COMPUTER SCIENCE

Computer Organization and Architecture

Cache Memory

Lecture_07

Vijay Agarwal sir

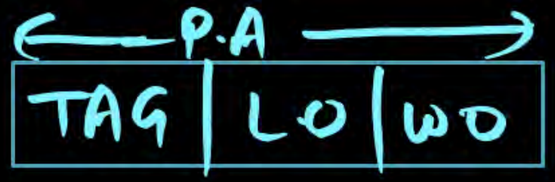
01 Replacement Algo & Updating Technique

02 Multi level Cache

Cache Memory

1. Memory org.
2. Mapping Technique
   1. Direct Mapping.
   2. Set Associative Mapping.
   3. Associative Mapping.
3. Replacement Algo.

① Direct Mapping.

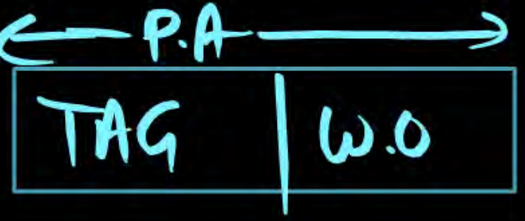


K MOD N = i

② Set Associative Mapping

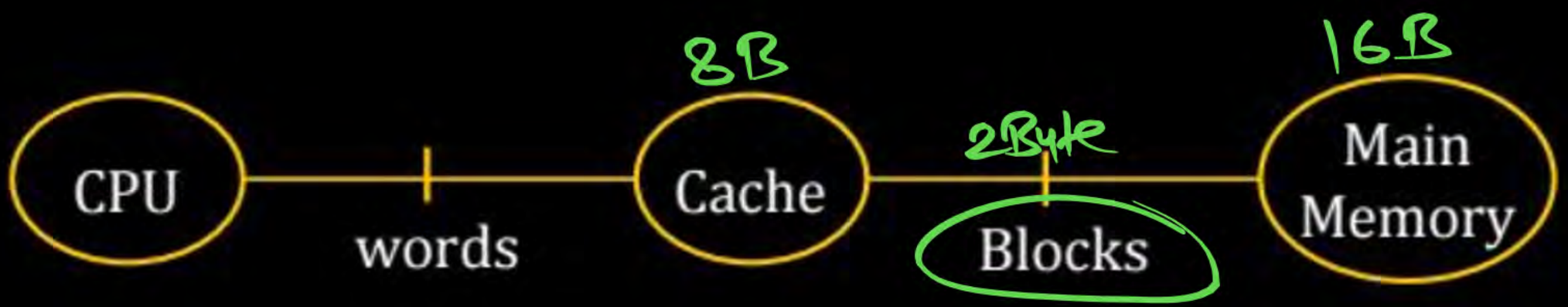
K MOD S = i

③ Associative | Fully Associative Mapping.



No Mapping function.

Memory Organization

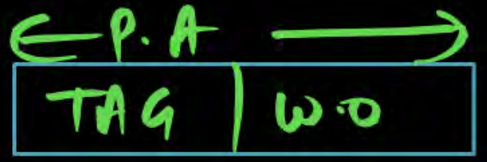


Mapping

The process of transfer the Data from Main Memory to Cache Memory is called mapping. There are 3 Type of Mapping Technique (In the Mapping One Complete Block transferred (Copied) from MM to CM.

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

Associative Mapping :



'No Mapping function'

Q.

Consider fully associative cache consists 8 Block & MM contain 128 Block & Request made by the CPU:

119, 84, 37, 0, 16, 0, 84, 120, 121, 93, 37, 0, 43, 39, 47, 48.

Calculate # of compulsory & Capacity miss?

0	119 43
1	84 39
2	37 47
3	0 48
4	16
5	120
6	121
7	93

Hit: 0, 84, 37, 0

#Hits: 4

Total Miss = 12

Replacement Ago.

1. Random
2. FIFO
3. LRU

   Direct Mapped Cache
4. 2 way set Associative LRU.
   N-way

Replacement Algorithm

When Cache is Full, then replacement algorithm are required to replace the exist cache block with new block.

In the CM design 3 type of replacement algorithm is used.

1) Random Algorithm
2) FIFO Replacement
3) LRU Replacement

In the random algorithm, any cache block can be replaced based on the random selection.

3 Type of Miss

1. Compulsory | First Reference | Cold start Miss.
2. Capacity miss
3. Conflict | Collision Miss.

Types of Misses

In the CM design 3 types of misses are present.

1) Compulsory miss - (Cold start miss / first reference miss)

   This miss will occur when the very first reference to the cache itself a miss. (when very first time Any MM Block brought from MM to CM)

2) Capacity Miss - This miss will occur when cache is full.

3) Conflict Miss (Collison miss / reference miss)

   This miss will occur when the too many blocks are placed into same cache line or same cache SET.

Q.1 Consider 4 block cache memory (initially empty) with the following MM block references.

7, 8, 10, 15, 7, 8, 16, 7, 8, 10

Identify the Hit Ratio using

(i) FIFO **(ii) LRU**

(iii) Direct Mapped cache **(iv) 2 - way Set Associative with LRU**

MM Request = 10

Cache line = 4.

7 → Compulsory Miss
8 → Compulsory Miss
10 → Compulsory Miss
15 → Compulsory Miss
7 → Conflict
8 → Conflict
16 → Compulsory Miss.
7
8
10] Conflict / capacity

Q.1 Consider 4 block cache memory (initially empty) with the following MM block references.

7, 8, 10, 15, 7, 8, 16, 7, 8, 10

Identify the Hit Ratio using

(i) FIFO

Q.1 Consider 4 block cache memory (initially empty) with the following MM block references.

7, 8, 10, 15, 7, 8, 16, 7, 8, 10

Identify the Hit Ratio using

(ii) LRU

Q.1 Consider 4 block cache memory (initially empty) with the following MM block references.

7, 8, 10, 15, 7, 8, 16, 7, 8, 10

Identify the Hit Ratio using

(iii) Direct Mapped cache

$\Rightarrow 3/10$

Hit Ratio = 0.3

Q.1 Consider 4 block cache memory (initially empty) with the following MM block references.

7, 8, 10, 15, 7, 8, 16, 7, 8, 10

Identify the Hit Ratio using

(iv) 2 - way Set Associative with LRU

#Hit = 4 Total MM Request = 10

Hit Ratio = $\frac{4}{10}$ = 0.4

LRU

Least Recently Used.

Q. Consider a small two-way set-associative cache memory, consisting of 4 blocks. For choosing the block to be replaced, use the least recently used (LRU) scheme. The number of cache misses for the following sequence of block addresses is 8, 12, 0, 12, 8

(a) 2 (b) 3 (c) 4 (d) 5

[GATE - 2004]

Soln

#CM lines = 4 2 way Set Associative

#SET = $4/2 = 2$ (SET No) S=2

k MOD S = i

k MOD 2 = i

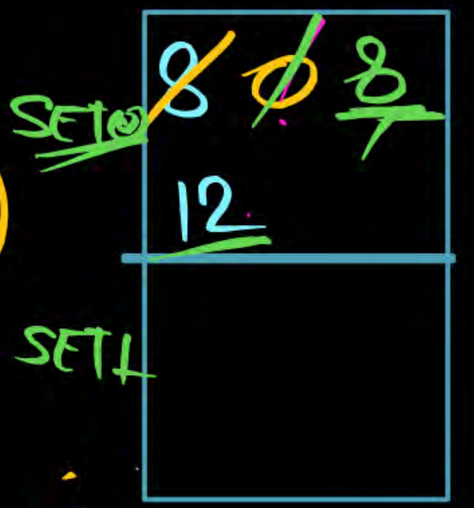
8 MOD 2 = 0 → M

12 MOD 2 = 0 → M

0 % 2 = 0 → M [Cache Set is full then Apply LRU Repl. Algo.]

