COMPUTER SCIENCE

Computer Organization and Architecture

Cache Memory

Lecture_03

Vijay Agarwal sir

TOPICS TO BE COVERED

01 Memory Access

02 Cache Memory

Memory Hierarchery

Cache memory

- Simultoneous Access
- Hierarchical Access

[L.O.R]

Locality of Reference

(i) Temporaral LOR

(ii) Spatial LOR

Type of Cache

Numerical.

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.

(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.

L.O.R

Medical Store

⇓

10pc Medicine

Only for 1 time u go Medical store

9pc Medicine is Near by you at ur Home.

$$\begin{array}{r} 42 \\ 29 \\ \hline \end{array}$$

Till Now: 71 GATE QUESTION

30 → Cache

10 → Disk & DMA

111+ GATE PYQ in class.

Type of Memory Org

1. Simultaneous Access Memory Org.

Last level Hit Ratio = 1

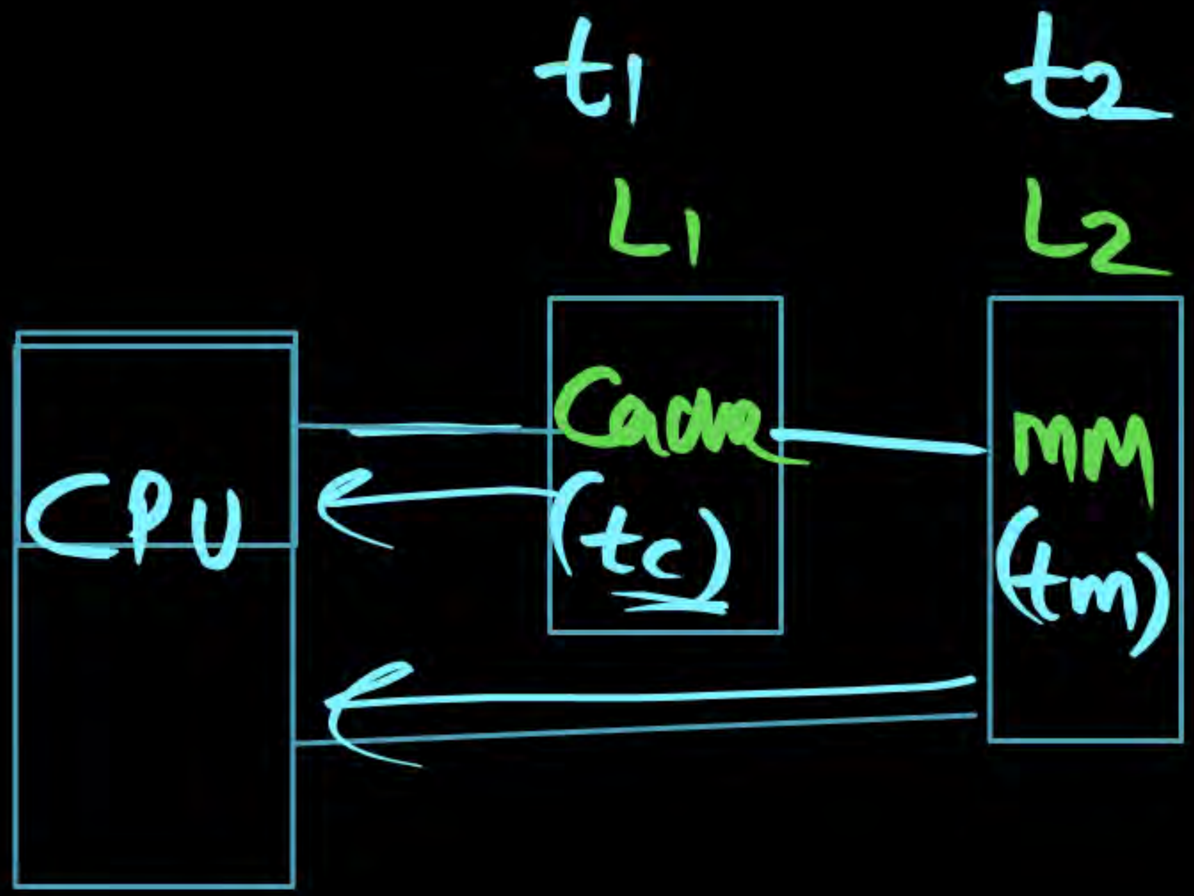
$\boxed{H_2 = 1}$

$$H_1 T_1 + (1 - H_1) \boxed{H_2} (T_2)$$

$$\boxed{T_{avg} = Ht_1 + (1-H)t_2}$$

if in term of Cache & Main Memory then Tavg

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



Type of Memory Org

1. Simultaneous Access Memory Org.

1 Word Access time = T_{avg}

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

(Data transfer Rate)

