COMPUTER SCIENCE

Computer Organization and Architecture

Cache Memory

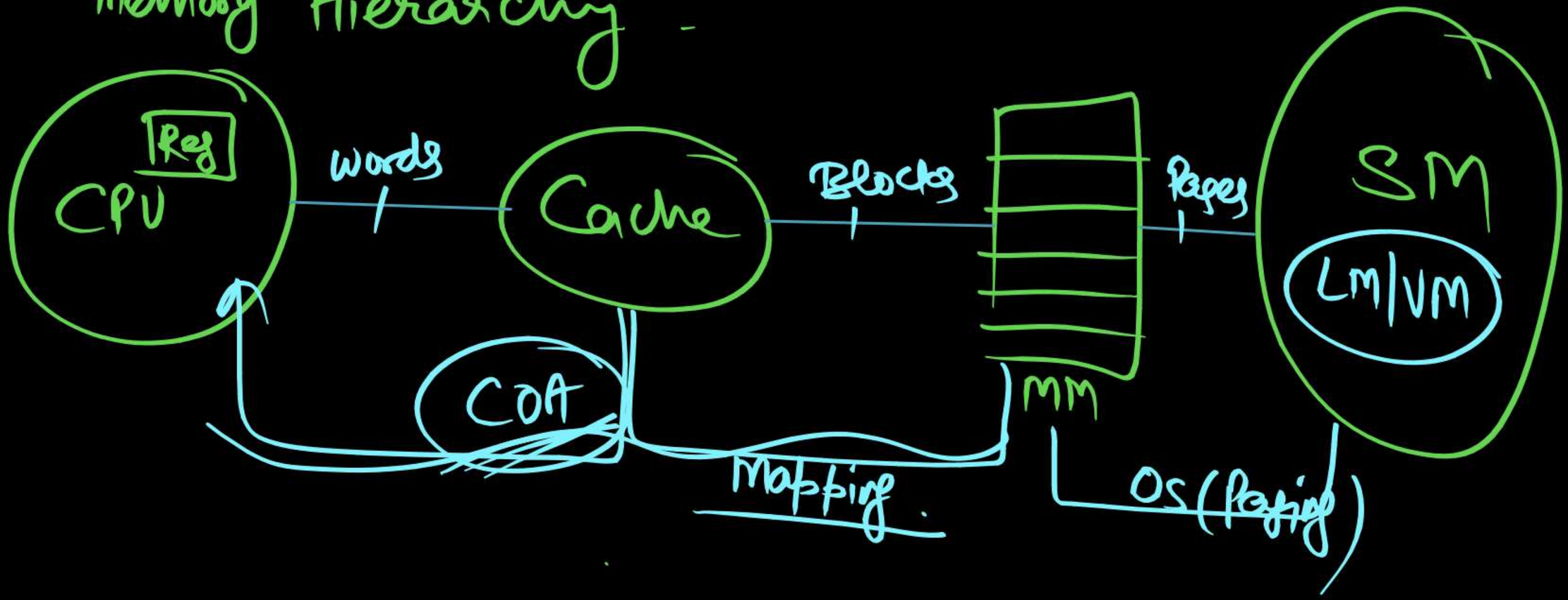
Lecture_02

Vijay Agarwal sir

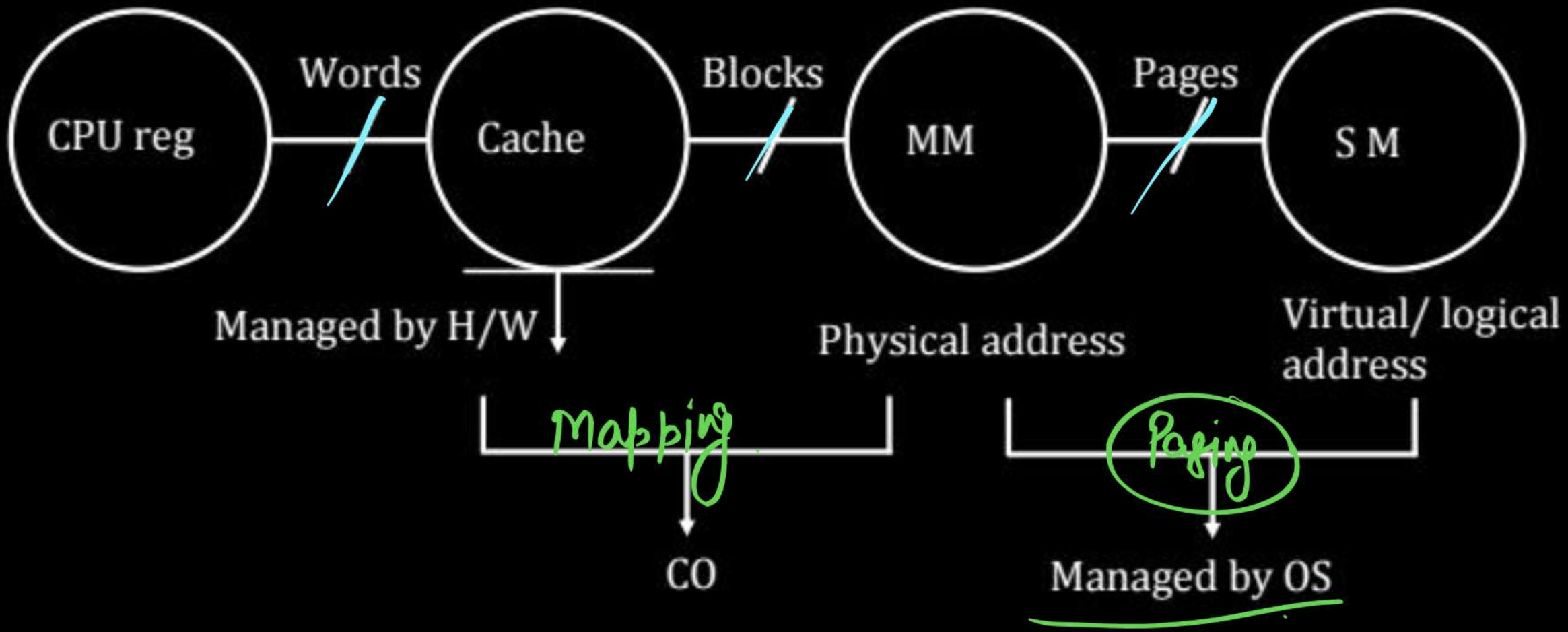
01 Memory Access

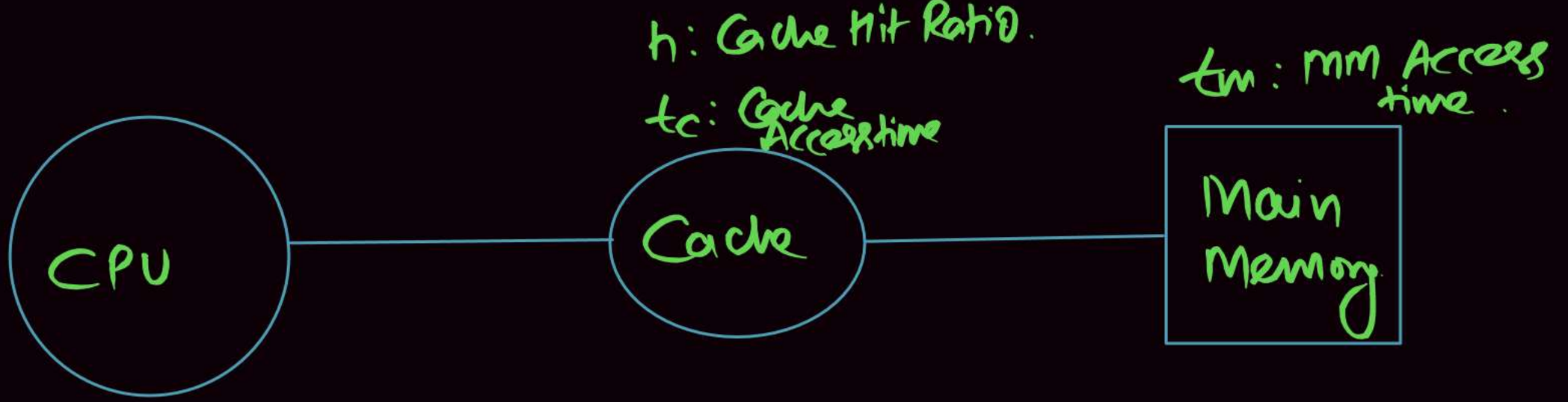
02 Cache Memory

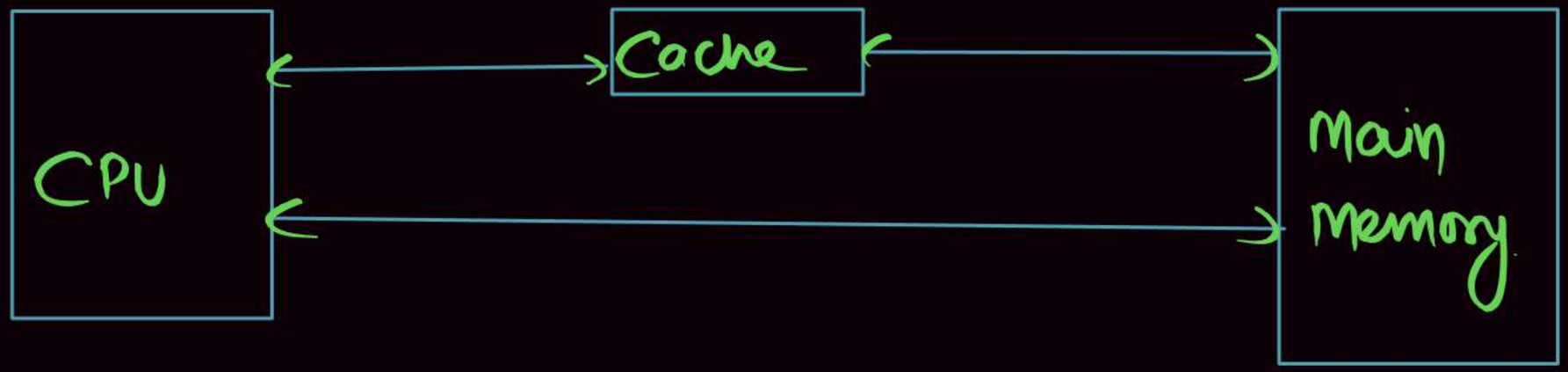
Memory Hierarchy -



Memory


CPU reg
Words
Cache
Blocks
MM
Pages
S M
Managed by H/W
Physical address
Virtual/ logical address
Mapping
Paging
CO
Managed by OS


h: Cache Hit Ratio.
tc: Cache Access time
tm: mm Access time.
CPU
Cache
Main Memory


CPU
Cache
Main
Memory

Average Memory Access time : [Tavg]

$$T_{avg} = H \times \text{Time taken by memory when there is a Hit} + (1-H) \times \text{Time Taken by memory when there is a miss.}$$

Hit

$$\text{Hit Ratio } [H] = \frac{\text{Total \# Hit}}{\text{Total \# Access}}$$

(H)

Q.1

Calculate the average Access time , when the CPU request for the memory 100 times, out of 100 times, 90 times hit & 10 Time miss. If time taken when there is a hit(Each hit) is 20ns & time taken when there is a Miss(Each Miss) is 150ns. ?

