CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

CACHE Organization

Lecture No.- 01

By- Vishvadeep Gothi sir

Recap of Previous Lecture

Topic Multiple Chips in Single Memory System

Topic DRAM Refresh

Topics to be Covered

Topic Associative Memory

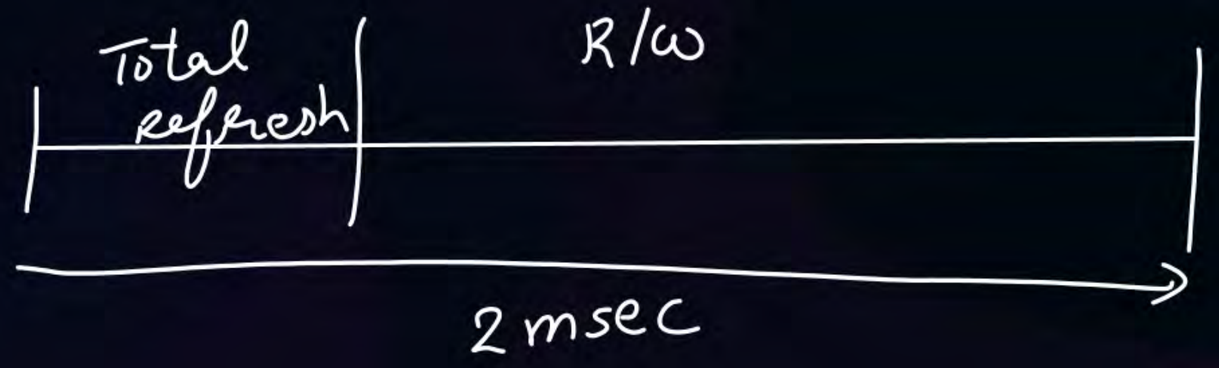
Topic Locality of Reference

Topic Cache Memory

[NAT]

GATE-2018

#Q. A 32-bit wide main memory unit with a capacity of 1 GB is built using 256M X 4-bit DRAM chips. The number of rows of memory cells in the DRAM chip is 2^{14}. The time taken to perform one refresh operation is 50 nanoseconds. The refresh period is 2 milliseconds. The percentage (rounded to the closet integer) of the time available for performing the memory read/write operations in the main memory unit is _____?



Total refresh time $= 2^{14} * 50\ ns$

$= 800 * 2^{10}\ ns$

$2^{10} = 1000$

$= 800\ \mu sec$

$= 0.8\ msec$

$2^{10} = 1024$

$= 819200\ ns$

$= 0.8192\ ms$

% of time Remaining for Read write $= \frac{2 - 0.8}{2} * 100\%$

$= 60\%$

$= \frac{2 - 0.8192}{2} * 100\%$

$= 59\%$

[NAT]

Total cells

$x = 128k$
$y = 2$

#Q. A DRAM chip of 256K×8 bits has x rows of cells with y cells in each row? If DRAM takes 20ns for 1 refresh and 2.56 milliseconds for entire chip refresh then the value of x , y is _____?