OR

Performance

Type of Memory Org

2. Hierarchical Access Memory Org.

$$\cancel{ht_c} + t_m + t_c - ht_m - \cancel{ht_c}$$

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

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

$\downarrow$ OR

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

$$\cancel{ht_c} + (1-h)t_m + t_c - \cancel{ht_c}$$

$$t_c + (1-h)t_m$$

t_c : Cache Mem Access time

t_m : Main Memory Access time

h : Cache Hit Ratio.

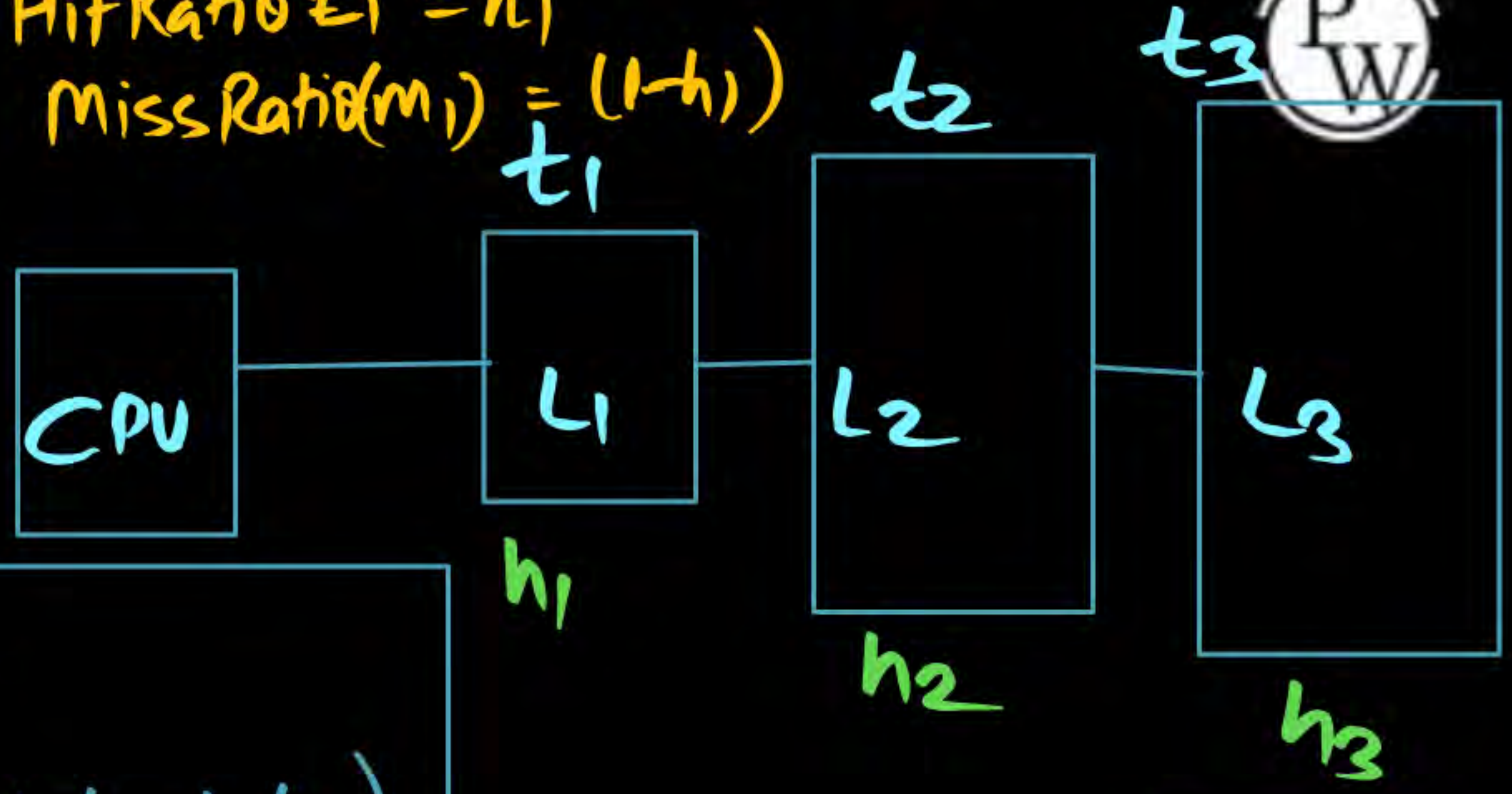
$(1-h)$: Cache Miss Ratio

Type of Memory Org

2. Hierarchical Access Memory Org.

Hit Ratio $t_1 = h_1$

Miss Ratio$(m_1) = (1-h_1)$



3 level

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

Last level hit Ratio always = 1

OR

Miss Rate

$$T_{avg} = h_1t_1 + \underline{m_1}\, h_2(t_2+t_1) + m_1\,\underline{m_2}\,(t_3+t_2+t_1)$$

m_i: miss Ratio of level i.