Total CPU Request = 100

times Hit = 90 ; Hit take time of 20ns

Time Miss = 10

When there is a miss then 150 nsec

Total time = $90 \times 20 + 10 \times 150$

$= 1800 + 1500$

Total Time = 3300 nsec

$T_{avg} = \frac{3300}{100} =$ 33 nsec

Total CPU Request = 100

Total # Hit = 90, Total Miss = 10

Hit Ratio [H] $= \frac{90}{100} = 0.9$

Miss Ratio $= \frac{10}{100} = 0.1$

OR

$(1-H) = 1 - 0.9 =$ 0.1

$T_{avg} = H \times$ Time taken by memory for each Hit $+ (1-H) \times$ Time taken by memory for each Miss

$\Rightarrow 0.9 \times 20 + (1-0.9) \times 150$

$= 18 + 15 =$ 33 nsec

Q.9

Calculate the average Access time , when the CPU request for the memory 100 times, out of 400 times, 300 times hit & 100 Time miss. If time taken when there is a hit(Each hit) is 20ns & time taken when there is a Miss(Each Miss) is 150ns. ?

Total CPU Request = 400

times Hit = 300 ; Hit take time of 20ns

Time Miss = 100

when there is a miss then 150 nsec

Total time = $300 \times 20 + 100 \times 150$

$= 6000 + 15000$

Total Time = 21000 nsec.

$T_{avg} = \frac{21000}{400} = \frac{210}{4} = 52.5$ nsec

Total CPU Request = 400

Total # Hit = 300, Total Miss = 100

Hit Ratio (H) $= \frac{300}{400} = 0.75$

Miss Ratio OR $(1-H) = 1 - 0.75 = 0.25$

$T_{avg} = H \times$ Time taken by memory for each Hit $+ (1-H) \times$ Time taken by memory for each Miss

$= 0.75 \times 20 + 0.25 \times 150$

$= 15 + 37.5 = 52.5$ Avg

Type of Memory Org

1. Simultaneous Access Memory Org.

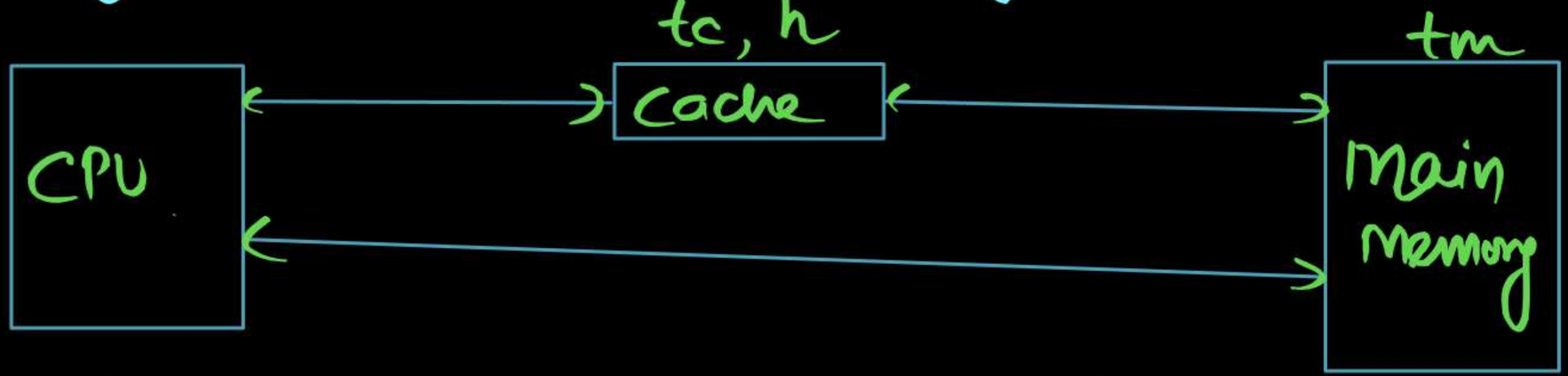
2. Hierarchical Access Memory Org.

Type of Memory Org

t_c : Cache Access time
h : Hit Ratio
t_m : Main Memory Access time.

1. Simultaneous Access Memory Org.

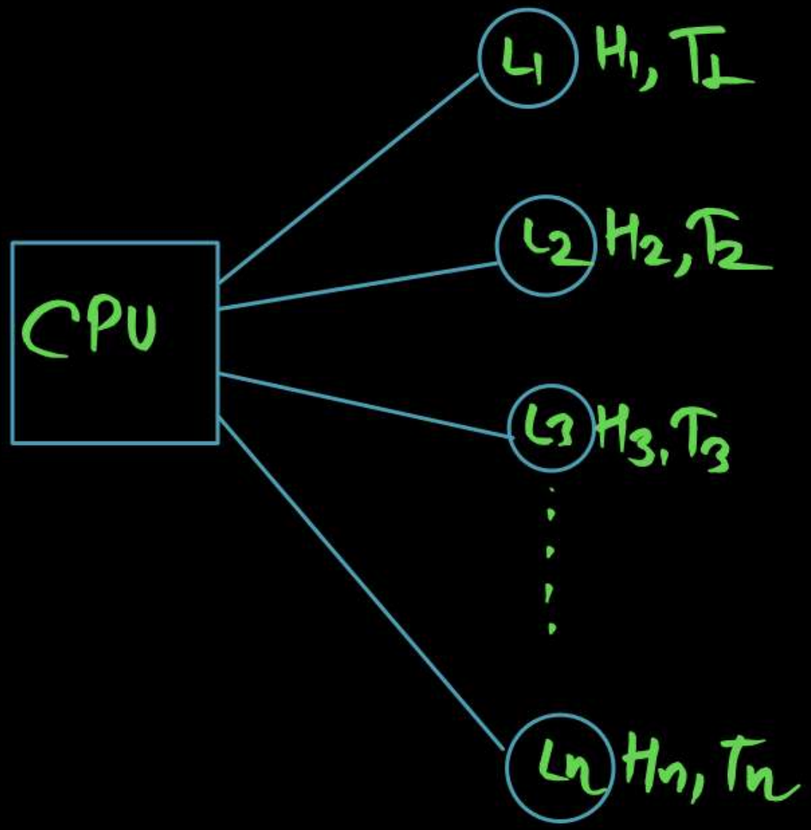
(Both Memory Access Simultaneously/Parallely)