$2.56\ ms = x * 20ns$

$x = \frac{2.56 * 10^{-3}\ sec}{20 * 10^{-9}\ sec}$

$= 0.128 * 10^{6}$

$= 128 * 10^{3}$

$= 128K$

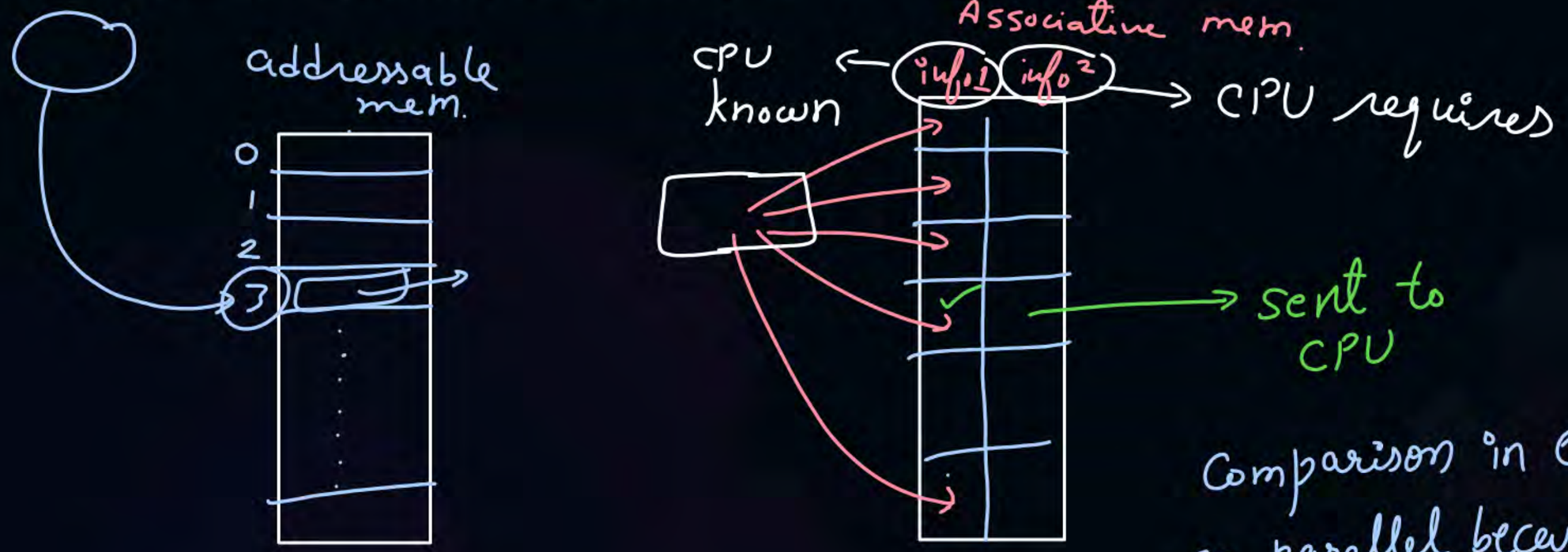
$128k * y = 256k$

$y = 2$

Topic : Associative Memory

↳ in this, cells do not have addresses

Known as content addressable memory also



Comparison in each cell is done in parallel because each cell has it's own comparison H/w.

⇒ mem becomes faster but expensive

⇒ Associative mem. is faster & more expensive as compared to SRAM.

⇒ Associative mem. is used for implementing cache, TLB.

