CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Cache Organization

Lecture No.- 07

By- Vishvadeep Gothi sir

Recap of Previous Lecture

Topic Cache Mapping Hardware

Topic Array Access with Cache

Topics to be Covered

Topic Array Access with Cache

Topic Multilevel Cache

[NAT]

Ans = $\underline{\underline{97.6}}$

#Q. Consider a direct mapped write back data cache of size 2KB with block size of 128 bytes. The cache is considered to be empty initially. The byte addressable main memory has size 1Mbytes. Further consider that there is an array A[30][25] with each element occupies 4 bytes. The base address of array is $(1A300)_{16}$. The array is accessed 4 times. And between the accesses, there is no any data cache changes happen. Hit ratio of cache for the array access is?

array size = $30 * 25 = 750$ elements $= 750 * 4 = 3000\,B$

no. of blocks for array $= \left\lceil \frac{3000\,B}{128\,B} \right\rceil = 24$

no. of blocks in cache $= \frac{2KB}{128B} = 2^4 = 16$

→ extra blocks $= 24 - 16$
$= 8$

no. of misses $= 24 + (2 * 8) + (2 * 8) + (2 * 8)$

$= 72$

Total mem. accesses $= 750 * 4 = 3000$

no. of hits $= 3000 - 72 = 2928$

hit ratio $= \frac{2928}{3000} * 100\%$

$= 97.6\%$

Topic : Questions on Cache and Array

- Cache Size = 32 bytes
- Block size = 4 bytes

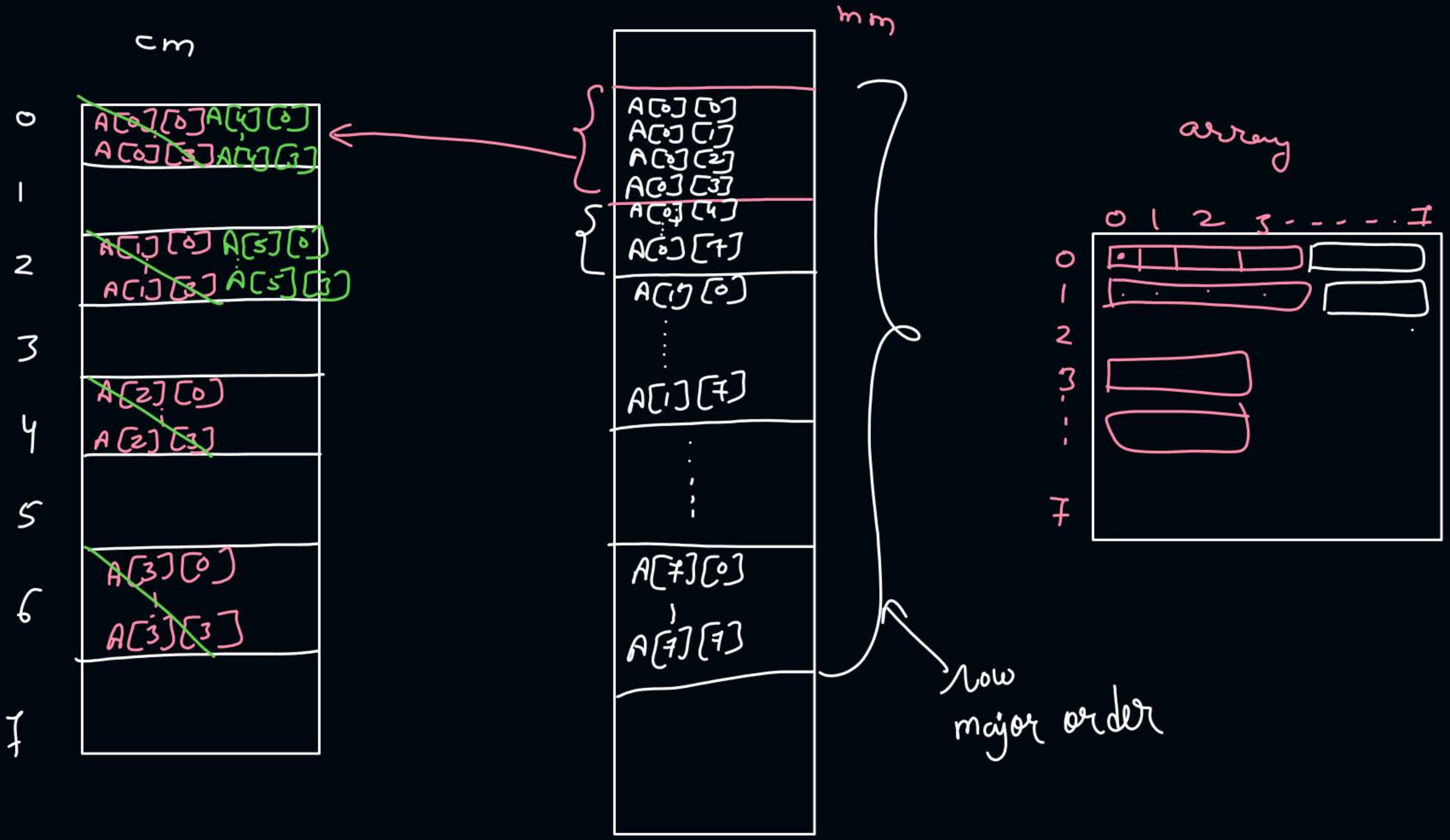
} no. of blocks in cache = $\frac{32B}{4B} = 8$