$$T_{avg} = h * t_c + (1-h)\, t_m$$

Type of Memory Org

1. Simultaneous Access Memory Org.

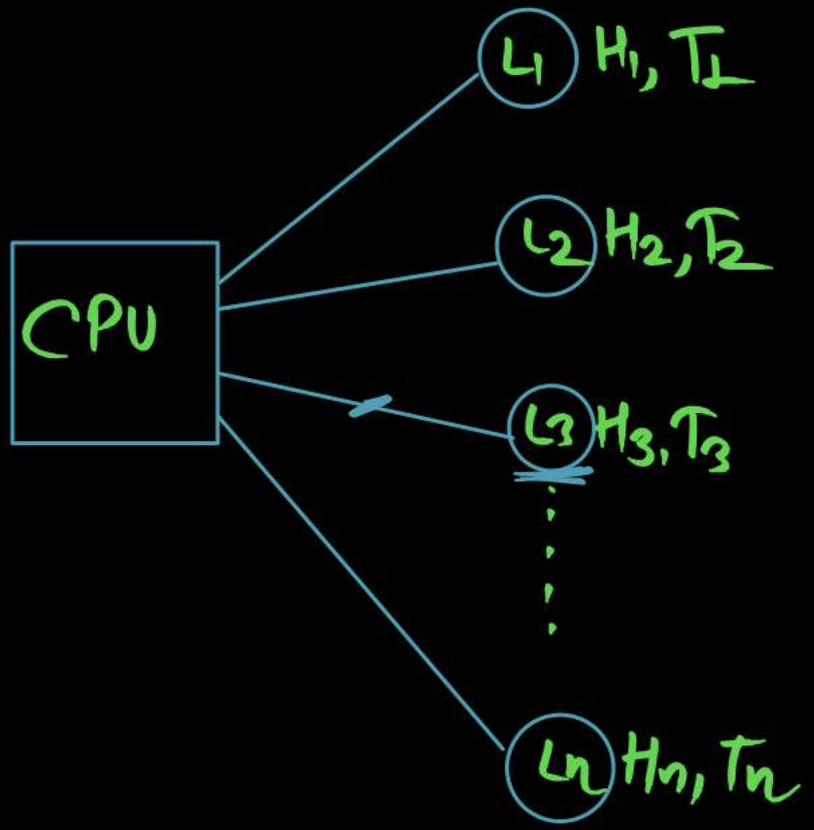


In the Simultaneous Access CPU is Communicating All Level of Memory Directly. [All the levels of memory Directly connected to the CPU] But Follow the sequence.

- When there is a Miss in L_1 & Hit in Level L_2 Memory then Directly Data is transfered from Level 2 memory to CPU. without copying into Level 1 Memory.

Type of Memory Org

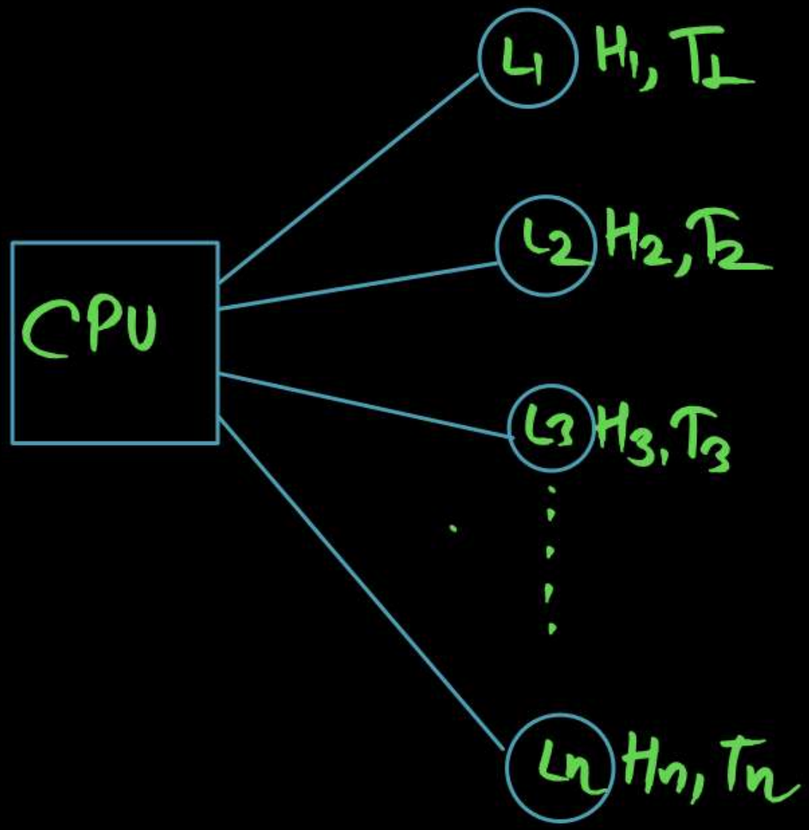
1. Simultaneous Access Memory Org.



- When there is a Miss in Level 1, Level 2 (L_2) but Hit in Level 3 memory then Directly Data given from L_3 memory to CPU Without Copying into Level 1 & Level 2 Memory.

