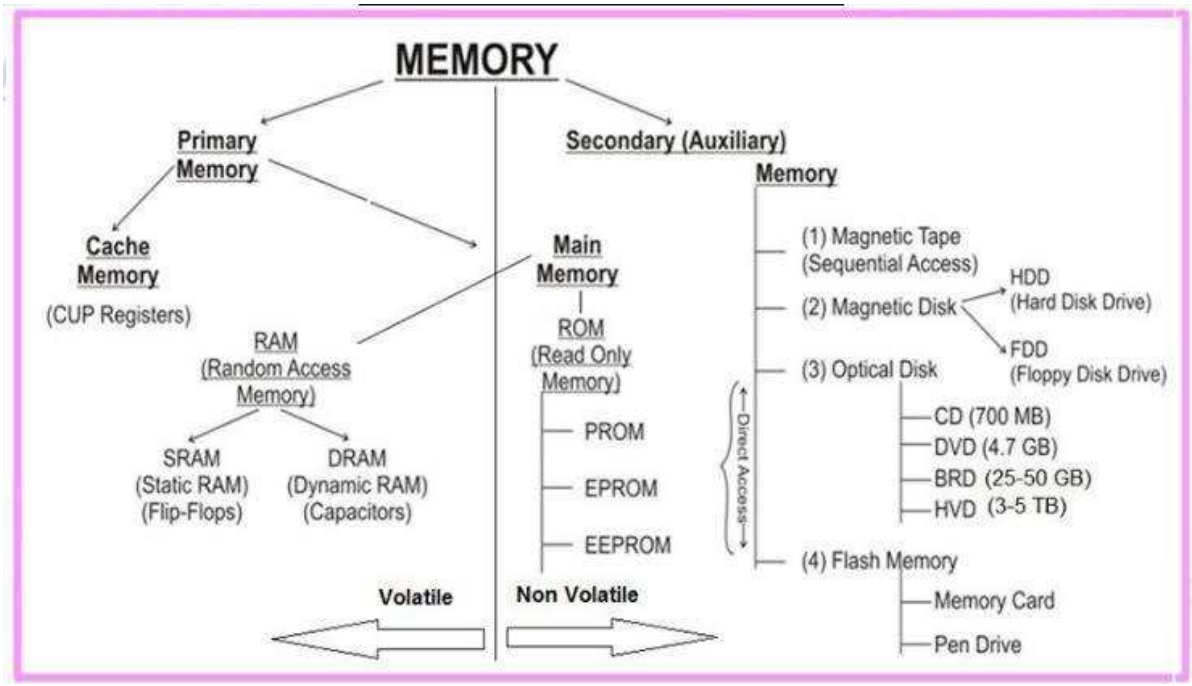
COMPUTER MEMORY

DR. PARIMITA SINGH
FACULTY, S.S. IN COMMERCE,
V. U, UJJAIN (M.P.)

A memory is just like a human brain. It is used to store data and instructions. Computer memory is the storage space in the computer, where data is to be processed and instructions required for processing are stored. 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.



- Primary Storage (memory), also known as main storage and it is the area in a computer in which data is stored for quick access by the computer's processor. The terms random access memory (RAM) and memory are often as synonyms for primary or main storage. Primary storage is volatile (that can change suddenly and unexpectedly)and can be contrasted with non-volatile secondary storage (retaining data even if there is a break in the power supply) also known as auxiliary storage.

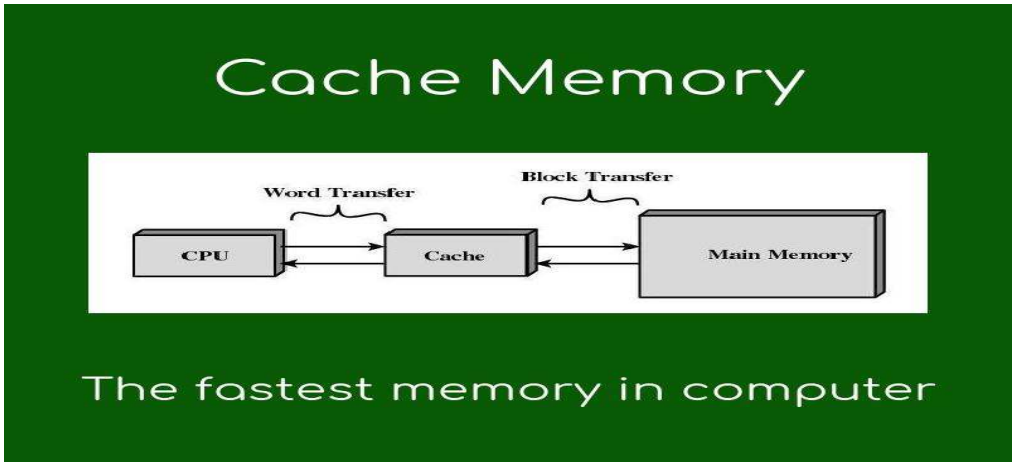
- Cache memory is a smaller, faster memory which stores copies of the data from frequently used main memory locations. A CPU cache is a hardware cache used by the central processing unit (CPU) of a computer to reduce the average time to access data from the main memory.

