COMPUTER ORGANIZATION

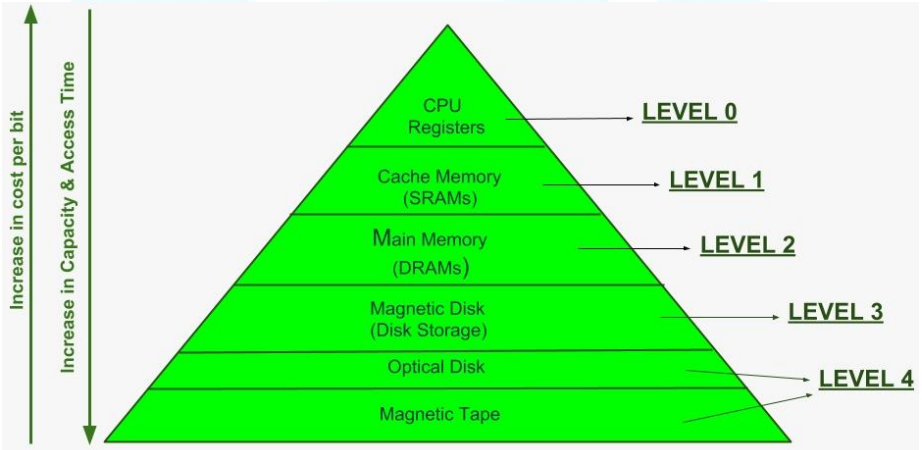
Memory – Hierarchy, organisation of RAM, types of RAM (SRAM, DRAM, SDRAM, and DDRAM).
Cache-operation, cache mapping, multilevel organization of cache (LI/LII, Primary/secondary).
Virtual memory page fault, TLB, segmentation – Multiple memory modules & interleaving.
Secondary storage – Disk-CDROM/DVD.

MEMORY

- Computer memory refers to the hardware device that are used to store and access data or programs on a temporary or permanent basis.

MEMORY HIERARCHY

- In the Computer System Design, Memory Hierarchy is an enhancement to organize the memory such that it can minimize the access time.
- The Memory Hierarchy was developed based on a program behaviour known as locality of references.
- The figure below clearly demonstrates the different levels of memory hierarchy :



MEMORY HIERARCHY DESIGN

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.

Characteristics of Memory Hierarchy Design

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.

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.

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.
- This results in lower performance of the system and thus, enhancement was required.
- This enhancement was made in the form of Memory Hierarchy Design because of which the performance of the system increases.
- One of the most significant ways to increase system performance is minimizing how far down the memory hierarchy one has to go to manipulate data.

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.

MEMORY ORGANISATION

- The memory is organized in the form of a cell; each cell is able to be identified with a unique number called address.
- Each cell is able to recognize control signals such as “read” and “write”, generated by CPU when it wants to read or write address.
- Whenever CPU executes the program there is a need to transfer the instruction from the memory to CPU because the program is available in memory.
- To access the instruction CPU generates the memory request.

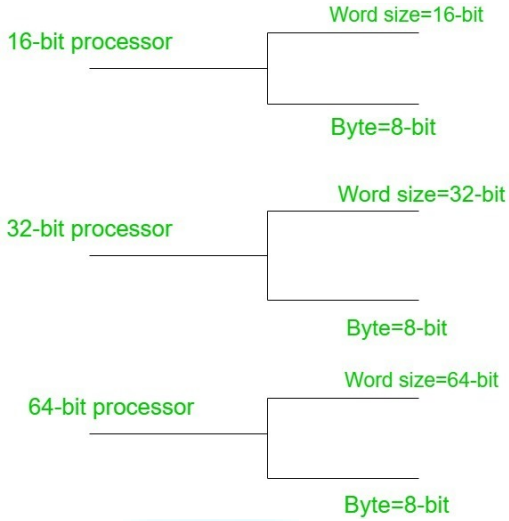
Memory Request:

- Memory request contains the address along with the control signals.
- For Example, when inserting data into the stack, each block consumes memory (RAM) and the number of memory cells can be determined by the capacity of a memory chip.
- With the number of cells, the number of address lines required to enable one cell can be determined.

Word Size:

- It is the maximum number of bits that a CPU can process at a time and it depends upon the processor.

- Word size is a fixed size piece of data handled as a unit by the instruction set or the hardware of a processor.



- Word size varies as per the processor architectures because of generation and the present technology, it could be low as 4-bits or high as 64-bits depending on what a particular processor can handle.
- Word size is used for a number of concepts like Addresses, Registers, Fixed-point numbers, and Floating-point numbers.

TYPES OF MEMORY

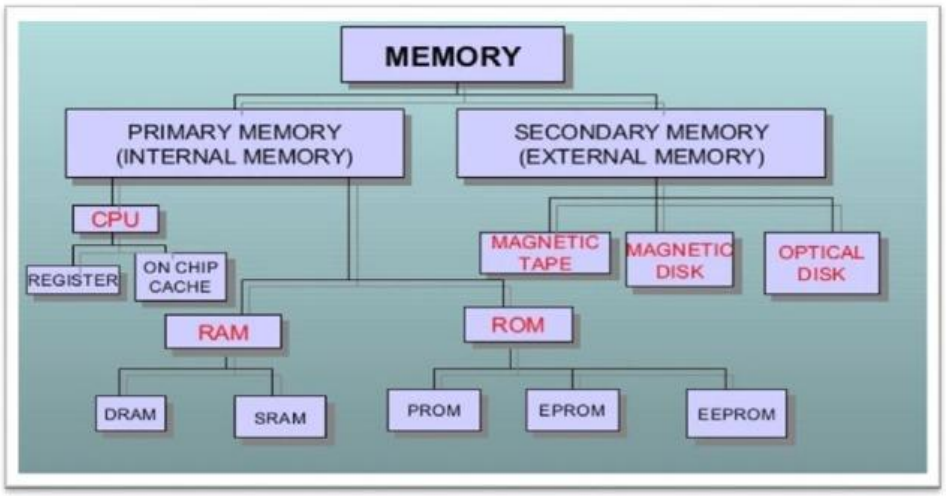
This Memory Hierarchy Design is divided into 2 main types:

External Memory or Secondary Memory –

- Comprising of Magnetic Disk, Optical Disk, and Magnetic Tape i.e. peripheral storage devices which are accessible by the processor via I/O Module.

Internal Memory or Primary Memory –

- Comprising of Main Memory, Cache Memory & CPU registers. This is directly accessible by the processor.



- 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 primarily of three types −

  - **Cache Memory**
  - **Primary Memory/Main Memory**
  - **Secondary Memory**

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.

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.

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.**

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.

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.

RANDOM ACCESS MEMORY (RAM)

- 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 volatile, i.e. data stored in it is lost when we switch off the computer or if there is a power failure.
- Hence, a backup Uninterruptible Power System (UPS) is often used with computers.
- RAM is small, both in terms of its physical size and in the amount of data it can hold.

RAM is of two types −

- **Static RAM (SRAM)**
- **Dynamic RAM (DRAM)**

Static RAM (SRAM)

- The word static indicates that the memory retains its contents as long as power is being supplied.
- However, data is lost when the power gets down due to volatile nature.
- SRAM chips use a matrix of 6-transistors and no capacitors.
- Transistors do not require power to prevent leakage, so SRAM need not be refreshed on a regular basis.
- There is extra space in the matrix; hence SRAM uses more chips than DRAM for the same amount of storage space, making the manufacturing costs higher.
- SRAM is thus used as cache memory and has very fast access.

Characteristic of Static RAM

-Long life
-No need to refresh
-Faster
-Used as cache memory
-Large size
-Expensive
-High power consumption

Dynamic RAM (DRAM)

- DRAM, unlike SRAM, must be continually refreshed in order to maintain the data.
- This is done by placing the memory on a refresh circuit that rewrites the data several hundred times per second.
- DRAM is used for most system memory as it is cheap and small.

- All DRAMs are made up of memory cells, which are composed of one capacitor and one transistor.

Characteristics of Dynamic RAM
-Short data lifetime
-Needs to be refreshed continuously
-Slower as compared to SRAM
-Used as RAM
-Smaller in size
-Less expensive
-Less power consumption

SDRAM (synchronous DRAM)

- SDRAM stands for Synchronous Dynamic Random Access Memory.
- It synchronizes itself with the computer's system clock.
- This makes it easy to manage faster, and the speed of the SDRAM measured in MHz instead of nanoseconds.
- The first commercial SDRAM was the Samsung KM48SL2000 memory chip, which had a capacity of 16 Mb.
- It was manufactured by Samsung Electronics using a complementary metal-oxide-semiconductor (CMOS) fabrication process in 1992 and mass-produced in 1993.
- By 2000, SDRAM had replaced virtually all other types of DRAM in modern computers, because of its greater performance and faster speed.
- It exploits the fact that most PC memory accesses are sequential and are designed to fetch all the bits in a burst as fast as possible.
- With SDRAM an on-chip burst counter allows the column part of the address to be incremented very rapidly which speeds up the retrieval of information in sequential reads considerably.
- The size of the block of memory location required is provided by the memory controller and the SDRAM chip supplies the bits as fast as the CPU can take them, using a clock to synchronize the timing of the memory chip to the CPU's system clock.
- Since its manufacture, SDRAM is modified and a new version is produced for better performance.

