COMPUTER SCIENCE

Computer Organization and Architecture

Cache Memory

Lecture_01

Vijay Agarwal sir

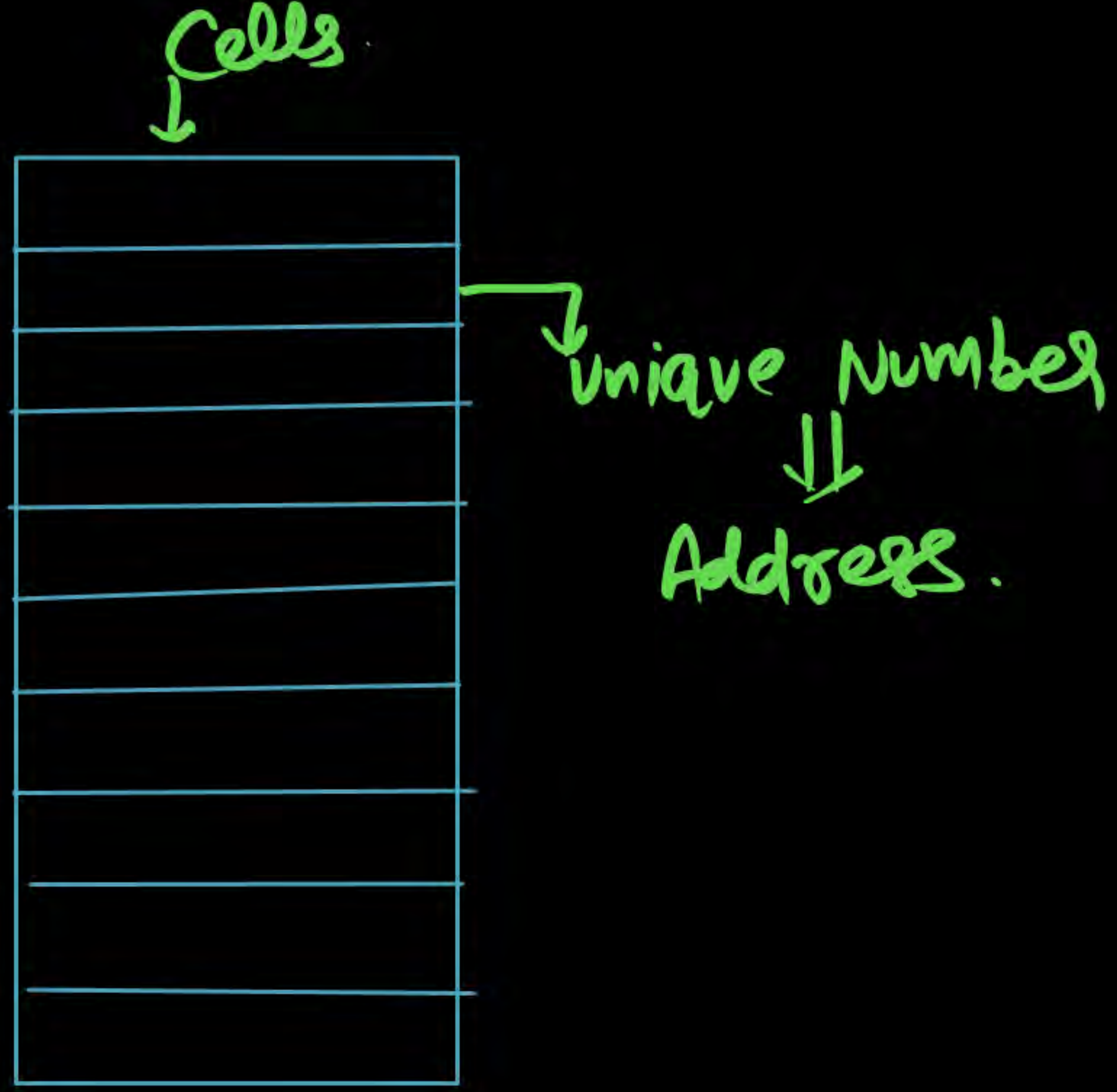
TOPICS TO BE COVERED

01 Memory Hierarchy

02 Cache Memory

1. Introduction of COA
2. Machine Instn & AM
3. Floating Point Representation
4. ALU Data Path & Control Unit
5. PIPE LINING.
6. CACHE Memory.

Memory



(Q) Memory 256KB then Address Size ?

Soln

256KB.

$2^{8}.2^{10} \times 8 \text{bit}$

$2^{18} \times 8 \text{bit}$

Address = 18 bit.

$2^1 = 2$

$2^2 = 4$

$2^3 = 8$

$2^4 = 16$

$2^5 = 32$

$2^6 = 64$

$2^7 = 128$

$2^8 = 256$

$2^9 = 512$

$2^{10} = 1024$ [1k]

$2^{10} = 1K$ (kilo)

$2^{20} = 1M$ (mega)

$2^{30} = 1G$ (Giga)

$2^{40} = 1T$ (Tera)

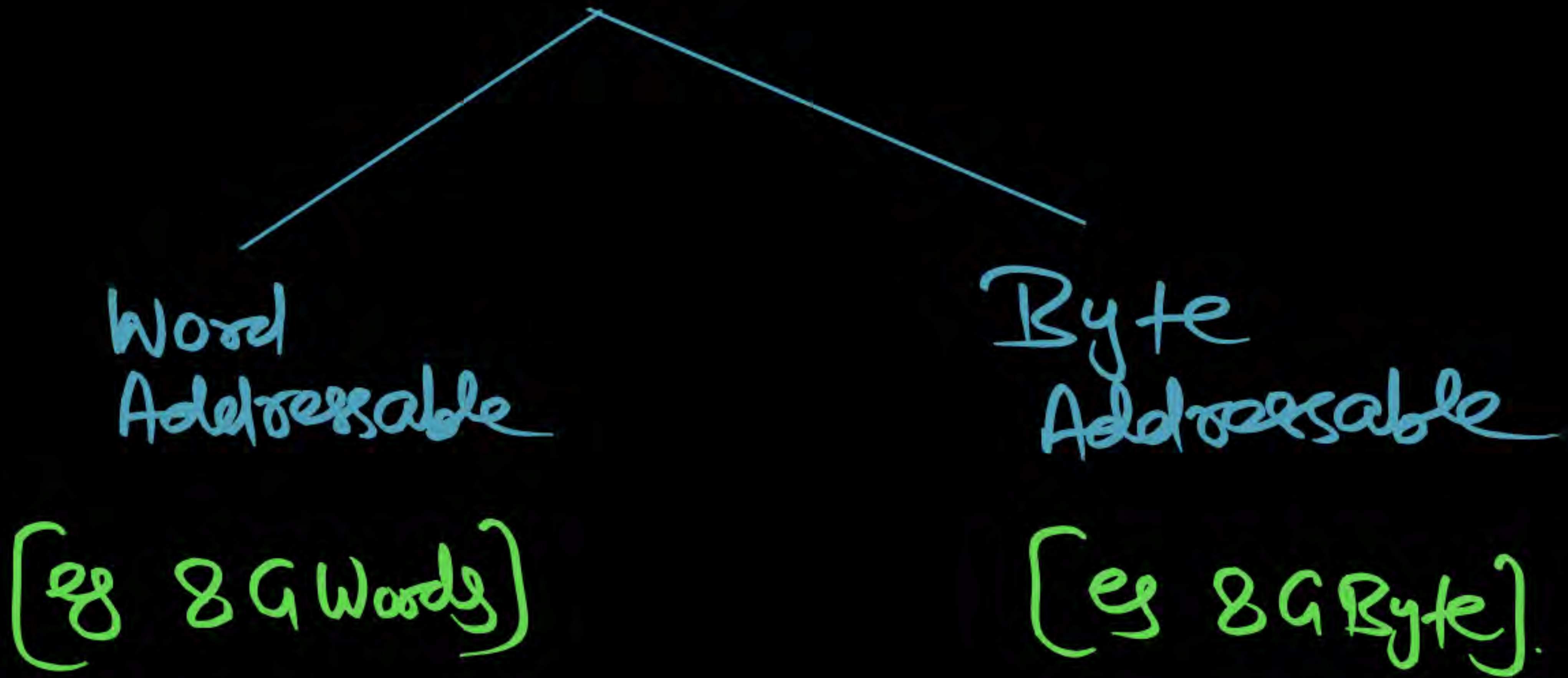
$2^{50} = 1P$ (Peta)

$2^{60} = 1E$ (Exa)

$1 msec = 10^{-3} sec$

$1 \mu sec = 10^{-6} sec$

$1 nsec = 10^{-9} sec$.