Type of Memory Org

1. Simultaneous Access Memory Org.



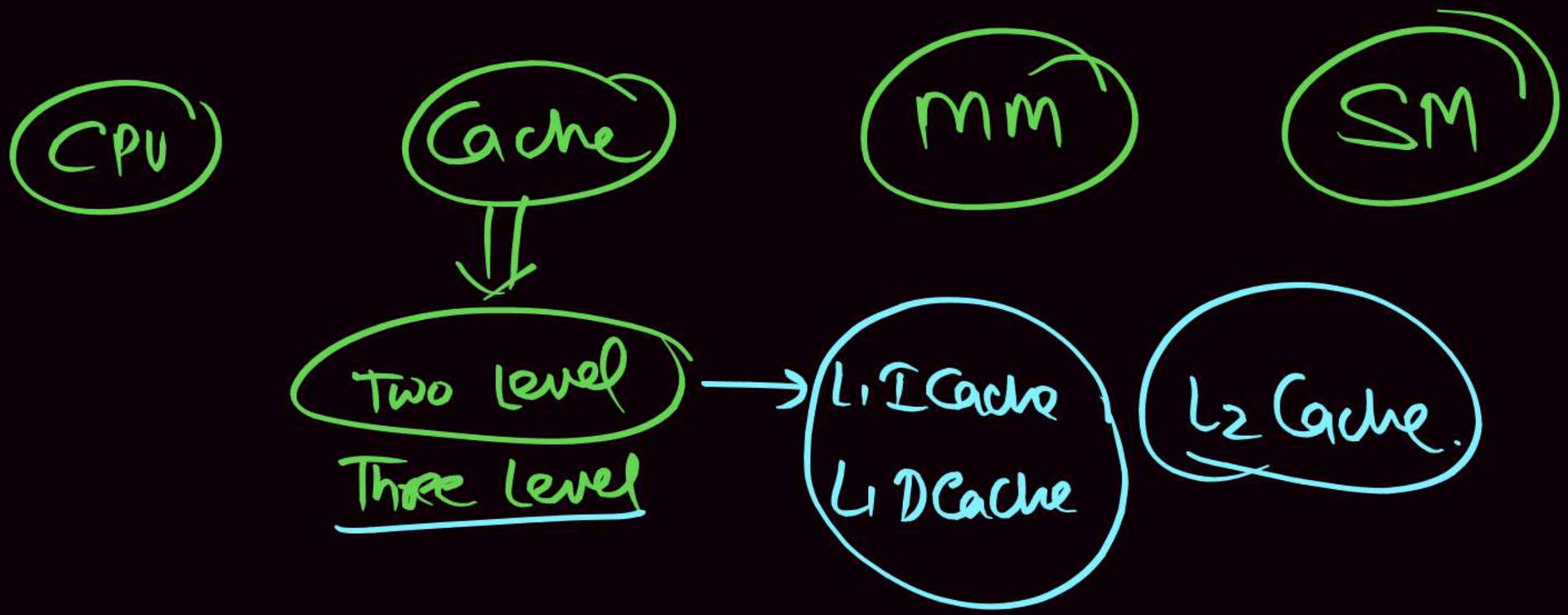
Here $H_1, H_2, H_3 \ldots H_n$ are Hit Ratio &

T_1, T_2, T_3 & T_n are Access time of Respective memory.

Time Required to Access (Read/write) 1 Word from the Memory is called T_{avg}.

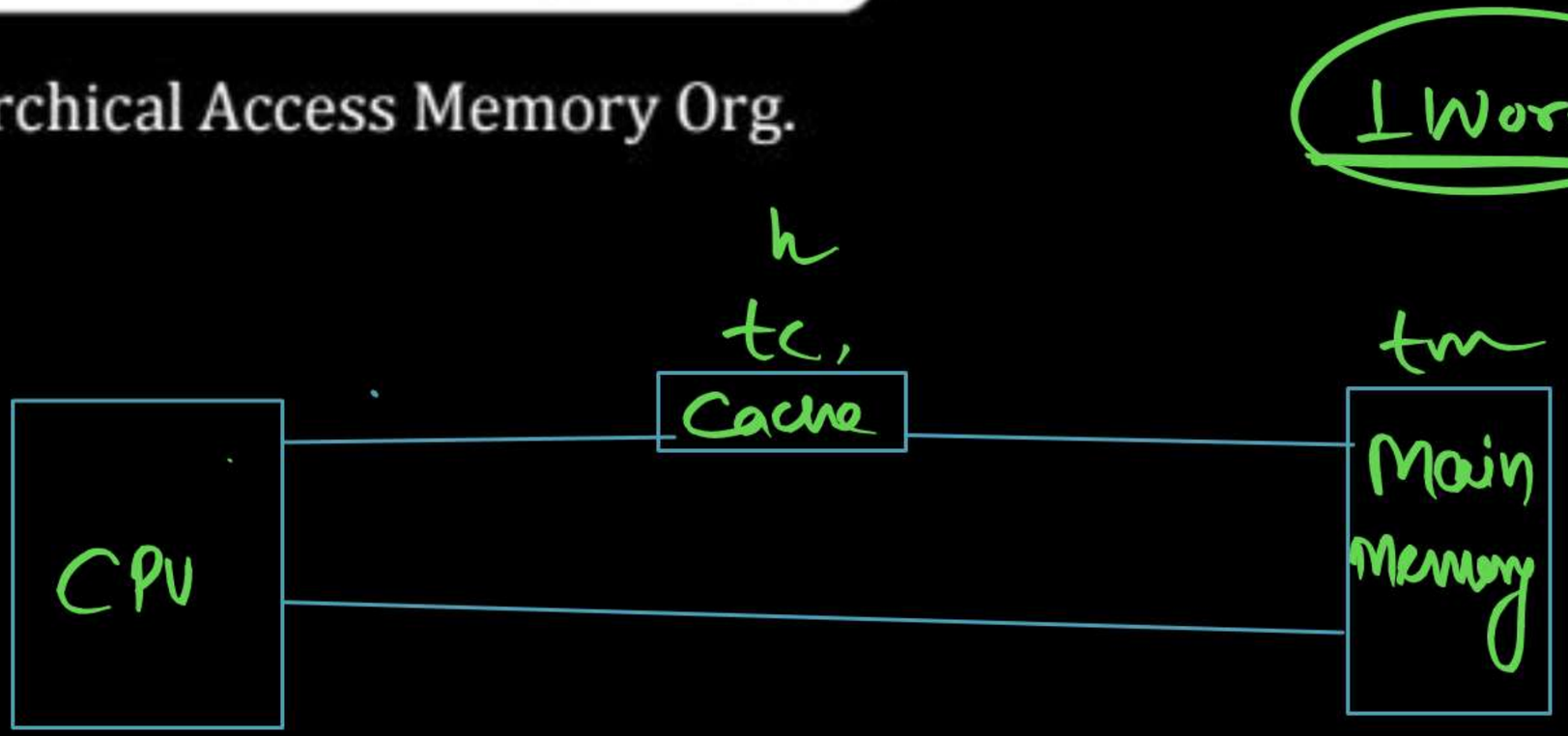
$$T_{avg} = H_1T_1 + (1-H_1)H_2T_2 + (1-H_1)(1-H_2)H_3T_3 + \ldots\ldots (1-H_1)(1-H_2)(1-H_3)\ldots(1-H_{n-1})H_nT_n.$$