Generations of SDRAM

- The list of various generations is:

  - SDR SDRAM
  - DDR SDRAM
  - DDR2 SDRAM
  - DDR3 SDRAM
  - DDR4 SDRAM
  - DDR5 SDRAM

Characteristics

- Speed: SDRAM has higher operation speed make it popular. SDRAM access time is 6 to 12 nanoseconds (ns)
- Clock: SDRAM uses one edge of the clock. DDR uses both edges of the clock.

- Data transfer: SDRAM sends signals once per clock cycle. DDR transfers data twice per clock cycle.

Advantages

- It is faster as compared to the other versions of RAM.
- It is more efficient, which is up to 4 times the performance of the other standard DRAMs.
- Has name suggest, it gets synchronized with the system clock.

Disadvantages

- It can't use with the older motherboards.
- It works in a single data rate, i.e., it can do only tasks per clock cycle.

DDR-RAM Double Data Rate Synchronous Dynamic Random-Access Memory

- DDR-RAM stands for Double Data Rate Synchronous Dynamic Random-Access Memory.
- These are the computer memory that transfers the data twice as fast as regular chips like SDRAM chips because DDR memory can send and receive signals twice per clock cycle as a comparison.
- They are widely used in applications that are demanding high speed, memory, for example, graphic cards that need to access a large amount of information in a very short time to achieve the best graphics processing efficiency to improve the gaming.

History of DDR RAM

- In 1997 Samsung was the first company that released the DDR memory prototype in 1997 and launched the first commercial DDR SDRAM chip in June 1998 which was followed by many other companies.
- The first retail PC motherboard using DDR SDRAM was released in August 2000.
- There are many versions of DDR RAM. The DDR was first released in 1998 followed by its successor version i.e. DDR 2 in 2003, DDR 3 in 2007, and DDR 4 in 2014.

Characteristic of DDR RAM

- Its primary advantage is the ability to fetch data on both the rising and falling edge of a clock cycle, doubling the data rate for a given clock frequency.
- For example, data transfer frequency is 200 MHz of DDR200 device, but its bus speed is 100 MHz
- DDR1, DDR2, and DDR3 are the types of DDR RAM memories that use the 2.5, 1.8, and 1.5V supply voltages respectively, thus it produces less heat and provides more efficiency in power management than older SDRAM chipsets, which uses 3.3V.

Advantages of DDR RAM

- It offers much faster speeds than SDRAM.
- Each generation of DDR are updated and following with its successor DDR2, DDR3, DDR4.

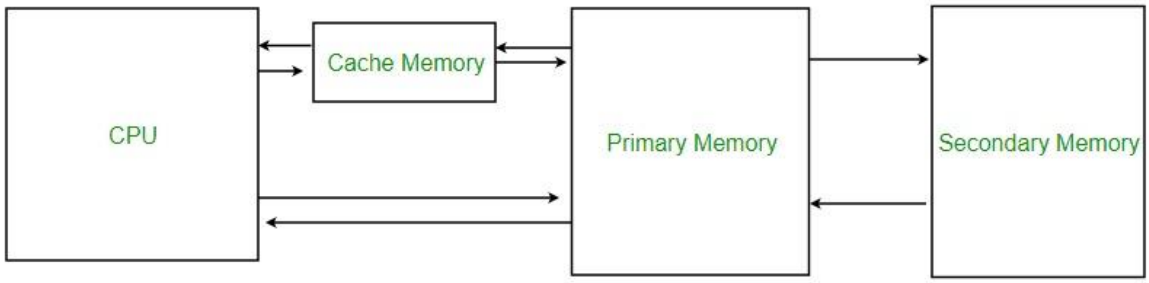
Disadvantages of DDR RAM

- It can't be used with old motherboards.
- Some machines support only slower RAM.
- They are not physically fit in memory slots due to its notches
- There pinouts are totally different from others.

ENTRI

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. I
- 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.
- Cache Memory is a special very high-speed memory.
- It is used to speed up and synchronizing with high-speed CPU.
- Cache memory is costlier than main memory or disk memory but economical than CPU registers.
- Cache memory is an extremely fast memory type that acts as a buffer between RAM and the CPU.
- It holds frequently requested data and instructions so that they are immediately available to the CPU when needed.
- Cache memory is used to reduce the average time to access data from the Main memory.
- The cache is a smaller and faster memory which stores copies of the data from frequently used main memory locations.
- There are various different independent caches in a CPU, which store instructions and data.
- The block diagram for a cache memory can be represented as:



The basic operation of a cache memory is as follows:

- When the CPU needs to access memory, the cache is examined. If the word is found in the cache, it is read from the fast memory.
- If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word.
- A block of words one just accessed is then transferred from main memory to cache memory. The block size may vary from one word (the one just accessed) to about 16 words adjacent to the one just accessed.
- The performance of the cache memory is frequently 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 the cache, it is in main memory and it counts as a miss.
- The ratio of the number of hits divided by the total CPU references to memory (hits plus misses) is the hit ratio.

CACHE MAPPING

- Cache mapping refers to a technique using which the content present in the main memory is brought into the memory of the cache.
- Three distinct types of mapping are used for cache memory mapping.

When cache hit occurs,

1. The required word is present in the cache memory.
2. The required word is delivered to the CPU from the cache memory.

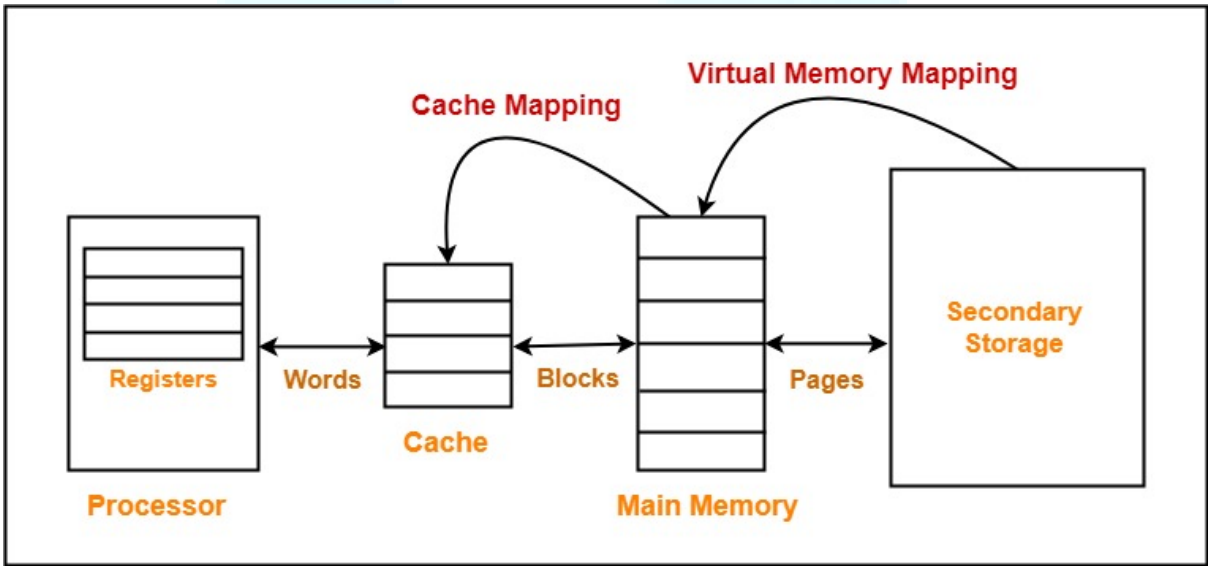
When cache miss occurs,

1. The required word is not present in the cache memory.
2. The page containing the required word has to be mapped from the main memory.
3. This mapping is performed using cache mapping techniques.

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:

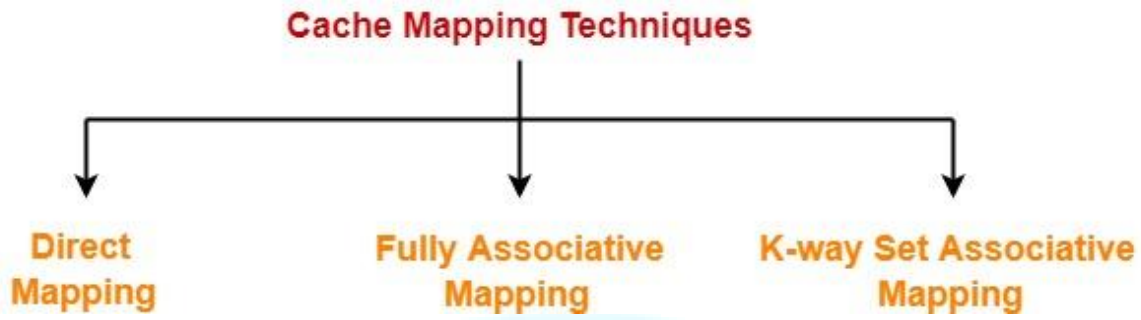


Important Note:

✓ 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.