Word Addressable

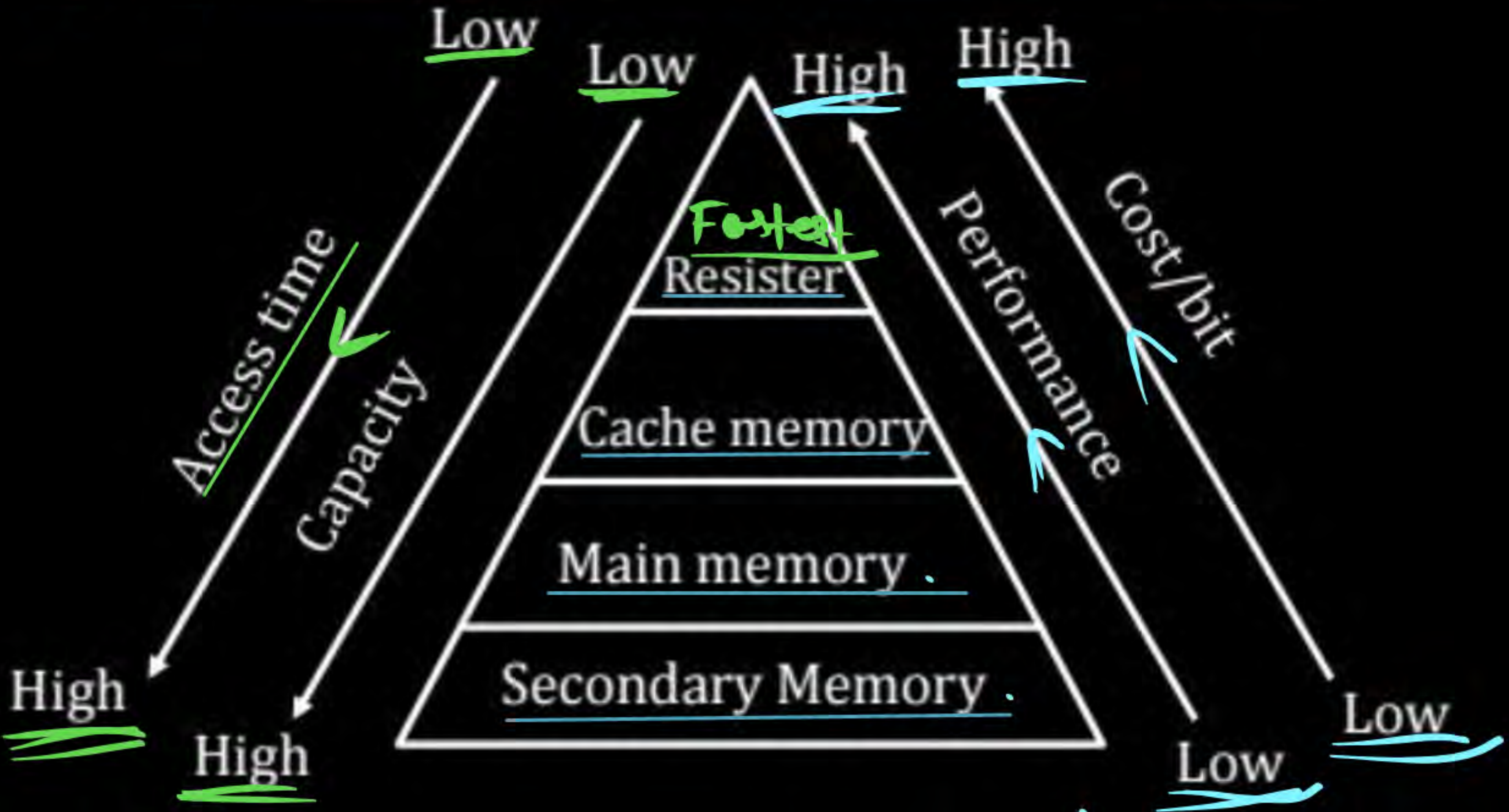
[eg 8G Words]

Byte Addressable

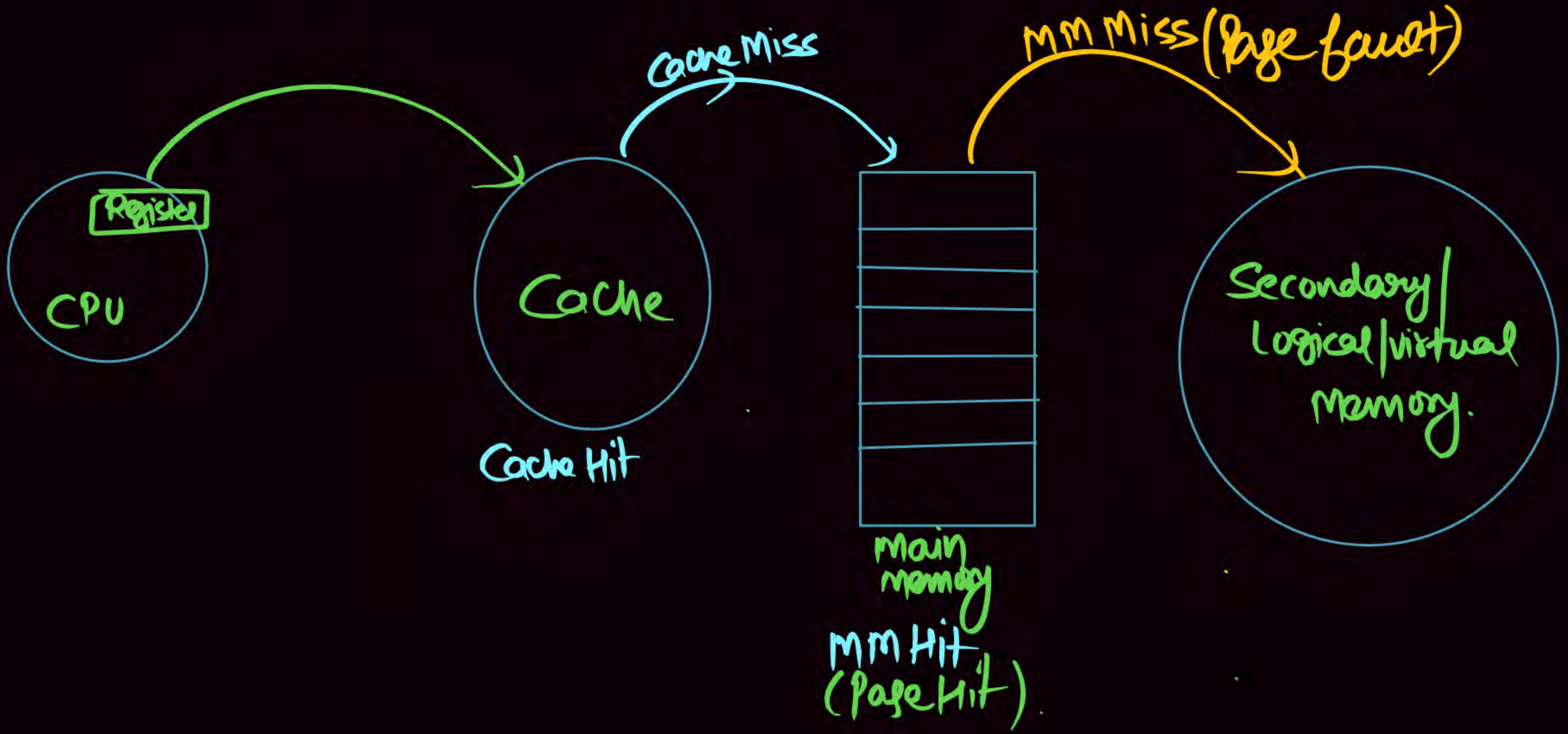
[eg 8G Byte].

