COMPUTER SCIENCE

Computer Organization and Architecture

Cache Memory

Lecture_02

Vijay Agarwal sir

TOPICS TO BE COVERED

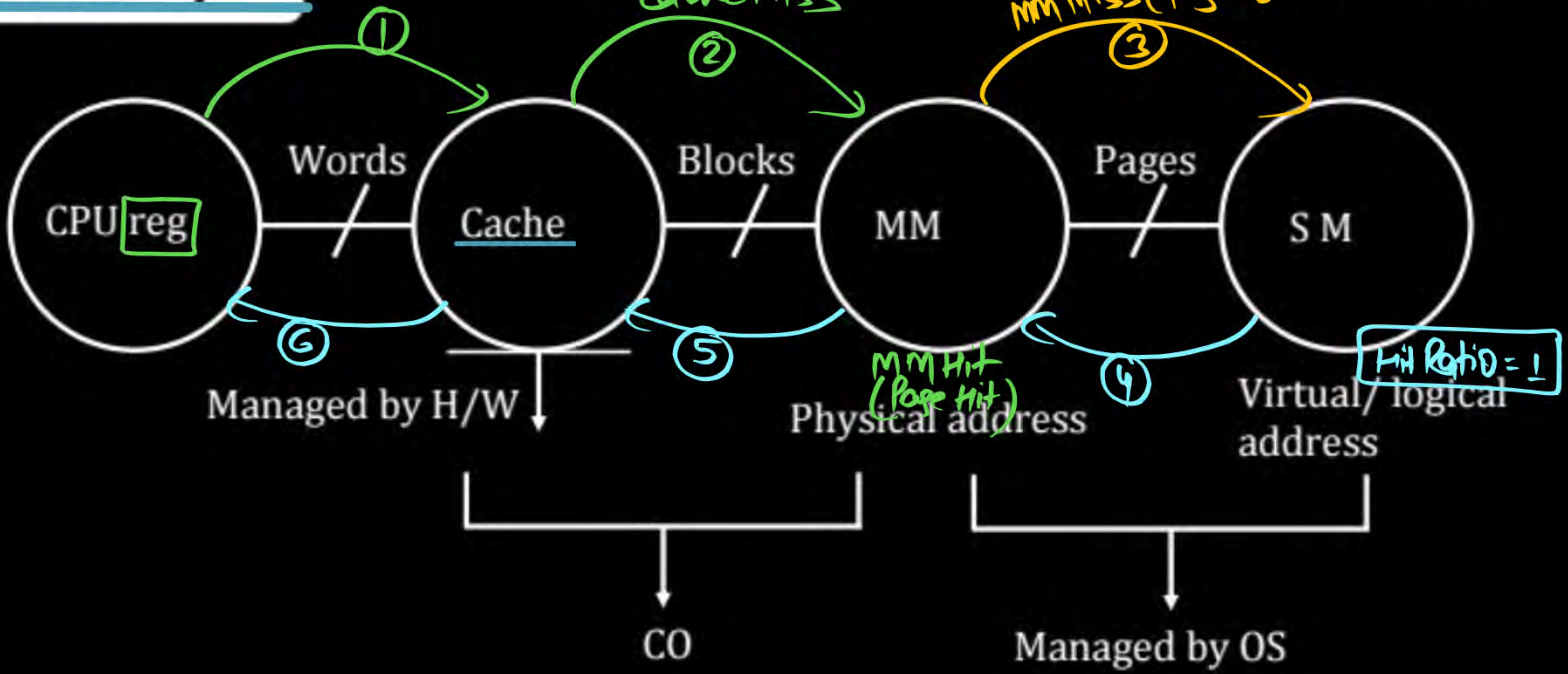
01 Memory Access

02 Cache Memory

Memory Hier.

Type of Access.