12 % 2 = 0 ⇒ HIT

8 % 2 = 0 Miss [Apply LRU]



Ans [C].

Types of Misses

In the CM design 3 types of misses are present.

1) Compulsory miss - (Cold start miss / first reference miss)

   This miss will occur when the very first reference to the cache itself a miss.

2) Capacity Miss - This miss will occur when cache is full.

3) Conflict Miss (Collison miss / reference miss)

   This miss will occur when the too many blocks are placed into same cache line or same cache SET.

How to Reduce Three Miss:

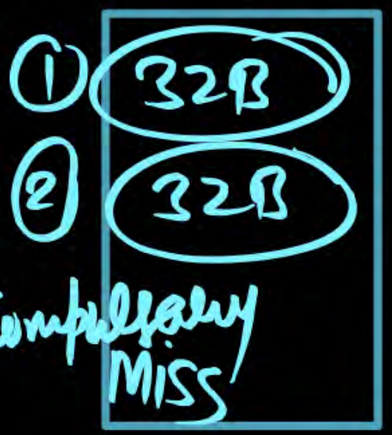
① Compulsory Miss : We can Reduce the Compulsory Miss by Increasing the BlockSize. [Larger Block size]

Parking wallah example Airport

③ If B.S = 16Byte

4 Compulsory Miss.

② If Block size = 32B



2 Compulsory Miss

1 Miss

Cache

① ← BlockSize = 64B

② Capacity Miss: Capacity Miss occur when Cache is Full.
If we Increase the Cache Size then Capacity Miss can be Reduced. (Increasing the Cache Size)

③ Conflict | Collision Miss: Conflict Miss can be Reduced by Increasing the Associativity.

Eg If 2 Way
→ 4 Way Set Associative
→ 8 Way Set Associative etc

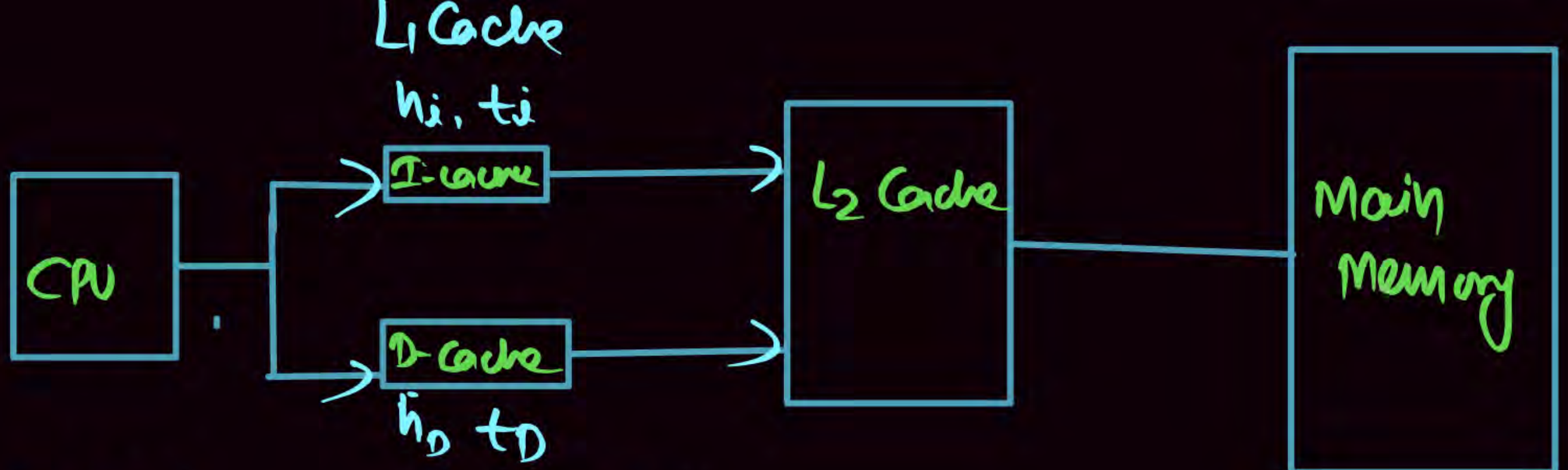
Multi-level Cache Concept

Multilevel-Cache

2 Level Cache
- L1 Cache
- L2 Cache

L1 Cache
- I-Cache
- D-Cache

Multilevel-Cache



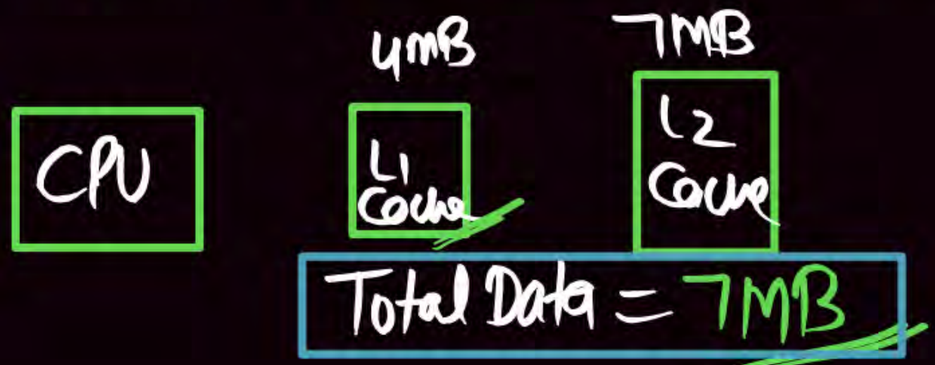
I-Cache : Instruction Cache
D-Cache : Data Cache.

Multilevel-Cache

Working Principal of Multilevel Cache

1. Principal of Inclusion
2. Principle of Exclusion.

1. Principal of Inclusion :



Data Present in L1 Cache, Must be Present in L2 Cache

OR

L1 Cache Content is Subset of L2 Cache Content

OR

L1 Contain Copy of L2 Data.

Multilevel-Cache

② Principal of Exclusion : Data Present in L1 Cache Must be Differ from L2 Cache.

Total Data = 4+7

Total Data = 11MB.

WHY Multilevel Cache Used

Sol^n

Level 1 Cache Access time = 10nsec.