CACHE MAPPING TECHNIQUES

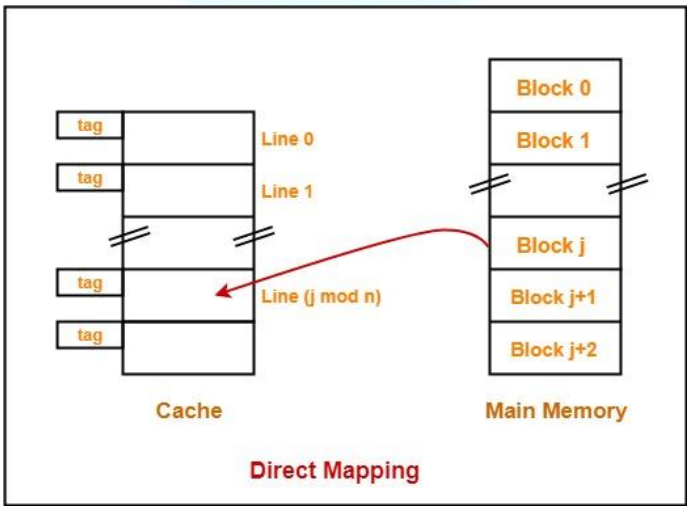


1. Direct Mapping-

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)

Example-
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.



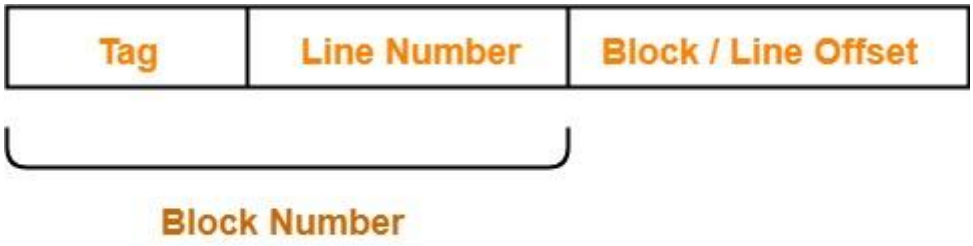
Need of Replacement Algorithm-

In direct mapping,

- ✓ There is no need of any replacement algorithm.
- ✓ This is because a main memory block can map only to a particular line of the cache.
- ✓ Thus, the new incoming block will always replace the existing block (if any) in that particular line.

Division of Physical Address-

In direct mapping, the physical address is divided as-



Division of Physical Address in Direct Mapping

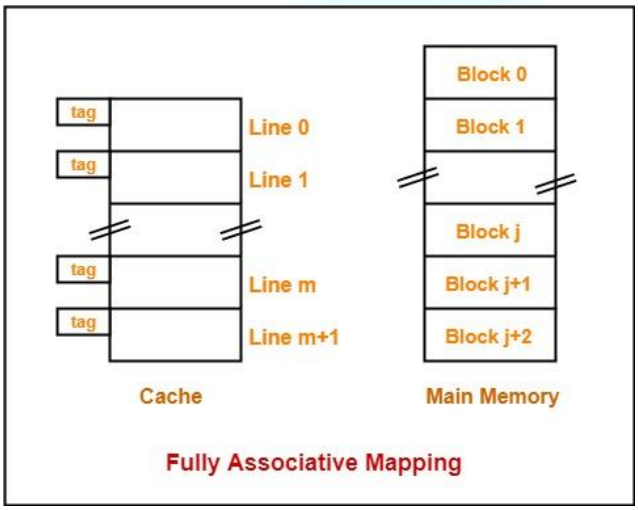
2. Fully Associative Mapping-

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.

Example-

Consider the following scenario-



Fully Associative Mapping

Here,

- ✓ 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.

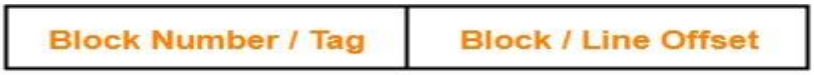
Need of Replacement Algorithm-

In fully associative mapping,

- ✓ A replacement algorithm is required.
- ✓ Replacement algorithm suggests the block to be replaced if all the cache lines are occupied.
- ✓ Thus, replacement algorithm like FCFS Algorithm, LRU Algorithm etc is employed.

Division of Physical Address-

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



Division of Physical Address in Fully Associative Mapping

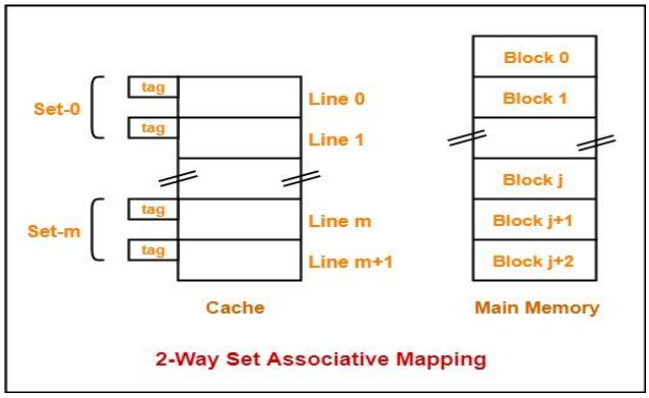
3. K-way Set Associative Mapping-

In k-way set associative mapping,

- ✓ 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)

Example-



2-Way Set Associative Mapping

Here,

- ✓ 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.

Need of Replacement Algorithm-

- ✓ Set associative mapping is a combination of direct mapping and fully associative mapping.
- ✓ It uses fully associative mapping within each set.
- ✓ Thus, set associative mapping requires a replacement algorithm.

Division of Physical Address-

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

Division of Physical Address in K-way Set Associative Mapping

Special Cases-

- ✓ 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.

Difference between Direct-mapping, Associative Mapping & Set-Associative Mapping:

Direct-mapping	Associative Mapping	Set-Associative Mapping
Needs only one comparison because of using direct formula to get the effective cache address.	Needs comparison with all tag bits, i.e., the cache control logic must examine every block's tag for a match at the same time in order to determine that a block is in the cache/not.	Needs comparisons equal to number of blocks per set as the set can contain more than 1 block.

Main Memory Address is divided into 3 fields: TAG, BLOCK & WORD. The BLOCK & WORD together make an index. The least significant TAG bits identify a unique word within a block of main memory, the BLOCK bits specify one of the blocks and the Tag bits are the most significant bits.	Main Memory Address is divided into 1 field: TAG & WORD.	Main Memory Address is divided into 3 fields: TAG, SET & WORD.
There is one possible location in the cache organization for each block from main memory because we have a fixed formula.	The mapping of the main memory block can be done with any of the cache block.	The mapping of the main memory block can be done with a particular cache block of any direct-mapped cache.
If the processor needs to access same memory location from 2 different main memory pages frequently, cache hit ratio decreases.	If the processor need to access same memory location from 2 different main memory pages frequently, cache hit ratio has no effect.	In case of frequently accessing two different pages of the main memory if reduced, the cache hit ratio reduces.
Search time is less here because there is one possible location in the cache organization for each block from main memory.	Search time is more as the cache control logic examines every block's tag for a match.	Search time increases with number of blocks per set.

MULTILEVEL CACHE ORGANISATION

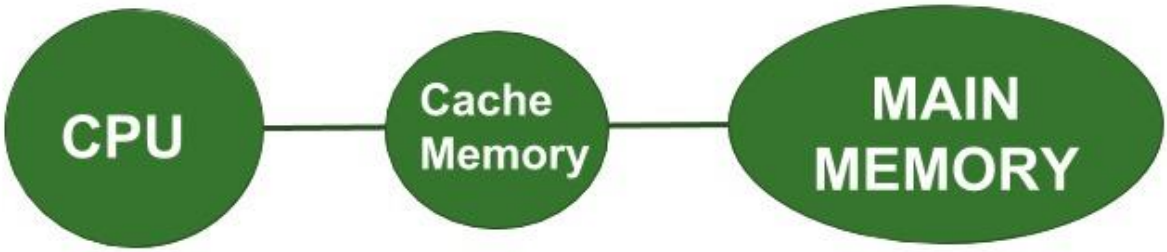
- Cache is a random access memory used by the CPU to reduce the average time taken to access memory.
- Multilevel Caches is one of the techniques to improve Cache Performance by reducing the "MISS PENALTY".
- Miss Penalty refers to the extra time required to bring the data into cache from the Main memory whenever there is a "miss" in the cache.
- For clear understanding let us consider an example where the CPU requires 10 Memory References for accessing the desired information and consider this scenario in the following 3 cases of System design :

Case 1: System Design without Cache Memory

- Here the CPU directly communicates with the main memory and no caches are involved.