Memory



CPU reg

Words

Cache

Blocks

MM

Pages

S M

Managed by H/W

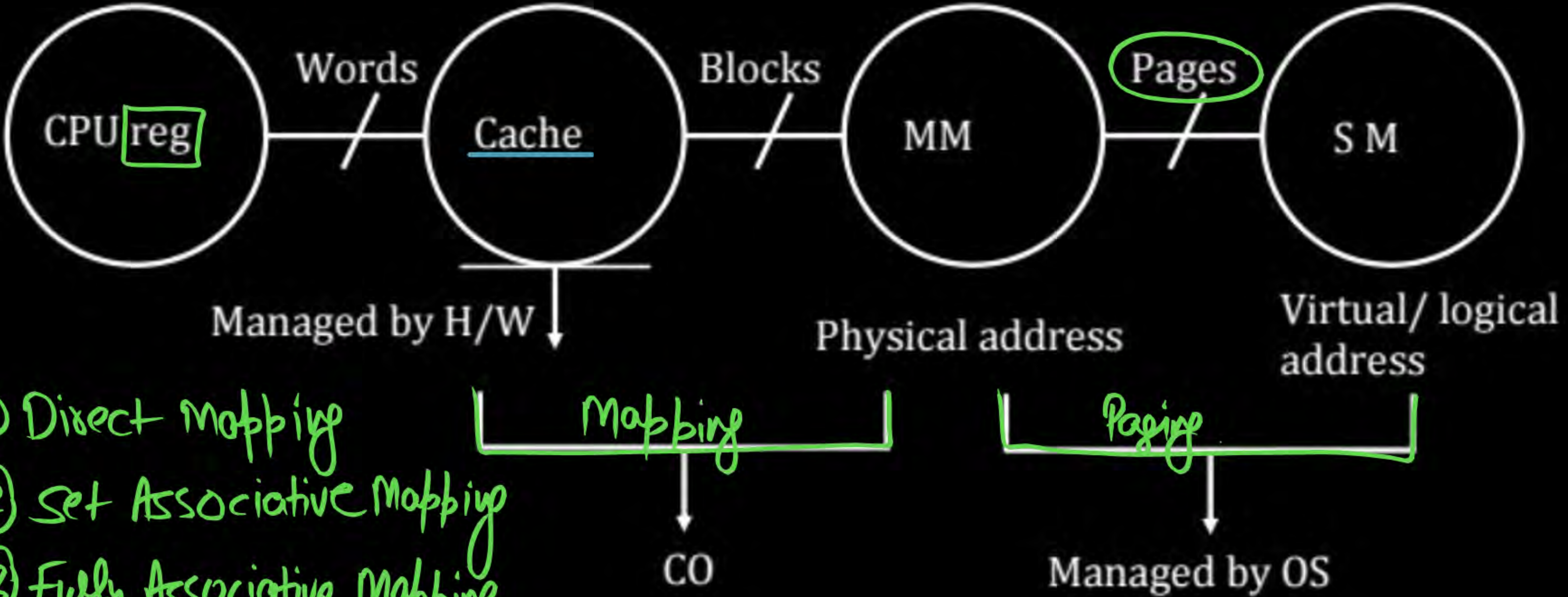
Physical address

Virtual/ logical address

CO

Managed by OS

Memory



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

Q. 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. ?

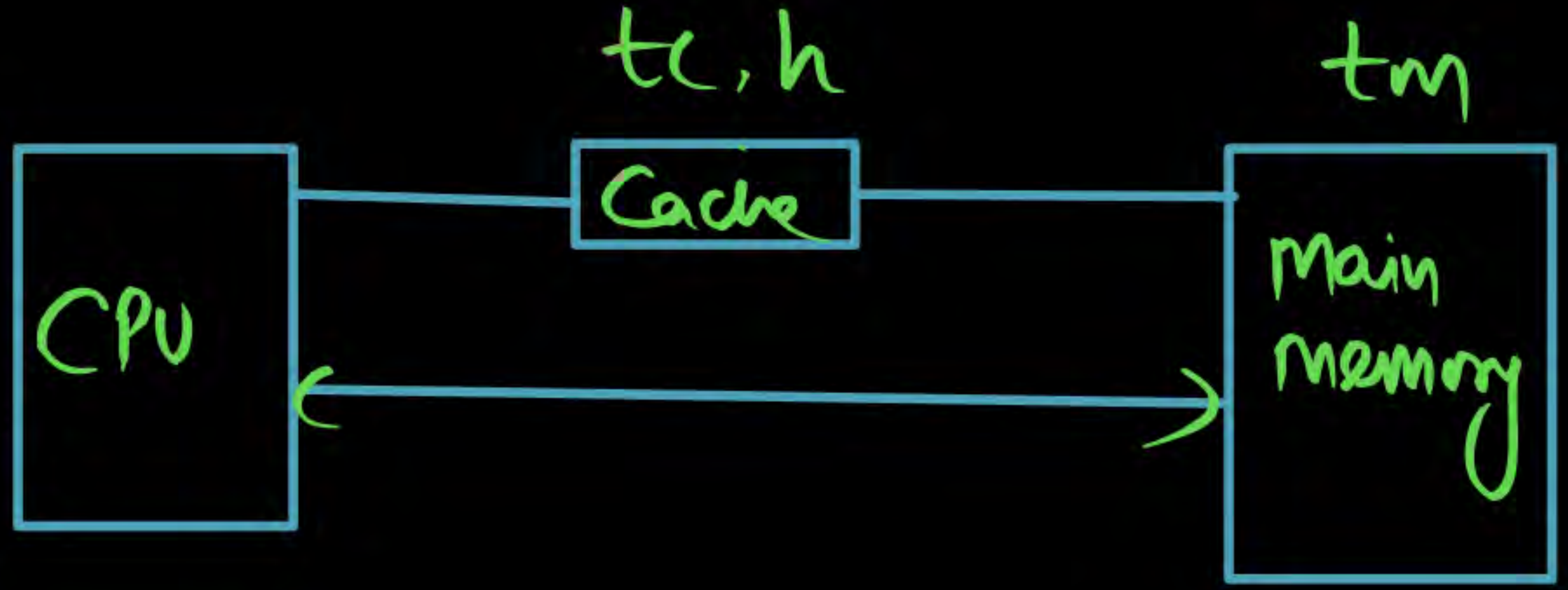
Type of Memory Org

1. Simultaneous Access Memory Org. : 2level

h : Cache Hit Ratio.

tc : Cache Access time

tm : MM Access time

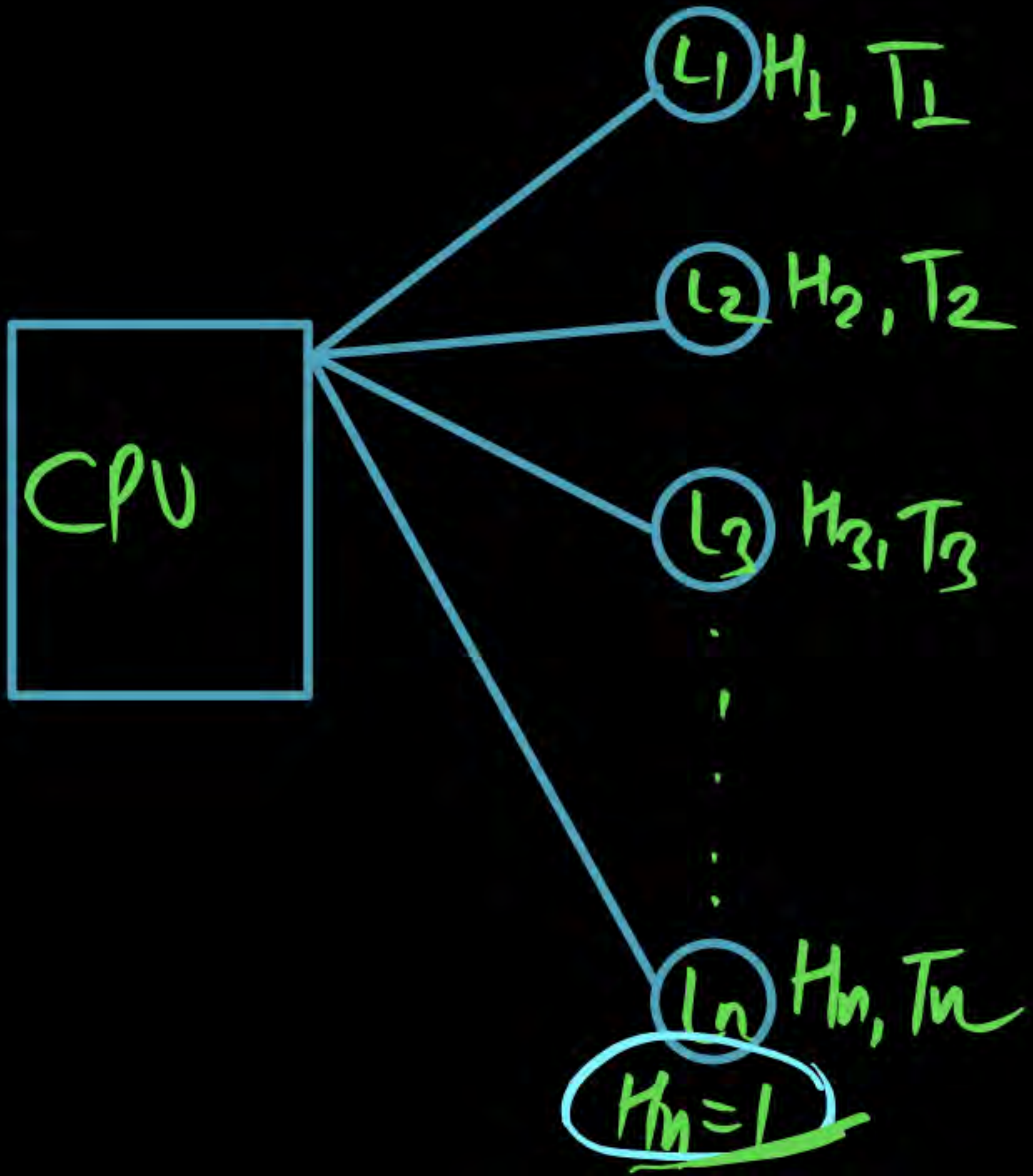


1 Word Access time Tavg.

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

Type of Memory Org

1. Simultaneous Access Memory Org.: n Level



$$\boxed{\begin{aligned} T_{avg} = {} & H_1T_1 + (1-H_1)H_2T_2 + (1-H_1)(1-H_2)H_3T_3 \\ & + \ldots (1-H_1)(1-H_2)(1-H_3)\ldots(1-H_{n-1})H_nT_n \end{aligned}}$$

Type of Memory Org

1. Simultaneous Access Memory Org.

1 Word Access time = T_{avg}.

Data transfer Rate. $\left(\#\text{Words}/\text{sec}\right)$ (Performance) (efficiency) $= \dfrac{1}{T_{avg}}$.

Remember

Avg Instn ET $= 5.51 \times 10^{-9}$ sec.