$H_n = 1$ Last level Hit Ratio = 1


CPU
Cache
MM
SM
Two Level
Three Level
L1 ICache
L1 DCache
L2 Cache

Type of Memory Org

2. Hierarchical Access Memory Org.



$$Tavg = h * tc + (1-h)(tm + tc)$$

$$= h\,tc + tm + tc - h\,tm - h\,tc$$

$$tc + (1-h)\,tm$$

EMAT [Tavg] $= tc + (1-h)\,tm$ in Hierarchal.

1 Word Access time

(Q) Hit Ratio = 80 % t_c = 20ns t_m = 100nsec

T_{avg} Using Hierarchical Access ?

$T_{avg} = h * t_c + (1-h)(t_m + t_c)$

$\Rightarrow .80 \times 20 + (1-0.8)[100+20]$

$= .80 \times 20 + .20[120]$

$16 + 24 =$ 40nsec

$T_{avg} = t_c + (1-h)\, t_m$

$= 20 + (1-0.8) * 100$

$\Rightarrow 20 + (0.2 \times 100)$

$= 20 + 20 =$ 40nsec

Hit (h) = 0.80

(m) Miss (1-h) = 0.20

$0.8 \times 20 + .20[20+100]$

$0.8 \times 20 + .20 \times 20 + .20 \times 100$

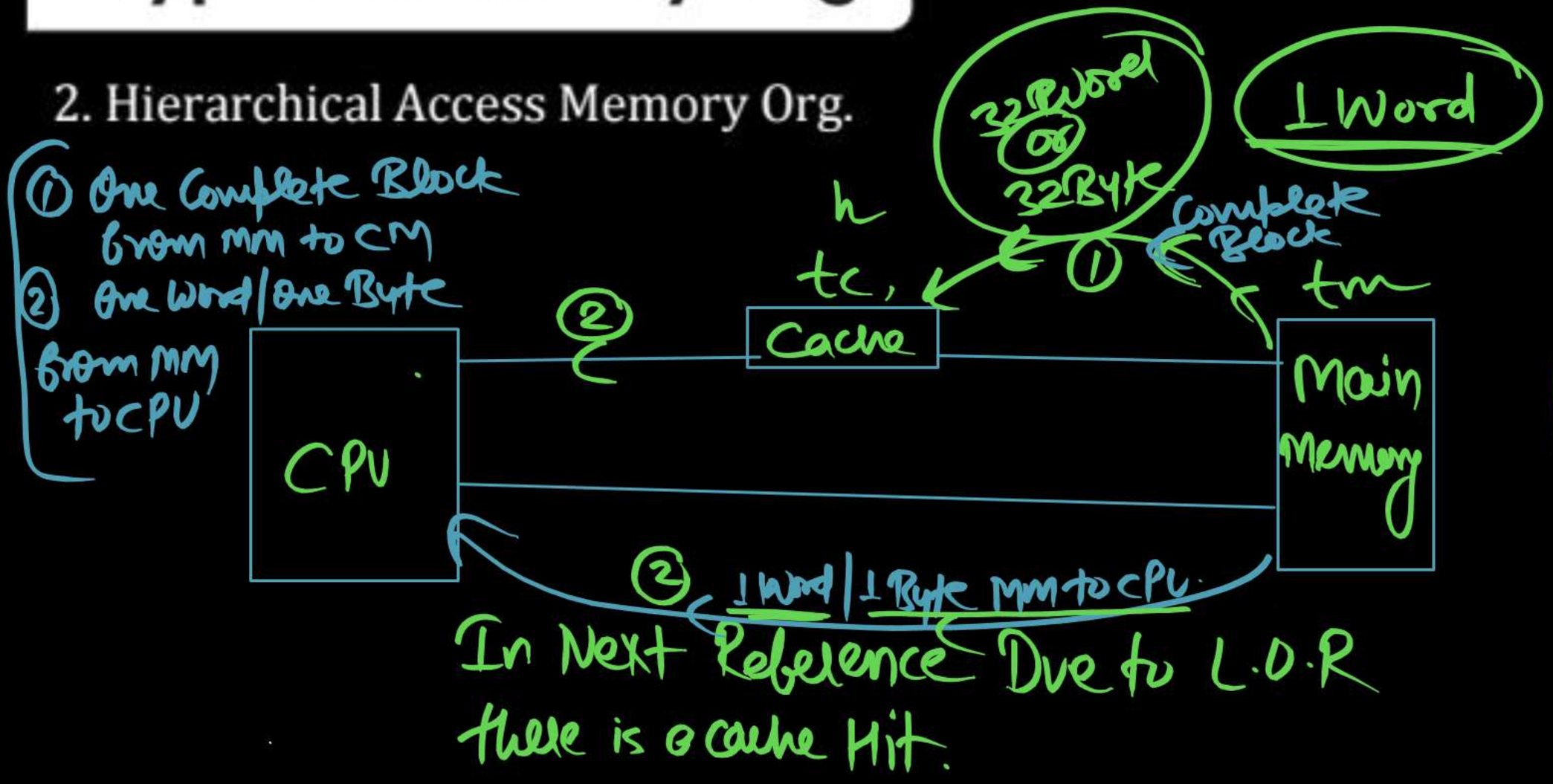
$1 \; (100\%) \times 20 + \underbrace{.20}_{(1-h)} \times \underbrace{100}_{t_m}$