- Array in main memory int A[8][8], each element is 1 bytes

array size = $8 * 8 = 64$ elements = $64 * 1 = 64B$

no. of blocks to store array = $\frac{64B}{4B} = 16$

no. of array elements per block = $\frac{4B}{1B} = 4$


cm
0
1
2
3
4
5
6
7
A[0][0] A[4][0]
A[0][3] A[4][3]
A[1][0] A[5][0]
A[1][3] A[5][3]
A[2][0]
A[2][3]
A[3][0]
A[3][3]
mm
A[0][0]
A[0][1]
A[0][2]
A[0][3]
A[0][4]
A[0][7]
A[1][0]
A[1][7]
A[7][0]
A[7][7]
row major order
array
0 1 2 3 - - - - - 7
0
1
2
3
7

array is accessed column-wise :-

CPU access	Hit/miss
A[0][0]	miss
A[1][0]	miss
A[2][0]	miss
A[3][0]	miss
A[4][0]	miss
A[5][0]	miss
⋮	⋮

no. of miss = no. of elements in array

= 64

no. of hits = 0

[NAT]

#Q. A CPU has a 32KB direct mapped cache with 128 byte-block size. Suppose A is two-dimensional array of size 512 × 512 with elements that occupy 8-bytes each. Consider the following two C code segments, P1 and P2.

P1 :

```
for (i = 0 ; i < 512 ; i ++)
{
    for ( j = 0 j < 512 j ++)
    {
        x += A[i][j] ;
    }
}
```

P2:

```
for ( i = 0 ; i < 512 i ++)
{
    for ( j = 0 j < 512 ; j ++)
    {
        x+= A[j] [i];
    }
}
```

[NAT]

#Q. P1 and P2 are executed independently with the same initial state, namely, the array A is not in the cache and i, j, x are in registers. Let the number of cache misses experienced by P_1 be M_1 and that for P_2 be M_2.

The value of M_1 is :

array size $= 512 * 512 = 2^{18} = 2^{18} * 8 = 2^{21}$ B

no. of blocks for array $= \frac{2^{21} B}{128 B} = 2^{14}$

no. of blocks in cm $= \frac{32 KB}{128 B} = 2^{8} = 256$

M_1 = no. of blocks to store array

$= 2^{14}$

$= 16384$

[NAT]

#Q. P1 and P2 are executed independently with the same initial state, namely, the array A is not in the cache and i, j, x are in registers. Let the number of cache misses experienced by P_1 be M_1 and that for P_2 be M_2.

The value of $\underline{M_2}$ is :

no. of miss = no. of elements in array

$= 2^9 * 2^9$

$= 2^{18}$

Ques) which of the following element access will replace the block of array element A[0][0] from cache ?