1 Instn ET $= 5.51 \times 10^{-9}$ sec.

In 1 Sec ⇒ How Many # Instn

In 1 sec $= \dfrac{1}{5.51 \times 10^{-9}}$ Instn/sec

$\Rightarrow \dfrac{1}{5.51} \times 10^{9}$ Instn/sec

$\Rightarrow \dfrac{1000 \times 10^{6}}{5.51} \Rightarrow 181.4 \times 10^{6}$

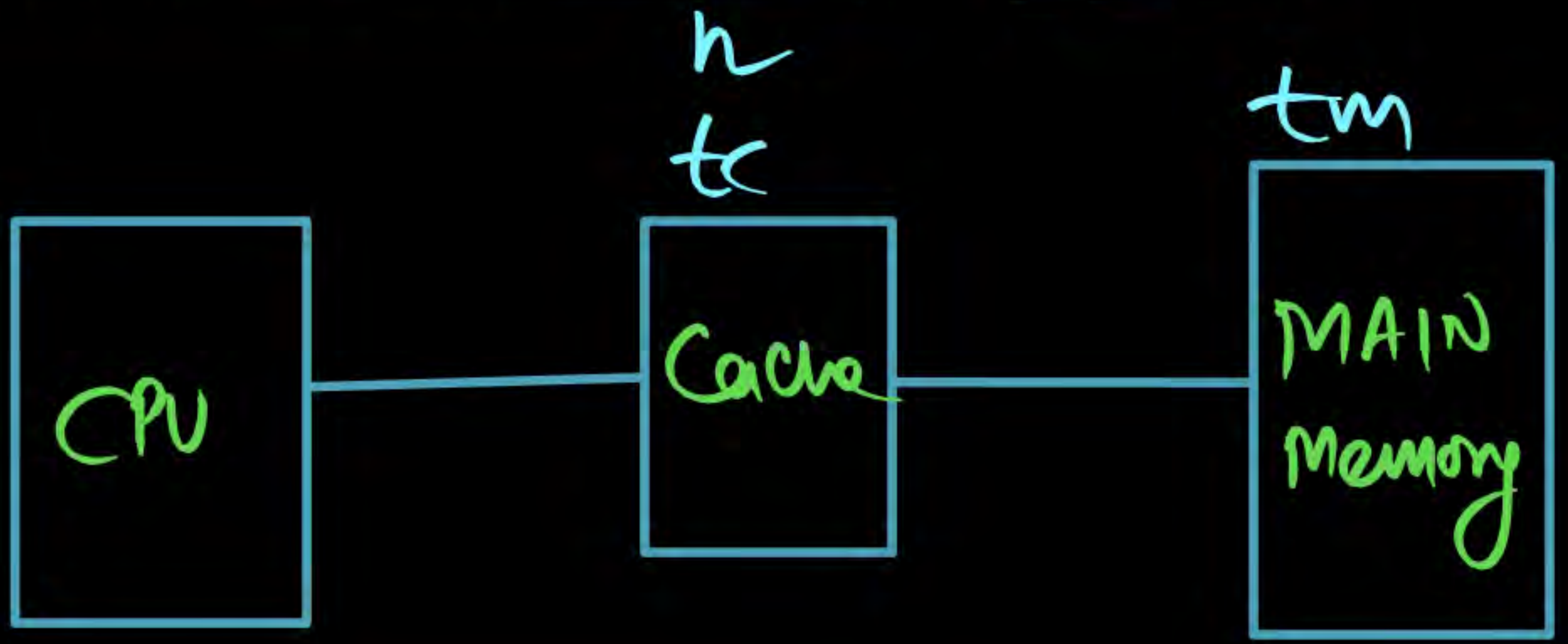
$= 181.4$ MIPS

Type of Memory Org

1. Simultaneous Access Memory Org.

Type of Memory Org

2. Hierarchical Access Memory Org. : 2 level



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

Hierarchical Access:

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

Hierarchical Access

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

$$\Rightarrow \cancel{h t_c} + t_m - h t_m + t_c - \cancel{h t_c}$$

$$\Rightarrow t_c + t_m - h t_m$$

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

eg Hierarchal Access $\underline{\text{Hit Ratio } 80\%}$, $t_c = 20ns$. MM $= 100nsec$

T_{avg}.

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

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

$\Rightarrow 16 + 0.2\,[120]$

$\Rightarrow 16+24$

$\boxed{T_{avg} = 40nsec}$ $\underline{\text{Ans}}$

$t_c + t_m$

$(1-0.8)\,[20+100]$

$\Rightarrow \underline{0.8 \times 20} + \underline{0.2}\,[20 + 100]$

$\Rightarrow \underline{0.8 \times 20} + \underline{0.2(20)} + 0.2(100)$

$\Rightarrow 20 + 0.2(100)$

$\Rightarrow 40 nsec$ $\underline{\text{Ans}}$

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

$\Rightarrow 20 + (1-0.8)\,100$

$= 20 + (0.2)\,100$

$\Rightarrow 20+20 \Rightarrow 40nsec$

$h = 0.8$

$(1-h) = 1-0.8$

$= 0.2$

Type of Memory Org

h_1: Hit Ratio in level

$(1-h_1)$: Miss Ratio in Level 1 (M_1)

$(1-h_2)$: Miss Ratio in Level (M_2)

2. Hierarchical Access Memory Org. 3 Level

$$\boxed{T_{avg} = h_1t_1 + (1-h_1)h_2(t_2+t_1) + (1-h_1)(1-h_2)h_3(t_3+t_2+t_1)}$$

$$\boxed{T_{avg} = h_1t_1 + M_1h_2(t_2+t_1) + M_1M_2h_3(t_3+t_2+t_1)}$$

M_1: Miss Ratio in level 1

M_2: Miss Ratio in Level 2.

Type of Memory Org

2. Hierarchical Access Memory Org. 3 level

$$T_{avg} = h_1 t_1 + (1-h_1) h_2 (t_2 + t_1) + (1-h_1)(1-h_2) h_3 (t_3 + t_2 + t_1)$$

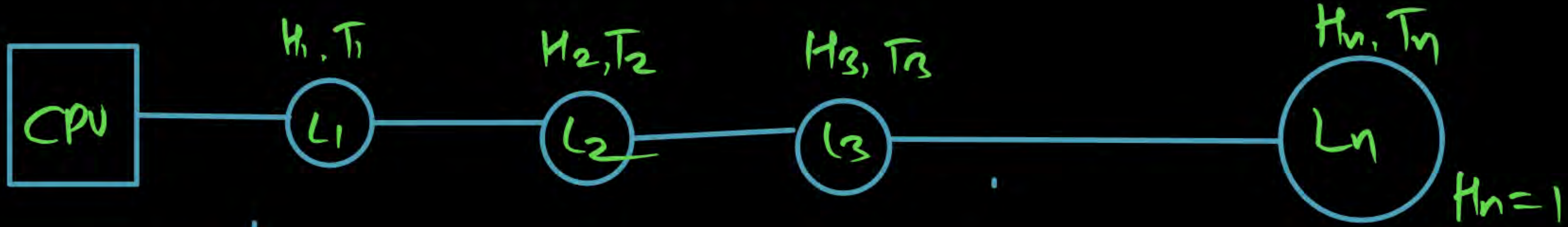
#words/sec $= \frac{1}{T_{avg}}$

(Performance)

(Data transfer Rate)

Type of Memory Org

2. Hierarchical Access Memory Org. : n level



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

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

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

11 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?

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 made to 100% then what is the Access time of L_1 & L_2 Memory?

→ Hit Ratio Not Effect levels Access time Only Tavg change.

$T_1 = 20ns$

$T_2 = 150ns$

$T_{avg} = 30ns$.

Simultaneous

$T_{avg} = ht_1 + (1-h)t_2$

$30 = h*20 + (1-h)150$

$30 = 20H + 150 - 150H$

$120 = 130H$

$H = \frac{120}{130} = 0.923$

$\Rightarrow \boxed{92.3\%}$ Ans

(ii) If Hit Ratio Made to 100% then Access time of L1 & L2 Memory.

$\left.\begin{array}{l} T_1 = 20ns \\ T_2 = 150ns \end{array}\right\}$ L1 & L2 Access Remain Same.

If Hit Ratio = 100% → $\boxed{h=1}$ $(1-h) = 0$

$T_{avg} = ht_1 + (1-h)t_2$

$\Rightarrow 1 \times 20 + \underbrace{(1-h)150}_{0}$

$T_{avg} = 20 ns$

Only Tavg change.

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

$T_1 = 20 \quad T_2 = 150, \quad T_{avg} = 30 \Rightarrow$ Now T_{avg} increased by 10% $\Rightarrow 30 + 10\%$ of 30

$\Rightarrow 30 + 3$

$= 33$ nsec

$$\boxed{T_{avg_{new}} = 33 \text{ nsec}}$$

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

$$33 = h * 20 + (1-h) 150$$

$$33 = 20H + 150 - 150H$$

$$130H = 117$$

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

$$\boxed{H = 0.9}$$

$$\boxed{\text{Now Hit Ratio} = 90\%}$$

Previous Hit Ratio = 92.3%

Hit Ratio Decreased by

$\boxed{2.3\%}$ ↓

$92.3 - 90 = 2.3\%$

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 14nsec Ans [GATE – 2015]

$$T_{avg} = h * 5 + (1-h) * 50$$

$$= 0.8 \times 5 + (1-0.8) \times 50$$

$$= 4 + 10$$

$$\boxed{T_{avg} = 14\,nsec}$$ Ans

NAT

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 14nsec .

(i) What is Data Transfer rate (performance) of this memory system (in words/sec)?

(ii) What is Bandwidth required of this memory system if word size is 8bit?

(i) Data transfer Rate $= \frac{1}{T_{avg}}$ words/sec

$\Rightarrow \frac{1}{14\times10^{-9}}$ Words/sec

$\Rightarrow \frac{1}{14}\times10^{9}$ Words/sec

$\Rightarrow \frac{1000}{14}\times10^{6}$ Words/sec

$\Rightarrow 71.42\times10^{6}$ Words/sec

$\Rightarrow$ 71.42 millions word Per sec.

$\Rightarrow$ 72 Millions word Per Sec.

72 Millions word Per Sec

(ii) <u>Bandwidth</u> ⇒ Word Size = 8 bits

⇒ 72 Million word Per sec ⇒ 72×10^6 word Per Sec.

⇒ $72 \times 10^6 \times$ 8 bits Per Sec ⇒ 72×8 Mbits/sec

⇒ $\boxed{576 \text{ M bits/sec}}$ ⇒ $\frac{576 \text{ Mbits}}{8 \text{ bits}}$ Byte [1Byte = 8bit]

OR

⇒ $\boxed{72 \text{ MBps}}$ Byte Ans

Cache Work on Locality of Reference.