In other words, Cache memory is a small-sized volatile computer memory that provides high-speed data access to a processor and stores frequently used computer programs, applications and data. Cache memory also called CPU memory, which is placed between random access memory (RAM) and a computer microprocessor. It

can be accessed quicker by microprocessor than regular RAM.It is designed to speed up the transfer of data and instructions. The data and instructions are retrieved from RAM when the CPU uses them for the first time. A copy of that data or instructions is stored in a cache. The next time the CPU needs that data or instructions, it first looks in a cache. If the required data is found there, it is retrieved from cache memory instead of main memory. It speeds up the working of CPU.

The purpose of cache memory is to store program instructions and data that are used repeatedly in the operation of programs or information that the CPU is likely to need next. The computer processor can access this information quickly from the cache rather than having to get it from the computer's main memory. Fast access to these instructions increases the overall speed of the program. A computer can have several different levels of cache memory. The level numbers refer to distance from CPU where Level 1 is the closest. All levels of cache memory are faster than RAM. The cache closest to CPU is always faster but generally costs more and stores less data than other levels of cache. The cache memory works according to various algorithms, which decide what information it has to store. These algorithms work out the probability to decide which data would be most frequently needed. This probability is worked out on the basis of past observations. In addition to hardware-based cache, cache memory also can be a disk cache, where a reserved portion on a disk stores and provides access to frequently accessed data from the disk. Cache memory generally tends to operate in a number of different configurations: direct mapping, fully associative mapping and set associative mapping. Direct mapping features blocks of memory mapped to specific locations within the cache, while fully associative mapping lets any cache location be used to map a block, rather than requiring the location to be pre-set. Set associative mapping acts as a halfway-house between the two, in that every block is mapped to a smaller subset of locations within the cache.

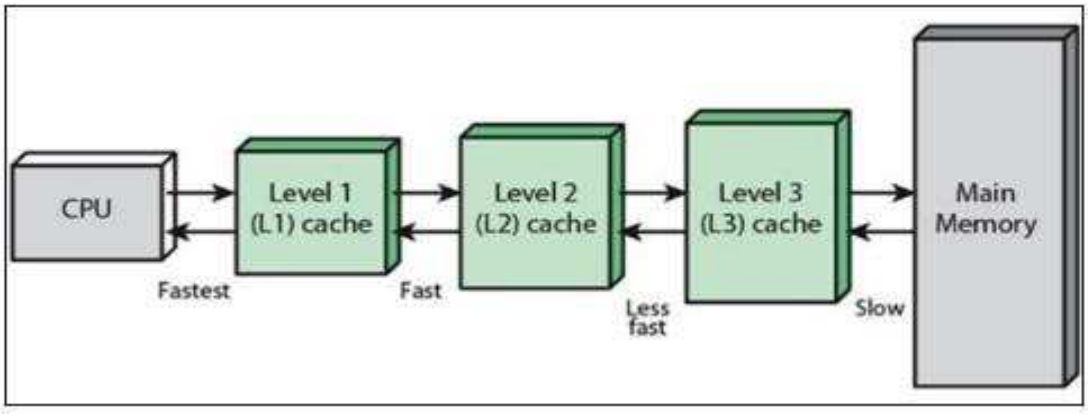
Types of Cache



Primary Cache (L1) - A primary cache is always located on the processor chip. This cache is small and its access time is comparable to that of processor registers.

Secondary Cache (L2) - Secondary cache is placed between the primary cache and the rest of the memory. It is referred to as the level 2 (L2) cache. Often, the Level 2 cache is also housed on the processor chip.