$$\boxed{h_3 = 1}$$

Last Level Hit Ratio = 1

m_1: Miss Ratio of level 1 $(1-h_1)$

m_2: Miss Ratio of level 2 $(1-h_2)$

Type of Memory Org

2. Hierarchical Access Memory Org.

1 Word Access time = Tavg

Data transfer Rate
(Performance)
efficiency
#Words/sec

$$= \frac{1}{T_{avg}.}$$

Remberber

5.51×10^{-9} sec

1 Instⁿ ET = 5.51 nsec

In 1 sec ⟶ # Instⁿ

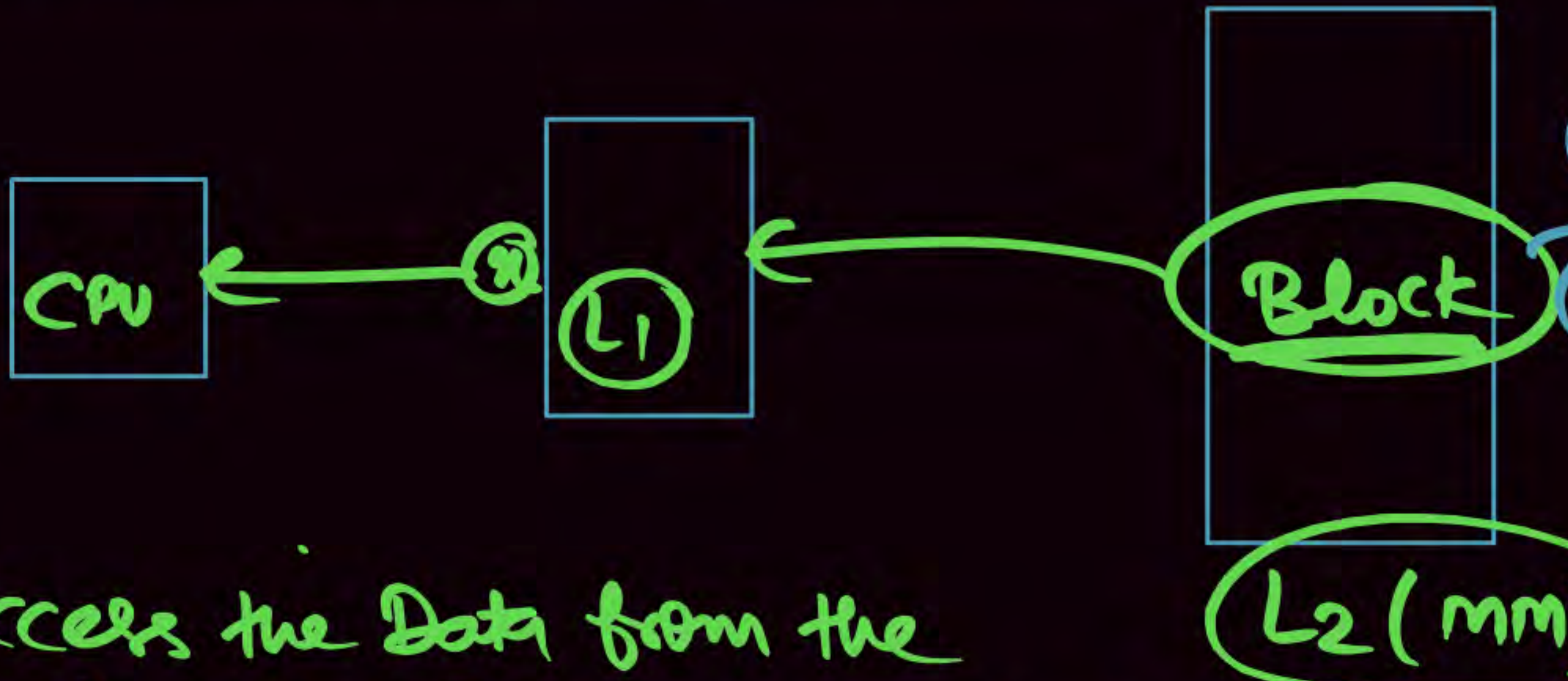
5.51×10^{-9} sec —— 1 Instⁿ

1 sec —— $\frac{1}{5.51} \times 10^{9}$ Instⁿ/sec

$$\frac{1000}{5.51} \times 10^{6}$$

= 181.4 MIPS

If Locality of Reference is Considered.

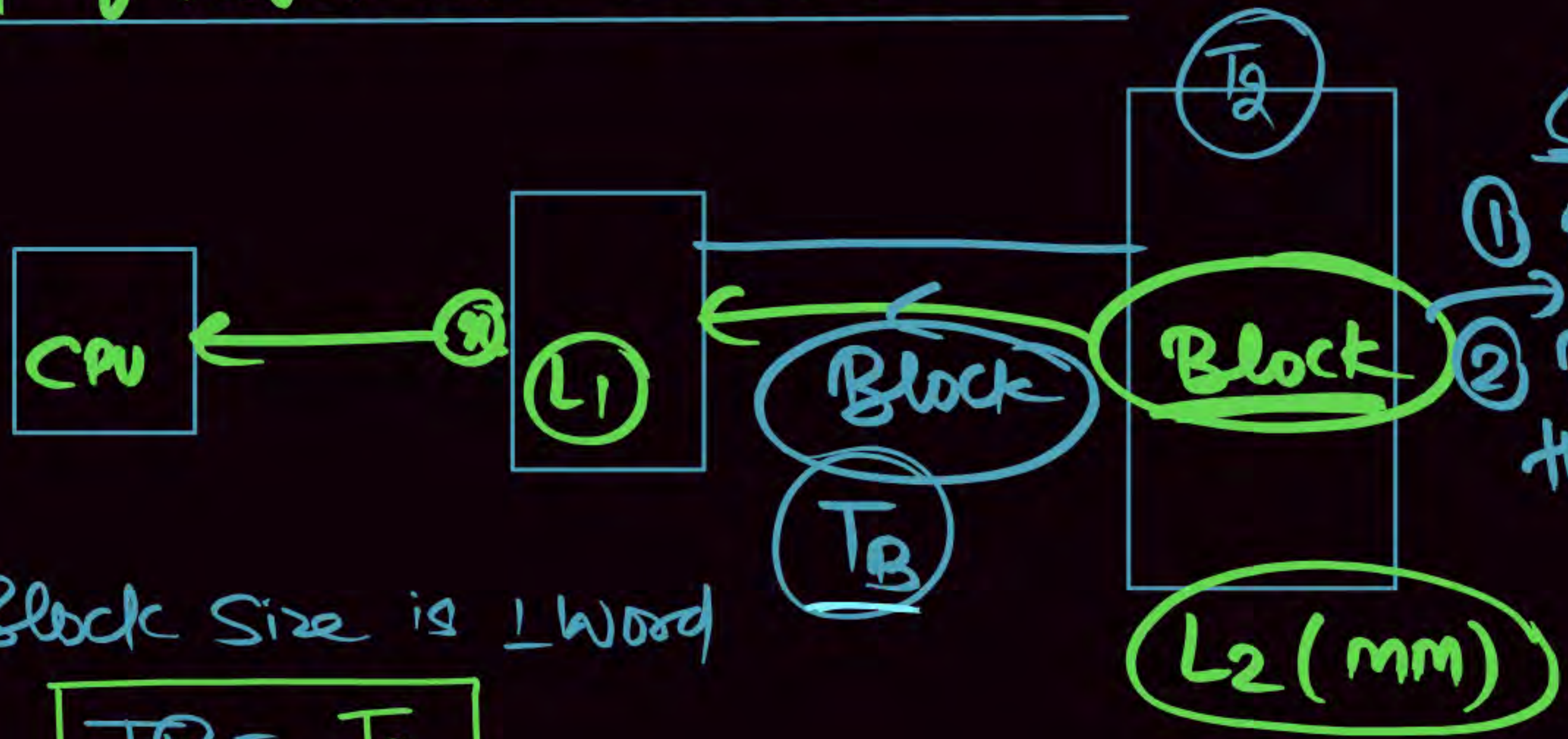


① Block Size 1 Word

② Block Size More than 1 Word (Assume n words)