Locality of Reference [LOR]

Accessing the Higher Level of Memory Data from the Level 1 Memory [Cache Memory] is Called Locality of Reference.
(Faster Memory)

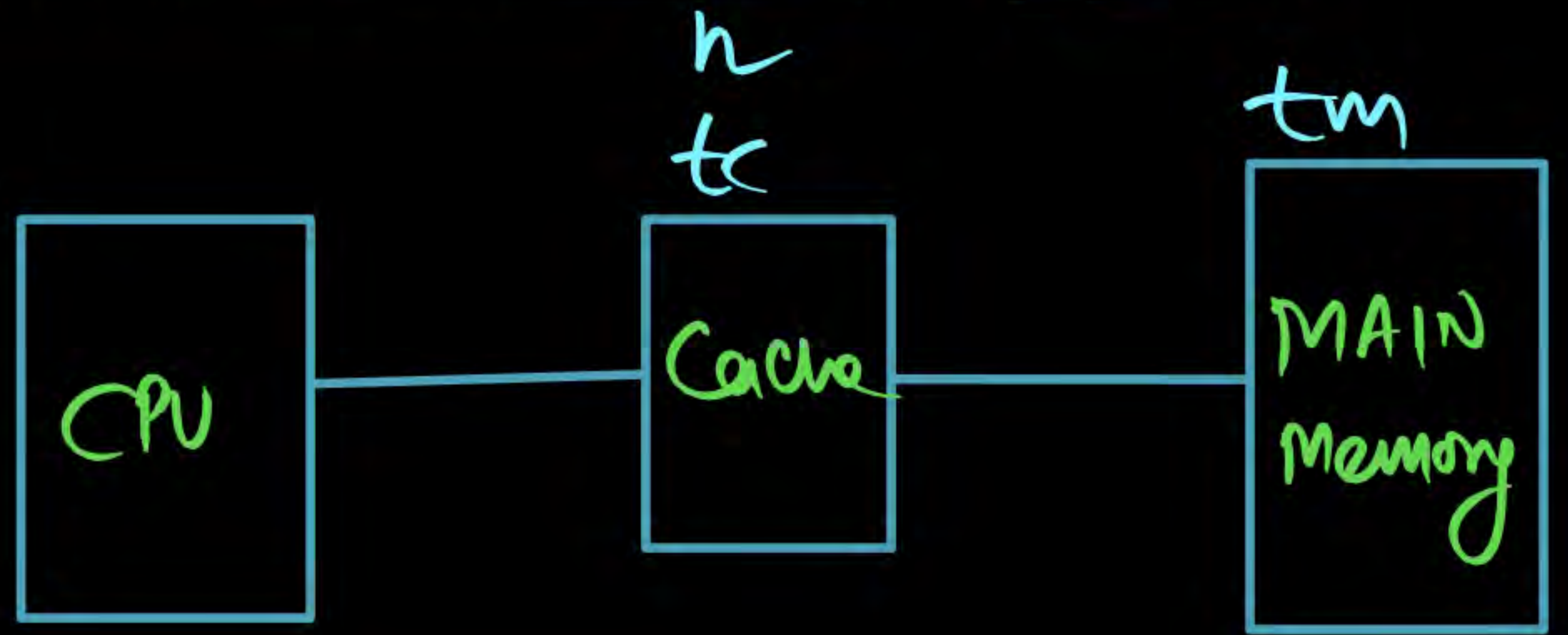
Locality of Reference [LOR]

✓ ① Temporal LOR

✓ ② Spatial LOR

Type of Memory Org

2. Hierarchical Access Memory Org. : 2 level

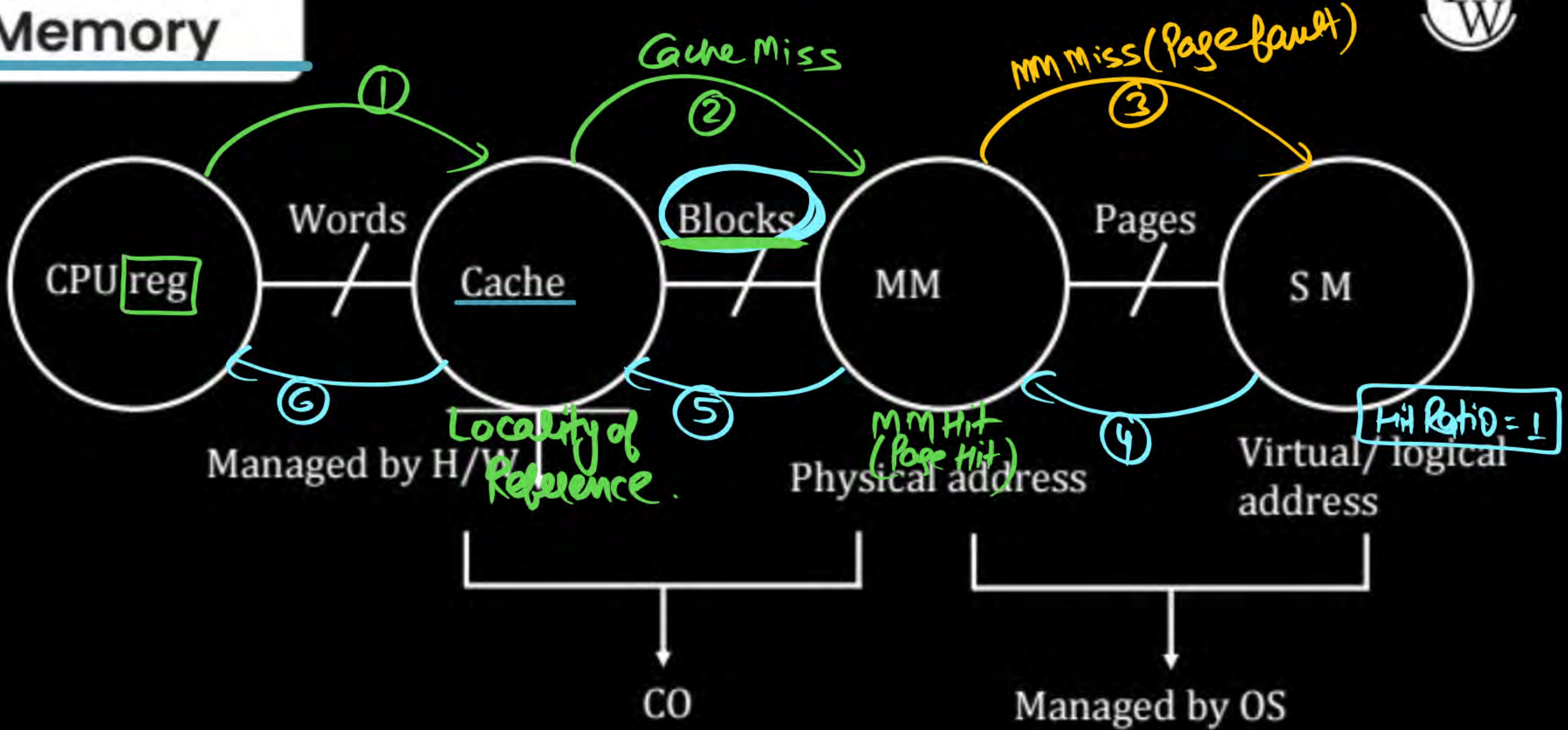


$$\boxed{T_{avg} = h*t_c + (1-h)[t_m+t_c)}$$

Hierarchical Access:

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

Memory



CPU reg

Words

Cache

Blocks

MM

Pages

S M

Managed by H/W

Physical address

Virtual/ logical address

CO

Managed by OS

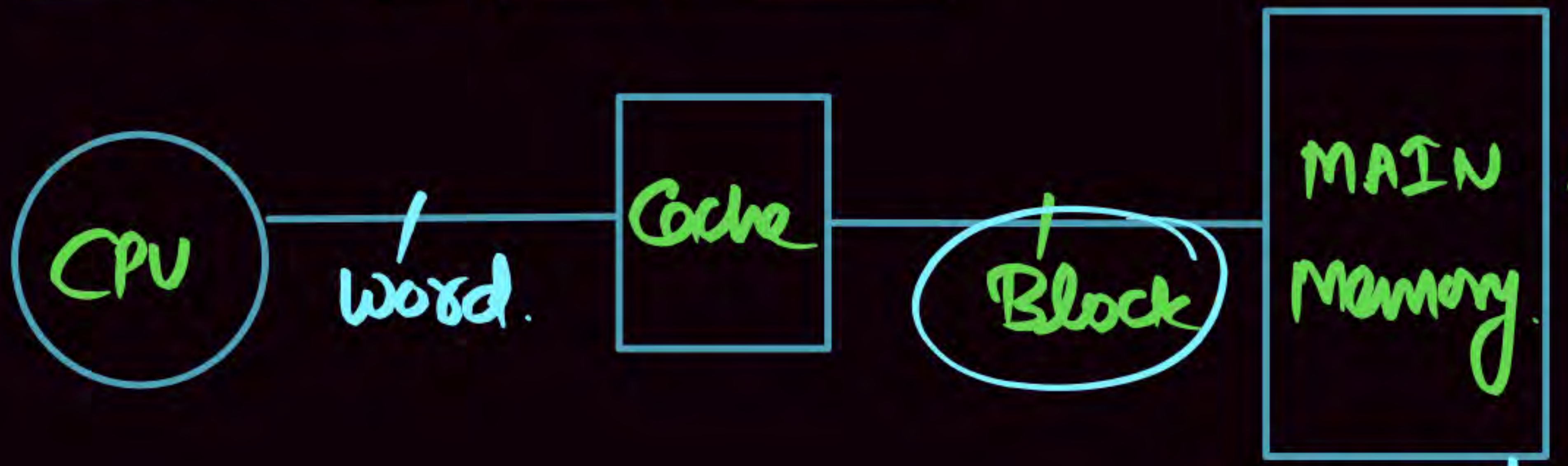
Non Technical eg

LOR: Medical Store

Only 1 Time go to Medical Store Buy One Patta of 10/20 Tablet.
(strips)