① $= t_c + (1-h)\, t_m$

$T_{avg} = t_c + (1-h)\, t_m$

Type of Memory Org

2. Hierarchical Access Memory Org.



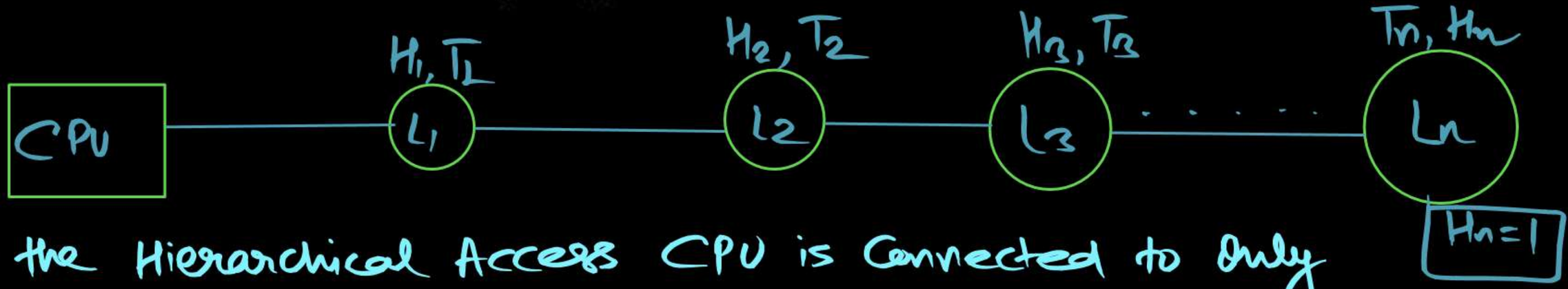
(Q) Block Size is 32 Byte / 32 Words But CPU Require Only 1 Word then ?

(Sol) In this Process Complete Block (32 Byte / 32 Words) Copied / Transfered from MM to Cache & that Respective 1 Word given to CPU [whichever CPU Demanded] [Locality of Reference]

Advantage So in the Next time when CPU Request either Same Word (or) Adjacent Word then that Request we find in Cache [Cache Hit]

Type of Memory Org

2. Hierarchical Access Memory Org.

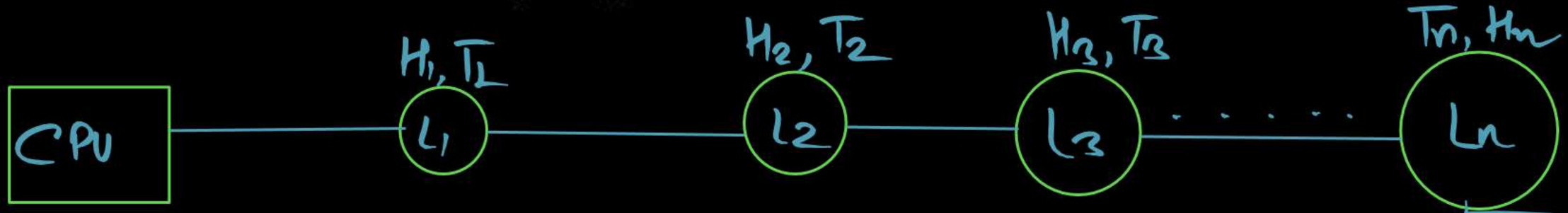


In the Hierarchical Access CPU is Connected to Only Level 1 Memory.

When there is a Miss in Level 1 Memory, & operation Hit in Level 2 Memory then Firstly Data Moves (copied) from L2 Memory to L1 Memory & then Level1 memory to CPU.

Type of Memory Org

2. Hierarchical Access Memory Org.



If there is a Miss in Level L_1 Memory & Level L_2 Memory But Hit in Level 3 Memory. then first Data Copied from L_3 Memory to L_2 Memory & Level2 Memory to L_1 Memory then L_1 Memory to CPU.

Type of Memory Org

2. Hierarchical Access Memory Org.

$$T_{avg} = H_1 T_1 + (1-H_1) H_2 (T_2+T_1) + (1-H_1)(1-H_2) H_3 (T_3+T_2+T_1)$$
$$\cdots + (1-H_1)(1-H_2)(1-H_3)\cdots(1-H_{n-1}) H_n (T_n+T_{n-1}+\cdots T_3+T_2+T_1)$$

Cache Work on Locality of Reference.

1. Temporal LOR
2. Spatial LOR.

Note

If in a Question Mention the keyword 'Hierarchical Access' or 'Levels' of Memory Accessing Hierarchical then Using Hierarchial Access.

Q. Calculate the average Access time with the cache access time 1ns, and main memory access time 100ns, Hit ratio 90%? Using Hierarchical Access?

Soln

$$T_{avg} = h * t_c + (1-h)(t_m + t_c)$$

$$\Rightarrow .90 \times 1 + (1-0.9)[100+1]$$

$$= 0.9 + 0.1(101)$$

$$= 0.9 + 10.1$$

$$= 11 \text{ nsec}$$ Ans

Q. In a 2 level memory, level 1 memory is 5 times faster than level 2. and its access time is 10ns < Average Access Time. Let level 1 Access time is 20ns, What is the hit ratio? Using simultaneous Access org?

$T_1 = 20$

$H_2 = 1$

Assume Access time of

$L_1 \Rightarrow T_1$

$L_2 \Rightarrow T_2$

$$5 = \frac{PL_1 \Rightarrow 1/T_1}{PL_2 \Rightarrow 1/T_2}$$

$$5 = \frac{T_2}{T_1}$$

$$\boxed{T_2 = 5T_1}$$

Here Given $T_1 = 20$

$T_2 = 5 * T_1 = 5 \times 20 \Rightarrow \boxed{T_2 = 100 nsec}$

$$\boxed{T_1 = T_{avg} - 10}$$

$$\boxed{T_{avg} = T_1 + 10}$$

$T_{avg} = T_1 + 10 = 20 + 10$

$\boxed{T_{avg} = 30 nsec}$

$T_{avg} = H T_1 + (1-H) T_2$

$30 = H \times 20 + (1-H)\ 100$

$30 = 20H + 100 - 100H$

$70 = 80H$

$H = \frac{70}{80} = 0.875$ Ans

87.5 % Ans

$$\text{Performance} \propto \frac{1}{ET}$$

Ram

10 Hours

SHYAM

5 Hours.

Same Work

SHYAM Performance is Fast.