Main Memory (L3) - The L3 cache is larger in size but also slower in speed than L1 and L2. In Multicore processors, each core may have separate L1 and L2, but all core share a common L3 cache. L3 cache double speed than the RAM.



Graphics processing chips often have a separate cache memory to the CPU, which ensures that the GPU can still speedily complete complex rendering operations without relying on the relatively high-latency system RAM.

- Secondarymemory is where programs and data are kept on a long-term basis. Common secondary storage devices are the hard disk and optical disks. The hard disk has enormous storage capacity compared to main memory. The hard disk is usually contained inside the case of a computer.

- Read-only memory (ROM) is a storage medium used in computers and other electronic devices. Data stored in ROM can only be modified slowly or with difficulty, or not at all.ROM is non-volatile and the contents are retained even after the power is switched off.

The types of ROM include PROM, EPROM and EEPROM.

1. PROM - (programmable read-only memory) is a memory chip on which data can be written only once. The difference between a PROM and a ROM (read-only memory) is that a PROM is manufactured as blank memory, whereas a ROM is programmed during the manufacturing process. To write data onto a PROM chip, you need a special device called a PROM programmer or PROM burner.
2. EPROM - (erasable programmable read-only memory) is a special type of PROM that can be erased by exposing it to ultraviolet light. EPROM, in full erasable programmable read-only memory, Form of computer memory that does not lose its content when the power supply is cut off and that can be erased and reused. EPROMs are generally employed

for programs designed for repeated use (such as the BIOS) but that can be upgradedwith a later version of the program.

Advantages of EPROM

Some of the advantages of EPROM are as follows:

- EPROM is non-volatile so it retains its memory even without power. So no external memory is required.
- EPROM is quite effective.
- EPROM is reprogrammable i.e. the data in the EPROM can be erased and reprogrammed.

Disadvantages of EPROM

Some of the disadvantages of EPROM are as follows:

- Transistors used in EPROM have a higher resistance.
- The EPROM needs UV light to erase the data. This can't be done using electrical signals.
- It is not possible to erase a particular byte of data in EPROM. The whole data is deleted.
- The static power consumption of EPROM is quite high.
- It takes some time to erase the data in EPROM. This is different than EEPROM where the data can be instantaneously erased.

3. EEPROM - (electrically erasable programmable read-only memory). EEPROM is a special type of PROM that can be erased by exposing it to an electrical charge. EEPROM is programmed and erased electrically. It can be erased and reprogrammed about ten thousand times. Both erasing and programming take about 4 to 10 ms (millisecond). In EEPROM, any location can be selectively erased and programmed. EEPROMs can be erased one byte at a time, rather than erasing the entire chip. Hence, the process of reprogramming is flexible but slow.

Benefits or advantages of EEPROM

Following are the benefits or advantages of EEPROM:

- ➡ The method of erasure is electrical and immediate.
- ➡ It is possible to erase entire contents of EEPROM as well as particular byte as per selection.
- ➡ It is very easy to program and erase the contents of EEPROM without removing it from board or test jig. The designers incorporate circuitry to program/erase the EEPROM in the board itself.
- ➡ To change the contents, additional equipments are not required.
- ➡ Electrical interfaces of different types viz. serial bus and parallel bus are available.
- ➡ It is possible to re-program EEPROM infinite number of times.

Drawbacks or disadvantages of EEPROM

Following are the disadvantages of EEPROM:

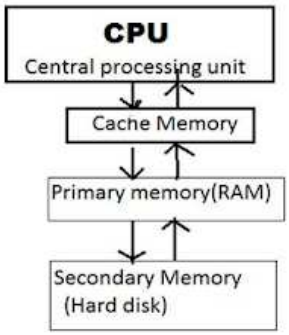
- ➡ EEPROM requires different voltages for erasing, reading and writing the data.
- ➡ EEPROM has limited data retention time period which is approx. 10 years for most of the devices.
- ➡ External serial EEPROM type requires long time to access. Hence it is advisable to select appropriate EEPROM type based on application of use.
- ➡ EEPROM devices are expensive compare to PROMs and EPROMs.

- ❖ Random Access Memory (RAM), allows the computer to store data for immediate manipulation and to keep track of what is currently being processed.RAM is referred to as volatile memory and is lost when the power is turned off.It also known as read/write memory as information can be read from and written onto it.

  The two main types of RAM are Static RAM and Dynamic RAM.