CPU Always Access the Data from the Cache (Faster/Level1) memory. If there is a Miss in Level1 Memory & Hit in Level 2 (MM (or) slower compare to L1) then One Complete Block is transfer from L2 Memory to L1 Memory & addressed word (which Request/Demand) by the CPU given from Faster (Level1/Cache] Memory.

If Locality of Reference is Considered.



Case

① Block Size 1 Word

② Block size More than 1 Word (Assume n words)

Case I: If Block Size is 1 Word

$$\boxed{TB = T_2}$$

Case II: If Block Size is n Words

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

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

For 2 Level

$$T_{avg} = h_1 t_1 + (1-h_1)\overset{\overset{1}{\uparrow}}{h_2}(t_2+t_1)$$

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

OR

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

Locality of Reference

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

OR

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

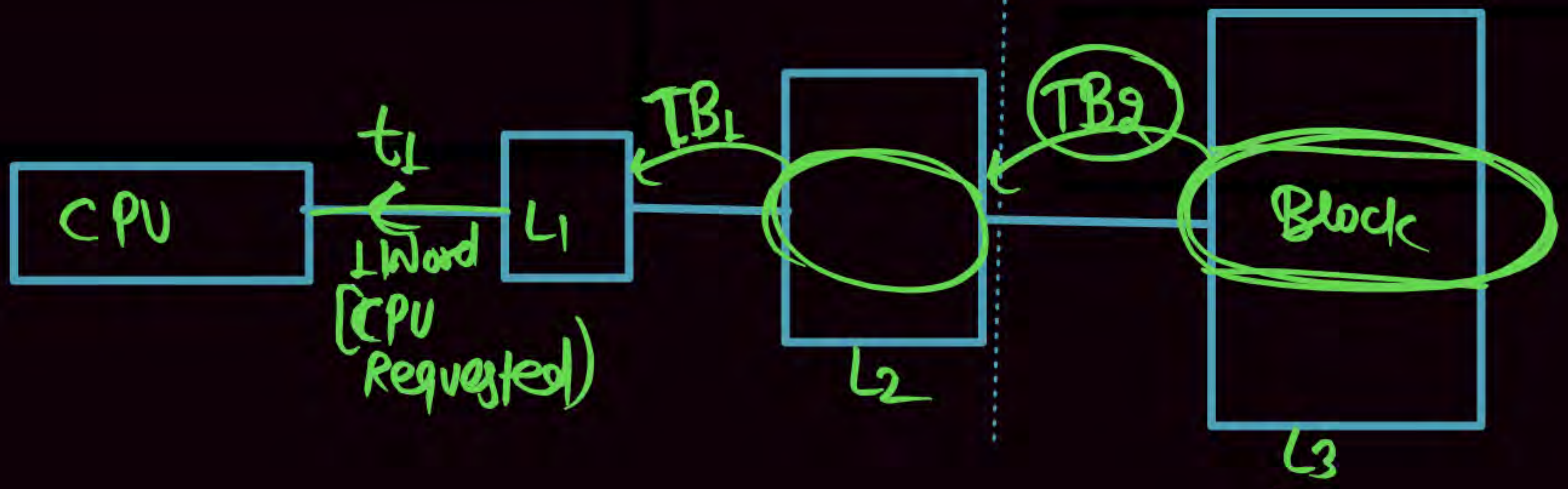
If for 3 Level

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

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

Locality of Reference.

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



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?

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

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

(iii) If there is a miss in level1 and Level 2 & hit in level3 .