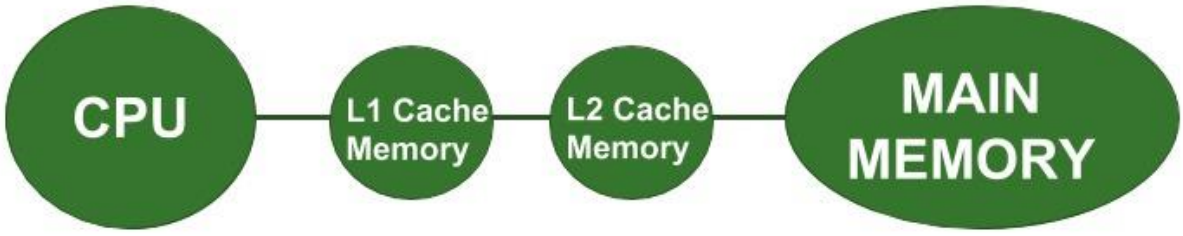
- In this case, the CPU needs to access the main memory 10 times to access the desired information.

Case 2: System Design with Cache Memory



- Here the CPU at first checks whether the desired data is present in the Cache Memory or not i.e. whether there is a "hit" in cache or "miss" in the cache.
- Suppose there is 3 miss in Cache Memory then the Main Memory will be accessed only 3 times.
- We can see that here the miss penalty is reduced because the Main Memory is accessed a lesser number of times than that in the previous case.

Case 3: System Design with Multilevel Cache Memory



- Here the Cache performance is optimized further by introducing multilevel Caches.
- As shown in the above figure, we are considering 2 levels Cache Design.
- Suppose there is 3 miss in the L1 Cache Memory and out of these 3 misses there is 2 miss in the L2 Cache Memory then the Main Memory will be accessed only 2 times.
- It is clear that here the Miss Penalty is reduced considerably than that in the previous case thereby improving the Performance of Cache Memory.

NOTE:

- We can observe from the above 3 cases that we are trying to decrease the number of Main Memory References and thus decreasing the Miss Penalty in order to improve the overall System Performance.
- Also, it is important to note that in the Multilevel Cache Design, L1 Cache is attached to the CPU and it is small in size but fast. Although, L2 Cache is attached to the Primary Cache i.e. L1 Cache and it is larger in size and slower but still faster than the Main Memory.

Effective Access Time = Hit rate * Cache access time+ Miss rate * Lower level access time

Average access Time For Multilevel Cache :(T_{avg})

Tavg = H1 * C1 + (1 – H1) * (H2 * C2 + (1 – H2) *M)

Where
H1 is the Hit rate in the L1 caches.
H2 is the Hit rate in the L2 cache.
C1 is the Time to access information in the L1 caches.
C2 is the Miss penalty to transfer information from the L2 cache to an L1 cache.
M is the Miss penalty to transfer information from the main memory to the L2 cache.

Example:
Find the Average memory access time for a processor with a 2 ns clock cycle time, a miss rate of 0.04 misses per instruction, a missed penalty of 25 clock cycles, and a cache access time (including hit detection) of 1 clock cycle. Also, assume that the read and write miss penalties are the same and ignore other write stalls.

Solution:

Average Memory access time (AMAT) = Hit Time + Miss Rate * Miss Penalty.

Hit Time = 1 clock cycle (Hit time = Hit rate * access time) but here Hit time is directly given so,

Miss rate = 0.04

Miss Penalty= 25 clock cycle (this is the time taken by the above level of memory after the hit)

So,
AMAT= 1 + 0.04 * 25
AMAT= 2 clock cycle

According to question 1 clock cycle = 2 ns

AMAT = 4ns

VIRTUAL MEMORY

- Virtual memory is a memory management technique where secondary memory can be used as if it were a part of the main memory.
- Virtual memory is a common technique used in a computer's operating system (OS).
- Virtual memory uses both hardware and software to enable a computer to compensate for physical memory shortages, temporarily transferring data from random access memory (RAM) to disk storage.

- Mapping chunks of memory to disk files enables a computer to treat secondary memory as though it were main memory.

- Today, most personal computers (PCs) come with at least 8 GB (gigabytes) of RAM. But, sometimes, this is not enough to run several programs at one time.

- This is where virtual memory comes in. Virtual memory frees up RAM by swapping data that has not been used recently over to a storage device, such as a hard drive or solid-state drive (SSD).
- Virtual memory is important for improving system performance, multitasking and using large programs.
- However, users should not overly rely on virtual memory, since it is considerably slower than RAM.
- If the OS has to swap data between virtual memory and RAM too often, the computer will begin to slow down -- this is called thrashing.
- Virtual memory was developed at a time when physical memory -- also referenced as RAM -- was expensive.
- Computers have a finite amount of RAM, so memory will eventually run out when multiple programs run at the same time.
- A system using virtual memory uses a section of the hard drive to emulate RAM. With virtual memory, a system can load larger or multiple programs running at the same time, enabling each one to operate as if it has more space, without having to purchase more RAM.

How virtual memory works

- Virtual memory uses both hardware and software to operate.
- When an application is in use, data from that program is stored in a physical address using RAM. A memory management unit (MMU) maps the address to RAM and automatically translates addresses.
- The MMU can, for example, map a logical address space to a corresponding physical address.
- If, at any point, the RAM space is needed for something more urgent, data can be swapped out of RAM and into virtual memory.
- The computer's memory manager is in charge of keeping track of the shifts between physical and virtual memory.
- If that data is needed again, the computer's MMU will use a context switch to resume execution.
- While copying virtual memory into physical memory, the OS divides memory with a fixed number of addresses into either page files or swap files.
- Each page is stored on a disk, and when the page is needed, the OS copies it from the disk to main memory and translates the virtual addresses into real addresses.
- However, the process of swapping virtual memory to physical is rather slow.
- This means using virtual memory generally causes a noticeable reduction in performance.
- Because of swapping, computers with more RAM are considered to have better performance.

Types of virtual memory

- A computer's MMU manages virtual memory operations.

- In most computers, the MMU hardware is integrated into the central processing unit (CPU).
- The CPU also generates the virtual address space. In general, virtual memory is either paged or segmented.
- Paging divides memory into sections or paging files.
- When a computer uses up its available RAM, pages not in use are transferred to the hard drive using a swap file.
- A swap file is a space set aside on the hard drive to be used as the virtual memory extension for the computer's RAM.
- When the swap file is needed, it is sent back to RAM using a process called page swapping.
- This system ensures the computer's OS and applications do not run out of real memory. The maximum size of the page file can be 1 ½ to four times the physical memory of the computer.
- The virtual memory paging process uses page tables, which translate the virtual addresses that the OS and applications use into the physical addresses that the MMU uses.
- Entries in the page table indicate whether the page is in RAM. If the OS or a program does not find what it needs in RAM, then the MMU responds to the missing memory reference with a page fault exception to get the OS to move the page back to memory when it is needed.
- Once the page is in RAM, its virtual address appears in the page table.
- Segmentation is also used to manage virtual memory. This approach divides virtual memory into segments of different lengths.
- Segments not in use in memory can be moved to virtual memory space on the hard drive.
- Segmented information or processes are tracked in a segment table, which shows if a segment is present in memory, whether it has been modified and what its physical address is.
- In addition, file systems in segmentation are only made up of segments that are mapped into a process's potential address space.
- Segmentation and paging differ as a memory model in terms of how memory is divided; however, the processes can also be combined.
- In this case, memory gets divided into frames or pages. The segments take up multiple pages, and the virtual address includes both the segment number and the page number.
- Other page replacement methods include first-in-first-out (FIFO), optimal algorithm and least recently used (LRU) page replacement.
- The FIFO method has memory select the replacement for a page that has been in the virtual address for the longest time.
- The optimal algorithm method selects page replacements based on which page is unlikely to be replaced after the longest amount of time; although difficult to implement, this leads to less page faults.
- The LRU page replacement method replaces the page that has not been used for the longest time in the main memory.

How to manage virtual memory

- Managing virtual memory within an OS can be straightforward, as there are default settings that determine the amount of hard drive space to allocate for virtual memory.
- Those settings will work for most applications and processes, but there may be times when it is necessary to manually reset the amount of hard drive space allocated to virtual

memory -- for example, with applications that depend on fast response times or when the computer has multiple hard disk drives (HDDs).