Main Memory Access time = 300ns

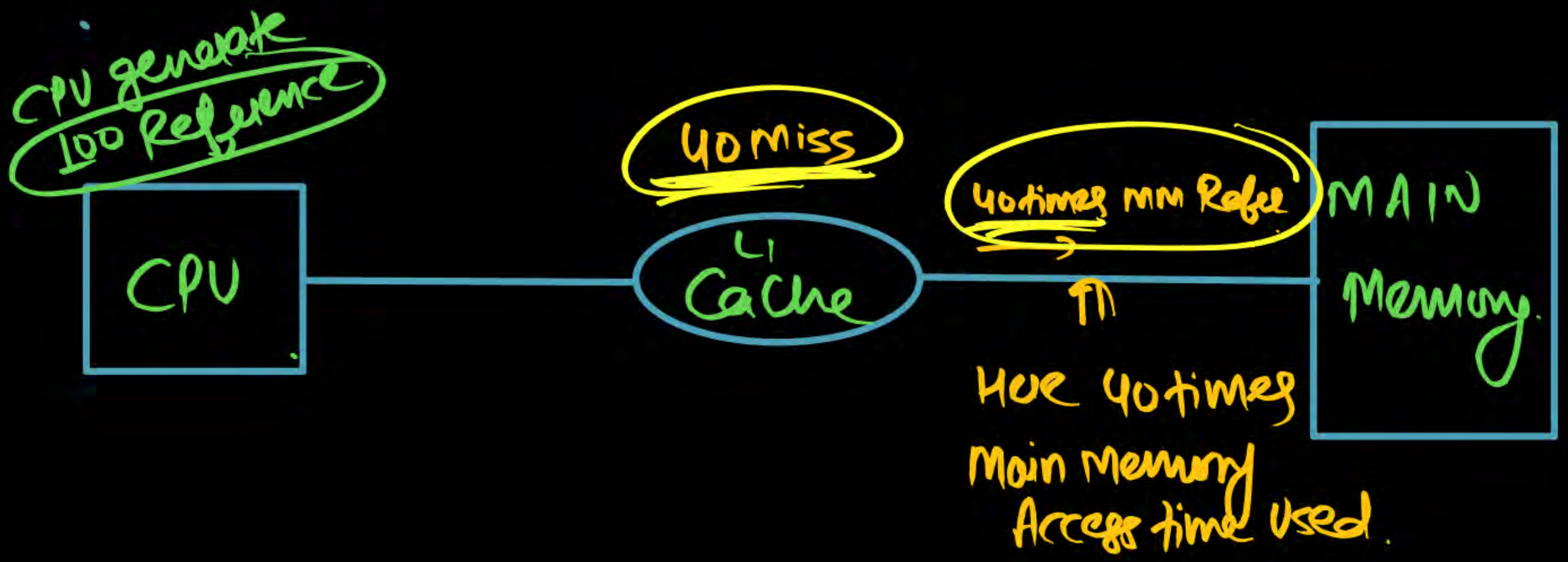
L1 Cache Access time = 10nsec

L2 Cache Access time = 15nsec

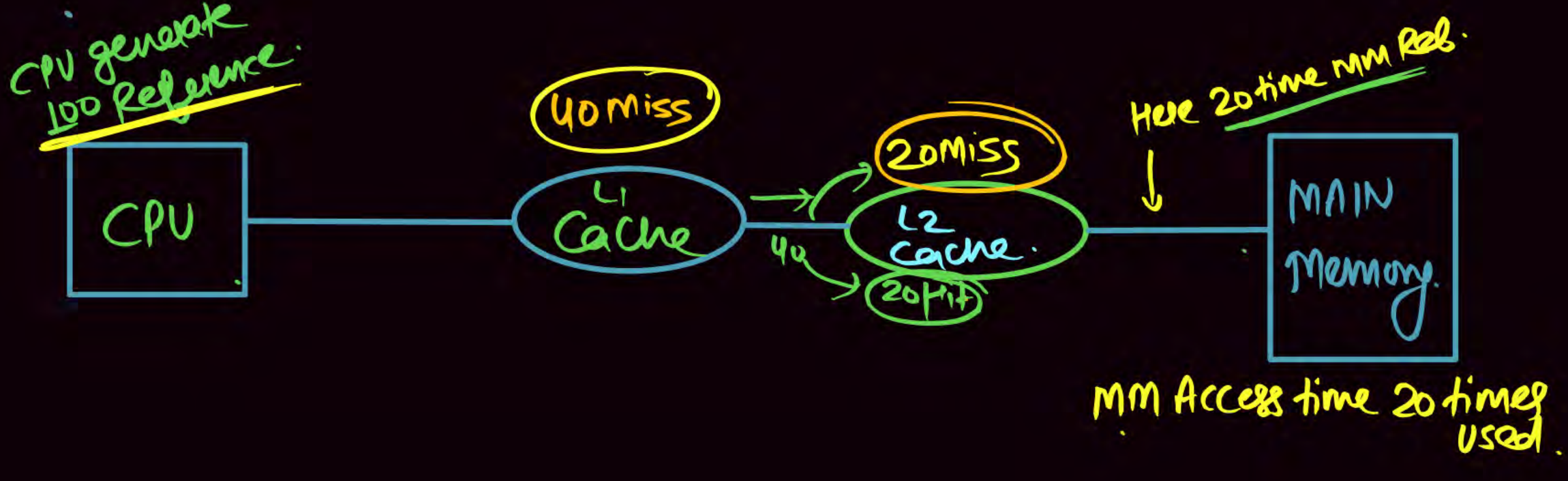
MM Access time = 300ns.

So Tavg will Reduce.

WHY Multilevel Cache Used



WHY Multilevel Cache Used



If Only L1 Cache then if there is Miss in Level 1 Cache then Directly Refer to Main Memory

(So Here Higher (More Number of) time MM Access So

Tavg Increase

But in 2 Level Cache. if Miss in Level 1 Cache then that Miss forward to level 2 Cache (so some Hit in level 2) then After Level 2 there is a Miss then we go Main Memory.

So Tavg will be Reduce (Bcz L1 & L2 Cache Access time very Low)

Multi level cache

- To reduce the miss penalty multi-level caches are used in the system design.
- The number of cycles required to transfer the data from higher levels to L_1 due to miss operation is called as miss penalty

2 Type of Miss Rate

① Local Miss Rate [LMR]

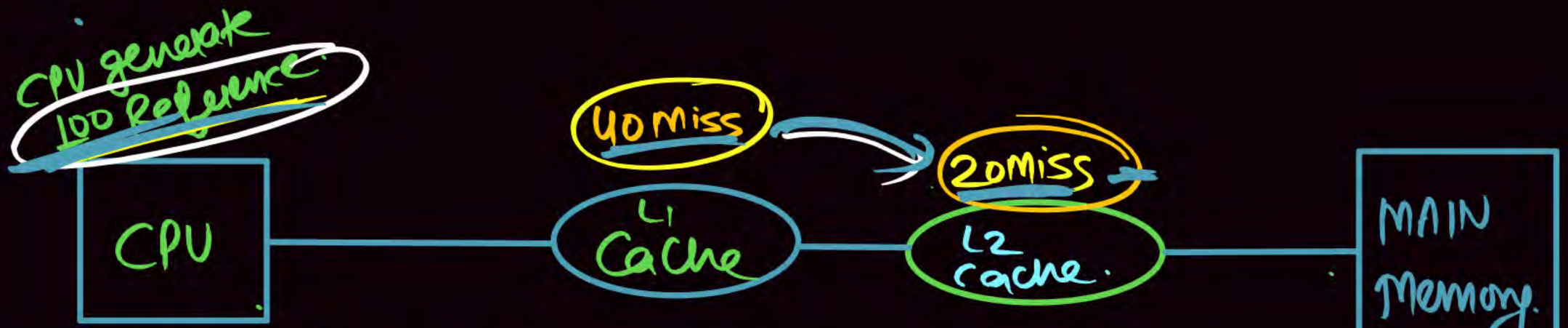
② Global Miss Rate [GMR]