So Next time if Same Tablet Require then No Need to go Medical store, take it from Home.

Locality of Reference:



Assume

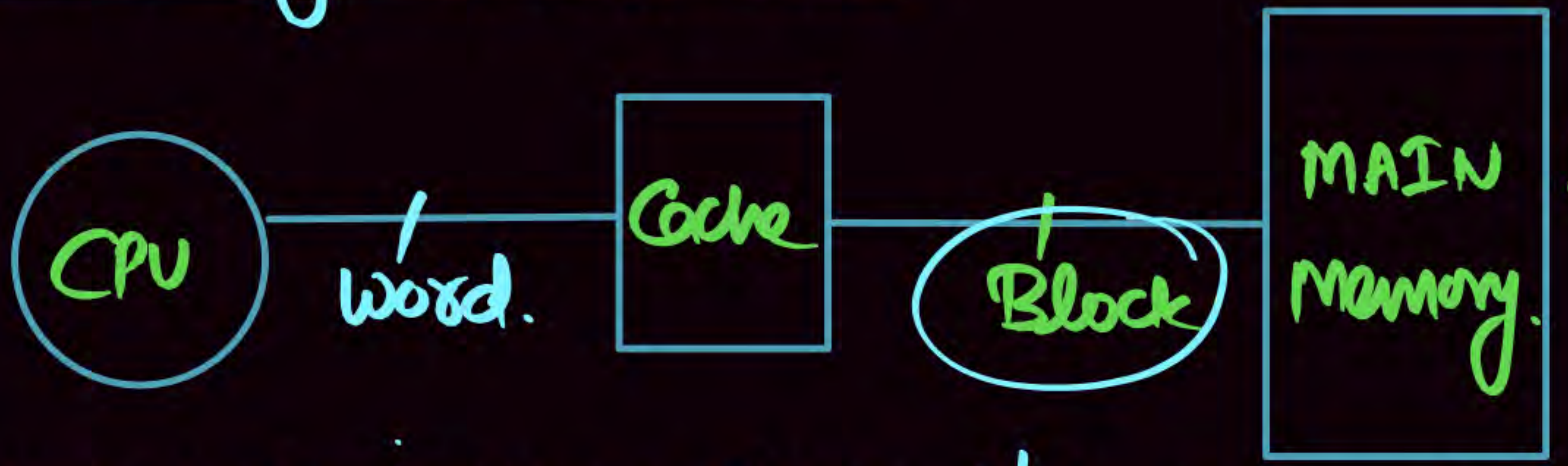
1 Block Size = 32 Byte | 32 Words

(32 Word | 32 Byte)

If Miss in Cache then 1 Complete Block is Transfered From Main Memory to Cache. the Respective Word (Which is Requested | Demanded by CPU) give from Cache Memory to CPU.

(Next Reference there is a Cache hit).

Locality of Reference:



1 Block Size = 32 Byte | 32 Words

Q Here 1 Block Size = 32 Word | 32 Byte & CPU Require Only 1 Word. ?

Sol^n In this Process 1 Complete Block of 32 Words | 32 Byte transfered from Main memory to Cache Memory. then Demanded (Requested) 1 Word transfered from Cache Memory to CPU.

Note So in the Next time When CPU Request Some Word or Adjacent Word then that Request Available in Cache. [Cache Hit].

Locality of Reference [LOR]

LOR [Locality of Reference]

Faster memory [Cache Memory]

□ Access the higher level of memory Data from level 1 Memory is called L.O.R,.

eg 32 word

(1) Temporal LOR

(2) Spatial LOR.

Cache

B5 ← M[2000]

M[2000] [x] memory location word Available in B5.

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

M[2000] location wallah word Again & Again

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.

Memory $(x+1)$ Wallah Word location which is Available in B_5.

Cache hit

Types of Cache

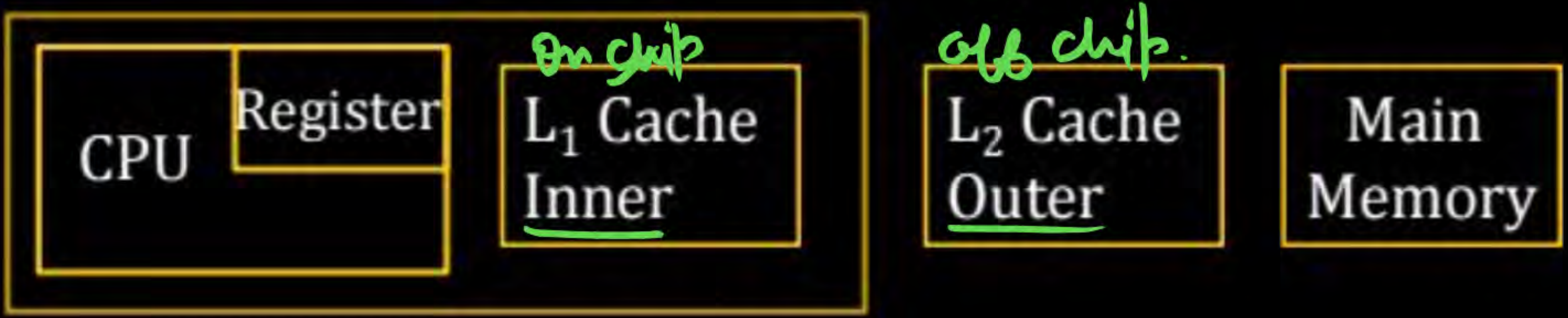
1) Unified Cache: Instruction & Data both are placed in Same Cache.

2) Split Cache: This Cache logically Divide into two parts

(i) Instruction Cache [I - cache]

(ii) Data Cache [D- cache]

