CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Cache Organization

Lecture No.- 08

By- Vishvadeep Gothi sir

Recap of Previous Lecture

Topic Fully Associative Mapping

Topic Set Associative Mapping

Topics to be Covered

- Topic: Block Replacement
- Topic: Cache Miss Penalty
- Topic: Types of Cache Miss

[MCQ]

no. of sets = $\frac{16}{4} = 4$ (mm block no. % 4)

#Q. Consider a 4-way set associative cache (initially empty) with total 16 cache blocks. The main memory consists of 256 block 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?

A 3

B 8

C 129

D ✓ 216

0	~~0~~ 48	~~4~~ 32	8	~~216~~ 92
1	1	133	129	73
2				
3	~~255~~ 155	3	159	63

[MCQ]

#Q. Consider a fully associative cache with 8 cache blocks (numbered 0–7) and the following sequence of memory block requests:

4,3,25,8,19,6,25,8,16,35,45,22,8,3,16,25,7

If LRU replacement policy is used, which cache block will have memory block 7?

0	1	2	3	4	5	6	7
~~4~~ 45	~~3~~ 22	25	8	~~19~~ 3	~~6~~ 7	16	35

A 4

B ✓✓ 5

C 6

D 7

Topic : Cache Miss Penalty

Time required to bring a missed block from main memory to cache

Topic : Cache Miss Penalty

Assume:

Cycles required to send address to memory : 1 cycle

Cycles required to access 1 main memory cell : 10 cycles

Cycles required to transfer 1 cell data to cache : 1 cycle

one time needed

Cache Block Size	Main memory cell size	Miss Penalty
4 bytes	1 byte	$1+(4*10)+(4*1)=45$ cycles
4 bytes	2 bytes	$1+(2*10)+(2*1)=23$ cycles
4 bytes	4 bytes	$1+(1*10)+(1*1)=12$ cycles

if CPU cycle time given

then miss penalty time = miss penalty * 1 cycle time
cycles

[NAT]

#Q. 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 system when the program running on the processor issues a series of read operations is __160__ $\times 10^6$ bytes/sec?

block size = 8 words = 8 * 4 = 32 bytes

Cache miss penalty = 1 + 3 + 8 = 12 cycles = $\frac{12 * 1}{60MHz} = \frac{1}{5}$ μsec

in $\frac{1}{5}$ μsec time, data = 32 bytes

in 1 μsec —11— $= \frac{32 \text{ bytes}}{1/5 \text{ μs}}$

$= 160 \text{ B/μs}$

in 1 sec —11— $= \frac{160 \text{ B}}{10^{-6} \text{ sec}}$

$= 160 * 10^6 \text{ B/sec}$

Topic : Types of Cache Misses

1. Cold or Compulsory Miss
2. Capacity Miss
3. Conflict Miss

Topic : Types of Cache Misses

1. Cold or Compulsory Miss

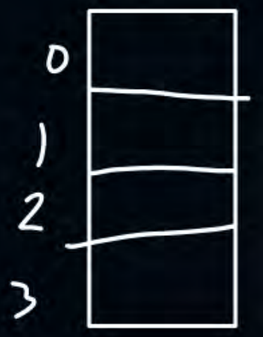
First time access of a block will always cause a miss

To reduce Cold misses: Increase block size

ex:-

mm size = 32 B
block size = 4 B
} 8 blocks in mm
⇓
8 Cold miss

mm size = 32B
block size = 8 B
} 4 blocks in mm
⇓
4 cold miss



Topic : Types of Cache Misses

2. Capacity Miss

If cache is full and hence miss occurs. (should not be the first time access of block)