  1. SRAM retains data as long as power is provided to the memory chip and need not be refreshed periodically. It is often used as CPU Cache memory. SRAM stands for Static Random Access Memory.
  2. The data on DRAM continues to move in and out of the memory as long as power is available and must be continually refreshed to maintain the data. DRAM stands for Dynamic Random Access Memory.

- ❖ Virtual memory is memory on the hard disk that the CPU uses as an extended RAM.



	Access Time	Storage Capacity	Cost per bit of storage
Primary memory	Faster	Smaller	High
Secondary memory	Slower	Higher	Low

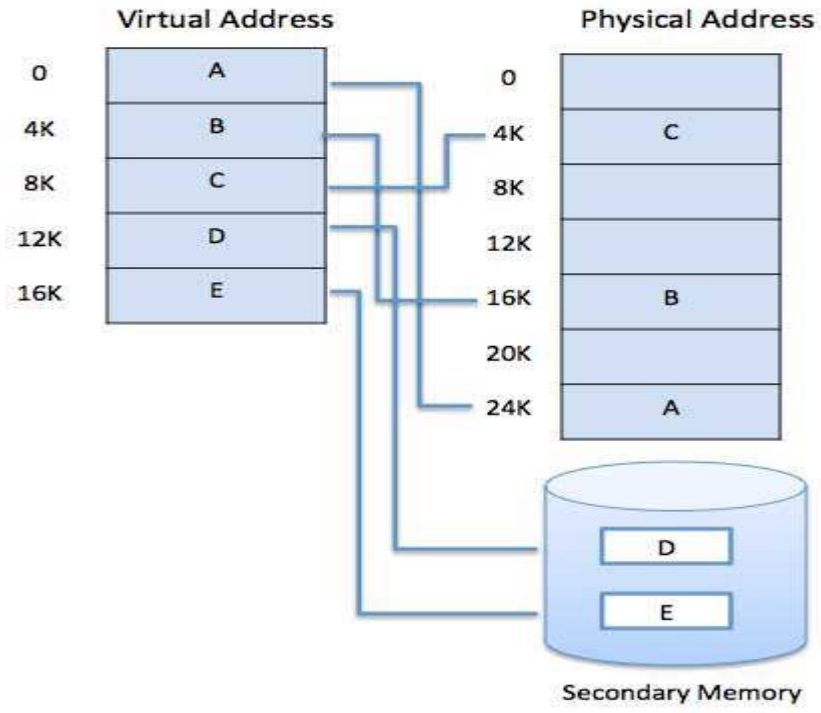
A computer can address more memory than the amount physically installed on the system. This extra memory is actually called virtual memory and it is a section of a hard disk that's set up to emulate the computer's RAM.

The main visible advantage of this scheme is that programs can be larger than physical memory. Virtual memory serves two purposes. First, it allows us to extend the use of physical memory by using disk. Second, it allows us to have memory protection, because each virtual address is translated to a physical address.

Following are the situations, when entire program is not required to be loaded fully in main memory.

- User written error handling routines are used only when an error occurred in the data or computation.
- Certain options and features of a program may be used rarely.
- Many tables are assigned a fixed amount of address space even though only a small amount of the table is actually used.
- The ability to execute a program that is only partially in memory would counter many benefits.
- Less number of I/O would be needed to load or swap each user program into memory.
- A program would no longer be constrained by the amount of physical memory that is available.
- Each user program could take less physical memory, more programs could be run the same time, with a corresponding increase in CPU utilization and throughput.

Modern microprocessors intended for general-purpose use, a memory management unit, or MMU, is built into the hardware. The MMU's job is to translate virtual addresses into physical addresses. A basic example is given below −