Memory Hierarchy

- Hierarchy design organize the system supported memory into 4 levels to minimize the Accessing times. They are:



$$\boxed{\text{Performance} \propto \frac{1}{\text{Execution Time}}.}$$

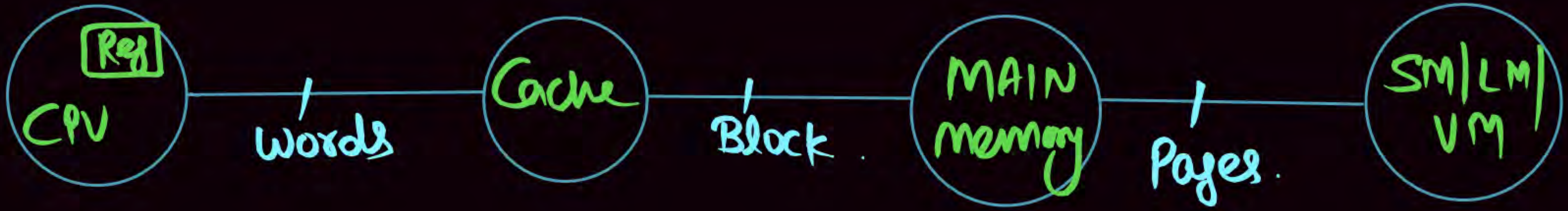

Register
CPU
Cache
Cache Hit
Cache Miss
Main memory
MM Hit
(Page Hit)
MM Miss (Page fault)
Secondary / Logical / virtual memory.

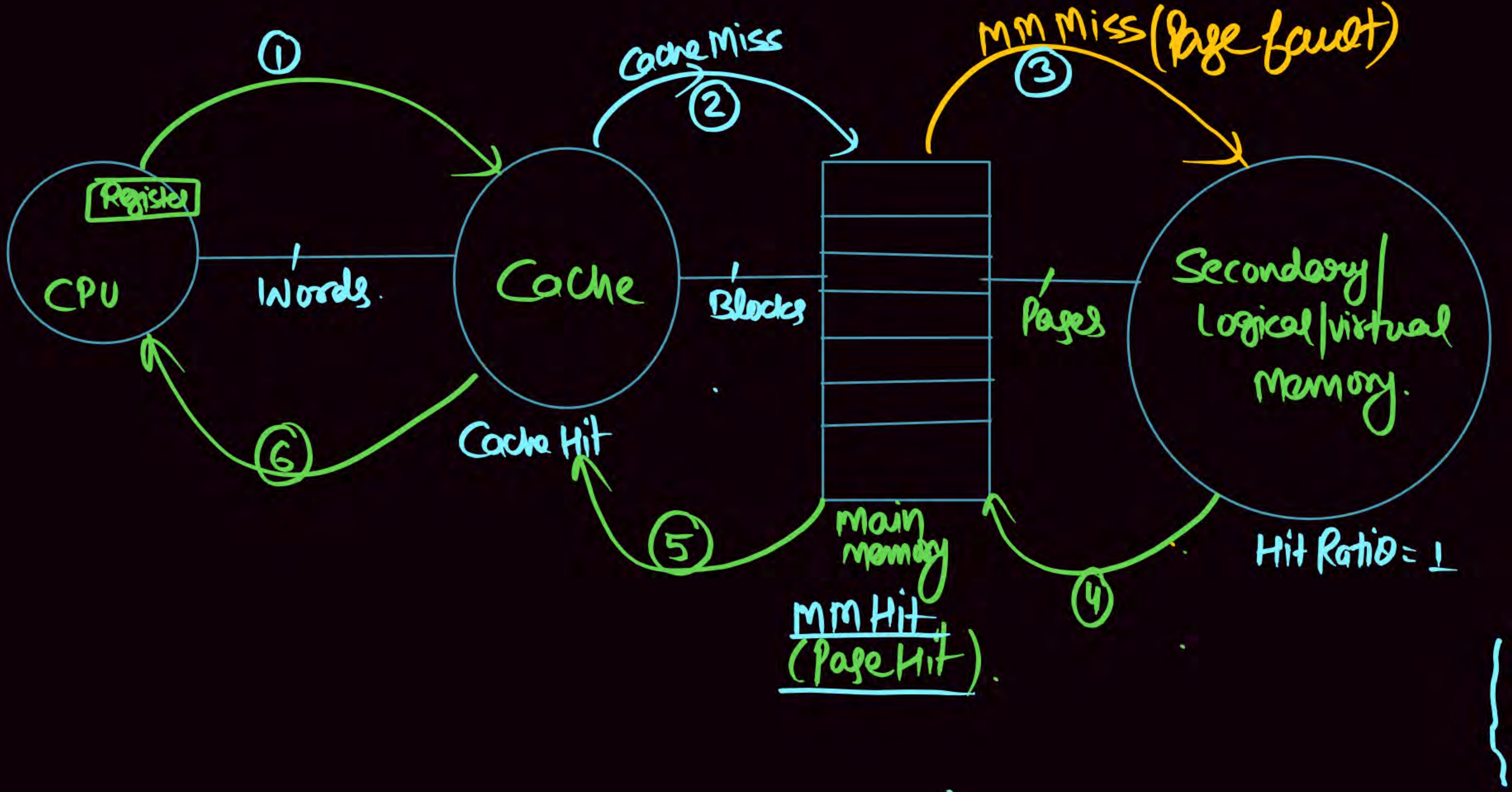
$$\boxed{\text{Hit Ratio} = \frac{\text{Number of Hit}}{\text{Total Number of Access}}}$$