To reduce Capacity misses: increase cm capacity

Topic : Types of Cache Misses

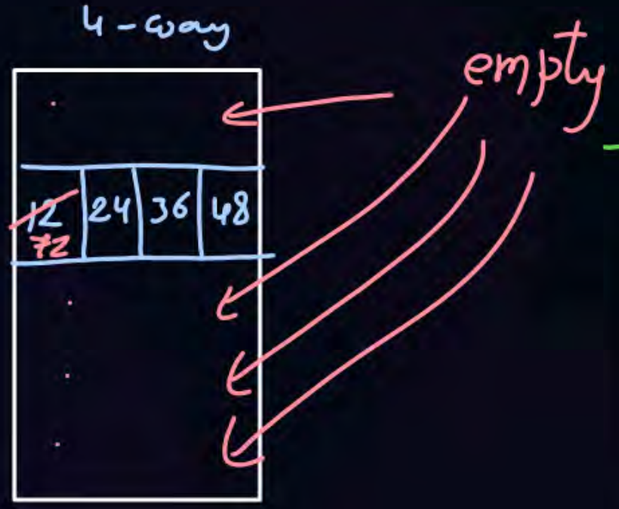
3. Conflict Miss

If cache set is full and ~~cache~~ miss occurs ~~due to tag mismatch~~ for a block.

↓

(miss should not be cold or capacity miss)

To reduce Conflict misses: increase associativity



no any conflict miss in fully associative cache ever

Topic : Example

- No. of blocks in cache = 4
- 2-way set associative cache
- LRU replacement policy
- CPU requests for main memory blocks

} no. of sets in cm = $\frac{4}{2} = 2$

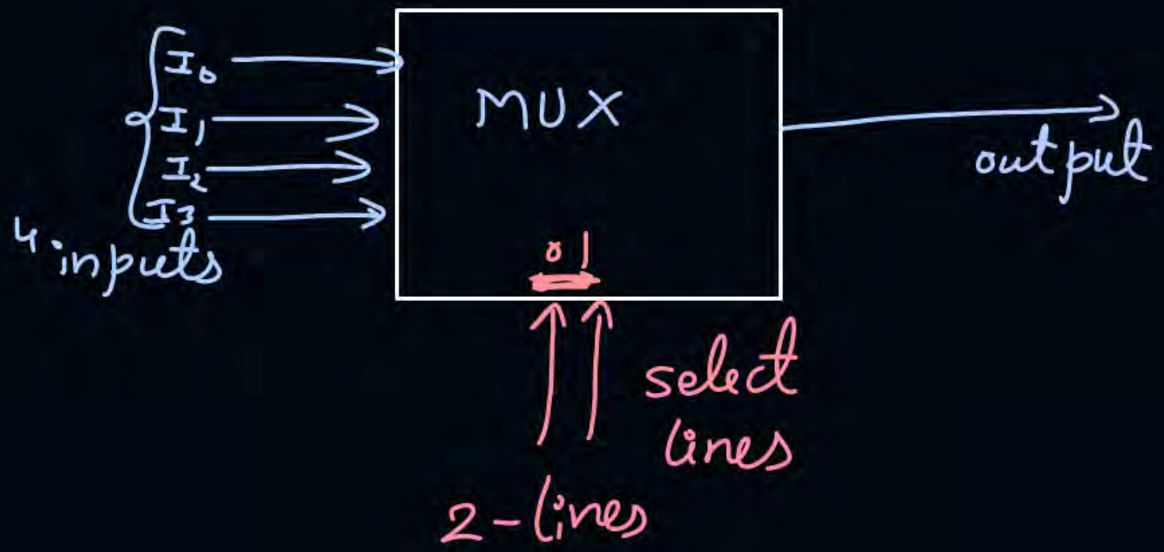
cm set no. = (mm block no.) % 2

	hit		hit		cold	cold	hit		hit	
0,	4,	0,	8,	0,	4,	1,	3,	1,	5,	1,	3
↓	↓		↓		↓				↓		↓
cold	cold		cold		conflict				cold		capacity

