherence Protocol:

There are two ways to maintain the coherence requirement.

- One method is to ensure that a processor has exclusive access to a data item before it writes that item. This style of protocol is called a *write invalidate protocol* because it invalidates other copies on a write. It is the most common protocol.
- Exclusive access ensures that no other readable or writable copies of an item exist when the write occurs: All other cached copies of the item are invalidated.

- The alternative to write invalidate is the *write broadcast* or *write update* mechanism. Here, all the cached copies are updated simultaneously.

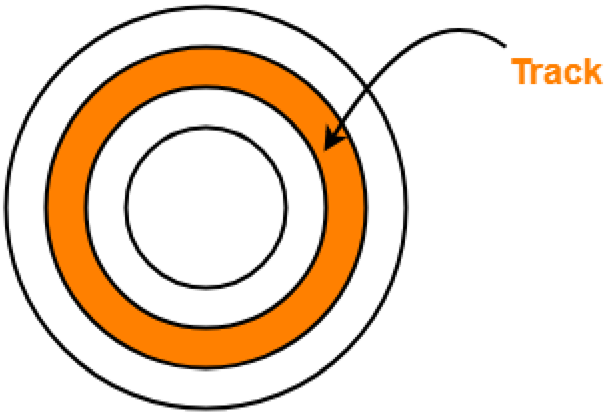
Secondary Storage

: Secondary storage devices are external storage devices For ex. Hard disk, CD,SDD, Pendrives, magnetic disk etc.

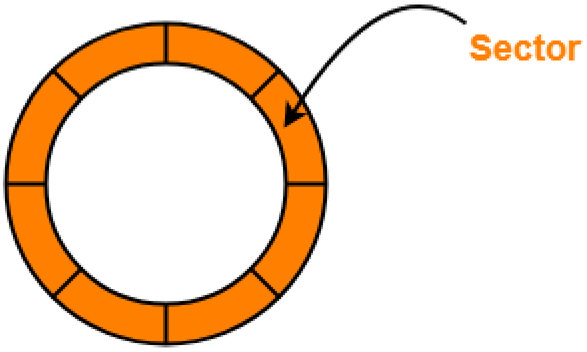
- Magnetic disk is a storage device that is used to write, rewrite and access data.
- It uses a magnetization process.

<u>Architecture-</u>

- The entire disk is divided into **platters**.
- Each platter consists of concentric circles called as **tracks**.
- These tracks are further divided into **sectors** which are the smallest divisions in the disk.