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

Locality of Reference OR

$$\boxed{T_{avg} = H_1 T_1 + (1 - H_1) H_2 (TB_1 + T_1) + (1 - H_1)(1 - H_2)(TB_2 + TB_1 + T_1)}$$

$$\boxed{\text{Hit} + \text{Miss} = 1}$$

80% Hit that means 20% Miss.

$$(i)\quad H_1 = 100\% \qquad H_1 = 1 \qquad (1-H_1) = \underline{\underline{0}}$$

$$T_{avg} = \underset{0}{\underline{H_1T_1}} + (\underbrace{1-H_1}_{0})H_2(T_2+T_1) + (1-H_1)(1-H_2)$$

$$T_3+T_2+T_1$$

(i) $\boxed{H_1 = 1}$ $(1-H_1) = 0$

$$\boxed{T_{avg} = T_L}$$ Ans

(ii) $\boxed{H_2 = 1}$ $\boxed{H_1 = 0}$

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

(iii) $H_1 = 0$ $H_2 = 0$ $\underline{\underline{H_3 = 1}}$

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

OR

$$T_{avg} = \underset{\sim\sim}{H_1T_1} + (1-H_1)\underline{\underline{H_2}}(TB_1+T_1) + (1-H_1)(1-H_2)(TB_2+TB_1+T_1)$$

$$\boxed{T_{avg} = T_L}$$ Ans

$$\boxed{T_{avg} = TB_1 + T_L}$$ Ans

first Block transfered from L_2 to L_1 (TB1) then Respective (Requested) Word given L_1 to CPU (T_L)

$$\boxed{T_{avg} = TB_2 + TB_1 + T_L}$$ Ans

Q. $h_2 = 1$

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 ?

Soln

$h = 90\%$

$\boxed{h = 0.9}$

$T_1 = 30\,\text{nsec}$

$T_2 = 250\,\text{ns/word}$

Block Size = 4 word

$TB_L = 4 \times 250 = 1000\,\text{nsec}$

(Complete T_2)

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

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

$27 + 103 = 130\,\text{nsec}$ Ans

OR

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

OR

$\Rightarrow 30 + (1-0.9)\,1000 \Rightarrow 30 + 100 = 130\,\text{nsec}$ Ans

OR

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]

$$T_{avg} = 0.80 \times 5 + (1-0.80)\,50$$

$$= 4 + (0.20)\,50$$

$$= 4 + 10$$

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

NAT (1)

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?

$T_{avg} = 14nsec$

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

$\Rightarrow \frac{1}{14 \times 10^{-9}}$ words/sec

$\Rightarrow \frac{1}{14} \times 10^{9}$ words/sec

$\Rightarrow \frac{1000}{14} \times 10^{6}$ words/sec

$\Rightarrow 71.42 \times 10^{6}$

$\Rightarrow 72 \times 10^{6}$ words/sec

$\Rightarrow$ 72 Millions words/sec

OR

71.42 Millions words/sec Ans

72 Millions word / sec.

Word Size = 8 bit

(ii) Bandwidth = $72 \times 10^6 \times \underline{8\,\text{bit}}$ / sec.

= 576 Mbits/sec

⇒ 72×10^6 Byte / sec

⇒ 72 MBps or 71.43 MBps.

Q.2 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]

$h = 0.8$ $\qquad$ $(1-h) = 0.2$

$(1-0.8) \nearrow$

Cache Access time (t_c) = 10nsec

Miss Penalty (MM Access time) $\overset{(MP)}{}$ t_m = 100nsec

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

$$= 10 + (1-0.8)\,100 = 10 + 0.2(100)$$

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

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

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

$$= 8 + 0.2(110)$$

$$= 8 + 22$$

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

Optimization: ⇒ T_{avg} & Miss Penalty Not Affected (Remain same)

$h_{new} = ?$

$$\boxed{t_{c_{new}} = 15\,\text{nsec}}$$

(MP) miss Penalty Not affected

ie M.P / t_m = $\underline{100\,\text{nsec}}$.

(Not affected) $T_{avg_{new}} = \underline{30\,\text{nsec}}$

$$\boxed{T_{avg} = h_{new} \times t_{c_{new}} + (1-h)(t_m + t_{c_{new}})}$$

$$30 = h_{new} \times 15 + (1-h_{new})(100+15)$$

$$30 = 15h_{new} + 115 - 115h_{new}$$

$$85 = 100\,h_{new}$$

$$h_{new} = \frac{85}{100} = 0.85$$ Ans

$$\boxed{T_{avg} = t_{c_{new}} + (1-h_{new})(t_m)}$$

$$30 = 15 + (1-h_{new})\,100$$

$$30 = 15 + 100 - 100h_{new}$$

$$85 = 100h_{new}$$

$$h_{new} = \frac{85}{100}$$

$$\boxed{h_{new} = 0.85}$$ Ans

NAT

③

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 13.5ns ____.