1. Local Miss Rate [LMR]
2. Global Miss Rate [GMR]

$$\text{Local Miss Rate} = \frac{\#\text{misses in the cache}}{\#\text{ accesses to that cache}}$$

$$\text{Global Miss Rate} = \frac{\#\text{misses in the cache}}{\text{Total \#CPU generated reference}}$$

Local & Global Miss Rate in Multilevel Cache.



$$LMR_{L1} = \frac{40}{100} = 0.4$$

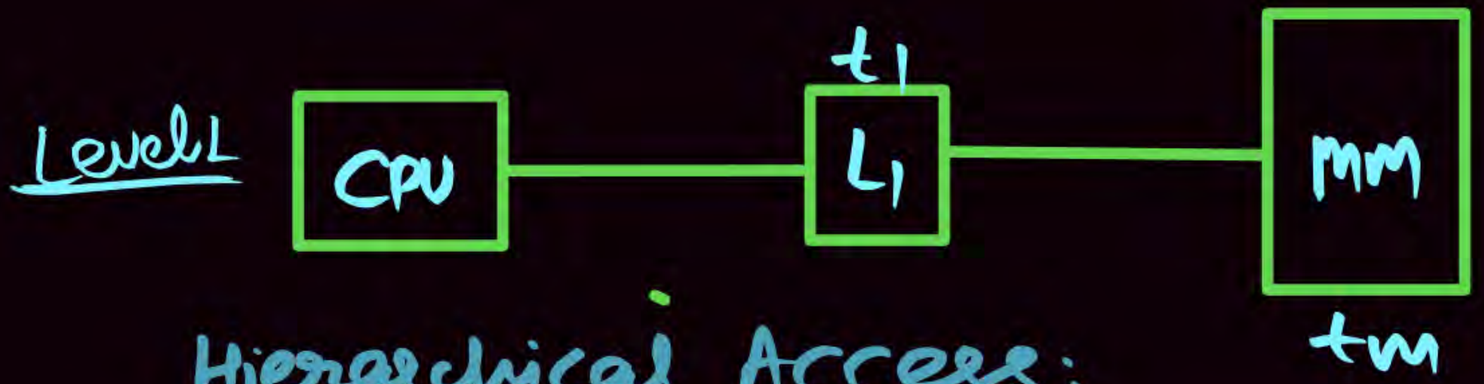
$$LMR_{L2} = \frac{20}{40} = 0.5$$

$$GMR_{L1} = \frac{40}{100} = 0.4$$

$$GMR_{L2} = \frac{20}{100} = 0.2$$

$\therefore$

Access time in Multilevel Cache



Hierarchical Access:

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

$$\Rightarrow T_{avg} = h t_1 + t_m + t_1 - h t_m - h t_1$$

$$= t_1 + t_m - h t_m$$

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

OR

$$T_{avg} = t_1 + (1-h)[t_m]$$

If 1 Level

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

OR

$$T_{avg} = t_1 + (1-h)\, t_m$$ ✓

If 2 Level

$$T_{avg} = h_1 t_1 + (1-h_1)\, h_2 (t_2 + t_1) + (1-h_1)(1-h_2)^{h_3'} [t_m + t_2 + t_1]$$

OR

$$T_{avg} = t_1 + (1-h_1)\left[t_2 + (1-h_2)\, t_m\right]$$

In Multi Level Cache

Tavg is Calculate in term of Hit time, Miss Rate, Miss Penalty.

If 1 Level

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

OR

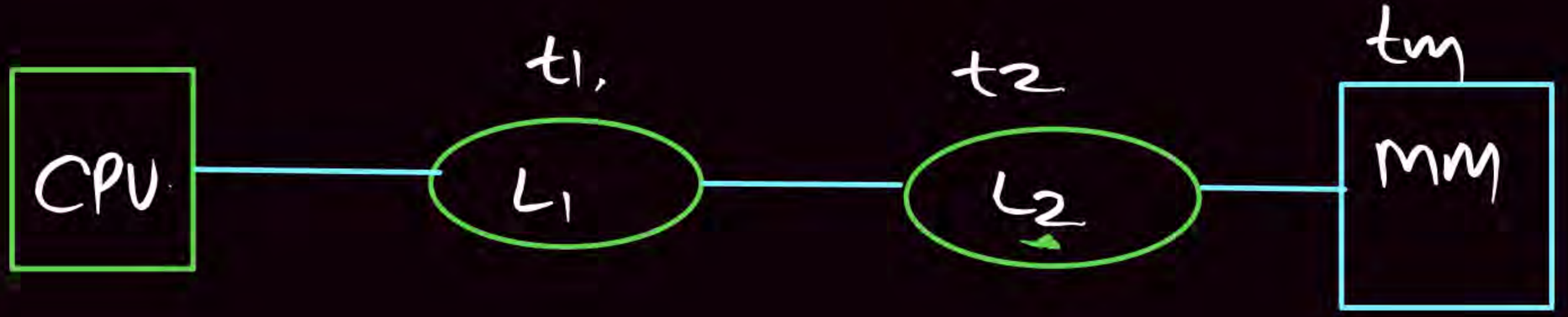
$$T_{avg} = t_1 + (1-h)\, t_m$$ ✓

If 2 Level

$$T_{avg} = h_1 t_1 + (1-h_1)\, h_2 (t_2 + t_1) + (1-h_1)(1-h_2)^{h_3'} [t_m + t_2 + t_1]$$

OR

$$T_{avg} = \underbrace{t_1}_{\text{HIT time in Level1}} + \underbrace{(1-h_1)}_{\text{miss Rate of Level1}} \left[\underbrace{t_2}_{\text{Hit time in Level2}} + \underbrace{(1-h_2)}_{\text{miss Rate in Level2}} \underbrace{t_m}_{\text{MM Access time}} \right]$$


CPU
t1
L1
t2
L2
tm
MM

Average access time of the memory is always calculated in terms of Hit time, miss rate and miss penalty is as follows:

$$T_{avg} = \text{Hit time } L_1 + (\text{Miss Rate } L_1 * \text{Miss penalty } L_1)$$ -(1)

$$\text{Miss penalty } L_1 = \text{Hit time } L_2 + (\text{Miss rate } L_2 * \text{Miss penalty } L_2)$$ -(2)

$$\text{Miss penalty } L_2 = \text{MM Access Time}$$ -(3)

(2) & (3) - Put into eq (1)

$$T_{avg} = \text{Hit time in } L_1 + \text{Miss Rate of } L_1 \left(\text{Hit time in } L_2 + \text{Miss Rate of } L_2 \left(\text{MM Access time} \right) \right)$$

$$\boxed{T_{avg} = t_1 + (1-h_1)\left[t_2 + (1-h_2)\, t_m\right]}$$

Same already we Derived. (Refer P.T.O)

If 1 Level

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

OR

$$T_{avg} = t_1 + (1-h)\, t_m$$ ✓

If 2 Level

$$T_{avg} = h_1 t_1 + (1-h_1)\, h_2 (t_2 + t_1) + (1-h_1)(1-h_2)\,(h_3)'\,[t_m + t_2 + t_1]$$

OR

$$T_{avg} = \underbrace{t_1}_{\text{HIT time in Level1}} + \underbrace{(1-h_1)}_{\text{miss Rate of Level1}} \left[\underbrace{t_2}_{\text{Hit time in Level2}} + \underbrace{(1-h_2)}_{\text{miss Rate in Level2}} \underbrace{t_m}_{\text{MM Access time}} \right]$$

Types of Misses

In the CM design 3 types of misses are present.