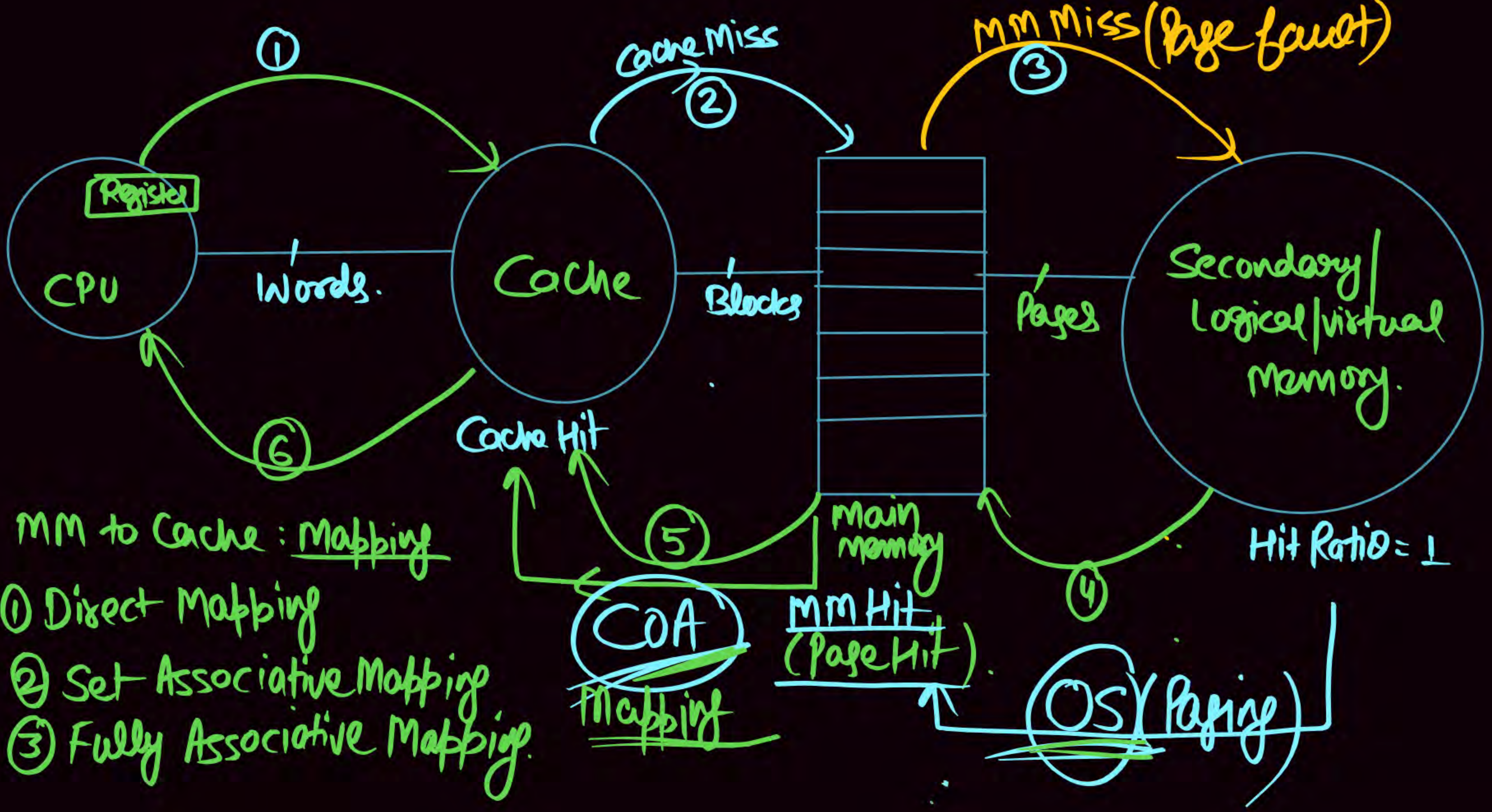
eg If Cache Hit Ratio 80%.

↳ ie out of 100, 80 Time there is Hit in Cache.

[Reference found in cache].


Reg
CPU
Words
Cache
Block
MAIN
memory
Pages
SM|LM|
VM


1
Register
CPU
Words.
Cache
Cache Miss
2
Blocks
MM Miss (Page fault)
3
Pages
Secondary/
Logical/virtual
Memory.
Cache Hit
6
5
Main memory
MM Hit
(Page Hit).
4
Hit Ratio = 1

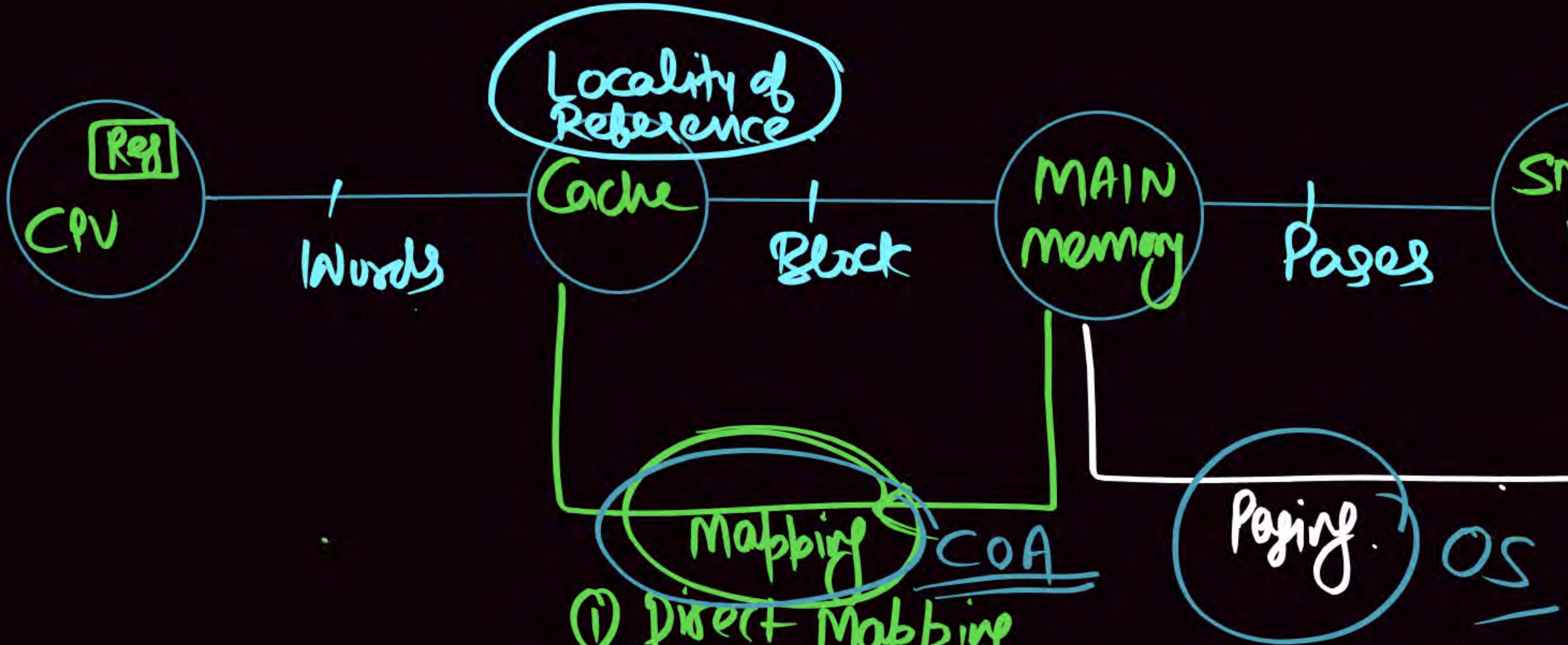


MM to Cache : Mapping

① Direct Mapping

② Set Associative Mapping

③ Fully Associative Mapping.



1. Direct Mapping
2. Set Associative
3. Associative | Fully Associative Mapping.

Memory

- CPU generate Request initally Refer to the Cache.

- If the Reference [Respective Data] found in the Cache that is Called Cache Hit [operation is Called Hit] then Respective Data give from Cache to CPU in the form of Words.

- If the Reference is Not found in the Cache, that is known as Cache Miss, then Reference forwarded to Main Memory.

Memory

- If the Reference found in the Main Memory then its called MM Hit or Page Hit. then Respective Data given from Main Memory to Cache in the form of blocks, & Cache to CPU in the form of words.

- If the Reference is Not found in the Main Memory that is called MM Miss or Page fault then Reference forwarded to secondary Memory.