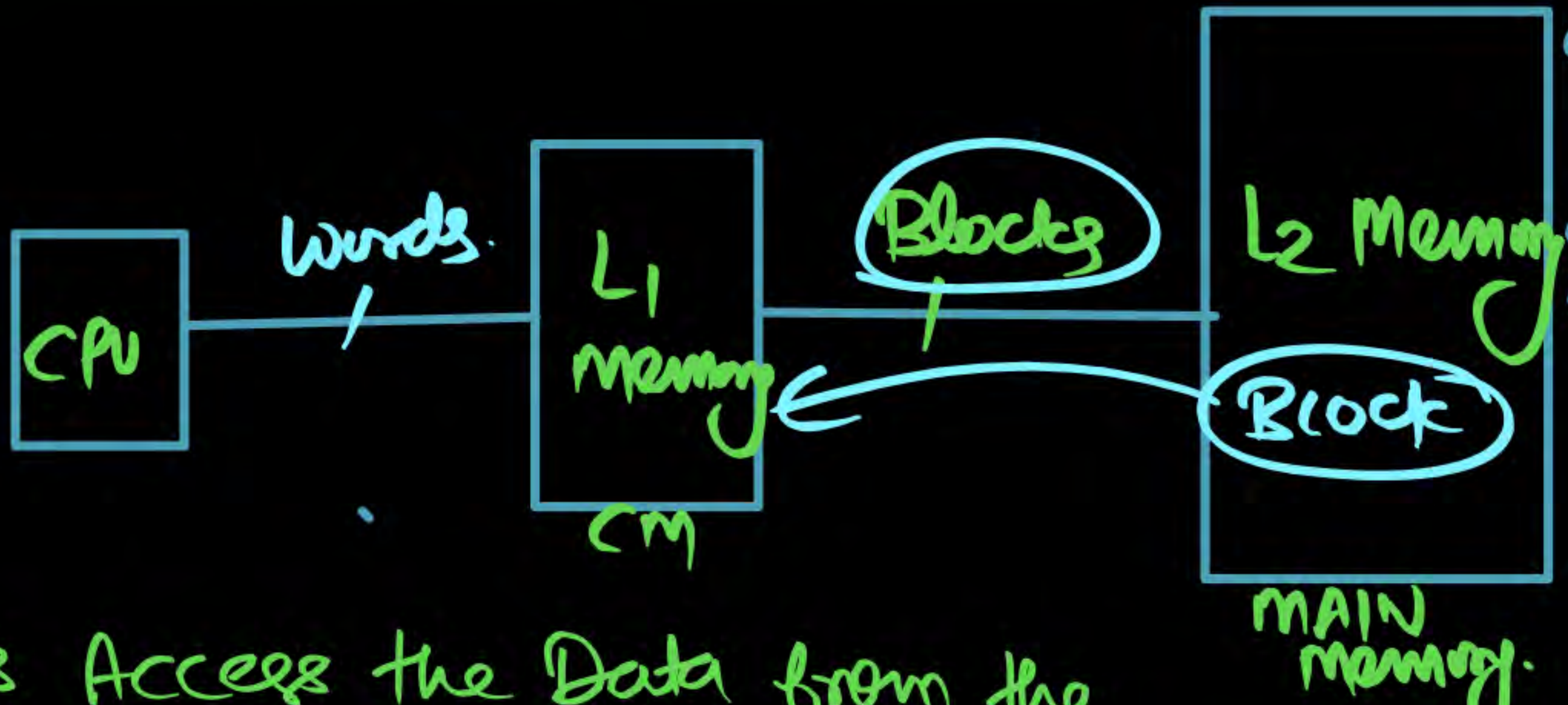
3) Multilevel Cache:



Size $L_1 < L_2$

Speed $L_1 > L_2$

2 Level of Memory (If Locality of Reference Included)

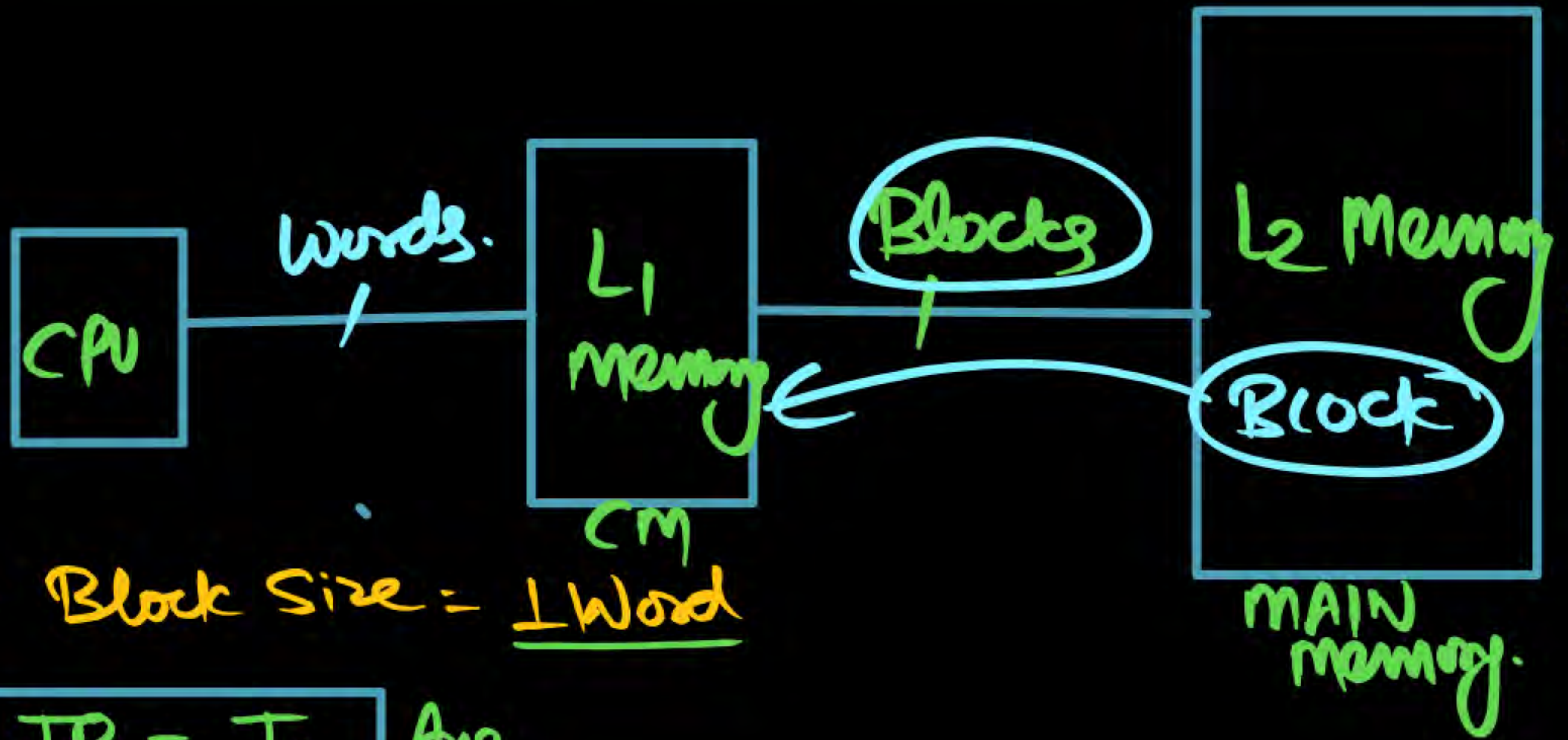


CASE ① Block size = 1 word

CASE Ⅱ Block size = n words.

CPU Always Access the Data from the faster Memory (Level1 | Cache Memory). If there is Miss in Level1 (Cache) Memory then One Complete Block is transfered From Level2 (Slow) Memory to Level1 [Cache] Memory & addressed Word (which is Requested | Demanded by CPU) that Respective Word given from Cache Memory to CPU.

2 Level of Memory (If Locality of Reference Included)



Case I : Block Size : 1 Word

CASE II : Block Size : n words

TB : Block Transfer time from L_2 Memory to L_1 Memory.