1) Compulsory miss - (Cold start miss / first reference miss)

   This miss will occur when the very first reference to the cache itself a miss.

2) Capacity Miss - This miss will occur when cache is full.

3) Conflict Miss (Collison miss / reference miss)

   This miss will occur when the too many blocks are placed into same cache line or same cache SET.

Home work

PYQ's & Home Work Questions

Q.1 Consider a Direct Mapping if the size of Cache memory is 512KB & Main Memory 512 KB & Cache line size (Block) is 64KB the calculate the number of bit required for

(i) P.A (ii) TAG (iii) B.O (iv) W.O

(v) #LINES (vi) TAG Memory Size.

Q.2 Consider a Direct Mapping, Cache size = 64 byte, Line Size = 8 Byte. MM = 256 Byte then #bits for P.A, TAG, L.O, W.O Tag memory size?

Q.3 Consider a Direct Mapping, Cache size = 128 KB, Line Size = 64 Byte. Main Memory is 1MB then what is the line number of physical address $(ABCDE)_{16}$?

Q.4 Consider a 2-way set associative if the size of cache memory is 512KB & Main Memory 512MB & Cache line size is 64KB then calculate the Number of bit Required for

1. (i) P.A (ii) TAG (iii) L.O (iv) W.O

(2) #lines (3) TAG memory size

Q.5 Consider a 2-way set associative Cache Size = 256 KB, Line size = 32 Byte, MM = 1MB, then what is the set number of Physical address $(ABCDE)_{16}$?

GATE PYQ's

NAT Q.1

An 8-way set associative cache of size 64 KB (1 KB = 1024 bytes) is used in a system with 32-bit address. The address is sub-divided into TAG, INDEX, and BLOCK OFFSET.

The number of bits in the TAG is 19 bits. [GATE-2023-CS: 2M]

Ans (19)

CM Size = $64\,KB = 2^{16}B$, 8 way Set Ass. MM = 2^{32} Byte.

Direct Mapping Tag bits = $\frac{MM\,Size}{CM\,Size} \Rightarrow \frac{2^{32}B}{2^{16}B} = 2^{16} \Rightarrow$ Tag = 16 bit

---

SET Associative Mapping

Tag bits = # Tag bits in Direct Mapping + $\log_2$ Nway

$\Rightarrow 16 + \log_2(8) \Rightarrow 16+3 =$ 19 bit Ans

OR

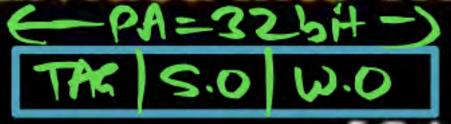
Alternate Method.

NAT Q.1

An 8-way set associative cache of size 64 KB (1 KB = 1024 bytes) is used in a system with 32-bit address. The address is sub-divided into TAG, INDEX, and BLOCK OFFSET.

The number of bits in the TAG is _____.

[GATE-2023-CS: 2M]



1st Approach try.

like Q.12

NAT Q.2

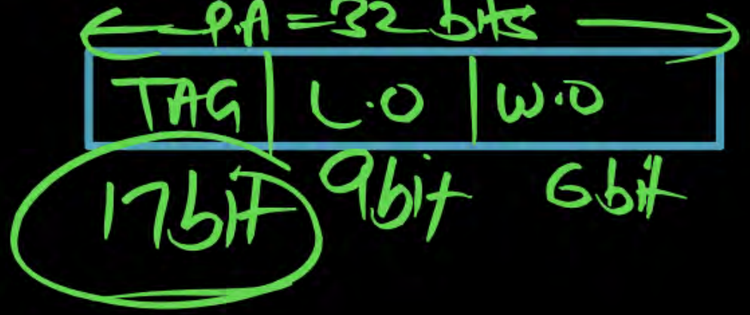
Consider a computer system with a byte-addressable primary memory of size 2^{32} bytes. Assume the computer system has a direct-mapped cache of size 32 KB (1 KB = 2^{10} bytes), and each cache block is of size 64 bytes.

The size of the tag field is __17__ bits.

W.O = 6bit

$$\#LINE = \frac{32KB}{64B} = \frac{2^{15}}{2^6} = 2^9$$

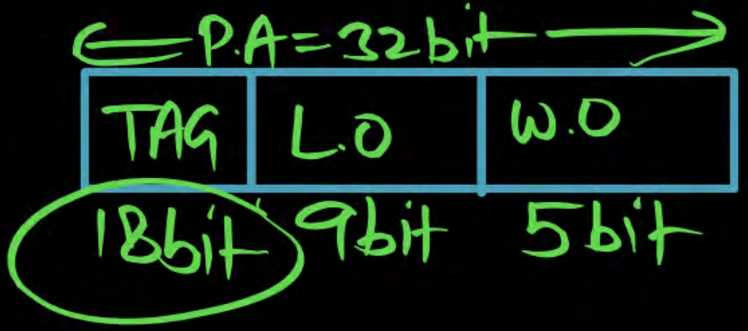
[GATE-2021(Set-1)-CS: 1M]



NAT Q.3

Consider a machine with a byte addressable main memory of 2^{32} bytes divided into blocks of size 32 bytes. Assume that a direct mapped cache having 512 cache lines is used with this machine. The size of the tag field in bits is 18 bit

[GATE-2017(Set-1)-CS: 2M]



MCQ Q.4

Consider a machine with a byte addressable main memory of 2^{20} bytes, block size of 16 bytes and a direct mapped cache having 2^{12} cache lines. Let the addresses of two consecutive bytes in main memory be $(E201F)_{16}$ and $(E2020)_{16}$. What are the tag and cache line address (in hex) for main memory address $(E201F)_{16}$?

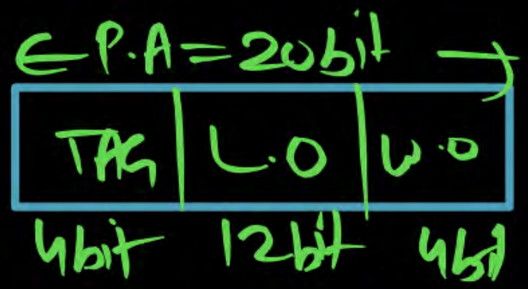
[GATE-2015(Set-3)-CS: 1M]

(A) E, 201

(B) F, 201

(C) E, E20

(D) 2, 01F



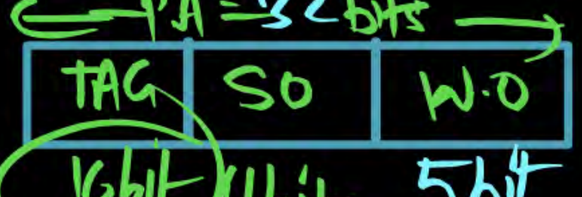
Q.5 [Common Data for this and next question]

A computer has a 256 Kbyte, 4-way set associative. Write back data cache with block size of 32 Bytes. The processor sends 32 bit address to the cache controller. Each cache tag directory entry contains, in addition to address tag, 2 valid bits, 1 modified bit and 1 replacement bit.