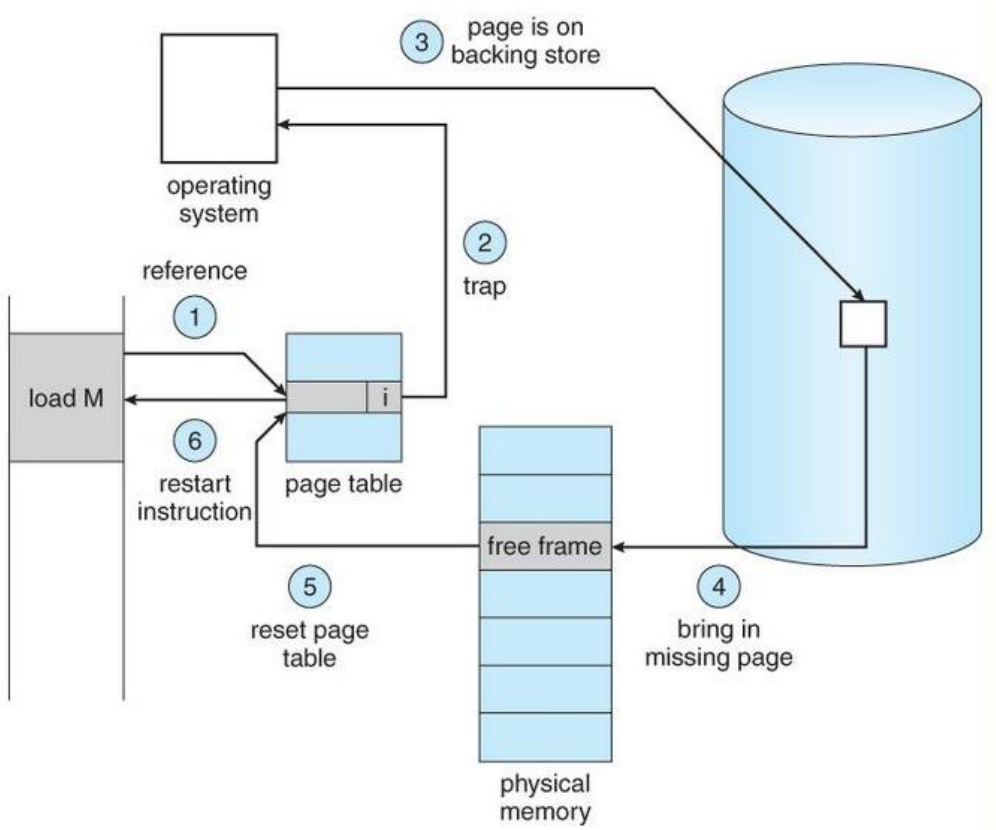
- When manually resetting virtual memory, the minimum and maximum amount of hard drive space to be used for virtual memory must be specified.
- Allocating too little HDD space for virtual memory can result in a computer running out of RAM.
- If a system continually needs more virtual memory space, it may be wise to consider adding RAM.
- Common OSes may generally recommend users not increase virtual memory beyond 1 ½ times the amount of RAM.
- Managing virtual memory differs by OS. For this reason, IT professionals should understand the basics when it comes to managing physical memory, virtual memory and virtual addresses.
- RAM cells in SSDs also have a limited lifespan.
- RAM cells have a limited number of writes, so using them for virtual memory often reduces the lifespan of the drive.

What are the benefits of using virtual memory?
The advantages to using virtual memory include:

- It can handle twice as many addresses as main memory.
- It enables more applications to be used at once.
- It frees applications from managing shared memory and saves users from having to add memory modules when RAM space runs out.
- It has increased speed when only a segment of a program is needed for execution.
- It has increased security because of memory isolation.
- It enables multiple larger applications to run simultaneously.
- Allocating memory is relatively inexpensive.
- It does not need external fragmentation.
- CPU use is effective for managing logical partition workloads.
- Data can be moved automatically.
- Pages in the original process can be shared during a fork system call operation that creates a copy of itself.
- In addition to these benefits, in a virtualized computing environment, administrators can use virtual memory management techniques to allocate additional memory to a virtual machine (VM) that has run out of resources. Such virtualization management tactics can improve VM performance and management flexibility.

PAGE FAULT

- A page fault occurs when a program attempts to access data or code that is in its address space, but is not currently located in the system RAM.
- So when page fault occurs then following sequence of events happens :



- The computer hardware traps to the kernel and program counter (PC) is saved on the stack. Current instruction state information is saved in CPU registers.
- An assembly program is started to save the general registers and other volatile information to keep the OS from destroying it.
- Operating system finds that a page fault has occurred and tries to find out which virtual page is needed. Sometimes hardware register contains this required information. If not, the operating system must retrieve PC, fetch instruction and find out what it was doing when the fault occurred.
- Once virtual address caused page fault is known, system checks to see if address is valid and checks if there is no protection access problem.
- If the virtual address is valid, the system checks to see if a page frame is free. If no frames are free, the page replacement algorithm is run to remove a page.
- If frame selected is dirty, page is scheduled for transfer to disk, context switch takes place, fault process is suspended and another process is made to run until disk transfer is completed.
- As soon as page frame is clean, operating system looks up disk address where needed page is, schedules disk operation to bring it in.
- When disk interrupt indicates page has arrived, page tables are updated to reflect its position, and frame marked as being in normal state.
- Faulting instruction is backed up to state it had when it began and PC is reset. Faulting is scheduled, operating system returns to routine that called it.
- Assembly Routine reloads register and other state information, returns to user space to continue execution.

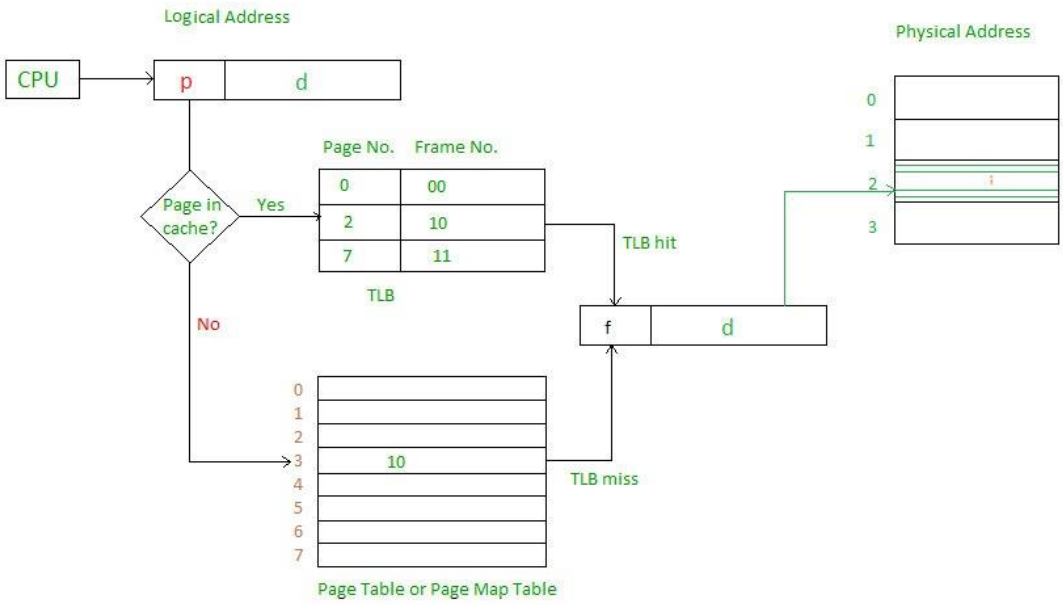
ENTRI

TRANSLATION LOOKASIDE BUFFER (TLB) IN PAGING

- For each process page table will be created, which will contain Page Table Entry (PTE).
- This PTE will contain information like frame number (The address of main memory where we want to refer), and some other useful bits (e.g., valid/invalid bit, dirty bit, protection bit etc).
- This page table entry (PTE) will tell where in the main memory the actual page is residing.
- Now the question is where to place the page table, such that overall access time (or reference time) will be less.
- The problem initially was to fast access the main memory content based on address generated by CPU (i.e. logical/virtual address).
- Initially, some people thought of using registers to store page table, as they are high-speed memory so access time will be less.
- The idea used here is, place the page table entries in registers, for each request generated from CPU (virtual address), it will be matched to the appropriate page number of the page table, which will now tell where in the main memory that corresponding page resides.
- Everything seems right here, but the problem is register size is small (in practical, it can accommodate maximum of 0.5k to 1k page table entries) and process size may be big hence the required page table will also be big (lets say this page table contains 1M entries), so registers may not hold all the PTE's of Page table. So this is not a practical approach.
- To overcome this size issue, the entire page table was kept in main memory. but the problem here is two main memory references are required:

  ✓ To find the frame number
  ✓ To go to the address specified by frame number

- To overcome this problem a high-speed cache is set up for page table entries called a Translation Lookaside Buffer (TLB).
- Translation Lookaside Buffer (TLB) is nothing but a special cache used to keep track of recently used transactions.
- TLB contains page table entries that have been most recently used.
- Given a virtual address, the processor examines the TLB if a page table entry is present (TLB hit), the frame number is retrieved and the real address is formed.
- If a page table entry is not found in the TLB (TLB miss), the page number is used as index while processing page table.
- TLB first checks if the page is already in main memory, if not in main memory a page fault is issued then the TLB is updated to include the new page entry.



Steps in TLB hit:

1. CPU generates virtual (logical) address.
2. It is checked in TLB (present).
3. Corresponding frame number is retrieved, which now tells where in the main memory page lies.