00	0	~~4~~ ~~8~~ 4
01	1	~~5~~ 5

0 % 2 = 0
4 % 2 = 0
8 % 2 = 0
1 % 2 = 1
3 % 2 = 1
5 % 2 = 1

MUX :—



Topic : Checking Hit/Miss in Direct Mapping

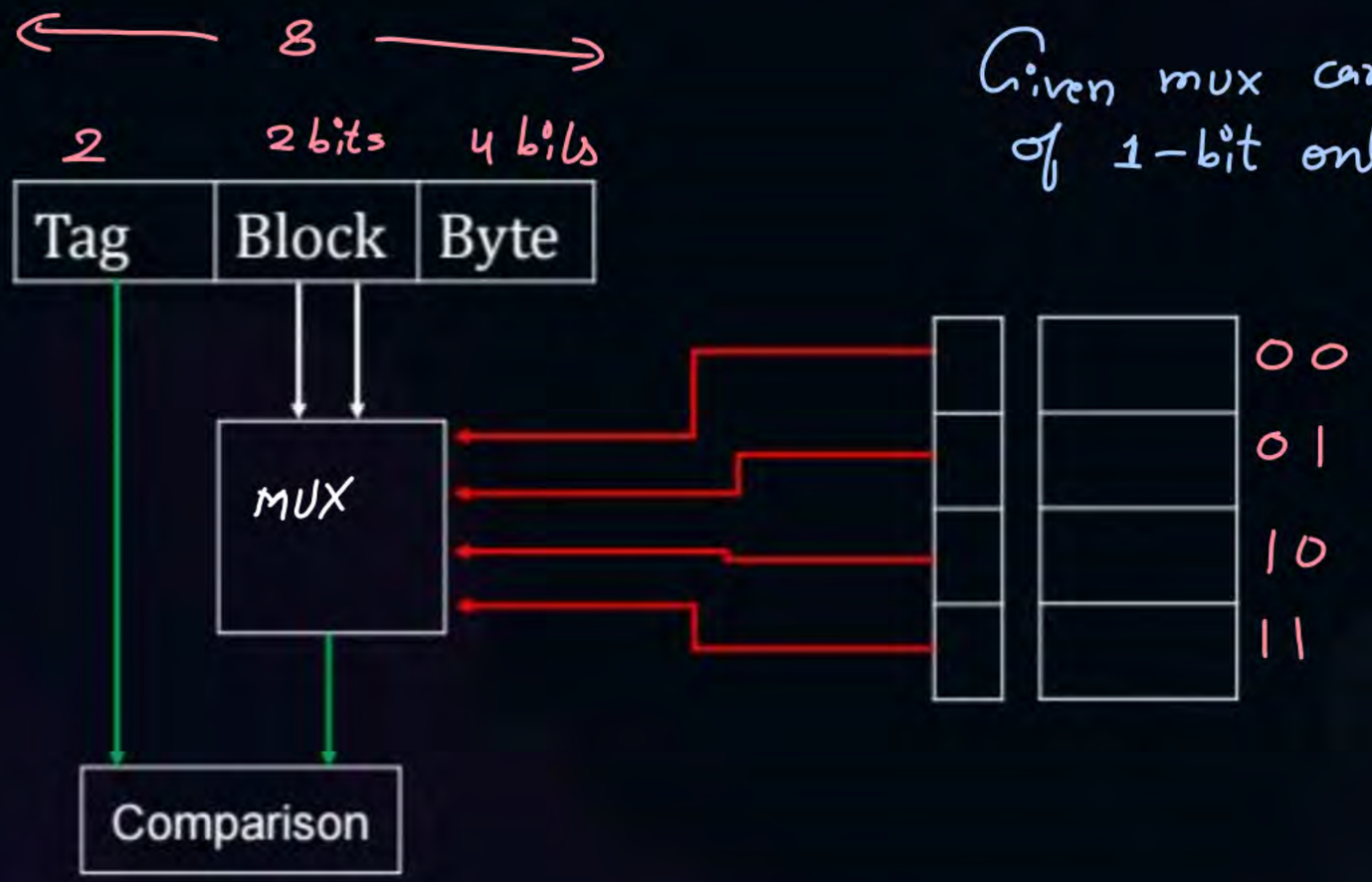
Topic : Hardware Implementation in Direct Mapping