[GATE-2020]

94%

$h = 0.94$

Block Size = 256 Byte

t_c (Cache Access) = 3 nsec

If Cache miss first word mm to Cache = 20 nsec

Each Subsequent word takes = 5 nsec.

1 Word Size = 64 bit ≃ 8 Byte

$$\#\text{Words} = \frac{256B}{8B} = \boxed{32 \text{ Words}}$$

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

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

$$= 2.82 + 0.06(178)$$

$$= \boxed{13.5 \text{ nsec}} \text{ Ans}$$

(Max Data transfer Rate)

Bandwidth = 10 MByte/sec

In 1 sec = 10×10^6 Byte Per Second

Q. If the cycle time of the Memory is 500 nsec.
then What is the Bandwidth ? (The Maximum rate which Memory
Can Access ——— Byte/sec ?)

Sol^n

500 nsec [cycle time] Access — 1 Byte

In 1 sec ——— $\frac{1}{500 \times 10^{-9}}$ Byte/sec

$\Rightarrow \frac{1}{500} \times 10^{9}$ Byte/sec $\Rightarrow \frac{1000}{500} \times 10^{6}$ Byte/sec

= 2 MBps Ans

or 16 Mbits/sec

NAT ④

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 ① first takes 1 cycle to accept the starting address of the block, it then ② takes 3 cycles to fetch all the eight words of the block, and ③ 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 160 $\times 10^6$ bytes/sec. [GATE-2019-CS: 2M]

Ans (160)

Cache Block Size = 8Words

1 Word Size = 4Byte

Clock Frequency = 60MHz

$$\boxed{\text{Cycle time} = \frac{1}{60\times10^6}\ \text{sec}}$$

Cache Block Size = 8 Words

⇒ 8 X 4 Byte

$$\boxed{\text{Cache Block Size} = 32\ \text{Byte}}$$

Transfer
1 Word / Cycle

So 8 Word ⇓ 8 Cycle

$$\text{Total Time Taken to Transfer Block} = \overset{\text{(Accept)}}{1\,\text{Cycle}} + \overset{\text{(Complete 8 words)}}{3\,\text{Cycle}} + \underline{8\,\text{Cycle}}$$

$$= \underline{12\,\text{Cycle}}$$

$$= 12 \times \frac{1}{60}\ \mu\text{sec} = \frac{1}{5}\times10^{-6}\ \text{sec}$$

Cycle time $= \frac{1}{60\times10^6}$ sec.