(A) A[4][0]

(B) [✓ A[8][0]]

(C) A[256][0]

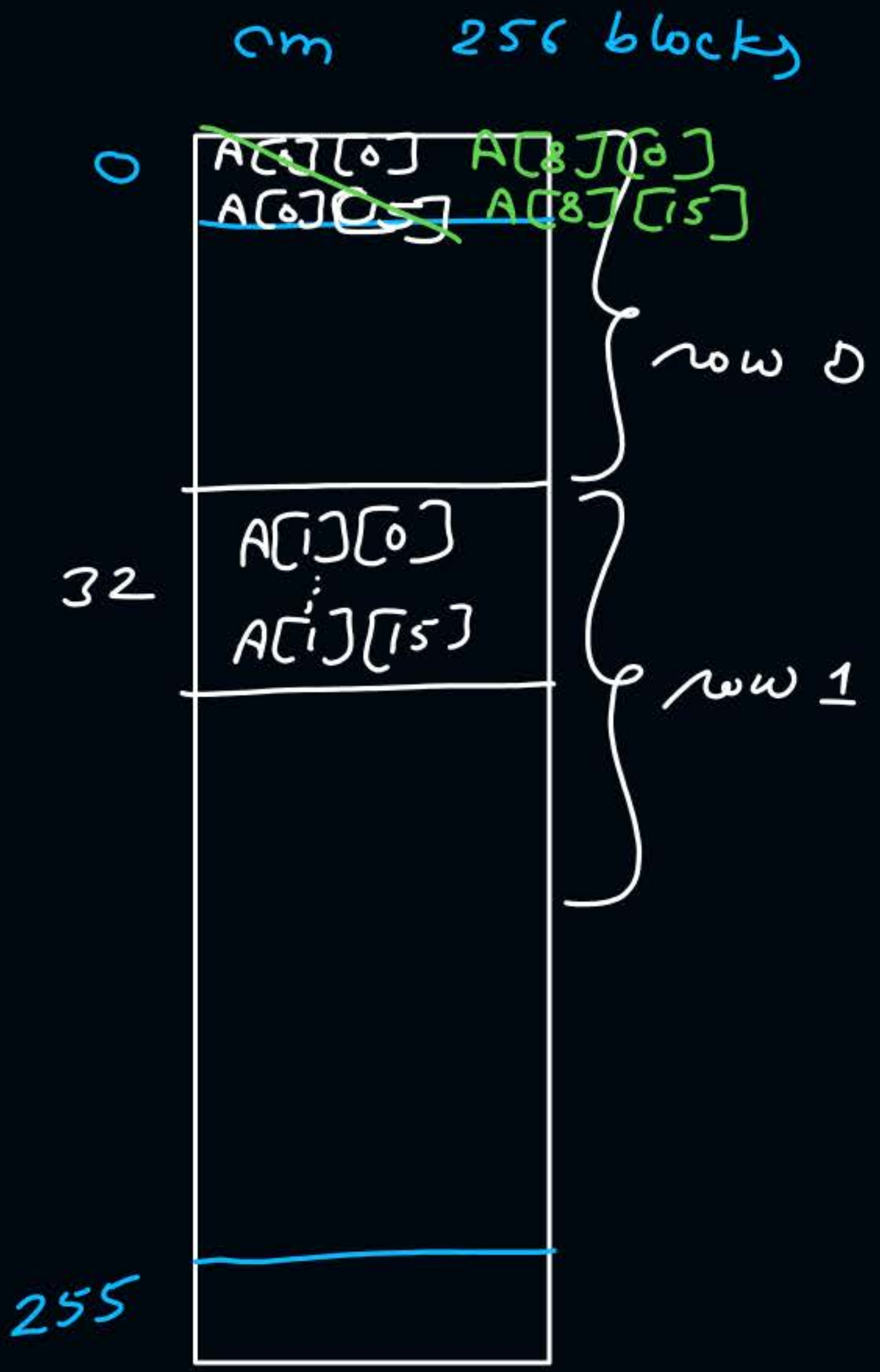
(D) A[511][0]

no. of elements per block $= \frac{128\,B}{8\,B} = 2^4 = 16$

no. of blocks for a row $= \frac{512}{16} = 2^5 = 32$

no. of rows can be stored in cache together $= \frac{256}{32}$

$= 8$


cm 256 blocks
0
~~A[0][0]~~
~~A[0][15]~~
A[8][0]
A[8][15]
row 0
32
A[1][0]
A[1][15]
row 1
255

[MCQ]

#Q. Consider two cache organizations. First one is 32 KB 2-way set associative with 32-bytes block size, the second is of same size but direct mapped. The size of an address is 32 bits in both cases. A 2-to-1 multiplexer has latency of 0.6 ns while a k-bit comparator has latency of $\frac{k}{10}$ns. The hit latency of the set associative organization is h_1 while that of direct mapped is h_2.

