Lecture-12 (Memory Hierarchy)

13 February 2024 13:00

Memory Storage ↑ Speed ↑ Latency ↓

at low cost

Cost :-

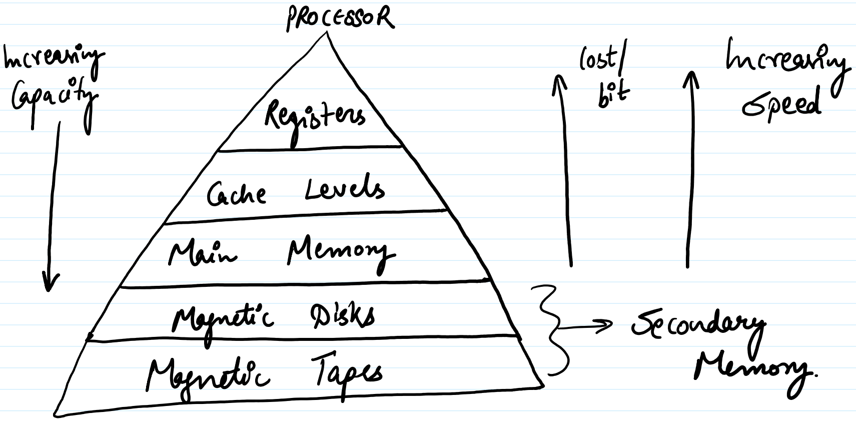
Static RAM > Dynamic RAM > Disk

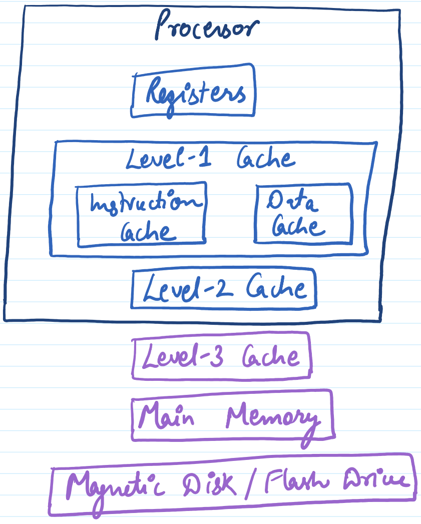
Possible Solution :-

Memory Hierarchy :-

Organization of memory in levels.

⇒ The memory is organized in such a way that faster technology is nearer to the processor

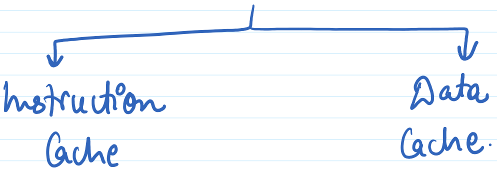




Memory Hierarchy

→ To speed up the processing time.

Cache Memory :- Small amount of memory used to store the frequently accessed data and instructions.



Main Memory :- memory used to store the operating system, applications, etc.

Locality of Reference

The program tends to reuse data and instructions used recently

90/10 Rule : 90% of the total execution time of program is spent only in 10% of the code.

Two dimensions of Locality of reference :-

① Temporal Locality (time)

→ If an item is referenced in memory, it will tend to be referenced again.

factorial of a number :-

```
fact = 1
for x=1 to N
  fact = fact * x;
```

⇒ instructions are being executed more frequently

② Spatial Locality (space)

→ If an item is referenced in memory, nearby items will tend to be referenced soon.

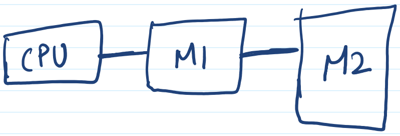
Accessing elements of an array :-

```
sum = 0
for k=1 to N
 sum = sum + A[k];
```

⇒ By copying the array into cache memory.

Performance of Memory Hierarchy :-

2-level Memory System :-



Ⓟ

L1 T_1, S_1, C_1, H_1 (Hit Ratio)

L2 T_2, S_2, C_2, H_2

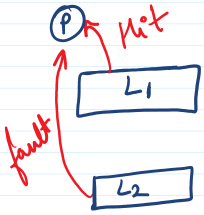
C_1, C_2 → Cost per bit of Memory M_1 and M_2

S_1, S_2 → Storage Capacity in bits of M_1 and M_2

$$C = \frac{C_1 S_1 + C_2 S_2}{S_1 + S_2}$$

(average cost per bit)

Default Hierarchy



T_1, S_1, C_1, H_1 (Hit Ratio),

T_2, S_2, C_2, H_2

Hit Ratio (H_1) → Probability that a logical address generated by processor refers to M_1.

$$H_1 + H_2 = 1$$

$$H_2 = 1 - H_1$$

Also known as miss-row for H_1

T_1, T_2 → Access time of M_1 and M_2

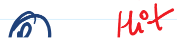
T_{avg} is the average time required by CPU to access a word in the memory.

$$\boxed{T_{avg} = \underbrace{H_1}_{\text{Hit Ratio.}} \times T_1 + \underbrace{(1-H_1)}_{\text{Miss Ratio}} \times \underbrace{T_2}_{\text{miss Panalty.}}}$$

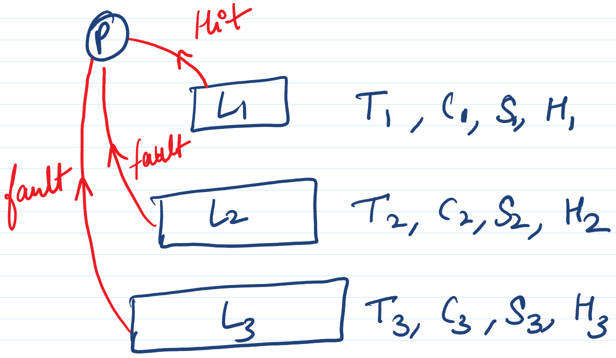
$$T_{avg} = H \times T_1 + (1-H)\underbrace{T_{miss}}_{\text{time required to handle the miss.}}$$

3-level memory system :-

Default Hierarchy



Default Hierarchy

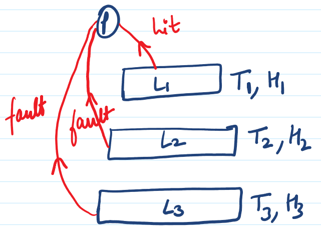


$$T_{avg} = H_1 \times T_1 + (1-H_1)H_2 \times T_2 + (1-H_1)(1-H_2)H_3 \times T_3$$

$$C_{avg} = \frac{C_1 S_1 + C_2 S_2 + C_3 S_3}{S_1 + S_2 + S_3}.$$

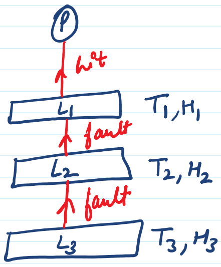
Default v/s Strict Hierarchy :-

Default Hierarchy



$$T_{avg} = H_1 \times T_1 + (1-H_1) H_2 \times T_2 + (1-H_1)(1-H_2) \times H_3 \times T_3$$

Strict Hierarchy



$$T_{avg} = H_1 \times T_1 + (1-H_1) H_2 (T_2 + T_1) + (1-H_1)(1-H_2)(T_3 + T_2 + T_1)$$

Question :-

Consider a 2-level m/m system, where the access time of level-1 & level-2 memories are 10ns and 150ns.

What is the average time if the L1 hit ratio is 90%?

____ ns.