Blocks in cache = 4

Block size = 16 bytes

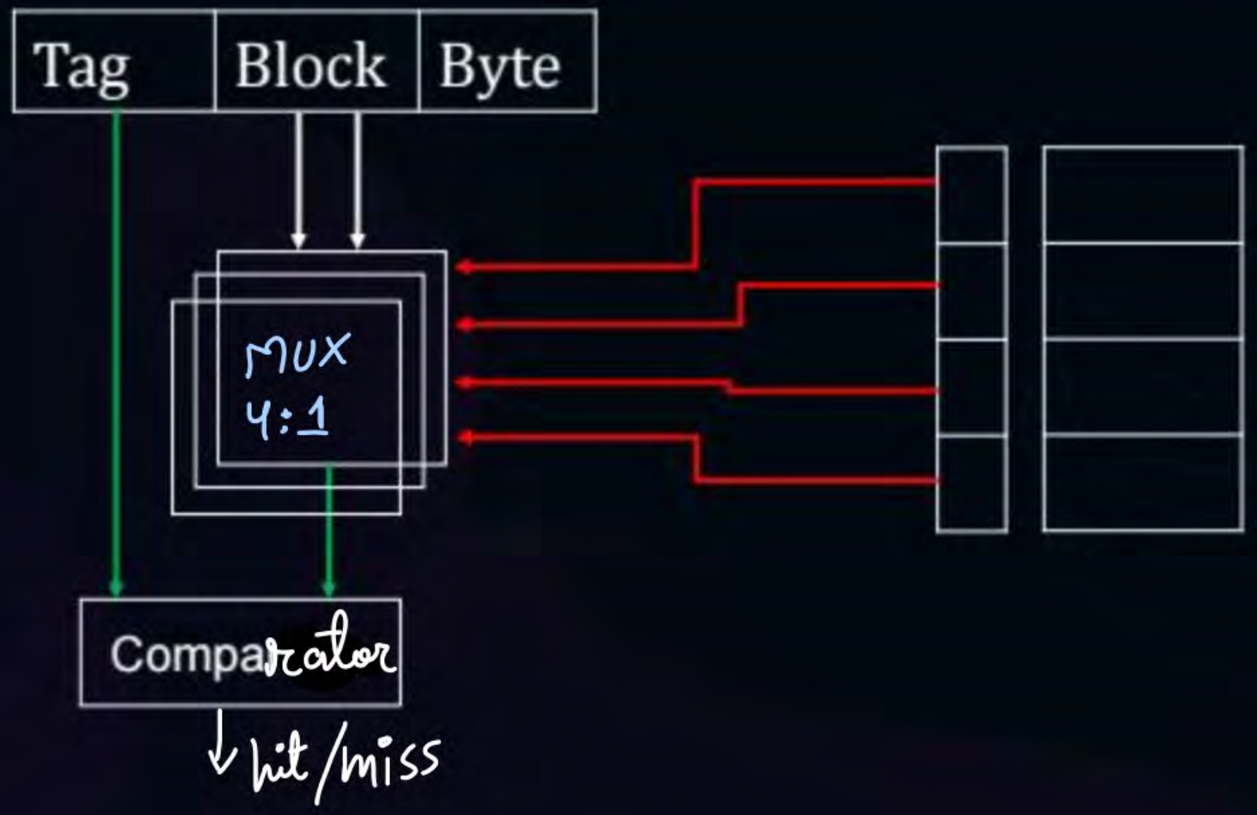
Main memory address = 8-bits

Topic : Hardware Implementation in Direct Mapping



Given mux can take each input of 1-bit only.

Topic : Hardware Implementation in Direct Mapping



H/W Requirements for direct mapping :-

1. Size of MUX needed for tag selection = No. of blocks in cache : 1
2. No. of MUX —||— = No. of tag bits
3. No. of comparators needed = 1
4. Size of —||— = No. of bits in tag