The number of bits in the tag field of am address is

(a) 11 (b) 14 (c) 16 (d) 27

[GATE - 2012: 2 Marks]



TAG = 32 - (11 + 5)
= 16 bit

4 way
CM Size = 256 KB
B.S = 32 Byte = 2^5 ⇒ W.O = 5 bit
P.A = 32 bit
#LINE = $\frac{CM}{B.S} = \frac{256KB}{32B} = \frac{2^{18}B}{2^5} = 2^{13}$ lines

#SET = $\frac{2^{13}}{2^2} = 2^{11}$ set
⇒ S.O = 11 bit

Q.6 [Common Data from previous question]

A computer has a 256 Kbyte, 4-way set associative. Write back data cache with block size of 32 Bytes. The processor sends 32 bit address to the cache controller. Each cache tag directory entry contains, in addition to address tag, 2 valid bits, 1 modified bit and 1 replacement bit. [GATE - 2012: 2 Marks]

The size of the cache tag directory is

(a) 160 Kbits (b) 136 Kbits

(c) 40 Kbits (d) 32 Kbits

Tag entry = 16+2+1+1 = 20 bit

Tag memory = #Lines × Tag bits

$\Rightarrow 2^{13} \times 20 \Rightarrow 2^3 \times 20 \times 2^{10}$ bit

$\Rightarrow 8 \times 20 \times K$ bit

$= 160$ Kbits

MCQ **Q.7**

An 8KB direct-mapped write back cache is organized as multiple blocks, each of size 32 bytes. The processor generates 32-bit addresses. The cache controller maintains the tag information for each cache block comprising of the following.

1Valid bit

1Modified bit

As many bits as the minimum needed to identify the memory block mapped in the cache.

What is the total size of memory needed at the cache controller to store meta-data (tags) for the cache? [GATE-2011-CS: 2M]

A 4864 bits

B 6144bits

C 6656 bits

D 5376 bits

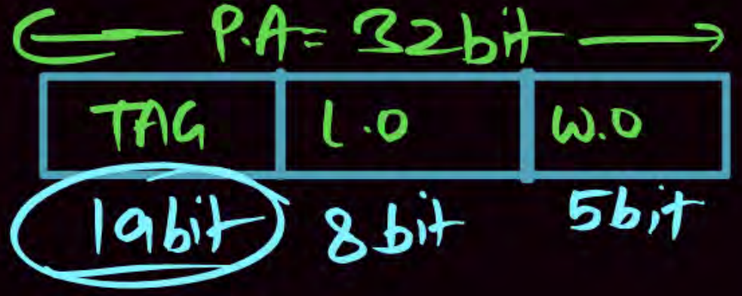
P.A = 32 bit

Block Size = 32B $= 2^5 B \Rightarrow W.O = 5bit$

Cache = 8KB

$\#LINES = \frac{CM\,size}{B.S} = \frac{8kB}{32B} = \frac{2^{13}B}{2^5 B} = 2^8$ Lines $= 256$ Lines

$L.O = 8bit$



TAG = 32 − (8 + 5)

= 19bit

Tag entry size = 19 + 1 + 1 = 21bit

Tag Memory = #Lines × Tag bits

= 256 × 21

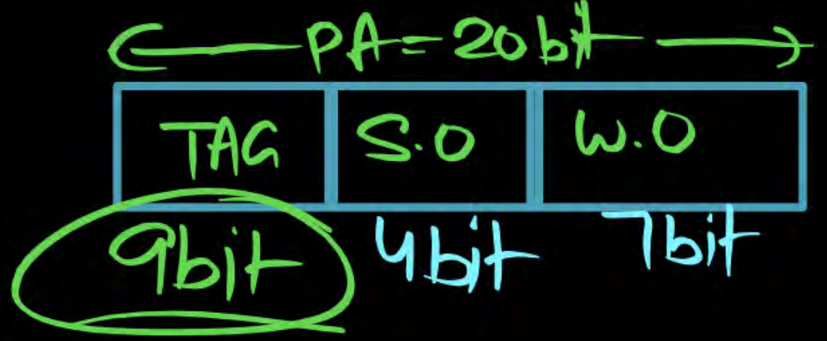
= 5376 bits Ans

Q.8 Q.9

Common Data for next two questions:

Consider a computer with a 4-ways set-associative mapped cache of the following characteristics: a total of 1 MB of main memory, a word size of 1 byte, a block size of 128 words and a cache size of 8 KB.

PA = 20 bit



$(2^{13}B)$

B.S = 128 Word

WordSize = 1 Byte

B.S = 128 × 1B = 128 Byte

W.O = 2^7B = 7bit

#LINE = $\frac{2^{13}B}{2^7B} = 2^6$ Line

#SET = $\frac{2^6}{2^2} = 2^4$ Set

S.O = 4bit

MCQ Q.8

While accessing the memory location 0C795H by the CPU, the contents of the TAG field of the corresponding cache line is

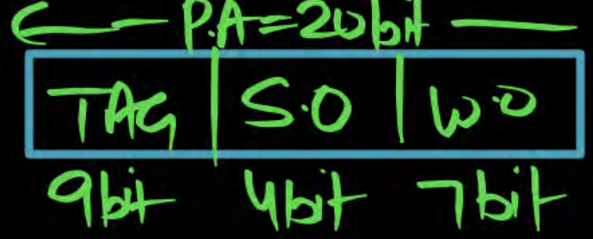
[GATE-2008-CS: 2M]

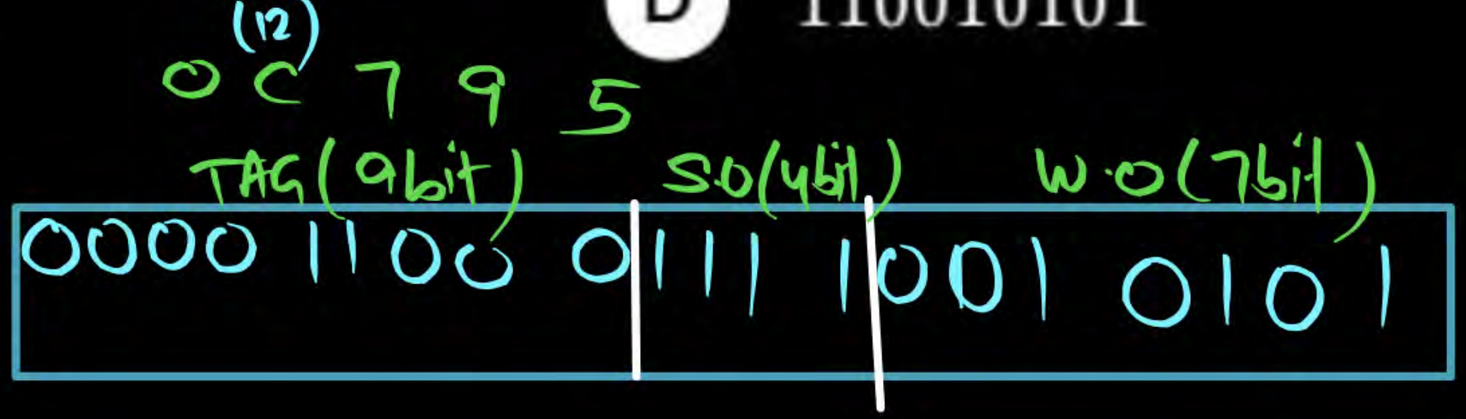
(A) 000011000

(B) 110001111

(C) 00011000

(D) 110010101





MCQ Q.9

The number of bits in the TAG, SET and WORD fields, respectively are:

[GATE-2008-CS: 2M]

A 7, 6, 7

B 8, 5, 7

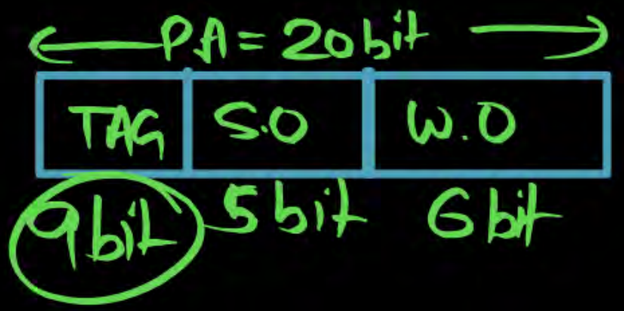
C 8, 6, 6

D 9, 4, 7

Q.10 Consider a 4-way set associative cache consisting of 128 lines with a line size of 64 words. The CPU generates a 20-bit address of a word in main memory. The number of bits in the TAG, SET and WORD fields are respectively.

(a) 9, 6, 5 (b) 7, 7, 6 (c) 7, 5, 8 (d) 9, 5, 6

[GATE - 2007]



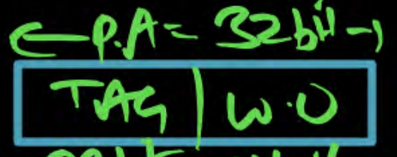
$\#SET = \frac{128}{4} = 32$

$\Downarrow$

S.O = 5 bit

MCQ **Q.11**

A certain processor uses a fully associative cache of size 16 kB. The cache block size is 16 bytes. Assume that the main memory is byte addressable and uses a 32-bit address. How many bits are required for the Tag and the Index fields respectively in the addresses generated by the processor?



[GATE-2019-CS: 1M]

A 24-bits and 0-bits

B 28-bits and 4-bits

C 24-bits and 4-bits

D 28-bits and 0-bits

NAT Q.12

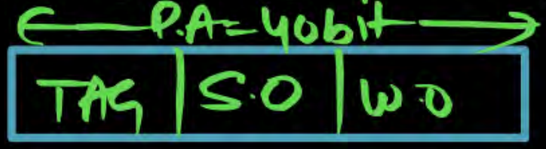
$\#LINES = \frac{CM\ Size}{Block\ Size}$

$\#SET = \frac{\#LINE}{N\text{-}Way}$

The width of the physical address on a machine is 40 bits. The width of the tag field In a 512 KB 8-way set associative cache is _____ bits.

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

P.A = 40 bit, Cache M. Size = 512 kB, 8 way set Associative



(40) P.A = TAG + S.O + W.O

$TAG + S.O + W.O = 40\ bits$ — (1)

Cache Size = #LINES × Block Size

Cache Size = #SET × Block Per Set (N-Way) × Block Size

512 kB = #SET × 8 Way × Block Size

$\#SET \times Block\ Size = 64\ kB\ (2^{16}\ B)$

$S.O + W.O = 16\ bit$ — (2) Put in eq (1)

$TAG + 16 = 40$

$TAG = 40 - 16$

$TAG = 24\ bit$ Ans

$$\#LINES = \frac{CM\,Size}{Block\,size}$$

$$\#SET = \frac{\#LINE}{N\text{-}way}$$

$$CM\,Size = \#LINES \times Block\,Size$$

$$\#LINES = \#SET \times N\text{-}way \quad (Block\ Per\ Set)$$

$$CM\,size \Rightarrow \#SET \times \underbrace{Block\ Per\ Set}_{(N\text{-}way)} \times Block\ size$$

Another Approach.

NAT **Q.12**

The width of the physical address on a machine is 40 bits. The width of the tag field In a 512 KB 8-way set associative cache is _____ bits.

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

Direct Mapping

$$\#Tag = \frac{mm\ Size}{CM\ Size} = \frac{2^{40}}{2^{19}} = 2^{21} \Rightarrow Tag = 21\ bit$$

8 way Set Associative

$$\text{Tag bit} = \#\text{Tag bit in Direct Map} + \log_2(8\ \text{way})$$

$$\Rightarrow 21 + \log_2(8)$$

$$= 21 + 3 = 24\ bit$$

Ans

NAT Q.13

A 4-way set-associative cache memory unit with a capacity of 16 KB is built using a block size of 8 words. The word length is 32 bits. The size of the physical address space is 4 GB. The number of bits for the TAG field is 20bit. [GATE-2014(Set-2)-CS: 1M]

4 GB → 2^{32} B, P.A = 32bit

Block Size = 8 word

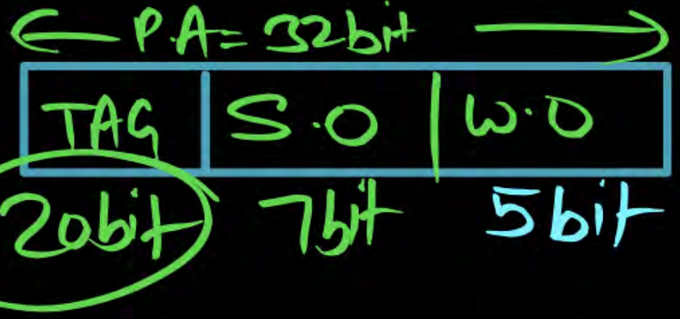
1 word Size = 32 bit (4 Byte)

Block Size = 8 × 4 B ⇒ 32 Byte (2^5 B)

W.O = 5bit

#LINE = $\frac{16\text{KB}}{32\text{B}} = \frac{2^{14}\text{B}}{2^5\text{B}} = 2^9$ lines

#SET = $\frac{2^9}{2^2} = 2^7$ Set ⇒ S.O = 7bit



NAT **Q.14**

A cache memory unit with capacity of N words and block size of B Words is to be designed. If it is designed as a direct mapped cache, the length of the TAG field is 10 bits. If the cache unit is now designed as a 16-way set-associative cache, the length of the TAG field is _____ bits.

[GATE-2017(Set-1)-CS: 2M]

Q.15 The main memory of a computer has 2 cm blocks while the cache has 2 c blocks. If the cache uses the set associative mapping scheme with 2 blocks per set, then block k of the main memory maps to the set. [GATE - 1999]

(a) (k mod m) of the cache
(b) (k mod c) of the cache
(c) (k mod 2 c) of the cache
(d) (k mod 2 cm) of the cache

MCQ **Q.16**

The size of the physical address space of a processor is 2^P bytes. The word length is 2^w bytes. The capacity of cache memory is 2^n bytes. The size of each cache block is 2^m words. For a K-way set-associative cache memory, the length (in number of bits) of the tag field is

[GATE-2018-CS: 2M]

A $P - N - \log_2 K$

B $P - N + \log_2 K$

C $P - N - M - W - \log_2 K$

D $P - N - M - W + \log_2 K$

MCQ **Q.17**

A computer system with a word length of 32 bits has a 16 MB byte-addressable main memory and a 64 KB, 4-way set associative cache memory with a block size of 256 bytes. Consider the following four physical addresses represented in hexadecimal notation.