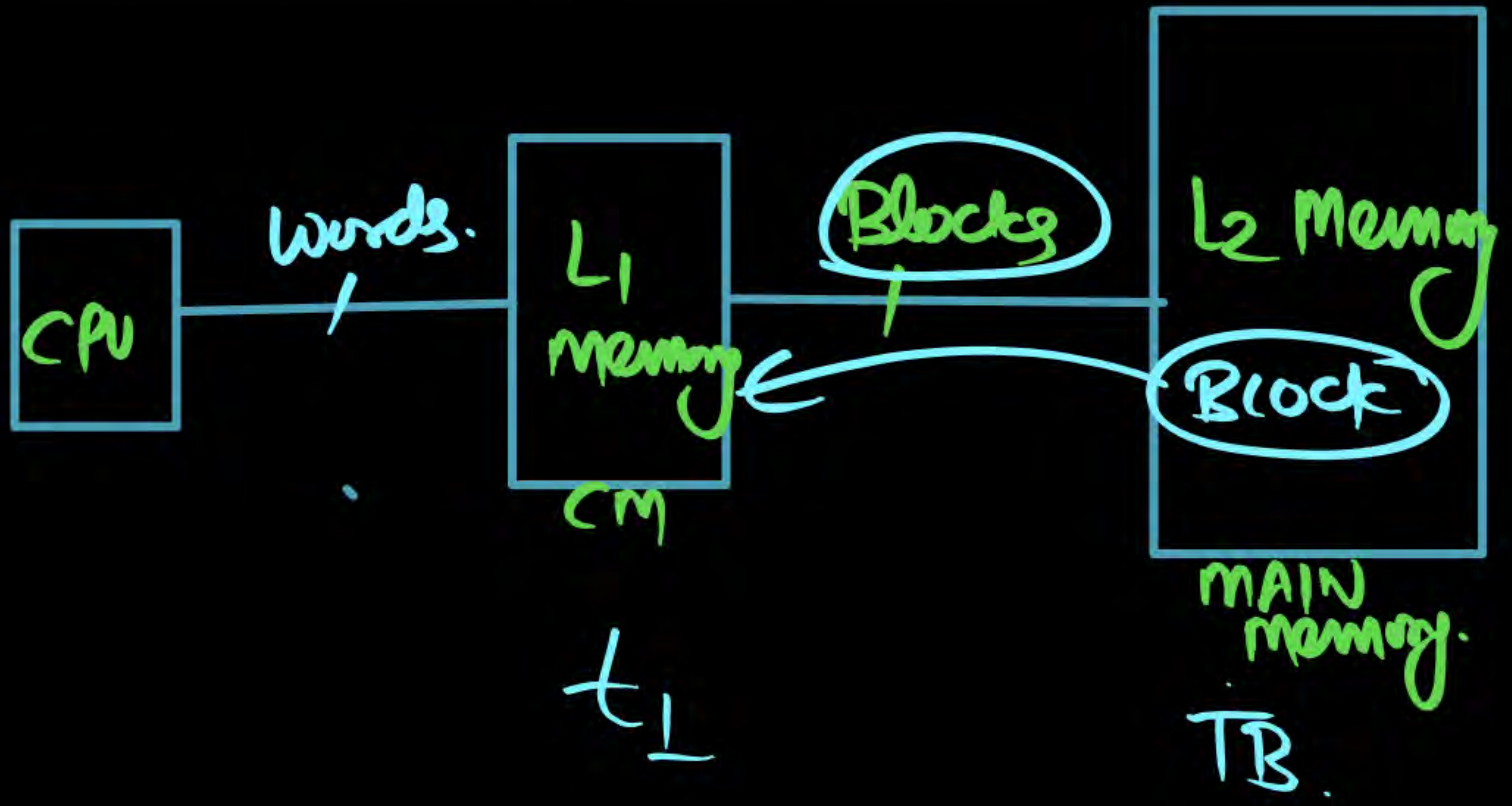
① CASE I : Block Size = 1 Word

$$\boxed{TB = T_2}$$ Ans

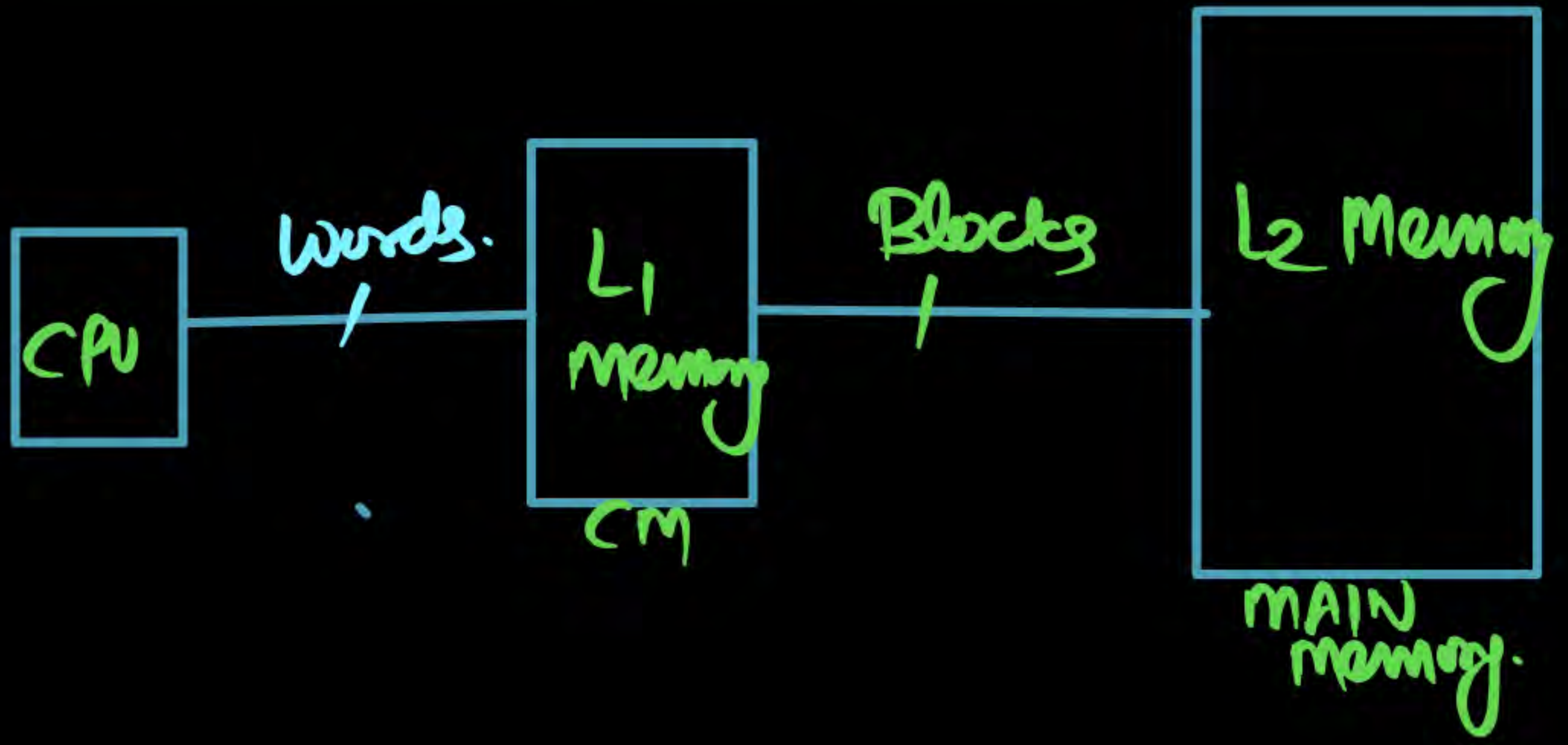
② CASE II : Block Size = n Words

$$\boxed{TB = n * T_2}$$ Ans

2 Level of Memory (If Locality of Reference Included)



2 Level of Memory (If Locality of Reference Included)



If. 2 Level Memory

$\boxed{h_2 = 1}$

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

OR

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

If Locality of Reference Included

$(TB = n * T_2)$

$$\boxed{T_{avg} = h t_1 + (1-h)\left[TB + T_1\right]}$$

OR

$$\boxed{T_{avg} = t_1 + (1-h) TB.}$$

If 3 Level Memory

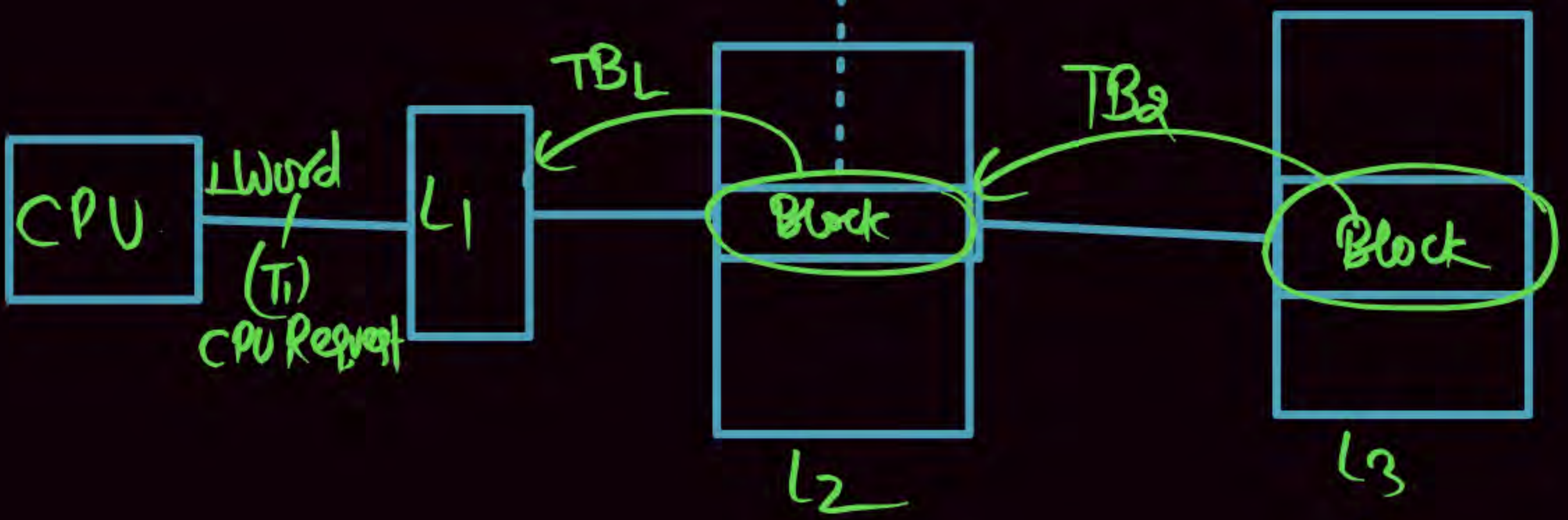
$T_{avg} = h_1 t_1 + (1-h_1) h_2 (t_2 + t_1) + (1-h_1)(1-h_2)(t_3 + t_2 + t_1)$