Steps in TLB miss:

1. CPU generates virtual (logical) address.
2. It is checked in TLB (not present).
3. Now the page number is matched to page table residing in main memory (assuming page table contains all PTE).
4. Corresponding frame number is retrieved, which now tells where in the main memory page lies.
5. The TLB is updated with new PTE (if space is not there, one of the replacement technique comes into picture i.e. FIFO, LRU or MFU etc).

Effective memory access time (EMAT)
TLB is used to reduce effective memory access time as it is a high speed associative cache.

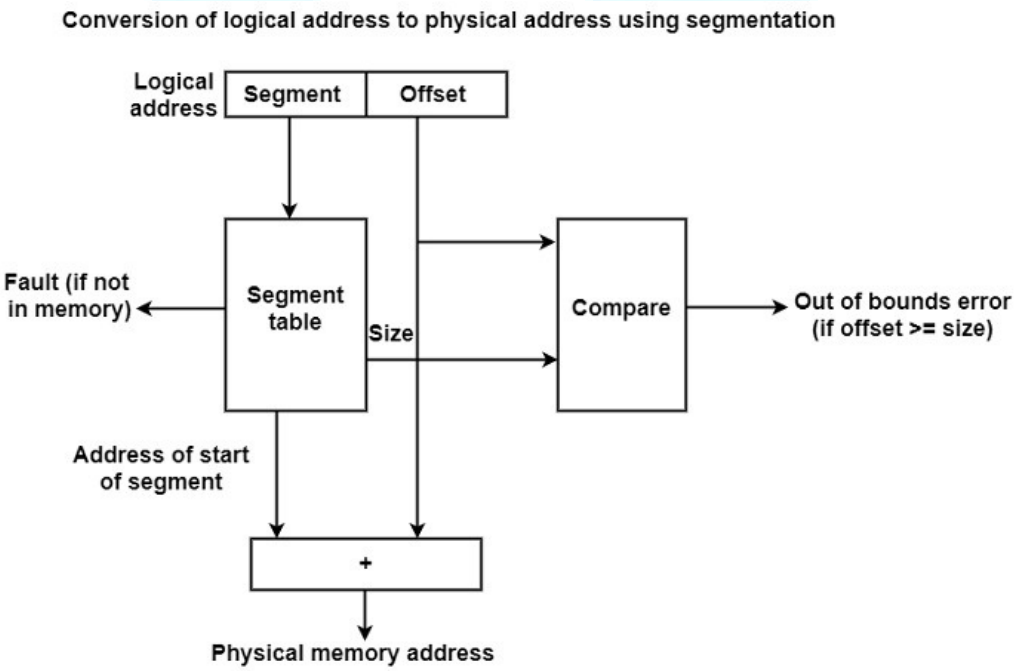
EMAT = h*(c+m) + (1-h)*(c+2m)

Where,
h = hit ratio of TLB
m = Memory access time
c = TLB access time

SEGMENTATION

- Segmentation is another approach of allocating memory that can be used rather than or in conjunction with paging.
- In its purest form, a program is broken into multiple segments, each of which is a self-contained unit, including a subroutine or data structure.
- Unlike pages, segments can vary in size.
- This requires the MMU to manage segmented memory somewhat differently than it would manage paged memory.
- A segmented MMU contains a segment table to maintain track of the segments resident in memory.
- A segment can create at one of many addresses and can be of any size; each segment table entry should contain the start address and segment size.
- Some system allows a segment to start at any address, while other limits the start address.
- One such limit is found in the Intel X86 architecture, which requires a segment to start at an address that has 6000 as its four low-order bits.
- A system with this limitation would not need to store these four bits in its segment table, since their value is implicit.

As shown in the figure, it shows a simplified address translation scheme for segmented memory.



- The offset is compared to the segment size.
- If the offset is higher than or the same as the segment size, signifying that the location is not a part of the segment, an error is created.
- If the offset has a true value, it is inserted to the beginning of the segment address to create the correct physical memory addresses.
- As with paging, a segmented MMU can also have a TLB to speed up the generation of segment start addresses and sizes.

- In paging, the page number is sent to the page table (and TLB) to produce a frame number.
- This value is concatenated with the offset to produce the physical address.
- In segmentation, the start address generated by the segment table or TLB is added to the offset, a process much more time-consuming than concatenation.
- Because segments can have different sizes, this method has both advantages and disadvantages as compared to paging.
- Consider the paging example for the relatively simple CPU.
- In the paged memory execution, each page is of size 4K.
- A program of size 4K + 1 would need the MMU to allocate two pages of memory, even though the second page would use only one of its 4K locations.
- This is referred to as internal fragmentation.
- If segmentation is used, a segment of exactly size 4K + 1 can be allocated, thus avoiding this problem.
- Segmentation has a problem known as external fragmentation.
- There are three segments resident in memory and 8K of free space.
- The free is subdivided such that no segment higher than 3K can be loaded into memory without changing or deleting one of the currently loaded segments.
- It introduces overhead by moving data back to the swap disk or relocating it in memory, either of which reduces system performance.

MEMORY MODULE

- A memory module is a circuit board with DRAM integrated circuits that are installed into the memory slot on a computer motherboard.
- Below is an image of a 512 MB DIMM computer memory module and the most common type of memory used today.
- Memory modules permit easy installation and replacement in electronic systems, especially computers such as personal computers, workstations, and servers.

Types of memory module include:

- TransFlash Memory Module
- SIMM, a single in-line memory module
- DIMM, dual in-line memory module

TRANSFLASH MEMORY MODULE

- The TransFlash product is an ultra small, semi-removable flash memory module based on the miniSD card and TriFlash designs for future mobile phone products, especially the transfer of personal content between TransFlash-enabled phones.
- Due to the ultra small size of the product, it is not intended to be handled or removed on a frequent basis. The TransFlash form factor will be promoted globally to other mobile phone manufacturers.

Advanced Features:

• Semi-removable Memory Module
• Expands memory capacity for mobile phones
• The world's smallest removable storage module
• Compatible with all TransFlash mobile devices
• The TransFlash memory module can be inserted into an adapter and used in other SD-enabled devices

Product Specifications:
• Dimensions: 15 x 11 x 1 mm (L x W x D)
• Capacity: 64, 128, or 256MB Flash Memory
• 5 year warranty

SIMM (SINGLE IN-LINE MEMORY MODULE)

- In the case of SIMM, the connectors are only present on the single side of the module and are shorted together.
- SIMMs is always used in matched-pairs.
- The maximum data storage offered by SIMM is 32-bit/cycle and voltage consumption is 5 volts.
- As technology evolved, SIMM became obsolete and was replaced by DIMM.

DIMM (DUAL IN-LINE MEMORY MODULE)

- DIMM has the row of connectors on both the sides (front and back) of the module and connectors are independent.
- This resulted in twice the capacity of DIMM with the same quantity of RAM hence supporting the 64-bit processors.
- While two SIMM sticks would be used in parallel for 64-bit data width (which is a disadvantage!).
- The voltage consumption of DIMM is 3.3 volts which are comparatively lower.
- It is not backward compatible i.e. it cannot be used on motherboards having SIMM slots.
- It is easier to replace damaged or corrupted RAM piece on DIMM.

- This proves that DIMM clearly outperforms SIMM in speed, latency, and power consumption. DIMM is generally available in 168, 184, 214 or 244 pins.

Classification of DIMM

- DIMM can be classified on the basis of buffer size and type of RAM:

DIMM classification based on buffer size:

1. **Unbuffered DIMM (UDIMM):**

- The system directly reads/writes from/to memory chip without validation hence increasing the electrical load on the motherboard but are very faster.

2. **Registered DIMM (RDIMM):**

- Uses register that buffers signals, hence increasing clock cycle but are more reliable.

DIMM classification based on the type of RAM:

1. **SDRAM (synchronous dynamic RAM) DIMM:**

   It was the first dynamic RAM to sync with the system clock. The refresh rate was much lower due to re-accessing data after the rising half cycle.

2. **SDR(single data rate) DIMM:**

   Single data rate means the packet of data is only accessed once per clock cycle. Serial data can be read via the serial data pins on the DIMM which enables the motherboard to auto configures to the exact type of DIMM installed.

3. **DDR(double data rate) DIMM:**

   Data packet is accessed twice each clock cycle. DDR DIMMs also use two notches on each side to enable compatibility with both low- and high-profile latched sockets.

4. **DDR2 DIMM:**

   The key difference between DDR and DDR2 is that in DDR2 the bus is clocked at twice the speed of the memory cells, so data can be transferred four times faster per memory cell cycle.

How to select proper DIMM?

- DIMM sizes vary from micro ATX to standard motherboards so no. of pins is an important factor.
- While purchasing a RAM stick, ratings are like 16 GB RAM can be 1X16GB (1 DIMM and 16 GB RAM each), 2X8GB (2 DIMM and 8 GB RAM each) or 4X4GB (4 DIMM and 4 GB RAM each).
- Operating frequency and maximum over clocking frequency should also be taken into consideration.

DIFFERENCE BETWEEN SIMM AND DIMM

Single In-Line Memory Modules (SIMM) is that the little circuit boards having notches wherever the RAM chips are fixed. SIMM connectors and therefore the slot situated on the motherboard are created of metal (gold or tin). In SIMM, Pins present in either facet are connected. There are two types of SIMM presents, one with 30 pins and another one is with 72 pins.