A1=0×42C8A4,A2=0×546888,A3=0×6A289C,A4=0×5E4880

Which one of the following is TRUE? [GATE-2020-CS: 2M]

(A) A1 and A3 are mapped to the same cache set.

(B) A2 and A3 are mapped to the same cache set.

(C) A3 and A4 are mapped to the same cache set.

(D) A1 and A4 are mapped to different cache sets.

NAT **Q.18**

Consider a set-associative cache of size 2 kb (1 KB = 2^{10} bytes) with cache block size of 64 bytes. Assume that the cache is byte - addressable and a 32-bit address is used for accessing the cache. If the width of the tag field is 22 bits, the associativity of the ache is _____.

[GATE-2021(set-2)-CS: 1M]

MCQ **Q.19**

Consider a 4-way set associative cache (initially empty) with total 16 cache blocks. The main memory consists of 256 blocks and the request for memory blocks is in the following order:

0, 255, 1, 4, 3, 8, 133, 159, 216, 129, 63, 8, 48, 32, 73, 92, 155

Which one of the following memory block will NOT be in cache if LRU replacement policy is used?

[GATE-2009-CS: 2M]

A 3

B 8

C 129

D 216

NAT Q.20

Consider a 2-way set associative cache with 256 blocks and uses LRU replacement. Initially the cache is empty. Conflict misses are those misses which occur due to contention of multiple blocks for the same cache set. Compulsory misses occur due to first time access to the block. The following sequence of accesses to memory blocks (0, 128, 256, 128, 0, 128, 256, 128, 1, 129, 257, 129, 1, 129, 257, 129) is repeated 10 times. The number of conflict misses experienced by the cache is _____.

[GATE-2017(Set-1)-CS: 2M]

2 Way Set Associative Cache

#Cache Line = 256

$$\#SET[S] = \frac{\#LINES}{N\text{-Way}} = \frac{256}{2} = 128$$

K MOD S = i

K MOD 128 = i

Only First time Conflict Miss

2 to 10 (2 to 10th Iteration) = 9 x x

MCQ **Q.21**

If the associativity of a processor cache is doubled while keeping the capacity and block size unchanged, which one of the following is guaranteed to be NOT affected?

[GATE-2014(Set-2)-CS: 2M]

A. Width of tag comparator

B. Width of set index decoder

C. Width of way selection multiplexer

D. Width of processor to main memory data bus

NAT **Q.22**

Consider a two-level cache hierarchy with L_1 and L_2 caches. An application incurs 1.4 memory accesses per instruction on average. For this application, the miss rate of L_1 cache is 0.1; the L_2 cache experiences on average, 7 misses per 1000 instructions. The miss rate of L_2 expressed correct to two decimal places is _____.

[GATE-2017(Set-1)-CS: 1M]

MCQ **Q.23**

In a two-level cache system, the access times of L_1 and L_2 caches are 1 and 8 clock cycles, respectively. The miss penalty from the L_2 cache to main memory is 18 clock cycles. The miss rate of L_1 cache is twice that of L_2. The average memory access time (AMAT) of this cache system is 2 cycles. The miss rates of L_1 and L_2 respectively are:

[GATE-2017(Set-2)-CS: 2M]

A 0.111 and 0.056

B 0.056 and 0.111

C 0.0892 and 0.1784

D 0.1784 and 0.0892

Q.24

Common Data for next two questions:

Consider a machine a 2-way set associative data cache of size 64Kbytes and block size 16 bytes. The cache is managed using 32 bit virtual addresses and the page size is 4 Kbytes. A program to be run on this machine begins as follows:

```
Double ARR [1024] [1024]
Int i, j;
/ * Initalize array ARR to 0.0 */
for (i = 0; i < 1024; i ++)
for (j = 0; j < 1024; j++)
ARR [i] [j] = 0.0;
```

The size of double 8 bytes. Array ARR is in memory starting at the beginning of virtual page 0×FF000 and stored in row major order. The cache is initially empty and no pre-fetching is done. The only data memory references made by the program are those to array ARR.

MCQ

The total size of the tags in the cache directory is

[GATE-2008-CS: 2M]

A 32 kbits

B 34 kbits

C 64 kbits

D 68 kbits

Q.25 [Common Data for this and next question]

Consider two cache organization. The first one is 32 KB 2-way set associative with 32-byte block size. The second one is of the 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 a latency of k/10 ns. The hit latency of the set associative organization is h_1 while that of the direct mapped one 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

[GATE - 2006: 2 Marks]

Q.26 [Common Data from previous question]

Consider two cache organization. The first one is 32 KB 2-way set associative with 32-byte block size. The second one is of the 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 a latency of k/10 ns. The hit latency of the set associative organization is h_1 while that of the direct mapped one is h_2. The value of h_2 is

(a) 2.4 ns (b) 2.3 ns (c) 1.8 ns (d) 1.7 ns

[GATE - 2006: 2 Marks]

Q.27 A computer system has a level - 1 instruction cache (1-cache), a level-1 data cache(D-cache) and a level-2 cache(L2-cache) with the following specifications.

	Capacity	Mapping method	Block size
1-cache	4K words	Direct mapping	4 words
D-cache	4K words	2-way set associative mapping	4 words
L2-cache	64K words	4-way set associative mapping	16 words

Capacity mapping method block size 1-cache 4K words direct mapping 4 words D-cache 4 k words 2 way set associative mapping 4 words L2-cache 64K words 4-way set associative mapping 16 words. The length of the physical address of a word in the main memory is 30 bits. The capacity of the tag memory in the 1-cache, D-cache & L2-cache is. Respectively,

(a) 1K×18-bit, 1K×19-bit, 4K×16-bit (b) 1K×16-bit, 1K×19-bit, 4K×18-bit

(c) 1K×16-bit, 512×18-bit, 1K×16-bit (d) 1K×16-bit, 512×18-bit, 1K×18-bit

[GATE - 2006: 2 Marks]

Q.28 Consider a small two-way set-associative cache memory, consisting of 4 blocks. For choosing the block to be replaced, use the least recently used (LRU) scheme. The number of cache misses for the following sequence of block addresses is 8, 12, 0, 12, 8

(a) 2 (b) 3 (c) 4 (d) 5

[GATE - 2004]

Q.29 Consider a system with 2 KB direct mapped data cache with a block size of 64 bytes. The system has a physical address space of 64 KB and a word length of 16 bits. During the execution of a program, four data words P, Q, R and S are accessed in that order 10 times (i.e., PQRSPQRS.....) Hence, there are 40 accesses to data cache altogether. Assume that the data cache is initially empty and no other data words are accessed by the program, The addresses of the first bytes of P, Q, R and S are 0×A248, 0×C28A, 0×CA8A and 0×A262, respectively. For the execution of the above program, which of the following statements is/are TRUE with respect to the data cache? [2022: MSQ 2M]

A. Every access to S is a hit.

B. Once P is brought to the cache it is never evicted.

C. At the end of the execution only R and S reside in the cache.

D. Every access to R evicts Q from the cache.

THANK YOU!