Translation lookaside buffer

Topic : Locality of Reference

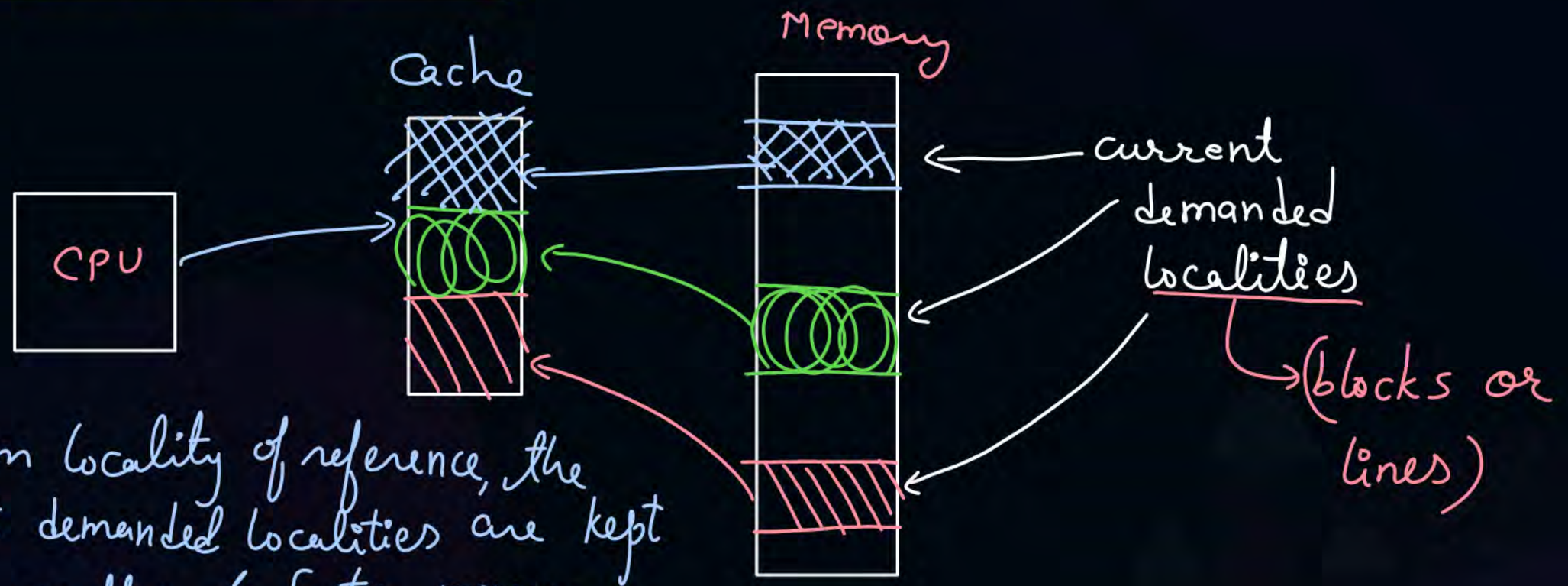
If CPU has requested one address for memory access, then that particular address or near by addresses will be accessed soon.

Topic : Locality of Reference

Types:

1. Spatial (according to space) ⇒ if CPU refers near by addresses soon.
2. Temporal (according to time) ⇒ —11 —— same address soon.

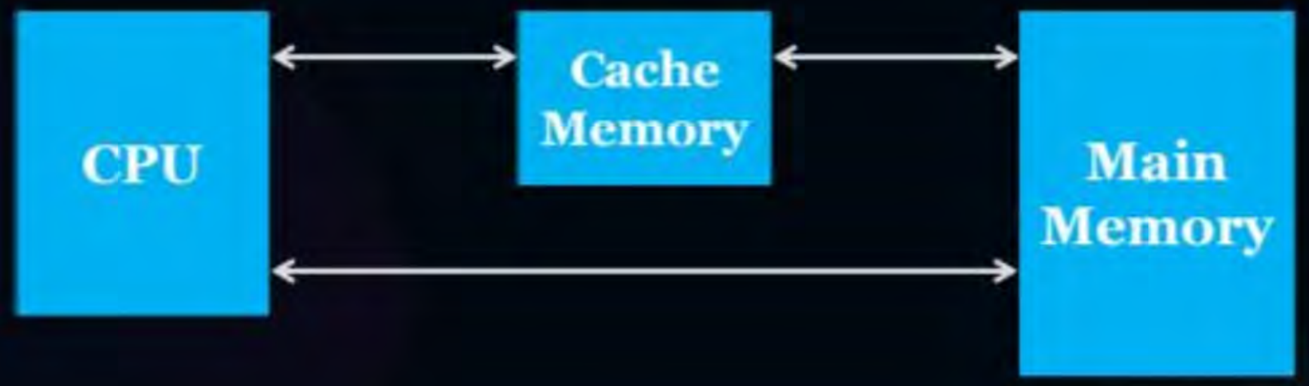
Topic : Cache Memory



Based on locality of reference, the current demanded localities are kept into a smaller & faster memory, called as Cache; to reduce CPU's mem. access time.

Topic : Cache Memory

Use of cache reduces average mem. access time.



Topic : Working of Cache Memory

1. Cache Hit ⇒ when CPU's demanded content is present in cache.
2. Cache Miss ⇒ —— " —— " —— is not —— " ——
3. Hit Ratio

$$\text{Hit ratio } (H \text{ or } h) = \frac{\text{no. of hits in cache}}{\text{Total no. of mem. references}}$$

$$\text{Miss ratio} = \frac{\text{no. of miss in cache}}{\text{Total no. of mem. references}} = (1 - H)$$

when there is a miss in cache (memory read), then demanded content is sent to CPU from main memory and in parallel to that a block (to which missed content belongs) is copied to cache.