The value of h_1 is:

A [✓ 2.4 ns]

B 2.3 ns

C 1.8 ns

D 1.7 ns

2-way

$\xleftarrow{\quad} 32 \xrightarrow{\quad}$

Tag	set	byte
18	9	5

$$\frac{18}{10} + 0.6 = 2.4 ns$$

Direct

$\xleftarrow{\quad} 32 \xrightarrow{\quad}$

Tag	block	byte
17	10	5

[MCQ]

#Q. Consider two cache organizations. First one is 32 KB 2-way set associative with 32-bytes block size, the second is of same size but direct mapped. The size of an address is 32 bits in both cases. A 2-to-1 multiplexer has latency of 0.6 ns while a k-bit comparator has latency of $\frac{k}{10}$ns. The hit latency of the set associative organization is h_1 while that of direct mapped is h_2.

The value of h_2 is:

$$\frac{17}{10} = 1.7 \text{ ns}$$

A 2.4 ns

B 2.3 ns

C 1.8 ns

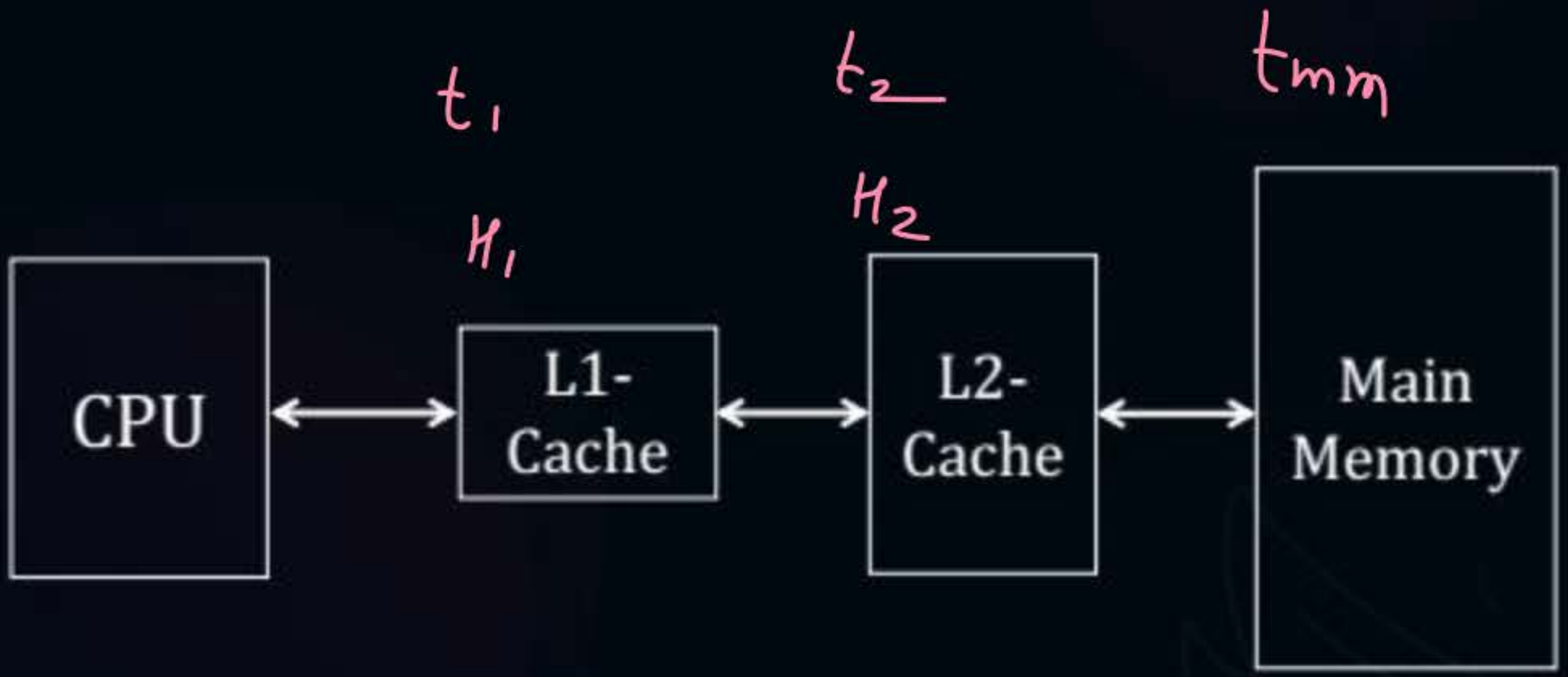
D [✓ 1.7 ns]