Q. Consider a system with 2 levels. Level 1 Access time is 20ns Level 2 Access time = 150ns $T_{avg} = 30$ using simultaneous Access.

(i) What is the Hit Ratio?

(ii) If the Hit Ratio is mode to 100% then what is the Access time of L_1 & L_2 Memory?

$T_1 = 20$ nsec

$T_2 = 150$ nsec

$T_{avg} = 30$ nsec.

Simultaneous Access

$T_{avg} = h*T_1 + (1-H)T_2$

$30 = h\times 20 + (1-H)150$

$30 = 20H + 150 - 150H$

$130H = 120$

$H = \frac{120}{130}$

$H = \frac{12}{13}$

Hit Ratio = 92.33% Ans

$H = 0.9233$ Ans

(ii) If Hit Ratio = 100% then

What's time of L_1 & L_2 memory

(Same) $(T_1)L_1 = 20$ns

$(T_2)L_2 = 150$ns

} Remain Same.

(Note) Hit Ratio does not Effect Level Access time it effect T_{avg}

$T_{avg} = h*t_1 + (1-h)t_2$

$= 1\times 20 + (1-1)150$

$= 0$

$T_{avg} = 20$ns Ans

Q. If the above Question if T_{avg} is increased by 10% then what is % of change in Hit Ratio?

$T_1 = 20ns$

$T_2 = 150ns$

$T_{avg} = 30ns$ — Increased by 10%

$T_{avg_{new}} = 30 + 10\%$ of $30 \Rightarrow 30 + 3$

$T_{avg_{new}} = 33$ nsec.

Simultaneous Access

$T_{avg_{new}} = h_{new} * t_1 + (1 - h_{new}) t_2$

$33 = h \times 20 + (1-h)\, 150$

$33 = 20H + 150 - 150H$

$117 = 130H$

$H = \frac{117}{130}$

$H = 0.9$

Hit Ratio = 90%

& earlier was 92.33%.

Hit Ratio Decreased by 2.33% Ans

Q.

Assume that for a certain processor, a read request takes 50 nanoseconds on a cache miss and 5 nanoseconds on a cache hit. Suppose while running a program, it was observed that 80% of the processor's read requests result in a cache hit. The average read access time in nanoseconds is _____. [GATE – 2015]

Hit Ratio = 80% = 0.8

When Hit time taken = 5ns

Miss Ratio = (1 − 0.8) = 0.2

When Miss time taken = 50nsec

$$T_{avg} = 0.8 \times 5 + 0.2 \times 50$$

$$4 + 10$$

$$T_{avg} = 14 \text{ nsec}$$

Ans

LOR [Locality of Reference]

- Access the higher level of memory Data from level 1 Memory is called L.O.R,.
  (1) Temporal LOR
  (2) Spatial LOR.

(Faster)

R ← M[2000]

Add: [2000]
Location Wallah Word which is available in Block No 85.

(1) Temporal LOR: means the same word in the same block is reference by the CPU in near future (Frequently)[Eg: LRU]

Or

Same data which access again and again then that type of data stored in Temporal LOR.

LOR [Locality of Reference]

(2) Spatial LOR means adjacent word in the same block is referenced by the CPU in a sequence.

(x+1) location wallah word which is available in Block No B5.

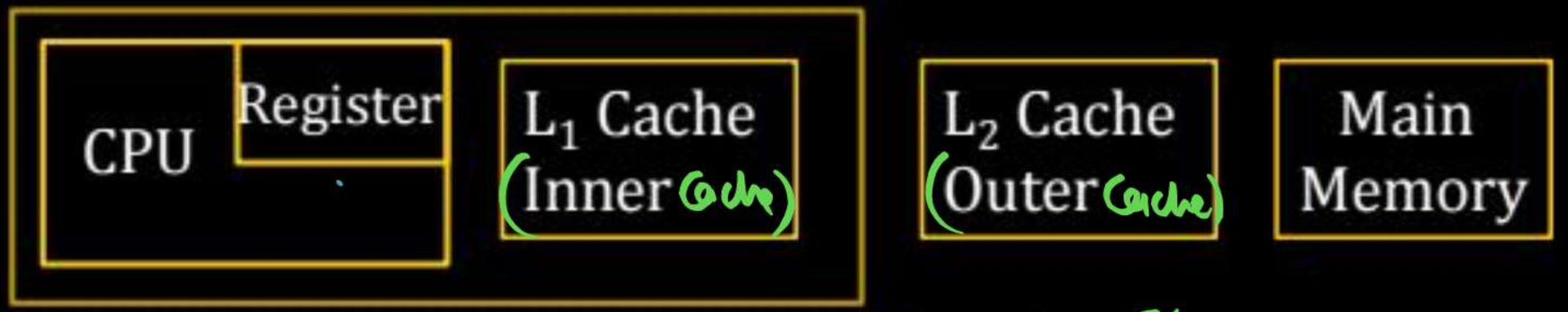
↓

Cache Hit

Bread
Milk
} ⇒ 1 Portion

Types of Cache

1) Unified Cache: Instruction & Data both are placed in Same Cache.
2) Split Cache: This Cache logically Divide into two parts
   (i) Instruction Cache [I - cache]
   (ii) Data Cache [D- cache]
3) Multilevel Cache:



Size $L_1 < L_2$
Speed $L_1 > L_2$

L2 > L1 Size
L1 > L2 speed

THANK YOU!