$h_3 = 1$ [Hit Ratio of last level Memory always].

$$\boxed{\begin{aligned} T_{avg} = {} & h_1 t_1 + (1-h_1) h_2 (t_2 + t_1) \\ & + (1-h_1)(1-h_2)(t_3 + t_2 + t_1) \end{aligned}}$$

If Locality of Reference included.

$$\boxed{\begin{aligned} T_{avg} = {} & h_1 t_1 + (1-h_1) h_2 (TB_1 + T_1) \\ & + (1-h_1)(1-h_2)(TB_2 + TB_1 + T_1) \end{aligned}}$$



Q. In a 3 level memory, level 1 memory Access time is T1 , level 2 memory Access time is T2(TB1) and level 3 memory Access time is T3(TB2). Hit ratio of level 1 is h1 and Hit ratio of level 2 is h2 What is the average Access time Using Hierarchical Access? $h_3 = 1$

(i) If there is a hit in level1(h1=100%). $\boxed{h_1 = 1}$

(ii) If there is a miss in level1 & hit in level2(h2=100%) . $\boxed{h_1 = 0}$ $\boxed{h_2 = 1}$

(iii) If there is a miss in level1 and Level 2 & hit in level3 . $\boxed{h_1 \& h_2 = 0}$ $\boxed{h_3 = 1}$

$$\boxed{T_{avg} = h_1 t_1 + (1-h_1)\, h_2 (t_2 + t_1) + (1-h_1)(1-h_2)(t_3 + t_2 + t_1)}$$

OR

$$\boxed{T_{avg} = h_1 t_1 + (1-h_1)\, h_2 (TB_1 + T_1) + (1-h_1)(1-h_2)[TB_2 + TB_1 + T_1)}$$

eq1

$$T_{avg} = \underline{h_1 t_1} + \underline{(1-h_1)}\, h_2 (t_2 + t_1) + \overset{h_3}{(}1-h_1)(1-h_2)(t_3 + t_2 + t_1)$$

Q.1

Soln 1

$h_1 = 100 \Rightarrow h_1 = 1$

$(1-h_1) = 0 \qquad 0 * x = 0$

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

Soln 2

$h_1 = 0$ & $h_2 = 100 \Rightarrow \boxed{h_2 = 1}$

$\boxed{1-h_2 = 0}$

$$\boxed{T_{avg} = T_2 + T_1}$$ Ans

If Locality of Reference Included.

$$T_{avg} = h_1 t_1 + \underline{(1-h_1)}\,[TB_1 + T_1] + \underline{(1-h_1)}(1-h_2)\overset{h_3}{[}TB_2 + TB_1 + T_1]$$

$h_1 = 100 \Rightarrow h_1 = 1$

$(1-h_1) = 0 \Rightarrow 0 * x = 0$

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

Soln 2 $h_1 = 0 \quad h_2 = 1 \quad (1-h_2) = 0$

$$\boxed{T_{avg} = TB_1 + T_1}$$ Ans.

$$\boxed{T_{avg} = h_1 t_1 + (1-h_1)\, h_2 (t_2 + t_1) + (1-h_1)(1-h_2)\, h_3 (t_3 + t_2 + t_1)}$$

Solⁿ $h_1 = 0$ $h_2 = 0$ $\boxed{h_3 = 1}$

$$\boxed{T_{avg} = T_3 + T_2 + T_1}$$

Ans

If Locality of Reference Included.

$$\boxed{T_{avg} = h_1 t_1 + (1-h_1) h_2 [TB_1 + T_1] + (1-h_1)(1-h_2)\, h_3 [TB_2 + TB_1 + T_1]}$$

Solⁿ $h_1 = 0$ $h_2 = 0$ $h_3 = 1$

$$\boxed{T_{avg} = TB_2 + TB_1 + T_1}$$

Ans

Q. In a 2 level memory, level 1 memory Access time is 30ns and level 2 memory Access time is 250ns/word. Hit ratio of level 1 is 90% . If there is a miss in level1 then 4word block must be transferred(moved) from level 2 into level1 and then addressed word is given to CPU. What is the average Access time ?

$h_1 = 90\% \Rightarrow h_1 = 0.9$

$t_1 = 30\text{nsec}$

$t_2 = 250\text{ns/word}$

$TB_1 = \underset{\text{word}}{4} \times 250\text{ ns/word}$

$$\boxed{T_{avg} = h_1 t_1 + (1-h_1)(TB_1 + T_1)}$$

$$\boxed{TB_1 = 1000\text{nsec}}$$

$\Rightarrow 0.9[30] + (1-0.9)[1000+30]$

$\Rightarrow 27 + 103$

$= 130\text{nsec}$

OR

$$\boxed{T_{avg} = T_1 + (1-h)\,TB_L}$$

$\Rightarrow 30 + (1-0.9)\,1000$

$\Rightarrow 30 + 100$

$$\boxed{T_{avg} = 130\text{nsec}}$$ 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 14nsec [GATE – 2015]

Already Done.

NAT

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 14ns

✓ (i) What is Data Transfer rate (performance) of this memory system (in words/sec)?

✓ (ii) What is Bandwidth required of this memory system if word size is 8bit?

Ans:

already Done in Today class Kindly Refer.

(i) 71.4 ⇒ 72 Million word Per Sec.

Soln (ii) 72 MBps.

Q. A cache memory that has a hit rate of 0.8 has an access latency 10 ns and miss penalty 100 ns. An optimization is done on the cache to reduce the miss rate. However, the optimization results in an increase of cache access latency to 15 ns, whereas the miss penalty is not affected. The minimum hit rate (rounded off to two decimal places) needed after the optimization such that it should not increase the average memory access time is ______.

[GATE 2022 : NAT]

$h = 0.8$, $t_c = 10$ nsec

Miss Penalty $(t_m) = 100$ nsec

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

$$\Rightarrow 0.8 \times 10 + (1-0.8)(100+10)$$

$$= 8 + 0.2(110)$$

$$= 8 + 22$$

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

OR

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

$$\Rightarrow 10 + (1-0.8) 100$$

$$= 10 + 20$$

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

Now optimization : $t_{c_{new}} = 15$ nsec.

t_m & t_{avg} Remain Same (Not Changed)

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

$$30 \Rightarrow h*15 + (1-h)(100+15)$$

$$30 = 15H + 115 - 115H$$

$$85 = 100H$$

$$H = \frac{85}{100}$$

$$\boxed{H = 0.85}$$ Ans

$$T_{avg_{new}} = t_{c_{new}} + (1-h)\, t_m$$

$$30 = 15 + (1-h)\,100$$

$$30 = 15 + 100 - 100H$$

$$85 = 100H$$

$$H = \frac{85}{100} = 0.85$$

$$\boxed{H = 0.85}$$ Ans

NAT

A direct mapped cache memory of 1 MB has a block size of 256 bytes. The cache has an access time of 3 ns and a hit rate of 94%. During a cache miss, it takes 20 ns to bring the first word of a block from the main memory, while each subsequent word takes 5 ns. The word size is 64 bits. The average memory access time in ns (round off to 1 decimal place) is ____.

[GATE-2020]

Block Size = 256 Byte.

1Word Size = 64 bit ≅ 8 Byte.

$$\text{Number of Words} = \frac{256\text{ Byte}}{8\text{ Byte}} = \boxed{32\text{ Words}}$$

⇓

1Word takes = 20nsec.

Remaining 31 Word = 5ns

Ist 1word

$$h * \underset{\text{time}}{\text{Cache}} + \underset{\text{miss}}{(1-h)} \left[\text{Cache} + \underset{20}{\text{MM}} + 31(5)\right]$$

$$T_{avg} = .94 \times 3 + (1-0.94)\left[3+20+31(5)\right]$$

$$\Rightarrow 2.82 + (0.06)(3+20+155)$$

$$\boxed{T_{avg} = 13.5\text{ nsec.}}$$

Ans

NAT

A certain processor deploys a single-level cache. The cache block size is 8 words and the word size is 4 bytes. The memory system uses a 60-MHz clock. To service a cache miss, the memory controller first takes 1 cycle to accept the starting address of the block, it then takes 3 cycles to fetch all the eight words of the block, and finally transmits the words of the requested block at the rate of 1 word per cycle. The maximum bandwidth for the memory requested block at the rate of 1 word per cycle. The maximum bandwidth for the memory system when the program running on the processor issues a series of read operations is ______$\times 10^6$ bytes/sec. [GATE-2019-CS: 2M]

THANK YOU!