Topic : Goal of Using Cache

1. Minimize Access Time → Use small cache
2. Maximize Hit Rate → Use larger cache
3. Minimize Miss Penalty

Use small cache } multilevel cache
Use larger cache }

Topic : Multilevel Cache



Topic : Average Access Time Multilevel Cache

Simultaneous :-

$$t_{avg} = H_1 t_1 + (1-H_1)\left[H_2\, t_2 + (1-H_2)\, t_{mm}\right]$$

or

$$= \underline{H_1} t_1 + \underline{(1-H_1)H_2}\, t_2 + \underline{(1-H_1)(1-H_2)}\, t_{mm}$$

Hierarchical :- (default)

$$t_{avg} = H_1 t_1 + (1-H_1)\left[H_2\,(t_1+t_2) + (1-H_2)\,(t_1+t_2+t_{mm})\right]$$

or

$$= \underline{H_1} t_1 + \underline{(1-H_1)H_2}\,(t_1+t_2) + \underline{(1-H_1)(1-H_2)}\,(t_1+t_2+t_{mm})$$

or

$$= t_1 + (1-H_1)\left[t_2 + (1-H_2)\, t_{mm}\right]$$

[NAT]

#Q. Consider a 3-level memory hierarchy with L1 cache, L2 cache and a main memory. The hit ratios of L1 is 90% and of L2 is 95%. The access times of L1, L2 and main memory are 15$\overset{ns}{\text{s}}$, 60ns and 350ns respectively. The average memory access time is ________ns?

$$t_{avg} = 15 + 0.1\left[60 + 0.05 * 350\right]$$

$$= 22.75 \text{ ns}$$

In prev Questn if only L1 and mm used

$t_{avg} = 15 + 0.1 * 350$

$= 50 ns$

In prev. Questn if only L2 and mm used.

$t_{avg} = 60 + 0.05 * 350$

$= 77.5 \ ns$

Topic : Probability of Access

fraction of time (Probability of access)

CPU gets content on L1 ⇒ H_1

—— " ———— L2 ⇒ $(1-H_1) * H_2$

—— " ———— mm ⇒ $(1-H_1) * (1-H_2)$

[NAT]

#Q. Consider a 3-level memory hierarchy with L1 cache, L2 cache and a main memory. The probability of access of L1 is 95%, of L2 is 4.5% and of main memory is 0.5%. The access times of L1, L2 and main memory are 10ns, 50ns and 400ns respectively. The average memory access time is ________ns?

$$= (0.95 * 10) + 0.045 * (10+50) + 0.005 * (10+50+400)$$

$$= 14.5 \text{ ns}$$

[NAT]

H.W.

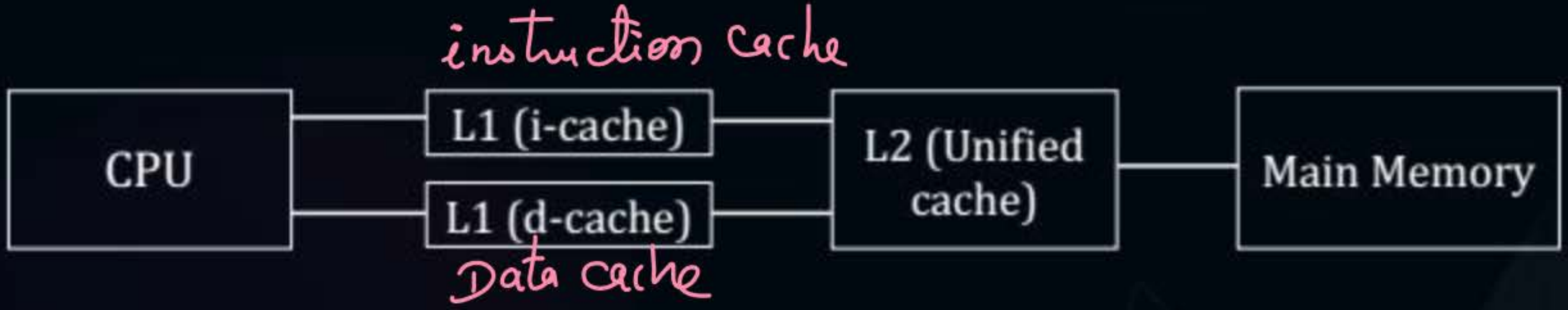
#Q. In the three-level memory hierarchy shown in the following table, p_i denotes the probability that an access request will refer to M_i.

