Cache Memory.

* What is memory?
the electronic holding space used to store information.
  - Store data and instructions for immediate use.

* Classification of memory:

- Primary
  - RAM
    - Static
    - Dynamic
  - ROM
    - PROM
    - EPROM
    - EEPROM
- Secondary
  - Hard Disk
  - Magnetic
  - Tape
  - CD, DVD
  - Disk
- Cache
- Register

* Characteristics of Computer Memory:-

:. Location:-

- Internal Memory.

Internal memory refers to the storage directly connected to the processor or the central processing unit (CPU).

. Main Memory

. Cache Memory.

- External Memory

also known as Secondary storage memory or Peripheral storage that are not directly connected to the processor.

. Hard Disk Drives.

. USB

. Memory Cards

:. Capacity :-

The amount of data that a memory can hold.

. Bit :- A smallest unit of data and can have two values 0 or 1

. Byte :- Byte a collection of 8 bits it's the smallest unit of storage. a small container that can hold 8 switches.

Internal: expressed in bytes or words 8, 16, 32 bits.

External: expressed in bytes.

∴ Unit of Transfer: (Word or Block)

The amount of data that can be moved between the memory to processor or I/O devices.

- The number of lines determines the unit of transfer.
- Unit of transfer is equal to the length of a word.
- Common word size 8, 16, 32 bits.
- A Block is a group of words that are transferd together
- more efficient data movment and reduces the number of sparate tranfer opreation.

∴ Methods of Access.

- Sequential Access.
  - linearly Search
  - Started from 1st location.
  - Increased the time.
  - Simple and Slow.
- Random Access:
  - Any location can be accessed directly.
  - Faster way to retrieve the data.
  - RAM. used in RAM. time same for any location.

;; Direct Access:-
Divided into blocks used by magnetic and optical discs.
- like Random Access
- Unlike Random Access.
  • In side block.

;; Performance:-
Factors for performance.
performance means how fast and well a system can accomplish its task.
• Access Time (Latency).
(RAM): the time it takes for read or write operation to be completed in the main memory.
Example: Box of toys.
- Non- RAM (HDD. SSD)
The read or write mechanism needs to physically move to the desired location before before data can be read or written from that location.
Example: BookShelp with many books
• Memory Cycle Time:
Access time + wait time
Time takes for the memory to complete on operation in a cycle.

• Transfer Rate.

The rate or speed at which data is transferred in or out of the memory. How quickly data can be read or write.

Example: bucket of water.

Short cycle time higher transfer rate.

Find: Transfer Rate:

• For Random-access memory

it is = $\frac{1}{\text{cycle time}}$

Problem: 01- Find read write time in RAM cycle = 5

$\frac{1}{5}$ = 0.2

• For non-Random access memory

$$T_n = T_A + \frac{n}{R}$$

Where: T_n = Average time to read or write n bits.

T_A = Average access time.

n = number of bits

R = Transfer in bits (bps)

n = 16, T_A = 0.001, T_n = 5

$R = \frac{16}{(5 - 0.001)}$ $R = \frac{16}{4.999}$ R = 3.2006 (bps)

∴ Physical Types.

1. Semiconductor Memory. (ICs)
   using integrated circuits
   Ex: RAM
2. Magnetic Memory.
   magnetic materials to store and retrieve data.
   Ex: HDD
3. Opitical Memory.
   Laser technology to read and write data
   Ex: (CDs, DVDs)

∴ Physical Characteristics:

1. Volatile memory
   It needs continous power supply to retain stored data
   Ex: RAM.
2. Non-Volatile memory.
   memory that retains stored data even the power is turned off. Ex: ROM
3. Erasable Memory
   memory that allows or data to be erased and rewritt
   Ex: USB. memory card. SSD
4. Non-erasable memory.
   memory that can't be erase or modified
   Ex: MROM (read only

* The Memory Hierarchy.
∴ Design Constraints on Computer's memory.
- How much
- How Fast
- How Expensive.

- Regesiter
- Cache
- main memory
- magnetic disk
- CD ROM
- DVD
- Blu Ray
- magnetic tape.

* Hit and Miss:-
• Hit:-
Processor request data and the data found in cache memory. It is Cache hit.

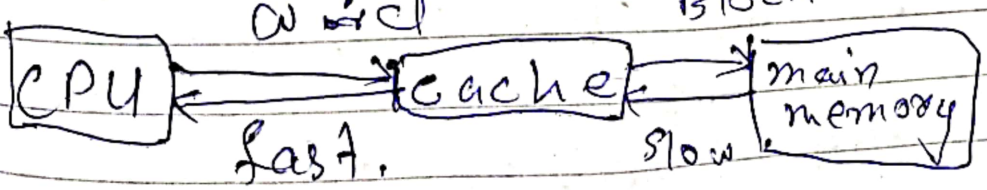
• Miss:- processor request dat but data is not found in cache. Cache miss occurs. It need to fetched from RAM.