---

while bringing missed block from mm to cache, if cache does not have empty space then an existing block from cache is replaced.

Topic : Average Memory Access Time

T_{avg} = H * time needed to access content in case of hit + (1 − H) * time needed to access content in case of miss ..... (1)

ex:-

CPU refers mem 200 times

no. of hits = 160

no. of miss = 40

Cache hit time = 20ns

Cache miss time = 100 ns

Total hit time = 160 * 20ns = 3200 ns

Total miss time = 40 * 100 = 4000 ns

Total time = 7200 ns

Avg mem. access = $\frac{7200\ ns}{200\ ns} = 36\ ns$

Hit ratio = $\frac{160}{200} = 0.8$

$T_{avg} = 0.8 * 20 + (1-0.8) * 100$

$= 16 + 20$

$= 36$ nsec

$$\frac{160 * 20 + 40 * 100}{200}$$

$$= \frac{160 * 20}{200} + \frac{40 * 100}{200}$$

$$= 0.8 * 20 + 0.2 * 100$$

Topic : Types of Cache Accesses

Simultaneous Access: (Parallel access)

Request for cache and main-memory are generated simultaneously

$$T_{avg} = H * t_{cm} + (1-H)\, t_{mm} \quad \text{.....---(2)}$$

H = Hit ratio of cache

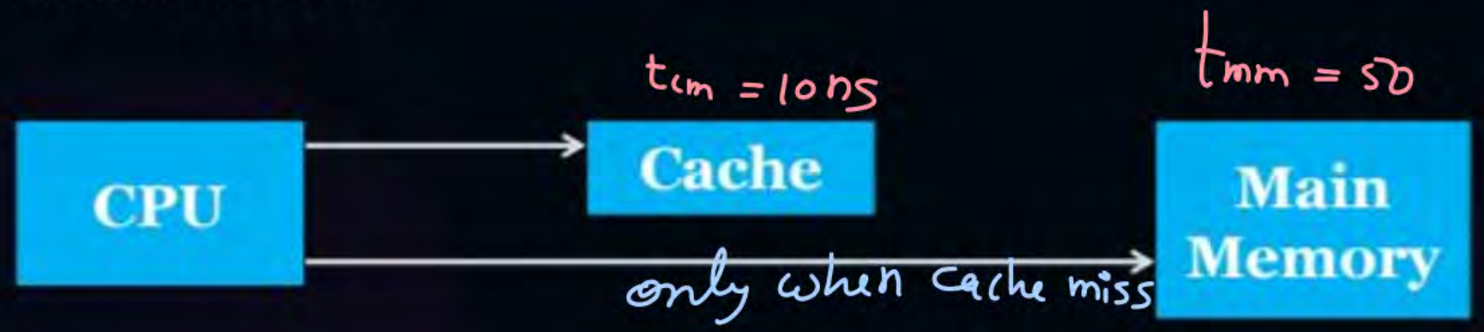
t_{cm} = cache access time

t_{mm} = main mem. access time

Topic : Types of Cache Accesses

Hierarchical Access: (serial access)

Only cache is accessed first



$$T_{avg} = H * t_{cm} + (1-H)(t_{cm} + t_{mm}) \quad \ldots\ldots\ldots\ldots \quad (3)$$

or

$$= t_{cm} + (1-H)\, t_{mm}$$

$$t_{avg} = H * t_{cm} + (1-H)(t_{cm} + t_{mm})$$

$$= \cancel{H * t_{cm}} + t_{cm} + t_{mm} - \cancel{H * t_{cm}} - H * t_{mm}$$

$$= t_{cm} + (1-H)\, t_{mm}$$

2 mins Summary

Topic Associative Memory

Topic Locality of Reference

Topic Cache Memory

Happy Learning

THANK - YOU