Hierarchy Level (M_i)	Access Time (t_i)	Probability of Access (p_i)	Page Transfer Time (T_i)
M_1	10^{-6}	0.99000	0.001 sec
M_2	10^{-5}	0.00998	0.1 sec
M_3	10^{-4}	0.00002	---

If a miss occurs at level M_i, a page transfer occurs from M_{i+1} to M_i and the average time required for such a page swap is T_i.

Calculate the average time t_A required for a processor to read one word from this memory system.

Topic : Dual Cache



Hier :-

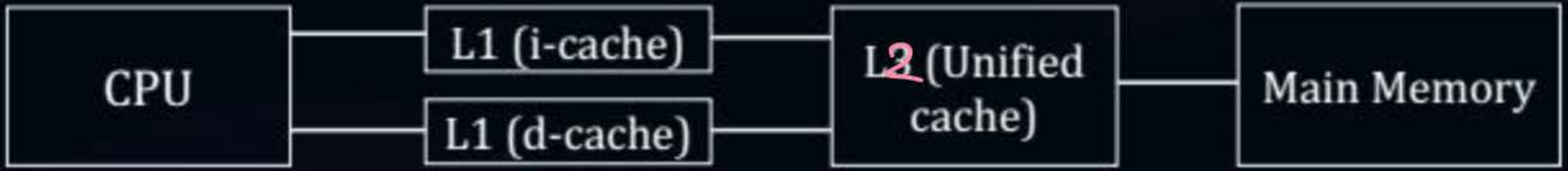
$$t_{avg\ inst^n} = H_{1i} * t_{1i} + (1 - H_{1i}) \left[H_2 (t_{1i} + t_2) + (1 - H_2)(t_{1i} + t_2 + t_{mm}) \right]$$

$$t_{avg\ data} = H_{1d} * t_{1d} + (1 - H_{1d}) \left[H_2 (t_{1d} + t_2) + (1 - H_2)(t_{1d} + t_2 + t_{mm}) \right]$$

$$t_{avg} = \%\ of\ inst^n\ access * t_{avg\ inst^n} + \%\ of\ data\ access * t_{avg\ data}$$

[NAT]

#Q. The multilevel memory hierarchy is given.



The hit ratio of $L1_i$, L~~2~~1_d, L3 and main memory are 0.8, 0.9, 0.95 and 1.0 respectively. The access times of respective memories are 10ns, 10ns, 50ns and 500ns. Among total memory references 60% of them are for data.

1. Average memory access time for only instructions access 25 ns
2. Average memory access time for only data access 17.5 ns
3. Average memory access time 20.5 ns

$$t_{avg}\ inst^n = 10 + 0.2\left[50 + 0.05 * 500\right] = 25\ ns$$

$$t_{avg}\ data = 10 + 0.1\left[50 + 0.05 * 500\right] = 17.5\ ns$$

$$t_{avg} = 0.4 * 25 + 0.6 * 17.5$$

$$= 20.5\ ns$$

[NAT] GATE- 2017

H.W.

#Q. The read access times and the hit ratios for different caches in a memory hierarchy are as given below:

Cache	Read access time (in nanoseconds)	Hit Ratio
I-cache	2	0.8
D=cache	2	0.9
L2-cache	8	0.9

The read access time of main memory in 90nanoseconds. Assume that the caches use the referred-word-first read policy and the write-back policy. Assume that all the caches are direct mapped caches. Assume that the dirty bit is always 0 for all the blocks in the caches. In execution of a program, 60% of memory reads are for instruction fetch and 40% are for memory operand fetch. The average read access time in nanoseconds (up to 2 decimal places) is ________?

[NAT]

#Q. Consider a program execution which has 36% instructions for load and store. The CPI without memory stalls is 2. The program experiences 2% miss for instruction cache and 4% miss for data cache. The cache miss penalty is 200 cycles.

1. CPI with memory stalls ⇒ 8.88

2. Performance gain (speed up) of perfect cache as compared to cache with stalls ⇒ 4.44

stalls for inst^n access $= 1 * 0.02 * 200 = 4$

stalls for data access $= 0.36 * 0.04 * 200 = 2.88$

$CPI = 2 + 4 + 2.88 = 8.88$

Speed up $= \frac{8.88}{2}$

$= 4.44$

Topic : Cache Inclusion Policies

mem blocks which are in L1 are also in L2 or not ?