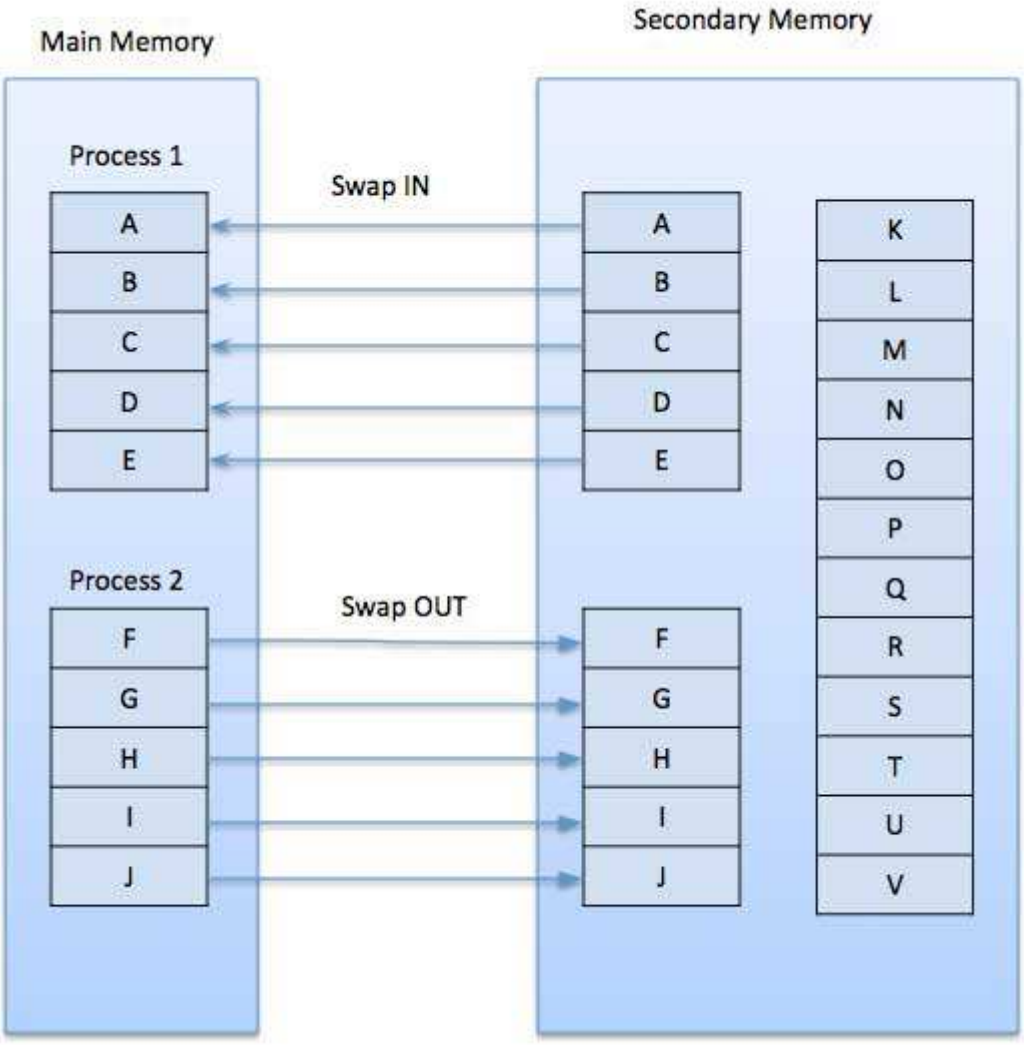


Virtual memory is commonly implemented by demand paging. It can also be implemented in a segmentation system. Demand segmentation can also be used to provide virtual memory.

Demand Paging

A demand paging system is quite similar to a paging system with swapping where processes reside in secondary memory and pages are loaded only on demand, not in advance. When a context switch occurs, the operating system does not copy any

of the old program's pages out to the disk or any of the new program's pages into the main memory Instead, it just begins executing the new program after loading the first page and fetches that program's pages as they are referenced.



While executing a program, if the program references a page which is not available in the main memory because it was swapped out a little ago, the processor treats this invalid memory reference as a page fault and transfers control from the program to the operating system to demand the page back into the memory.

Advantages

Following are the advantages of Demand Paging −

- Large virtual memory.
- More efficient use of memory.
- There is no limit on degree of multiprogramming.

Disadvantages

- Number of tables and the amount of processor overhead for handling page interrupts are greater than in the case of the simple paged management techniques.

Page Replacement Algorithm

Page replacement algorithms are the techniques using which an Operating System decides which memory pages to swap out, write to disk when a page of memory needs to be allocated. Paging happens whenever a page fault occurs and a free

page cannot be used for allocation purpose accounting to reason that pages are not available or the number of free pages is lower than required pages.

When the page that was selected for replacement and was paged out, is referenced again, it has to read in from disk, and this requires for I/O completion. This process determines the quality of the page replacement algorithm: the lesser the time waiting for page-ins, the better is the algorithm.