Dual In-Line Memory Module (DIMM), additionally has metal connectors almost like SIMM, however either of the perimeters of the connective doesn't admit the opposite. DIMM supports 64 bit channel for data transferring while SIMM supports only maximum of 32 bit

channel. There are three type of DIMM presents which are used by modern motherboard, one with 168 pins and second one is with 184 pins and third one is 240 pins.

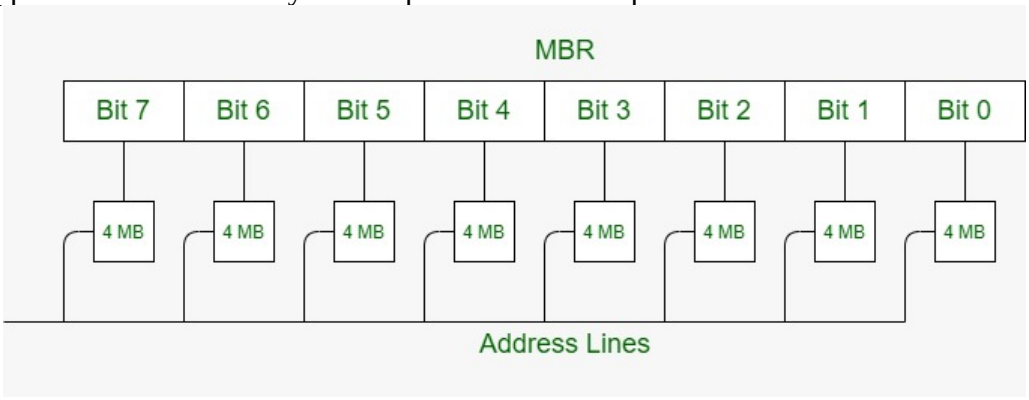
SIMM have just one usable aspect because of having only one set of the instrumentality whereas DIMM have completely different signal pins at either side that square measure usable and doesn't believe the opposite side.

Let's see that the difference b/w SIMM and DIMM:

SL NO	SIMM	DIMM
1.	In SIMM, Pins present in either facet are connected.	DIMM pins are freelance.
2.	SIMM supports 32 bit channel for data transferring.	DIMM supports 64 bit channel for data transferring.
3.	SIMM consumes 5 volts of power.	DIMM consumes 3.3 volts of power.
4.	SIMM provides the storage 4 MB to 64 MB.	DIMM provides the storage 32 MB to 1 GB.
5.	The classic or most common pin configuration of the SIMM module is 72 pins.	The foremost common pin configuration of the DIMM module is 168 pins.
6.	SIMMs are the older Technology.	DIMMs are the replacement of the SIMMs.

INTERLEAVING

- Memory Interleaving is an abstraction technique which divides memory into a number of modules such that successive words in the address space are placed in the different module.
- Suppose a 64 MB memory made up of the 4 MB chips as shown in the below



- We organize the memory into 4 MB banks, each having eight of the 4 MB chips. The memory thus has 16 banks, each of 4 MB.
- 64 MB memory = 2^26, so 26 bits are used for addressing.
- 16 = 2^4, so 4 bits of address select the bank, and 4 MB = 2^22, so 22 bits of address to each chip.

In general, an N-bit address, with N = L + M, is broken into two parts:

1. L-bit bank select, used to activate one of the 2^L banks of memory, and
2. M-bit address that is sent to each of the memory banks.

When one of the memory banks is active, the other (2L – 1) are inactive. All banks receive the M-bit address, but the inactive one does not respond to it.

Classification of Memory Interleaving:

Memory interleaving is classified into two types:

1. **High Order Interleaving –**

- In high-order interleaving, the most significant bits of the address select the memory chip.
- The least significant bits are sent as addresses to each chip. One problem is that consecutive addresses tend to be in the same chip.
- The maximum rate of data transfer is limited by the memory cycle time.
- It is also known as Memory Banking.

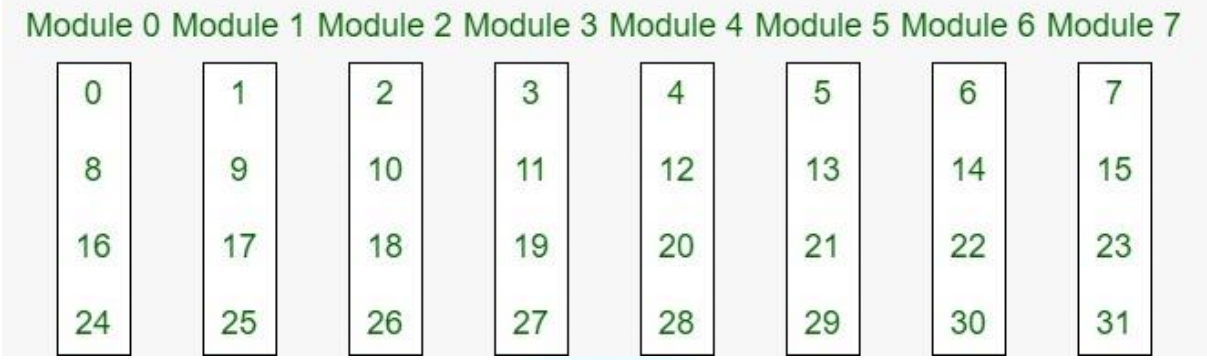
Address bits	25-22	21-0
Use	Bank Select	Address to the chip

Module 0	Module 1	Module 2	Module 3	Module 4	Module 5	Module 6	Module 7
0	4	8	12	16	20	24	28
1	5	9	13	17	21	25	29
2	6	10	14	18	22	26	30
3	7	11	15	19	23	27	31

2. **Low Order Interleaving –**

- In low-order interleaving, the least significant bits select the memory bank (module). In this, consecutive memory addresses are in different memory modules.
- This allows memory access at much faster rates than allowed by the cycle time.

Address bits	25-4	3-0
Use	Address to the chip	Bank Select



SECONDARY STORAGE

- A secondary storage device refers to any non-volatile storage device that is internal or external to the computer.
- It can be any storage device beyond the primary storage that enables permanent data storage.
- A secondary storage device is also known as an auxiliary storage device, backup storage device, tier 2 storage, or external storage.
- These devices store virtually all programs and applications on a computer, including the operating system, device drivers, applications and general user data.
- The Secondary storage media can be fixed or removable.
- Fixed Storage media is an internal storage medium like a hard disk that is fixed inside the computer.
- A storage medium that is portable and can be taken outside the computer is termed removable storage media.

- The main advantage of using secondary storage devices is:

❖ In Secondary storage devices, the stored data might not be under the direct control of the operating system. For example, many organizations store their archival data or critical documents on secondary storage drives, which their main network cannot access to ensure their preservation whenever a data breach occurs.

❖ Since these drives do not interact directly with the main infrastructure and can be situated in a remote or secure site, it is unlikely that a hacker may access these drives unless they're physically stolen.

NEED OF SECONDARY STORAGE

- Computers use main memory such as random access memory (RAM) and cache to hold data that is being processed.
- However, this type of memory is volatile, and it loses its data when the computer is switched off.

- General-purpose computers, such as personal computers and tablets, need to store programs and data for later use.
- That's why secondary storage is needed to keep programs and data long term.
- Secondary storage is non-volatile and able to keep data as long term storage.
- They are used for various purposes such as backup data used for future restores or disaster recovery, long-term archiving of data that is not frequently accessed, and storage of non-critical data in lower-performing, less expensive drives.
- Without secondary storage, all programs and data would be lost when the computer is switched off.

CHARACTERISTICS OF SECONDARY STORAGE DEVICES

- These are some characteristics of secondary memory, which distinguish it from primary memory, such as:

1. It is non-volatile, which means it retains data when power is switched off
2. It allows for the storage of data ranging from a few megabytes to petabytes.
3. It is cheaper as compared to primary memory.
4. Secondary storage devices like CDs and flash drives can transfer the data from one device to another.

TYPES OF SECONDARY STORAGE DEVICE

Here are the two types of secondary storage devices.

- **FIXED STORAGE**
- **REMOVABLE STORAGE.**

1. Fixed Storage

- Fixed storage is an internal media device used by a computer system to store data. Usually, these are referred to as the fixed disk drives or Hard Drives.
- Fixed storage devices are not fixed.
- These can be removed from the system for repairing work, maintenance purposes, and also for an upgrade, etc.
- But in general, this cannot be done without a proper toolkit to open up the computer system to provide physical access, which needs to be done by an engineer.
- Technically, almost all data, i.e. being processed on a computer system, is stored on some built-in fixed storage device.