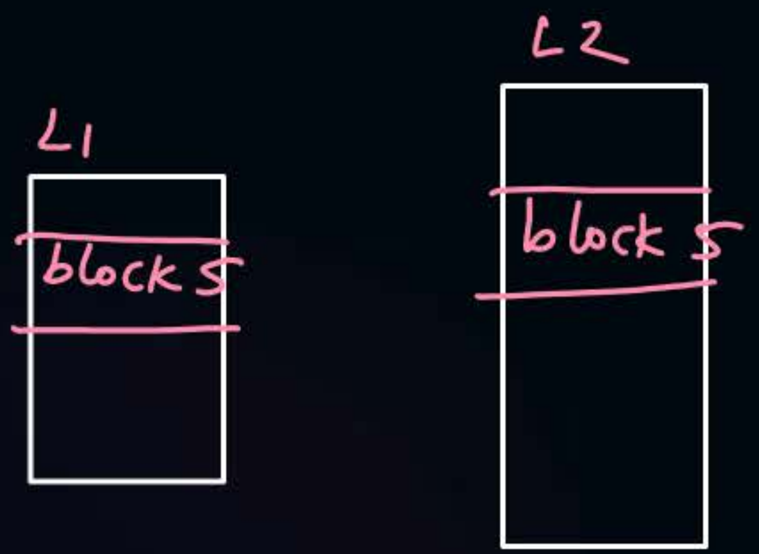
- → Inclusive
- → Exclusive

Topic : Inclusion Policy

all m.m. blocks which are in L1, must be present in L2.

⇓

L2 is inclusive of L1.



(content of L1) ∩ (content of L2) ⇒ (content of L1)

Value Inclusion policy:- values in blocks in L1 will always be consistent with blocks in L2.

Topic : Inclusion Policy

CPU wants to read a value x from mem.

1. Hit in L1 CPU reads x from L1.

2. Miss in L1 & Hit in L2 CPU reads x from L2.
   CPU copies missed block from L2 to L1.
   If any block is evicted (replaced) from L1 then there is no any role of L2.

3. Miss in L1 & Miss in L2 CPU reads x from mm
   CPU copies missed block from mm to L2, then from L2 to L1.
   If any block is replaced from L2 then L2 sends a back validation signal to L1 to make this block invalid in L1.

Topic : Exclusion Policy

any mm blocks which is present in L1 should not be present in L2.

$$(\text{content of } L1) \cap (\text{content of } L2) \Rightarrow \phi$$

L2 is called as victim cache:-

because L2 contains only victim (replaced) blocks of L1.

Topic : Exclusion Policy

1. Hit in L1 CPU reads x from L1

2. Miss in L1 & Hit in L2 CPU reads x from L2

   CPU moves (remove and copy) the missed block from L2 to L1.

   If any block is evicted from L1 then it is moved to L2.

3. Miss in L1 & Miss in L2 CPU reads x from mm

   CPU copies the missed block from mm to L1.

   If any block is evicted from L1 then it is moved to L2.

[MCQ GATE-2008]

#Q. For inclusion to hold between two cache levels L1 and L2 in a multi-level cache hierarchy, which of the following are necessary?

I. L1 must be write-through cache

II. L2 must be a write-through cache

III. The associativity of L2 must be greater than that of L1

IV. The L2 cache must be at least as large as the L1 cache

(A) IV only

(B) I and IV only

(C) I, II and IV only

(D) I, II, III and IV

[MCQ GATE-2021]

#Q. Assume a two-level inclusive cache hierarchy, L1 and L2, where L2 is the larger of the two. Consider the following statements.

S1: Read misses in a write through L1 cache do not result in writebacks of dirty lines to the L2

S2: Write allocate policy must be used in conjunction with write through caches and no-write allocate policy is used with writeback caches.

Which of the following statements is correct?