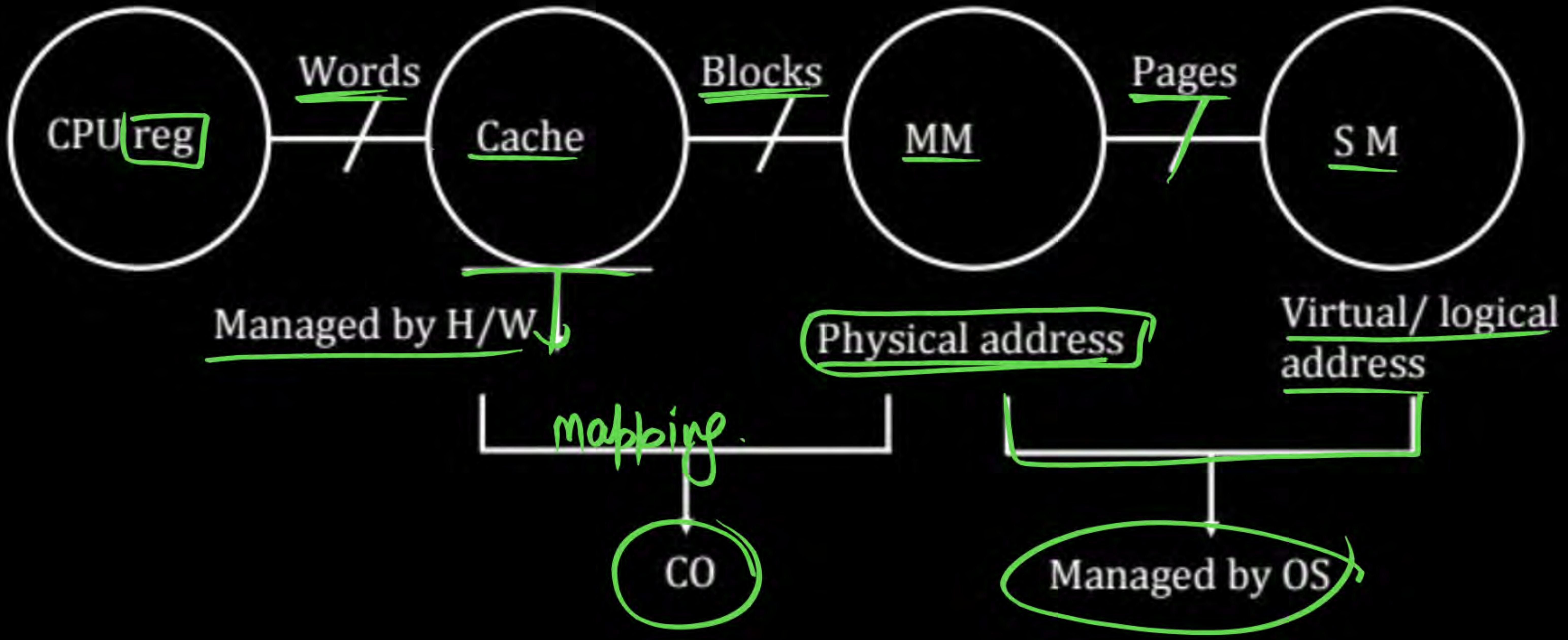
Note Secondary Memory is the Last Level of Memory in Which Hit Ratio always '1'.

Memory

So Respective Data is transfered from Secondary Memory to Main Memory in the form of Pages, Main memory to Cache Memory in the form of Blocks, then Cache memory to CPU in the form of Words.

- The Process of transfer the Data from Main memory to Cache Memory is Called 'Mapping'.

Memory


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

Average Memory Access time [Tavg]

$$\boxed{T_{avg} = \underset{[H]}{\text{Hit}} \times \begin{array}{c}\text{Time taken by}\\ \text{memory when}\\ \text{there is a Hit}\end{array} + (1-H)\begin{bmatrix}\text{Time taken by}\\ \text{memory when}\\ \text{there is a Miss}\end{bmatrix}}$$

Q.1 Consider CPU generate 100 Request. out of 100 ⇒ 90 times Hit & 10 times Miss. when there is Hit then time taken is 20nsec, & when there is a Miss then time taken is 150nsec Calculate the T_{avg}?

Soln Total CPU Request = 100

Hit = 90 Times

Miss = 10 Times

Hit will takes = 20nsec

Miss will takes = 150nsec

$$\text{Total Time} = 90 \times 20 + 10 \times 150$$

$$= 1800 + 1500$$

100 Request

$$\boxed{\text{Total Time} = 3300\text{nsec}}$$

$$T_{avg} = \frac{3300}{100} = 33\text{ns Ans}$$

Total CPU Request = 100.

Hit = 90 times Miss = 10 Times

$$\text{Hit Ratio} = \frac{90}{100} = 0.9$$

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

or

$$\text{Miss Ratio} = 1 - 0.9 \; (1-H)$$
$$= 0.1$$

$$T_{avg} = H \times \text{Time Taken when there is a Hit} + (1-H)\,[\text{Time taken when there is a Miss}]$$

$$\Rightarrow 0.9 \times 20 + 0.1(150)$$

$$\Rightarrow 18 + 15$$

$$\boxed{T_{avg} = 33\text{nsec}}$$ Ans

Q.2 ③ Consider CPU generate 400 Request. out of 300 times Hit 100 times Miss. when there is Hit then time taken is 20nsec, & when there is a Miss then time taken is 150nsec

Calculate the T_{avg}?

Soln Total CPU Request = 400

Hit = 300

Miss = 100

Hit will takes = 20nsec

Miss will takes = 150nsec

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

$= 6000 + 15000$

for 400 Req: $\boxed{\text{Total Time} = 21000\text{ns}}$

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

Total CPU Request = 400

Hit = 300 Miss = 100

Hit Ratio = $\frac{300}{400} = 0.75$

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

$= 0.25$

$\frac{100}{400} = 0.25$

or

T_{avg} = H × Time Taken when there is a Hit + $(1-H)$ [Time taken when there is a Miss]

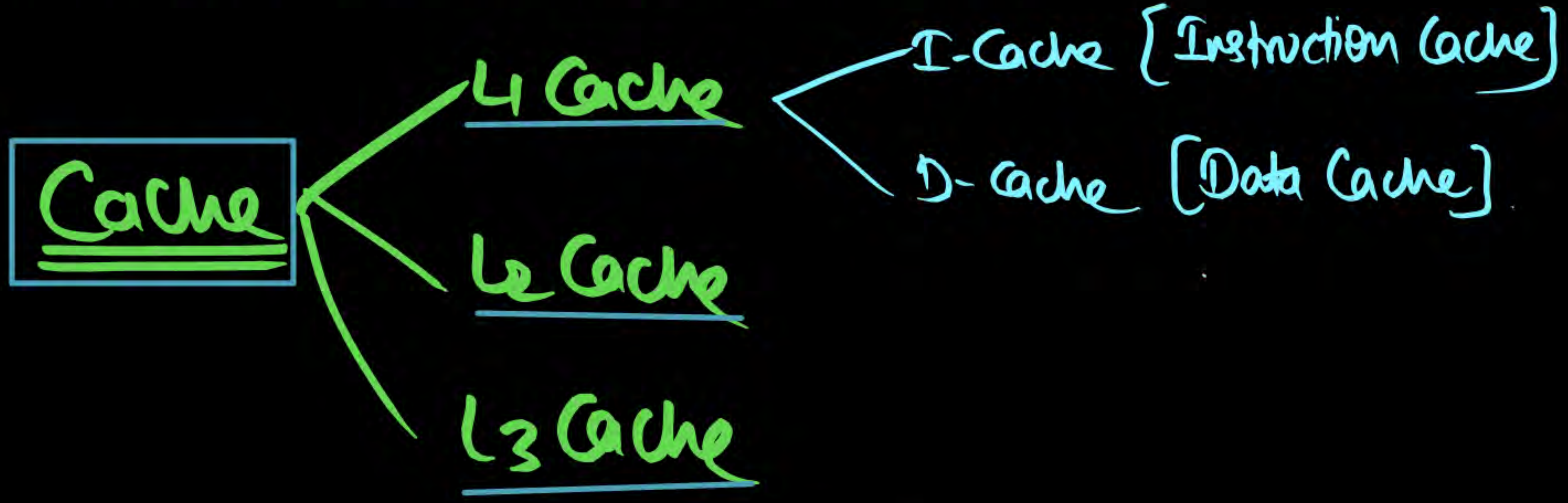
$\Rightarrow .75 \times 20 + .25 \times 150$

$\Rightarrow 15 + 37.5$

$\boxed{T_{avg} = 52.5\text{ns}}$ Ans

Memory

CPU ↔ Cache ↔ MM ↔ SM



Type of Memory Org

✓ 1. Simultaneous Access Memory Org.