A page replacement algorithm looks at the limited information about accessing the pages provided by hardware, and tries to select which pages should be replaced to minimize the total number of page misses, while balancing it with the costs of primary storage and processor time of the algorithm itself. There are many different page replacement algorithms. We evaluate an algorithm by running it on a particular string of memory reference and computing the number of page faults,

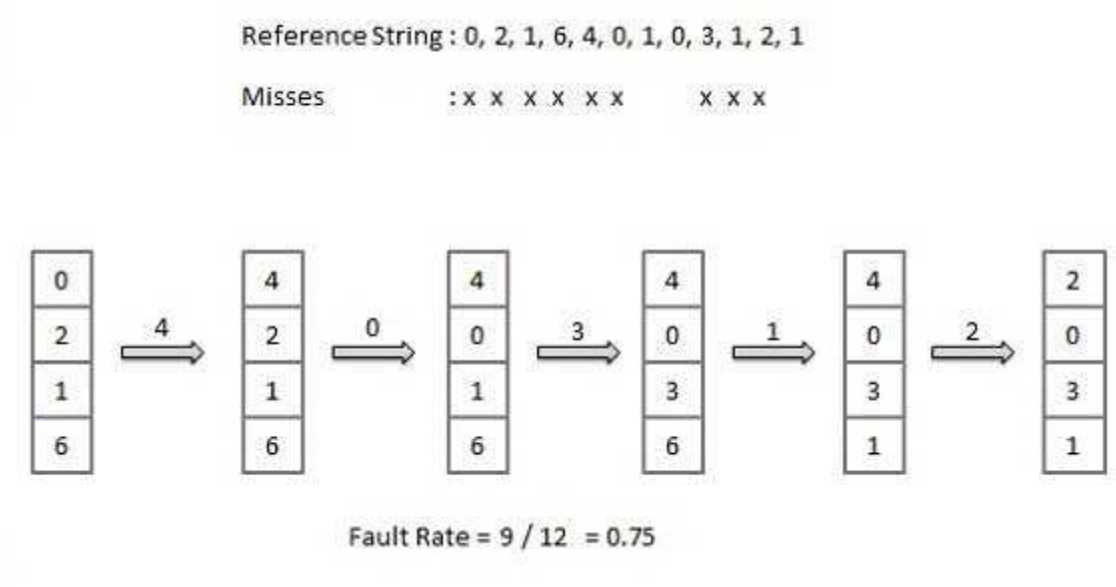
Reference String

The string of memory references is called reference string. Reference strings are generated artificially or by tracing a given system and recording the address of each memory reference. The latter choice produces a large number of data, where we note two things.

- For a given page size, we need to consider only the page number, not the entire address.
- If we have a reference to a page p, then any immediately following references to page p will never cause a page fault. Page p will be in memory after the first reference; the immediately following references will not fault.
- For example, consider the following sequence of addresses − 123,215,600,1234,76,96
- If page size is 100, then the reference string is 1,2,6,12,0,0

First In First Out (FIFO) algorithm

- Oldest page in main memory is the one which will be selected for replacement.
- Easy to implement, keep a list, replace pages from the tail and add new pages at the head.



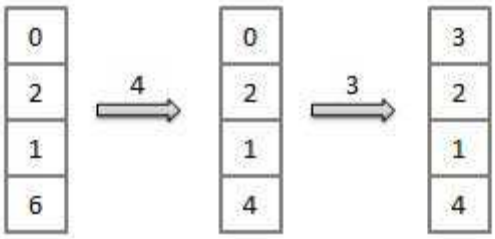
Optimal Page algorithm

- An optimal page-replacement algorithm has the lowest page-fault rate of all algorithms. An optimal page-replacement algorithm exists, and has been called OPT or MIN.

- Replace the page that will not be used for the longest period of time. Use the time when a page is to be used.

Reference String : 0, 2, 1, 6, 4, 0, 1, 0, 3, 1, 2, 1

Misses : x x x x x x



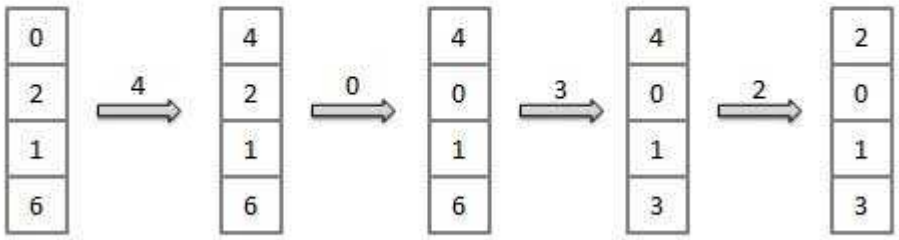
Fault Rate = 6 / 12 = 0.50

Least Recently Used (LRU) algorithm

- Page which has not been used for the longest time in main memory is the one which will be selected for replacement.
- Easy to implement, keep a list, replace pages by looking back into time.