(A) S1 is true and S2 is false

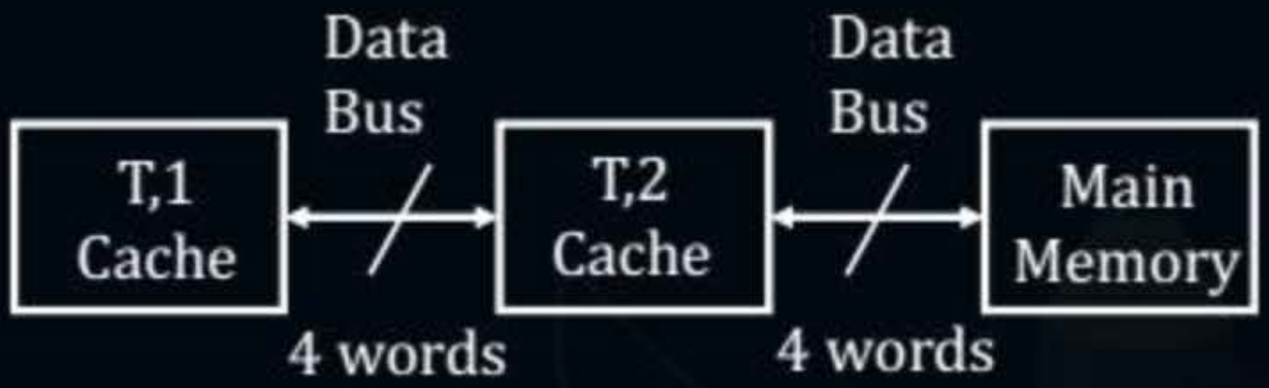
(B) S1 is false and S2 is true

(C) S1 is true and S2 is true

(D) S1 is false and S2 is false

[MCQ]

#Q. A computer system has an L1 cache, an L2 cache, and a main memory unit connected as shown below. The block size in L1 cache is 4 words. The block size in L2 cache is 16 words. The memory access times are 2 nanoseconds, 20 nanoseconds and 200 nanoseconds for L1 cache, L2 cache and the main memory unit respectively.



When there is a miss in L1 cache and a hit in L2 cache, a block is transferred from L2 cache to L1 cache. What is the time taken for this transfer?

A 2 nanoseconds

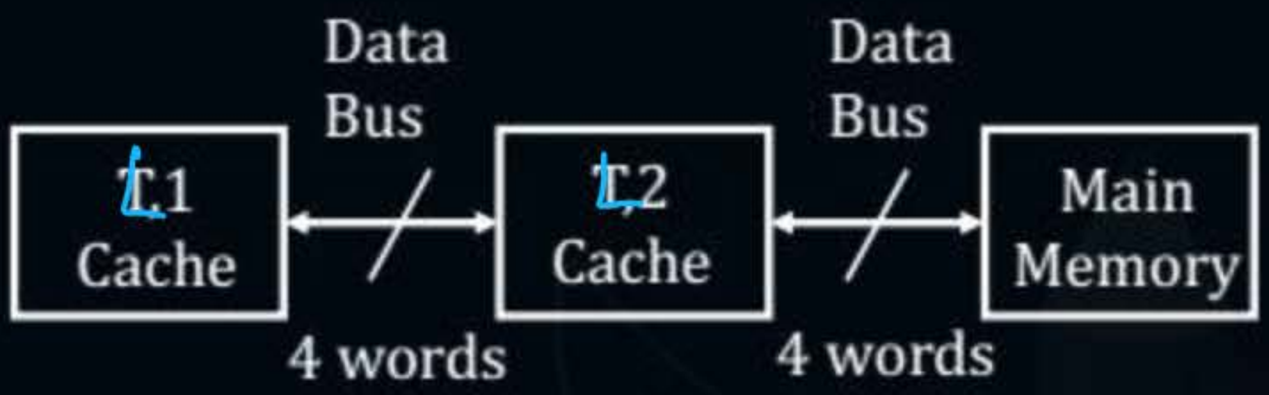
B 20 nanoseconds

C 22 nanoseconds

D 88 nanoseconds

[MCQ]

#Q. A computer system has an L1 cache, an L2 cache, and a main memory unit connected as shown below. The block size in L1 cache is 4 words. The block size in L2 cache is 16 words. The memory access times are 2 nanoseconds, 20 nanoseconds and 200 nanoseconds for L1 cache, L2 cache and the main memory unit respectively.



When there is a miss in both L1 cache and L2 cache, first a block is transferred from main memory to L2 cache, and then a block is transferred from L2 cache to L1 cache. What is the total time taken for these transfers?

A 222 nanoseconds

B 888 nanoseconds

C 902 nanoseconds

D 968 nanoseconds

2 mins Summary

Topic Array Access with Cache

Topic Multilevel Cache

Happy Learning

THANK - YOU