✓ 2. Hierarchical Access Memory Org.

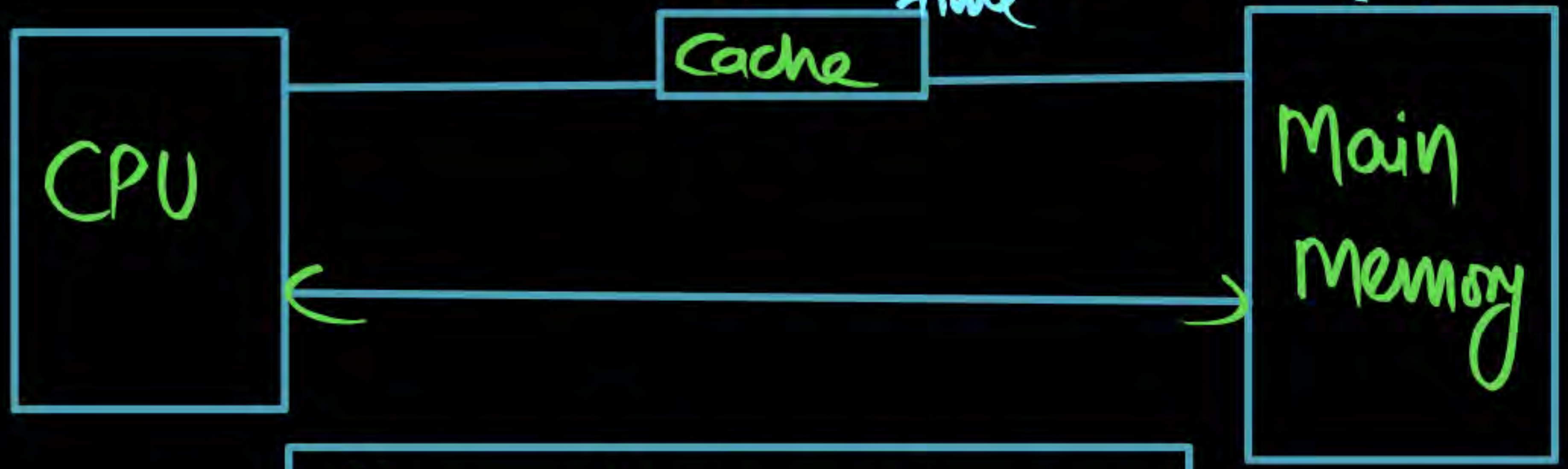
Type of Memory Org

1. Simultaneous Access Memory Org. : (Both Memory Access Simultaneously Parallely)

h : Hit Ratio.