* Principle of locality of Reference.
  - A programs has arrays. loops
  - It makes use Frequently accessed instructions are in faster memory
  - It improves the hit ratio..

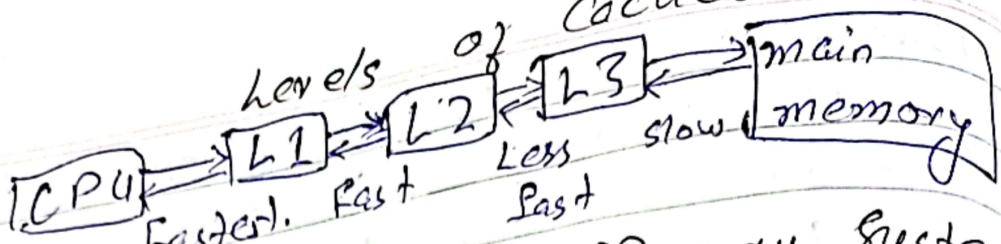
* Cache Memory:- Principles.
  a small and fast memory unit that is located between the processor and the main memory.
  * Large memory. Less expensive lower speed.

* Processor wants W or R from memory.
  - Check is made in the cache.
  - If available:
    word delived to the processor.
  - If not available:
    A block of word read from memory and the required word is delivered.

why block:
  principle of locality of reference.



Levels of Cache:-



* Structure of main Memory System.

- N is the length of each word.
  → memory location uniquely identified.
  $2^n - 1$ words can be addressed
  → estimat overall capacity.
  → memory addresses for storing and retriving.
- memory can be into block each block Contains a certain number of words. orginized

* Structure of Cache System.
  Cache is divided into smaller units called lines.
- tag is like a unique identifier for each line in the cache Requested data by Comparing the memory address with the tags of the lines.

- Control bits in Line:
  modified bit: weather the data modified since the data is loaded.

• Line Size:
  • excluding the control and tag bits is the length of line.
  • Smaller line sizes more afficient. 32 bits (4 words 1 byte 8 one word)
• Tag is usually a memory address which identifies which block of memory is residing in the cache line.
• We can not map one-one (one line to one block)

• Cache:
  connect to the processor via: data address and control line.
• Data and address lines also attach to data and address buffers.
  These connected to system bus to reach main memory.
• Hit: Data and address disabled.
  Communication bt processor and cache.

- miss: loaded from main memory on data buffer available on system bus. word transferred to cache and processor.

* Elements of Cache Design.

- cache Address:

Modern C support virtual memory.

- O.S provides virtual address space to each program. then the virtual address is generated by the program it self during its execution.
- MMU Translate the V.A to P.A:

- Cache Address.
  - Cache be placed between processor and MMU.
  - Cache be Placed between Processor and main memory.

- Logical cache (virtual)
  - Store data using virtual addresses.
  - Directly accessed via processor.
  - No need to go throg MMU.
  - Faster.
- Physical Cache:
  - Stores data using physical address
  - Slow.
- Cache size.
  Should minimum as possible.
  larger cach; large circuit
  slower access.

* Mapping Function:
  Fewer lines more blocks.
  An - Algorithm.
- Which block accupies which cache line.
  (mapping Algorithms)
  - Direct.
  - Associative.
  - Set Associative.

* Replacement Algorithms:
- Cache is filled.
- new block is brought
- Existing block must be replaced.

- LRU (Least Recently Used): (queue) a use bit is associated with each block in the cache. The use bit keeps track of whether a particular block has been accessed. (Timestap) longest time ago.

- FIFO (First in First Out): Block in the cache should be replaced based on the order of their arrival.

- LeFU (Least Frequently Used) A Counter is associated with each block in the cache. This Counter keeps track of the number of times a particular block has been referenced or accessed. (Prioritised)

* Write Policy:.
When a block is to be replaced.
no changes = Override
Single change = main memory must be updated.

. Write Through:
When ever write operation is performed that data is simultenously written to both cache and main memory. main memory data is all ways update.
DISAD: Traffic of writting.

. Write Back:
Updates are only made in cache when updates are made.
Dirty bit Checked..

* Multilevel Caches.
Internal Cache (L1)
External Cache (L2, L3)
reduced external bus activity.
* Unified vs split Caches.
- Unified cache:
Data and Instruction are Stored same Cache.
- Split Cache: Data and Instruction Stored on diffrent Cache.