$\underline{12 \text{ Cycle}} \longrightarrow 32$ Byte

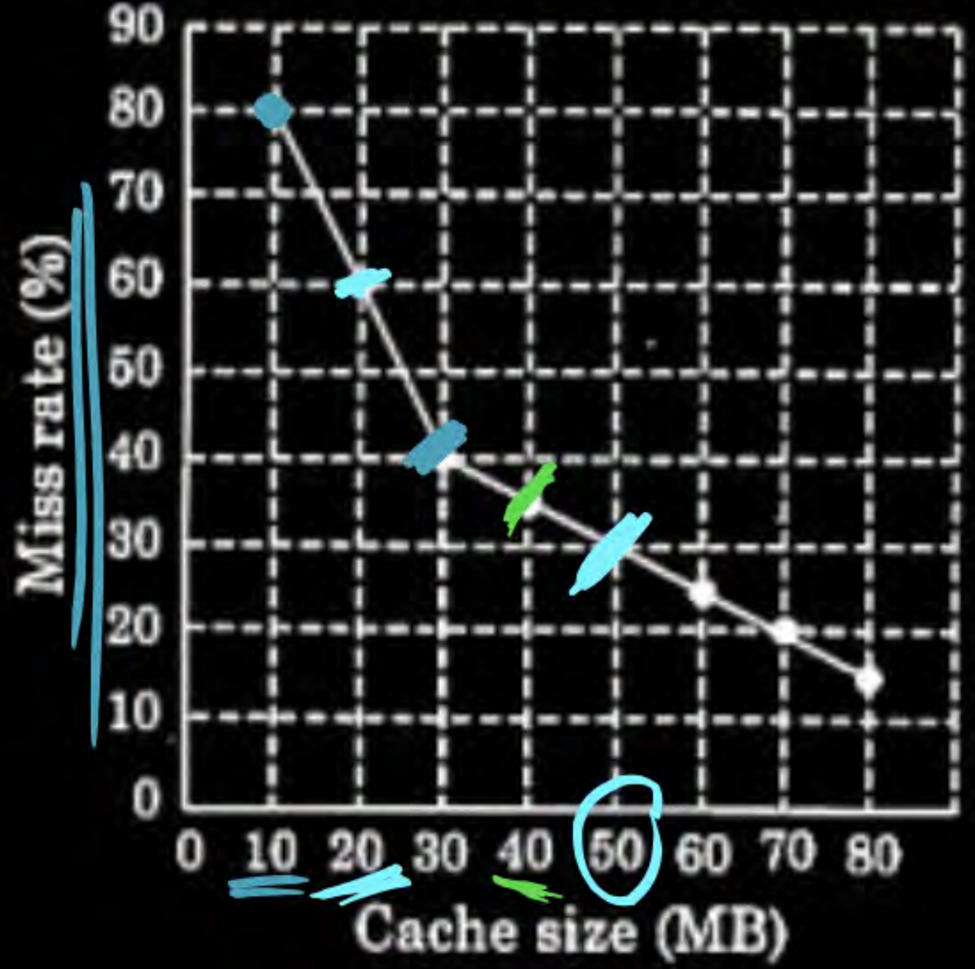
$12 \times \frac{1}{60\times10^6}$ sec —— 32 Byte

In $\underline{1 \text{ Sec}}$ —— $\dfrac{32 \text{ Byte}}{12 \times \frac{1}{60\times10^6}} \Rightarrow \dfrac{32B \times \overset{5}{\cancel{60}} \times 10^6}{\cancel{12}}$ Byte/sec

$= 160\times10^6$ Byte/sec $\underline{\text{Ans}}$

NAT 5

A file system uses an in-memory cache to cache disk blocks. The miss rate of the cache is shown in the figure. The latency to read a block from the cache is 1 ms and to read a block from the disk is 10 ms. Assume that the cost of checking whether a block exists in the cache is negligible. Available cache sizes are in multiples of 10 MB.



The smallest cache size required to ensure an average read latency of less than 6 ms is 30 MB. Ans

[GATE-2016(Set2)-CS: 2M]

Cache Access time [tc] = 1 msec.

Main Memory Access time(tm] = 10msec. Hit Rate = 0.2.

When Cache Size 10MB then Miss Rate 80% [0.8)

$T_{avg} = h*t_c + (1-h)(t_m+t_c) \Rightarrow 0.2 + 0.8(11)$

$= 0.2\times1 + 0.8(10+1). \quad = 0.2 + 8.8 = \boxed{9msec}$

Case II When Cache Size = 20MB

then Miss Rate $(1-h) = 60\%$ (0.6)

$h = 0.4$

$$T_{avg} = h * t_c + (1-h)(t_m + t_c) = 0.4 \times 1 + 0.6[10+1] \Rightarrow 0.4 + 6.6$$

$\boxed{T_{avg} = 7\,msec}$ When Cache Size is 20MB.

Case III: CacheSize = 30MB then Miss = 40% Hit = 0.6

$$T_{avg} = ht_c + (1-h)(t_m + t_c) \Rightarrow 0.6 \times 1 + 0.4(\overset{1+10}{11}) = 0.6 + 4.4 = 5\,msec$$

$\boxed{T_{avg} = 5\,msec}$ When Cache Size is 30MB.

↳ smallest

Case IV : Cache Size = 40MB then Miss Rate 35%. Hit = 0.65

$$T_{avg} = 0.65 \times 1 + 0.35(11)$$

$$\boxed{T_{avg} = 4.5\,\text{msec}}$$ when Cache is 40MB

Case V : When Cache Size = 50MB then Miss = 0.3 Hit = 0.7

$$T_{avg} = 0.7 \times 1 + 0.3(11)$$
$$= 0.7 + 3.3$$

$$\boxed{T_{avg} = 4\,\text{msec}}$$ when Cache is 50MB

THANK YOU!