t_c : Cache Access time

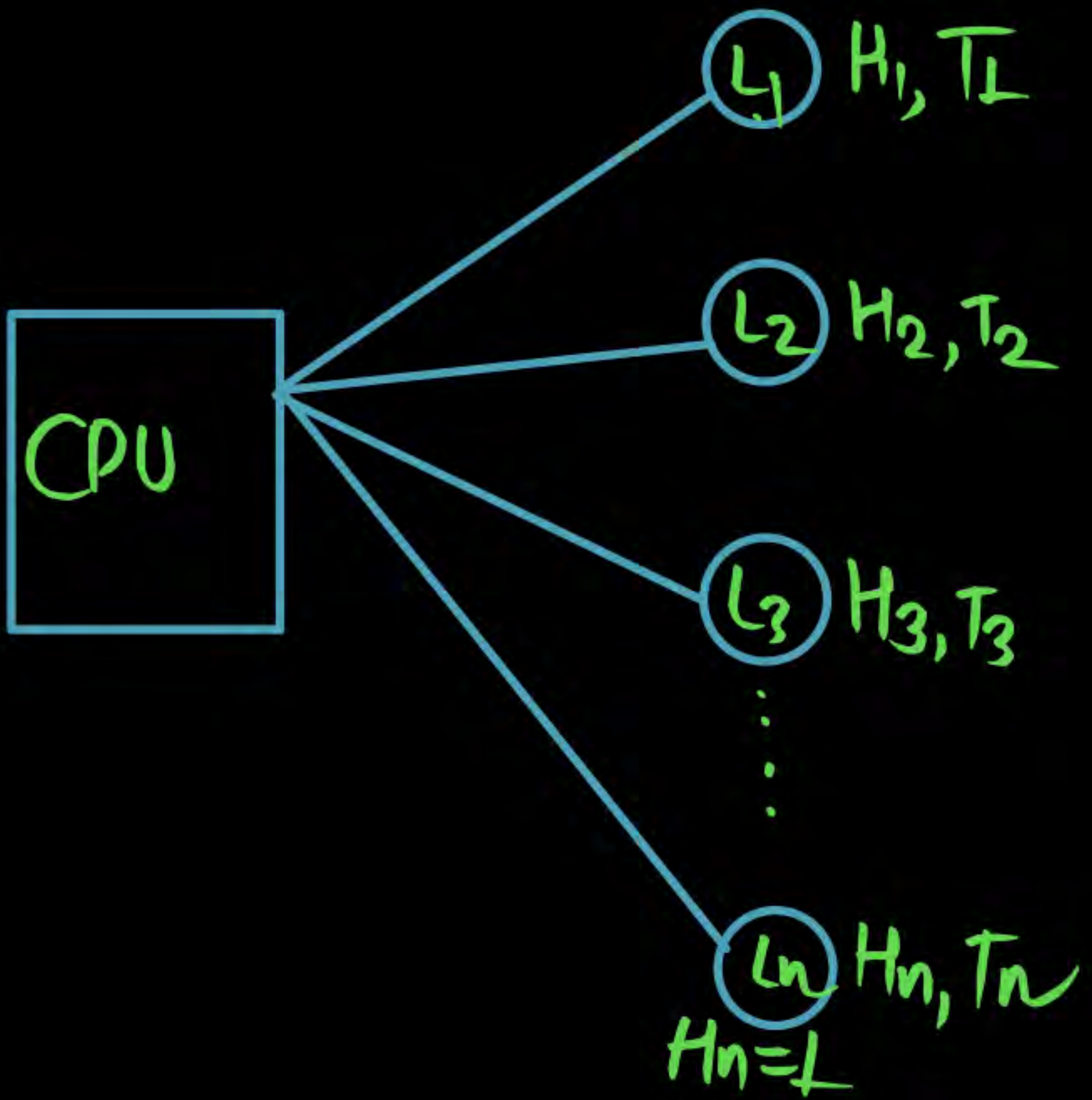
t_m : Main Memory Access time



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

Type of Memory Org

1. Simultaneous Access Memory Org.

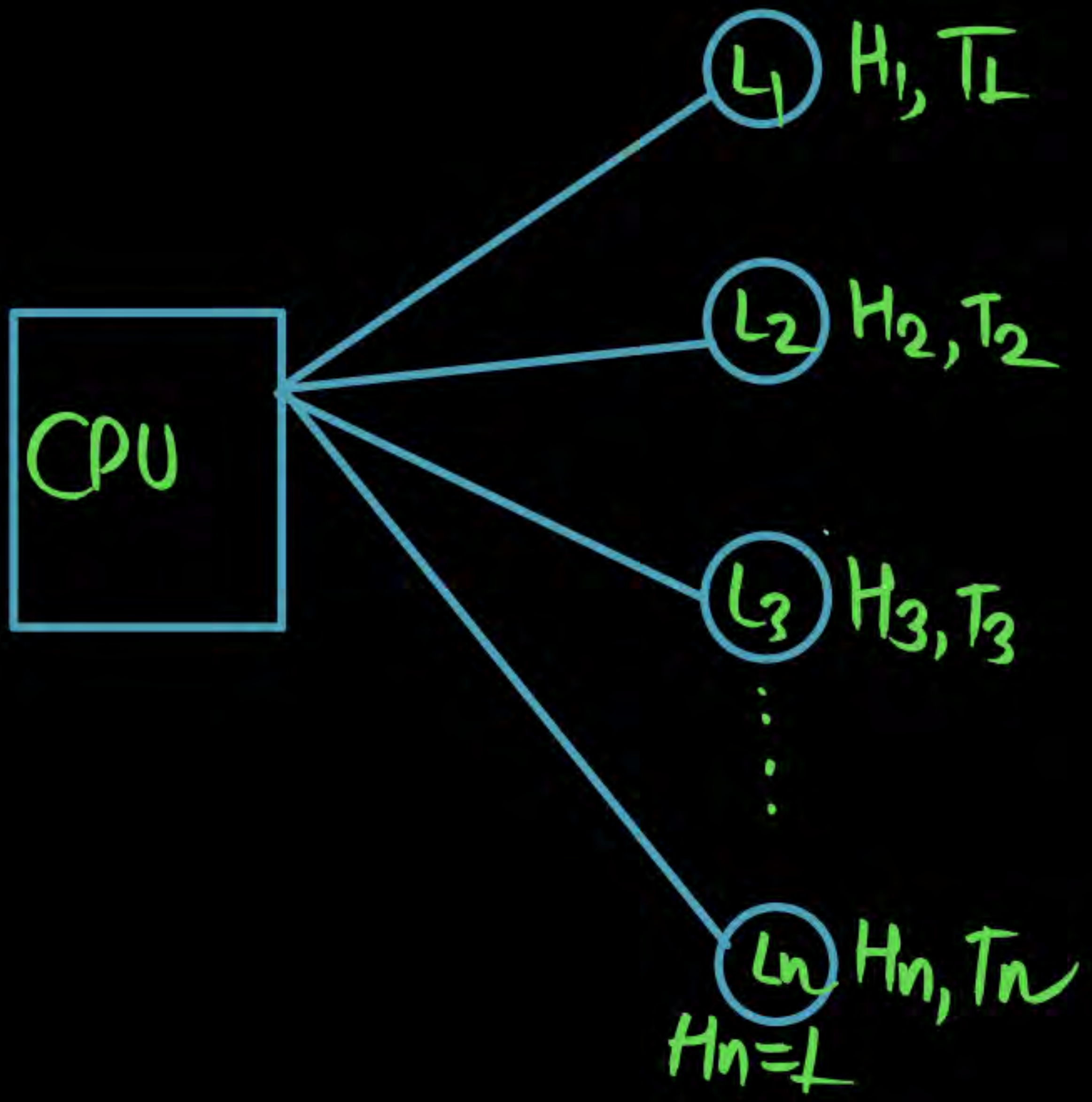


In the Simultaneous Access org ALL the level of Memory is Directly connected to CPU. But follow in Sequence (Acces in a sequence)

- when there is a Miss in level 1 Memory then Reference forword to level 2 Memory. when there is a Hit in Level 2 Memory then Directly Data is transfered from level 2 to CPU without copying into level 1 Memory.

Type of Memory Org

1. Simultaneous Access Memory Org.



When there is a Miss in Level1 & Level2 Memory & Hit in Level 3 Memory then Directly Data is transfered from Level3 Memory to CPU Without Copying into Level1 & Level 2 Memory.

Type of Memory Org

1. Simultaneous Access Memory Org.

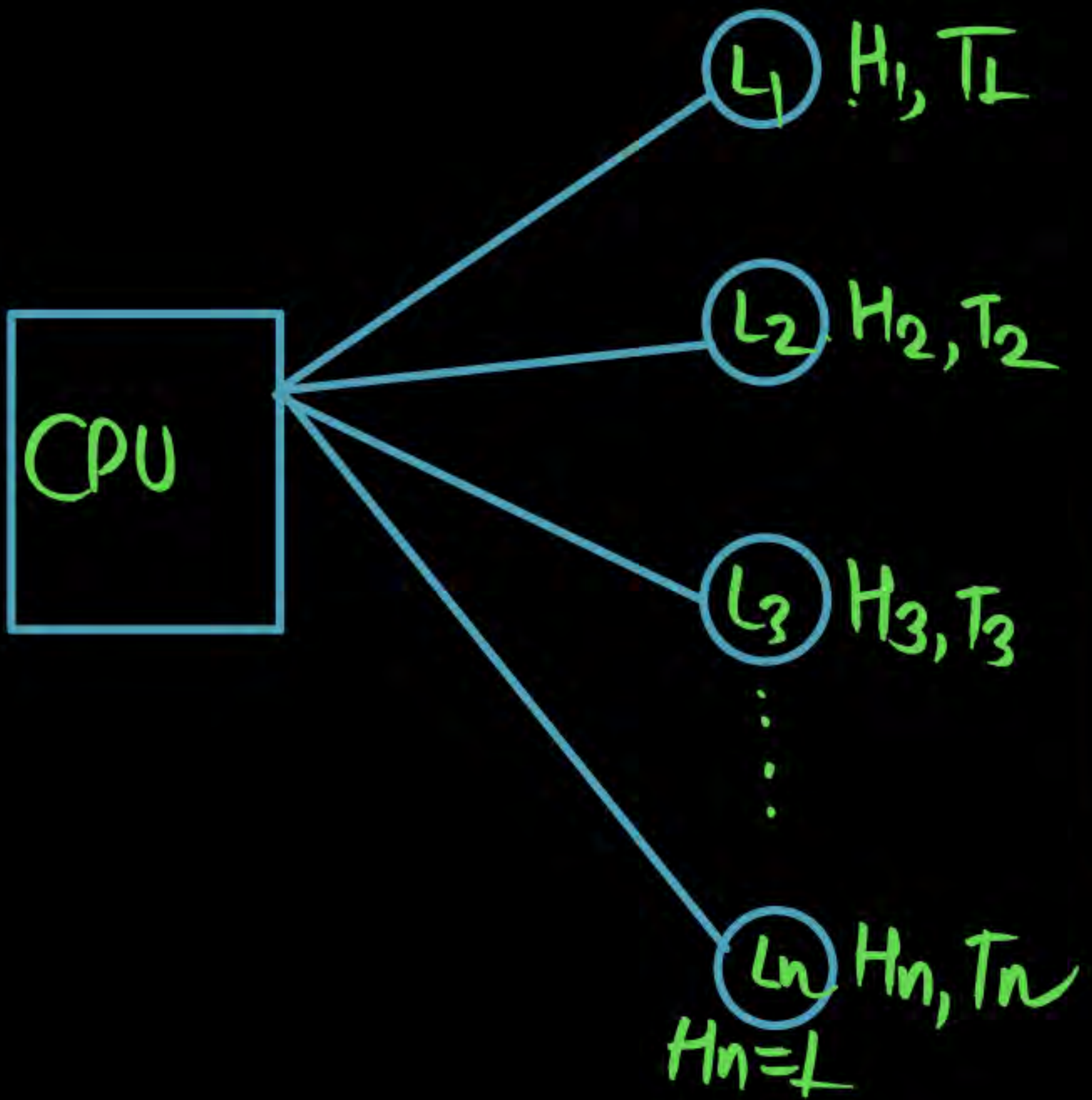
h_i: Hit Ratio of Level i

$(1-H_i)$: Miss Ratio of level i.