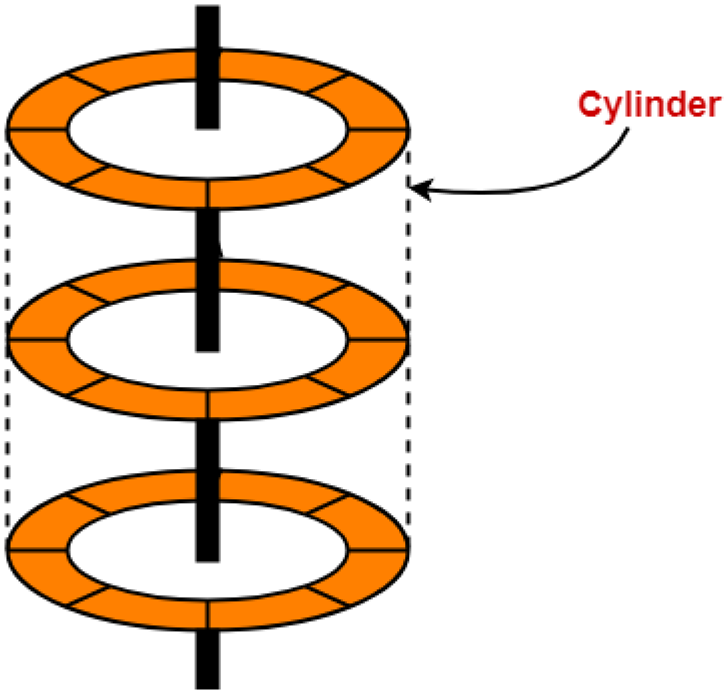


Disk divided into tracks

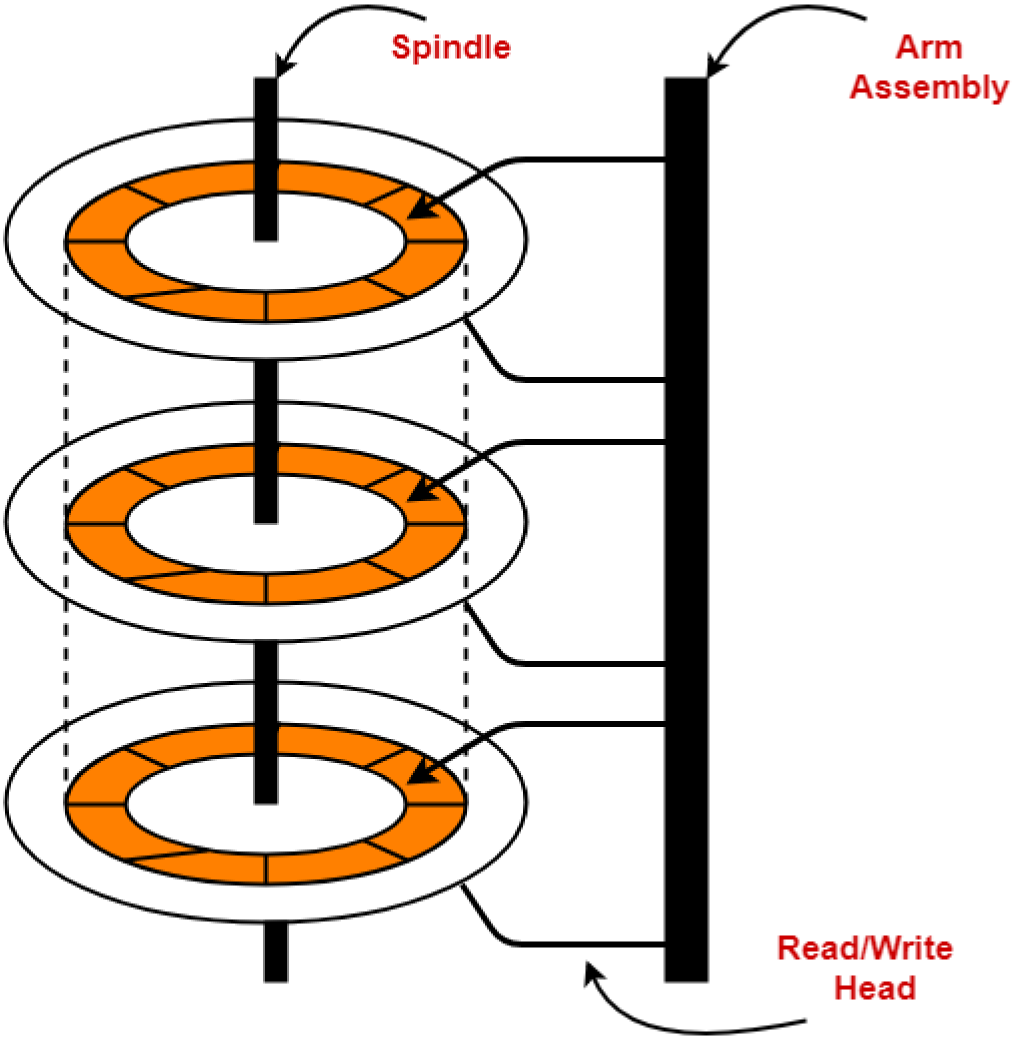


Track divided into sectors

- A **cylinder** is formed by combining the tracks at a given radius of a disk pack.



- There exists a mechanical arm called as **Read / Write head**.
- It is used to read from and write to the disk.
- Head has to reach at a particular track and then wait for the rotation of the platter.
- The rotation causes the required sector of the track to come under the head.
- Each platter has 2 surfaces- top and bottom and both the surfaces are used to store the data.
- Each surface has its own read / write head.



Disk Performance Parameters-

The time taken by the disk to complete an I/O request is called as **disk service time** or **disk access time**.

Components that contribute to the service time are-

1. Seek time
2. Rotational latency
3. Data transfer rate
4. Controller overhead
5. Queuing delay

Seek Time-

- The time taken by the read / write head to reach the desired track is called as **seek time**.
- It is the component which contributes the largest percentage of the disk service time.
- The lower the seek time, the faster the I/O operation.

Rotational Latency-

- The time taken by the desired sector to come under the read / write head is called as **rotational latency**.
- It depends on the rotation speed of the spindle.

Average rotational latency = 1 / 2 x Time taken for full rotation

Data Transfer Rate-

- The amount of data that passes under the read / write head in a given amount of time is called as **data transfer rate**.
- The time taken to transfer the data is called as **transfer time**.

It depends on the following factors-

1. Number of bytes to be transferred
2. Rotation speed of the disk
3. Density of the track
4. Speed of the electronics that connects the disk to the computer

Queuing delay

The time spent waiting for the disk to become free is called as **queuing delay.**

Capacity Of Disk Pack-

Capacity of a disk pack is calculated as-

Capacity of a disk pack

= Total number of surfaces x Number of tracks per surface x Number of sectors per track x Storage capacity of one sector

Track Capacity-

Capacity of a track is calculated as-

Capacity of a track

= Recording density of the track x Circumference of the track

Data Transfer Rate-

Data transfer rate is calculated as-

Data transfer rate

= Number of heads x Bytes that can be read in one full rotation x Number of rotations in one second

OR

Data transfer rate

= Number of heads x Capacity of one track x Number of rotations in one second