Reference String : 0, 2, 1, 6, 4, 0, 1, 0, 3, 1, 2, 1

Misses : x x x x x x x x



Fault Rate = 8 / 12 = 0.67

Page buffering algorithm

- To get a process start quickly, keep a pool of free frames.
- On page fault, select a page to be replaced.
- Write the new page in the frame of free pool, mark the page table and restart the process.
- Now write the dirty page out of disk and place the frame holding replaced page in free pool.

Least frequently Used (LFU) algorithm

- The page with the smallest count is the one which will be selected for replacement.
- This algorithm suffers from the situation in which a page is used heavily during the initial phase of a process, but then is never used again.

Most frequently Used (MFU) algorithm

- This algorithm is based on the argument that the page with the smallest count was probably just brought in and has yet to be used.

Hard Disk Drive (HDD)

A hard disk drive (HDD) is a non-volatile computer storage device containing magnetic disks or platters rotating at high speeds. It is a secondary storage device used to store data permanently, random access memory (RAM) being the primary memory device. Non-volatile means data is retained when the computer is turned off.

A hard disk drive is also known as a hard drive. Data is stored on a hard drive in binary code; using 0s and 1s. The information is spread out on the magnetic layer of the disk(s) and are read or written by the read heads that "float" above the surface thanks to the layer of air produced by the ultra-fast rotation of the disk.

In writing mode, an electrical current travels via the heads and modifies the surface of the electric field by inscribing a 0 or a 1. In read mode, the process is reversed: the magnetic field transmits an electrical current to the read head, and this signal is then translated into a digital signal readable by the computer.

The hard drive, which typically provides storage for data and applications within a computer, has four key components inside its casing -- the platter (for storing data), the spindle (for spinning the platters), the read/write arm (for reading and writing data) and the actuator (for controlling the actions of the read/write arm). Only the most technically proficient IT professionals should attempt to work on the components inside a hard drive.

Platters

The platters are the circular discs inside the hard drive where the 1s and 0s that make up your files are stored. Platters are made out of aluminium, glass or ceramic

and have a magnetic surface in order to permanently store data. On larger hard drives, several platters are used to increase the overall capacity of the drive. Data is stored on the platters in tracks, sectors and cylinders to keep it organized and easier to find.

The Spindle

The spindle keeps the platters in position and rotates them as required. The revolutions-per-minute rating determines how fast data can be written to and read from the hard drive. A typical internal desktop drive runs at 7,200 RPM, though faster and slower speeds are available. The spindle keeps the platters at a fixed distance apart from each other to enable the read/write arm to gain access. (ref 1+3)

The Read/Write Arm

The read/write arm controls the movement of the read/write heads, which do the actual reading and writing on the disk platters by converting the magnetic surface into an electric current. The arm makes sure the heads are in the right position based on the data that needs to be accessed or written; it's also known as the head arm or actuator arm. There is typically one read/write head for every platter side, which floats 3 to 20 millionths of an inch above the platter surface.

Actuator

The actuator or head actuator is a small motor that takes instructions from the drive's circuit board to control the movement of the read/write arm and supervise the transfer of data to and from the platters. It's responsible for ensuring the read/write heads are in exactly the right place at all times.

Other Components

As well as the casing on the outside of the hard disk that holds all of the components together, the front-end circuit board controls input and output signals in tandem with the ports at the end of the drive. No matter what the type of drive, it has one port for a power supply and one port for transferring data and instructions to and from the rest of the system.

Memory can also be categorized on the basis of their material:

- Semiconductor memory:-such as RAM, ROM, EPROM, and flash memory.
- Magnetic memory:-such as hard disk, floppy disk and magnetic tapes.
- Optical memory:-such as computer disk, DVD and blue-ray disk.

Reference:-
https://www.tutorialspoint.com
https://www.geekboots.com
https://computereducator.blogspot.com
https://www.britannica.com
https://www.getdroidtips.com
https://www.rfwireless-world.com
https://www.techopedia.com
https://smallbusiness.chron.com