Here H_1 H_2 H_3 . . . H_n are Hit Ration f

T_1, T_2, T_3 . . . . . T_n are the Access of Respective level memory.

The Time to Required to Access [Read|write] 1 Word from Memory

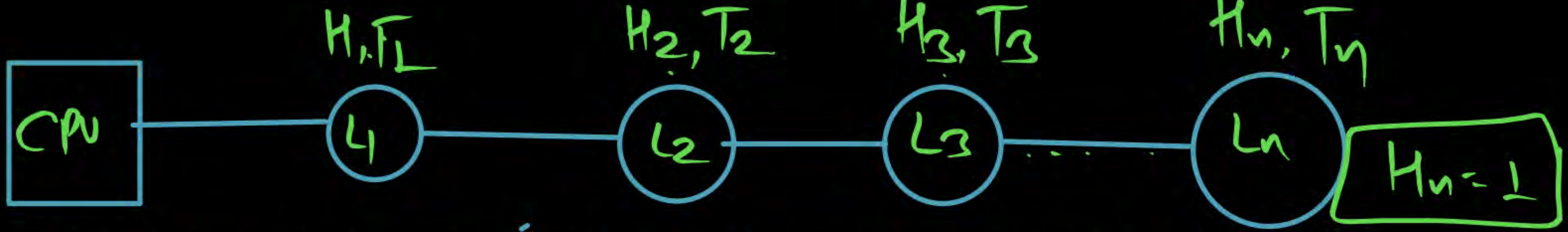


$$\boxed{T_{avg} = h_1t_1 + (1-h_1)h_2t_2 + (1-h_1)(1-h_2)h_3t_3 + \dots \dots (1-h_1)(1-h_2)(1-h_3)\dots\dots(1-h_{n-1})H_n t_n.}$$

$\boxed{H_n = 1}$ Last Level Hit Ratio = 1

Type of Memory Org

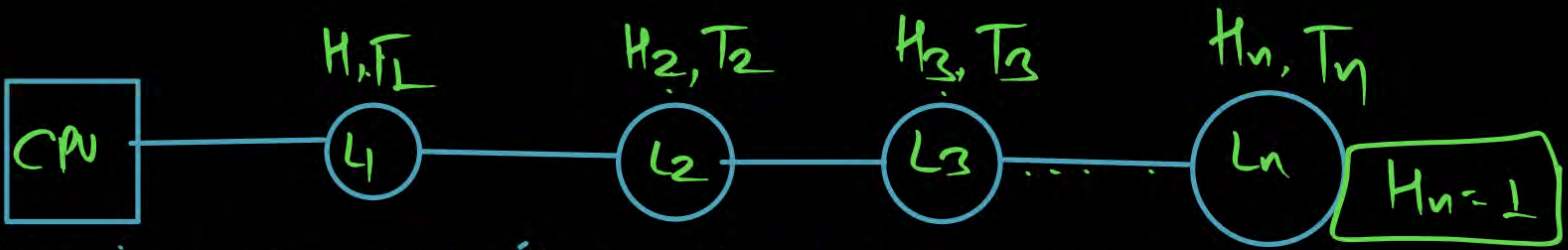
2. Hierarchical Access Memory Org.



- In the Hierarchical Access CPU is Communication with only Level 1 Memory.
- If there is a Miss in Level 1 Memory & Hit in Level 2 Memory then first Data is transfered from Level 2 Memory to Level 1 Memory then Level 1 Memory to CPU.

Type of Memory Org

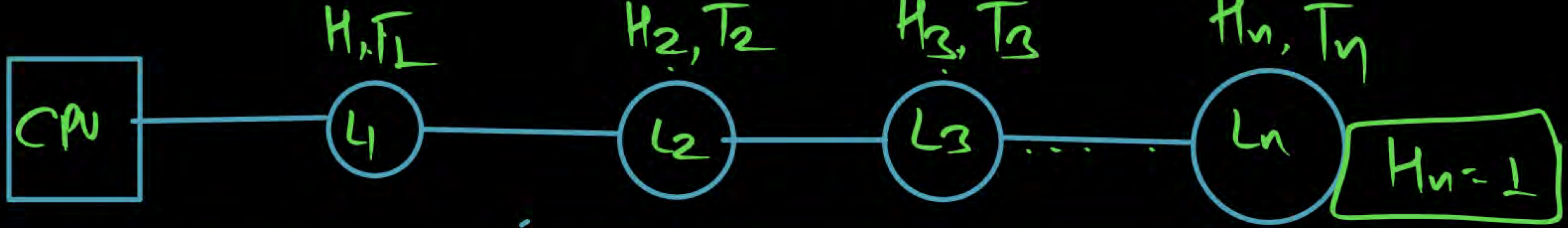
2. Hierarchical Access Memory Org.



When there is a Miss in Level1 & level2 memory & Hit in level 3 memory then firstly Data is transfered from Level 3 [L3] Memory to Level 2 [L2] Memory then Level2 [L2] Memory to Level1 [L1] Memory then from Level1 [L1] Memory to CPU.

Type of Memory Org

2. Hierarchical Access Memory Org.



$$T_{avg} = H_1T_1 + (1-H_1)H_2(T_2+T_1) + (1-H_1)(1-H_2)H_3(T_3+T_2+T_1) + \cdots + (1-H_1)(1-H_2)(1-H_3)\cdots(1-H_{n-1})H_n(T_n+T_{n-1}+\cdots T_3+T_2+T_1)$$

$\boxed{H_n = 1}$ last level Hit Ratio $= 1$.

Note

If in a Question Metioned the Keyword 'Hierarchical Access' or Level of Access or Hierarchical Meaning then Using Hierarchical 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

$t_c = 1 \text{nsec}$ $\quad h = 90\%$

$t_m = 100 \text{nsec}$ $\quad h = 0.9$

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

$$\Rightarrow 0.9 \times 1 + (1-0.9)(100+1)$$

$$= 0.9 + (0.1)(101)$$

$$= 0.9 + 10.1$$

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

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?

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

$L_1 \Rightarrow T_1$

$L_2 \Rightarrow T_2$

Sol^n

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

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

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

$$T_2 = 5 \times 20$$

$$\boxed{T_2 = 100\text{nsec}}$$

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

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

$$T_1 = 20\text{ns}$$

$$T_{avg} = T_1 + 10 \Rightarrow 20 + 10 \quad \boxed{T_1 = 30\text{nsec}}$$

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

$$30 = H * 20 + (1-H)100$$

$$30 = 20H + 100 - 100H$$

$$80H = 70$$

$$H = 70/80 = 0.875$$

$$\Rightarrow 87.5\% \ \underline{\text{Ans}}$$

Q.

~~H.W~~

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

PART I

(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?

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

PART II

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 14 nsec Ans [GATE – 2015]

Hit Ratio = 80 = 0.8

Miss Ratio = (1 − 0.8) = 0.2

When there is hit
Time taken = 5.

When there is a
Miss timetaken = 50

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

$$= 4 + 10$$

$$= 14\,\text{nsec} \text{ (Ans)}$$

Ans (14).

THANK YOU!