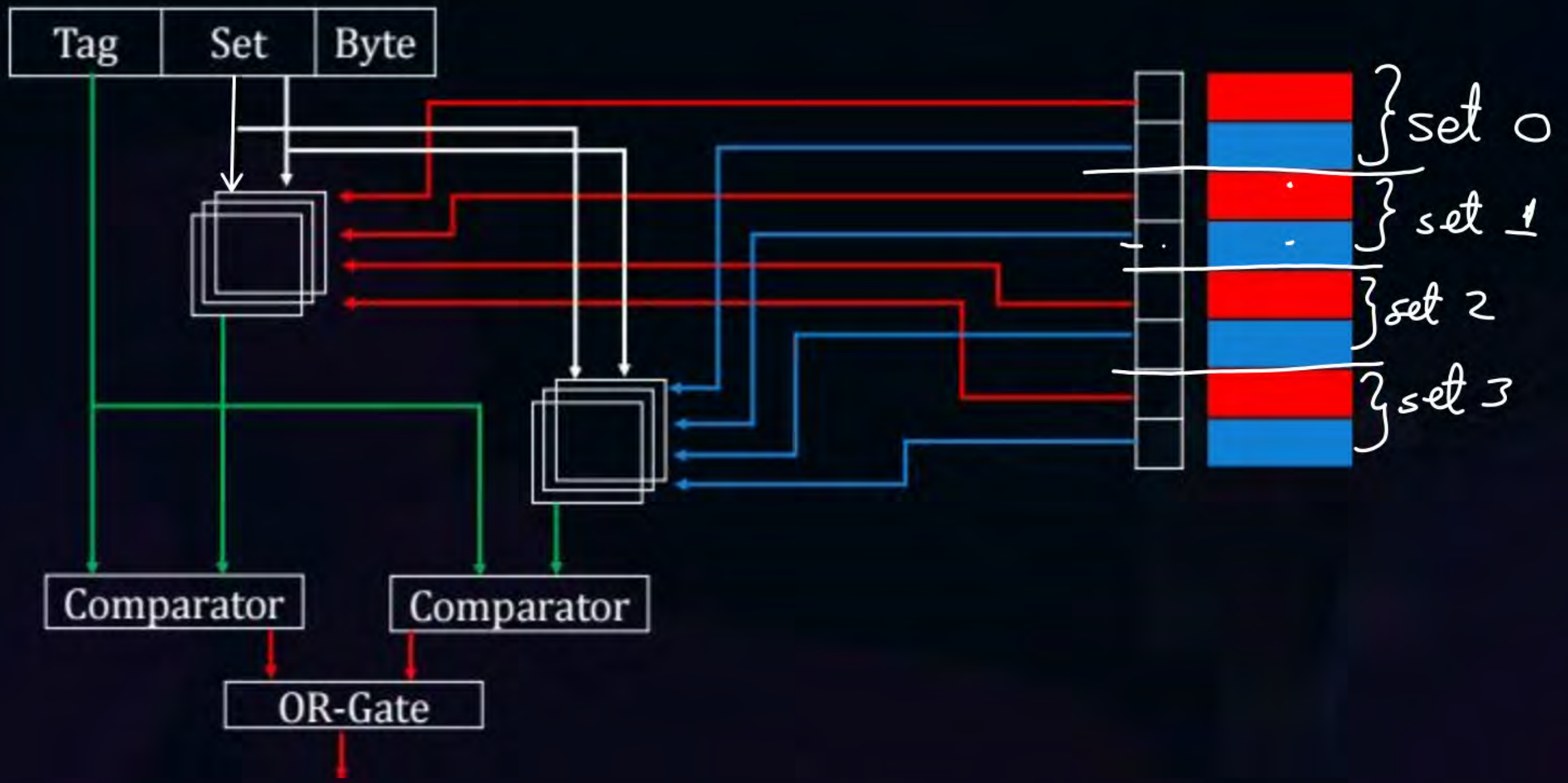
Topic : Hit Latency in Direct Mapping

Hit Latency → time to check hit/miss in cache

= 1 MUX delay + 1 Comparator delay

Topic : Checking Hit/Miss in Set Associative Mapping

Topic : Hardware Implementation in Set Associative Mapping



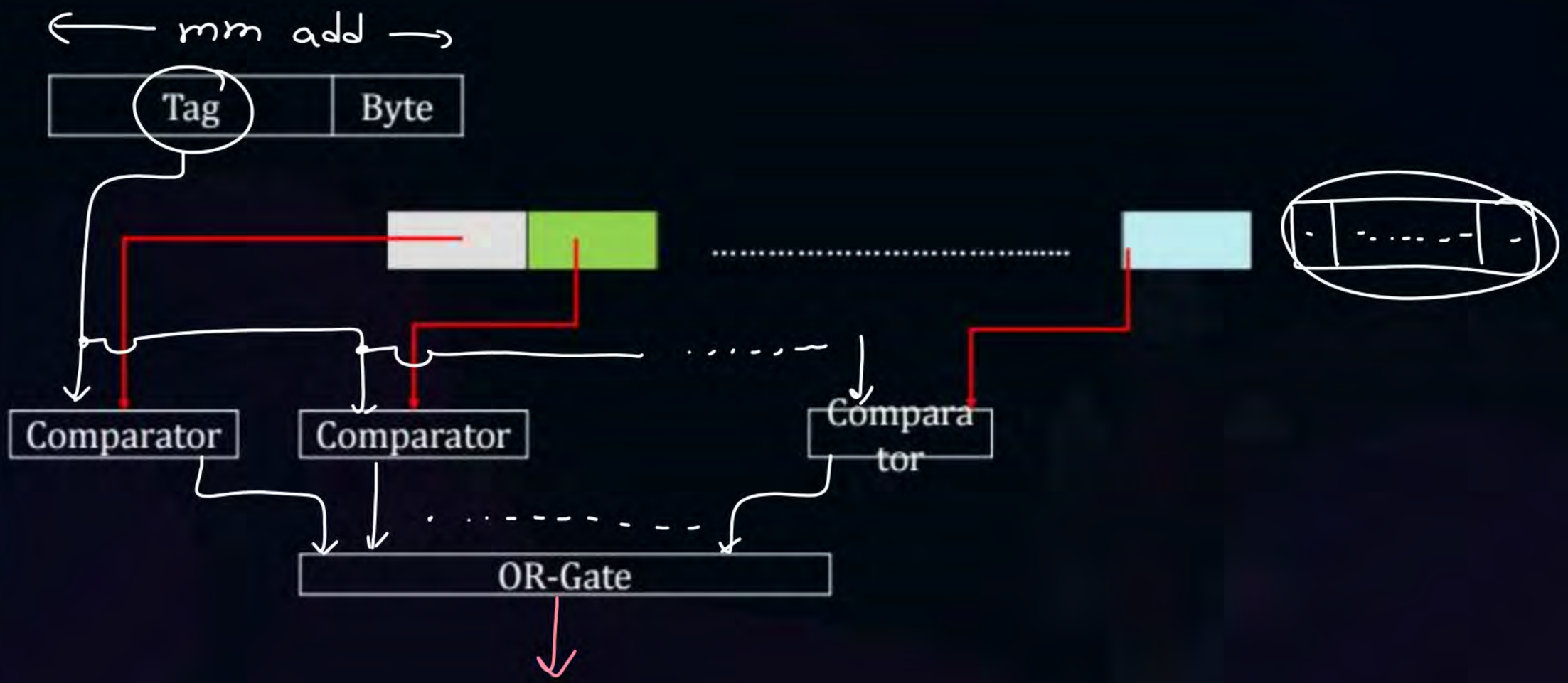
H/W Requirements for k-way set associative cache:-

1. No. of MUX needed for tag selection = k * no. of tag bits
2. Size — " — " — = no. of sets in cache : 1
3. No. of comparators needed = k
4. Size of each comparator = No. of bits in tag
5. No. of OR-gates needed = 1
6. No. of inputs in OR-gate = k

Topic : Hit Latency in Set Associative Mapping

= MUX delay for tag selection + 1 comparator delay + 1 OR gate delay

Topic : Checking Hit/Miss in Fully Associative Mapping



H/w Requirements of fully ass. cache :-

1. No. of comparators needed = no. of blocks in cache
2. Size of —11—— = no. of bits tag
3. No. of OR-gate = 1
4. No. of inputs in OR-gate = no. of blocks in cache

Topic : Hit Latency in Fully Associative Mapping

= Comparator delay + OR gate delay

2 mins Summary

Topic Block Replacement

Topic Cache Miss Penalty

Topic Types of Cache Miss

Happy Learning

THANK - YOU