- We have the following types of fixed storage:

  - Internal flash memory (rare)
  - SSD (solid-state disk) units
  - Hard disk drives (HDD)

2. Removable Storage

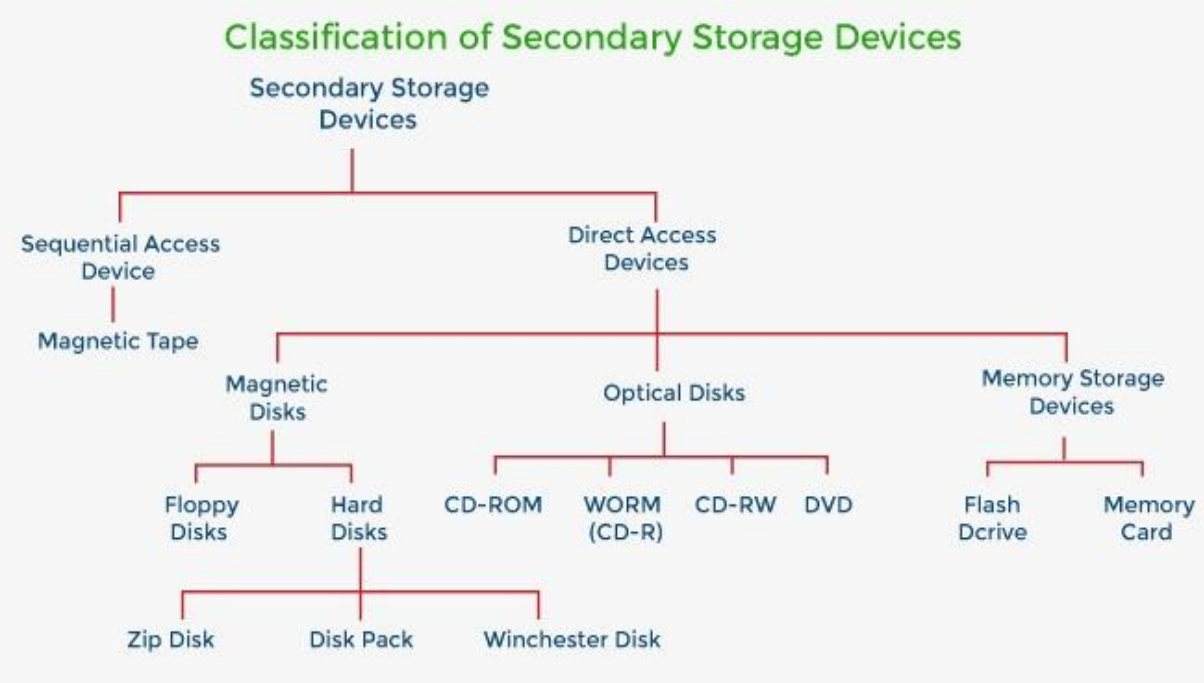
- Removable storage is an external media device that is used by a computer system to store data.
- Usually, these are referred to as the Removable Disks drives or the External Drives.
- Removable storage is any storage device that can be removed from a computer system while the system is running.
- Examples of external devices include CDs, DVDs, Blu-ray disk drives, and diskettes and USB drives.
- Removable storage makes it easier for a user to transfer data from one computer system to another.
- The main benefit of removable disks in storage factors is that they can provide the fast data transfer rates associated with storage area networks (SANs).

    - We have the following types of Removable Storage:

- ✓ **Optical discs (CDs, DVDs, Blu-ray discs)**
- ✓ **Memory cards**
- ✓ **Floppy disks**
- ✓ **Magnetic tapes**
- ✓ **Disk packs**
- ✓ **Paper storage (punched tapes, punched cards)**

CLASSIFICATION OF SECONDARY STORAGE DEVICES

The following image shows the classification of commonly used secondary storage devices.



Sequential Access Storage Device

- It is a class of data storage devices that read stored data in a sequence.
- This is in contrast to random access memory (RAM), where data can access in any order, and magnetic tape is the common sequential access storage device.

Magnetic tape:

- It is a medium for magnetic recording, made of a thin, magnetisable coating on a long, narrow strip of plastic film.
- Devices that record and play audio and video using magnetic tape are tape recorders and videotape recorders.
- A device that stores computer data on magnetic tape is known as a tape drive.
- It was a key technology in early computer development, allowing unparalleled amounts of data to be mechanically created, stored for long periods, and rapidly accessed.

Direct Access Storage Devices

- A direct-access storage device (DASD) is another name for secondary storage devices that store data in discrete locations with a unique address, such as hard disk drives, optical drives and most magnetic storage devices.

1. **Magnetic disks:**

- A magnetic disk is a storage device that uses a magnetization process to write, rewrite and access data.
- It is covered with a magnetic coating and stores data in the form of tracks, spots and sectors. Hard disks, zip disks and floppy disks are common examples of magnetic disks.

(A) Floppy Disk:

- A floppy disk is a flexible disk with a magnetic coating on it, and it is packaged inside a protective plastic envelope.
- These are among the oldest portable storage devices that could store up to 1.44 MB of data, but now they are not used due to very little memory storage.

(B) Hard Disk Drive (HDD):

- Hard disk drive comprises a series of circular disks called platters arranged one over the other almost ½ inches apart around a spindle.
- Disks are made of non-magnetic material like aluminium alloy and coated with 10-20 nm magnetic material.
- The standard diameter of these disks is 14 inches, and they rotate with speeds varying from 4200 rpm (rotations per minute) for personal computers to 15000 rpm for servers.
- Data is stored by magnetizing or demagnetizing the magnetic coating.
- A magnetic reader arm is used to read data from and write data to the disks. A typical modern HDD has a capacity in terabytes (TB).

2. **Optical Disk:**

- An optical disk is any computer disk that uses optical storage techniques and technology to read and write data. It is a computer storage disk that stores data digitally and uses laser beams to read and write data.

(I) CD Drive:

- CD stands for Compact Disk.
- CDs are circular disks that use optical rays, usually lasers, to read and write data.
- They are very cheap as you can get 700 MB of storage space for less than a dollar.
- CDs are inserted in CD drives built into the CPU cabinet.
- They are portable as you can reject the drive, remove the CD and carry it with you.

- There are three types of CDs:

CD-ROM (Compact Disk - Read Only Memory):

- The manufacturer recorded the data on this CDs. Proprietary Software; audio or video are released on CD-ROMs.

CD-R (Compact Disk - Recordable):

- The user can write data once on the CD-R. It cannot be deleted or modified later.

CD-RW (Compact Disk - Rewritable):

- Data can repeatedly be written and deleted on these optical disks.

(II) DVD Drive:

- DVD stands for digital video display.
- DVD is an optical device that can store 15 times the data held by CDs.
- They are usually used to store rich multimedia files that need high storage capacity.
- DVDs also come in three varieties –
    - **Read-only**
    - **Recordable**
    - **Rewritable.**

(III) Blu Ray Disk:

- Blu Ray Disk (BD) is an optical storage media that stores high definition (HD) video and other multimedia files.
- BD uses a shorter wavelength laser than CD/DVD, enabling the writing arm to focus more tightly on the disk and pack in more data.
- BDs can store up to 128 GB of data.

3. Memory Storage Devices:

- A memory device contains trillions of interconnected memory cells that store data.
- When switched on or off, these cells hold millions of transistors representing 1s and 0s in binary code, allowing a computer to read and write information.
- It includes USB drives, flash memory devices, SD and memory cards, which you'll recognize as the storage medium used in digital cameras.

(i) Flash Drive:

- A flash drive is a small, ultra-portable storage device.
- USB flash drives were essential for easily moving files from one device to another.
- Flash drives connect to computers and other devices via a built-in USB Type-A or USB-C plug, making one a USB device and cable combination.
- Flash drives are often referred to as pen drives; thumb drives, or jump drives.
- The terms USB drive and solid-state drive (SSD) are also sometimes used, but most of the time, those refer to larger, not-so-mobile USB-based storage devices like external hard drives.
- These days, a USB flash drive can hold up to 2 TB of storage.
- They're more expensive per gigabyte than an external hard drive, but they have prevailed as a simple, convenient solution for storing and transferring smaller files.

Pen drive has the following advantages in computer organization, such as:

- **Transfer Files:** A pen drive is a device plugged into a USB port of the system that is used to transfer files, documents, and photos to a PC and vice versa.
- **Portability:** The lightweight nature and smaller size of a pen drive make it possible to carry it from place to place, making data transportation an easier task.
- **Backup Storage:** Most of the pen drives now come with the feature of having password encryption, important information related to family, medical records, and photos can be stored on them as a backup.
- **Transport Data:** Professionals or Students can now easily transport large data files and video, audio lectures on a pen drive and access them from anywhere. Independent PC technicians can store work-related utility tools, various programs, and files on a high-speed 64 GB pen drive and move from one site to another.

(ii)Memory card:

- A memory card or memory cartridge is an electronic data storage device used for storing digital information, typically using flash memory.
- These are commonly used in portable electronic devices, such as digital cameras, mobile phones, laptop computers, tablets, PDAs, portable media players, video game consoles, synthesizers, electronic keyboards and digital pianos, and allow adding memory to such devices without compromising ergonomy, as the card is usually contained within the device rather than protruding like USB flash drives.