CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Cache Organization

Lecture No.- 1

By- Vishvadeep Gothi sir

Recap of Previous Lecture

Topic Multiple Chips in Single Memory System

Topic DRAM Refresh

Topic Associative Memory

Topics to be Covered

Topic Locality of Reference

Topic Cache Memory

[NAT]

GATE- 2018 Ans ⇒ 59 to 60

#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 _____?

one chip refresh time = $2^{14} * 50$ ns

$2^{10} = K$

$2^{10} = 1024$

$16 * 1K * 50$ ns

800 usec

$= 0.8$ ms

$= 16 * 1024 * 50$ ns

$= 0.8192$ ms

% of R/w remaining $= \frac{2 - 0.8}{2} * 100\%$ | $\frac{2 - 0.8192}{2} * 100\%$

$= 60\%$ | $= 59\%$

---

32 - bits wide main mem ⇒ when CPU generates add. it expects 32 bits content per address

no. of addresses $= \frac{1GB}{4B} = 2^{28}$

expected mem $= 2^{28} \times 4B$ or $2^{28} \times 32$ bits

$$\text{no. of chips} = \frac{2^{28} \times 32 \text{ bits}}{256\text{M} \times 4 \text{ bits}} = 8 \text{ chips}$$

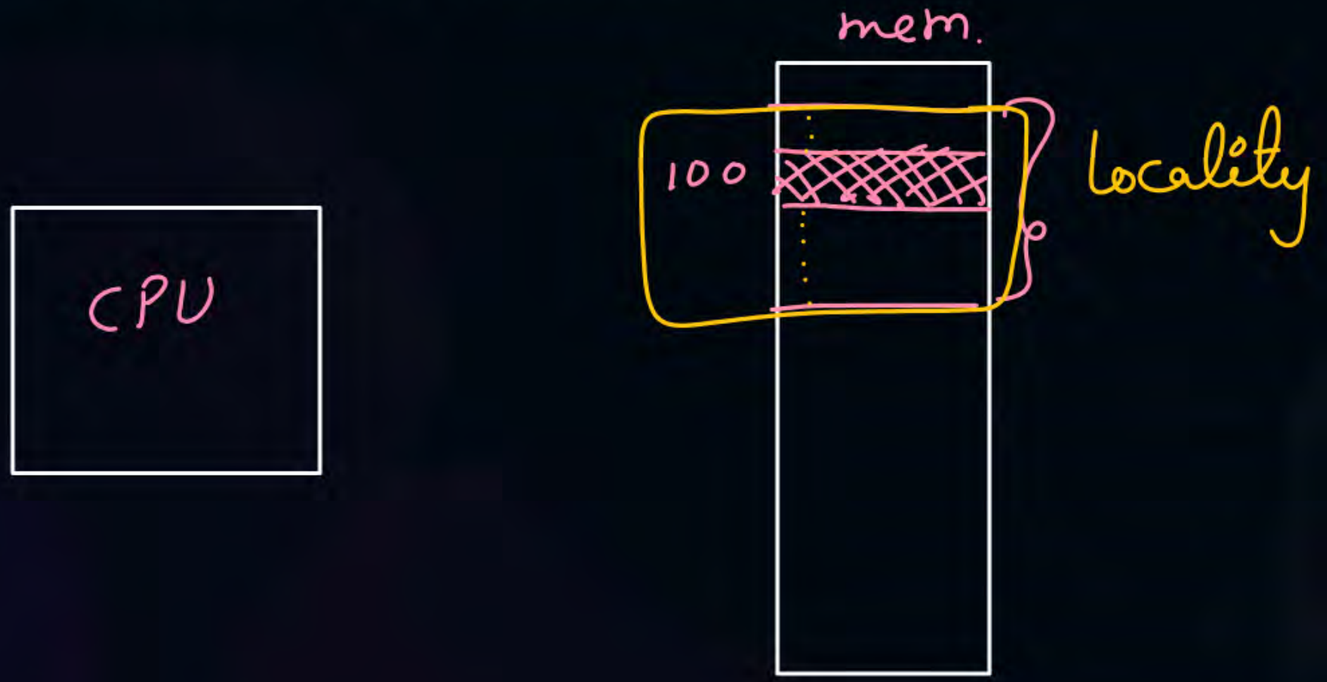
$2^{28} = 8$

$\Downarrow$

all horizontal

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.

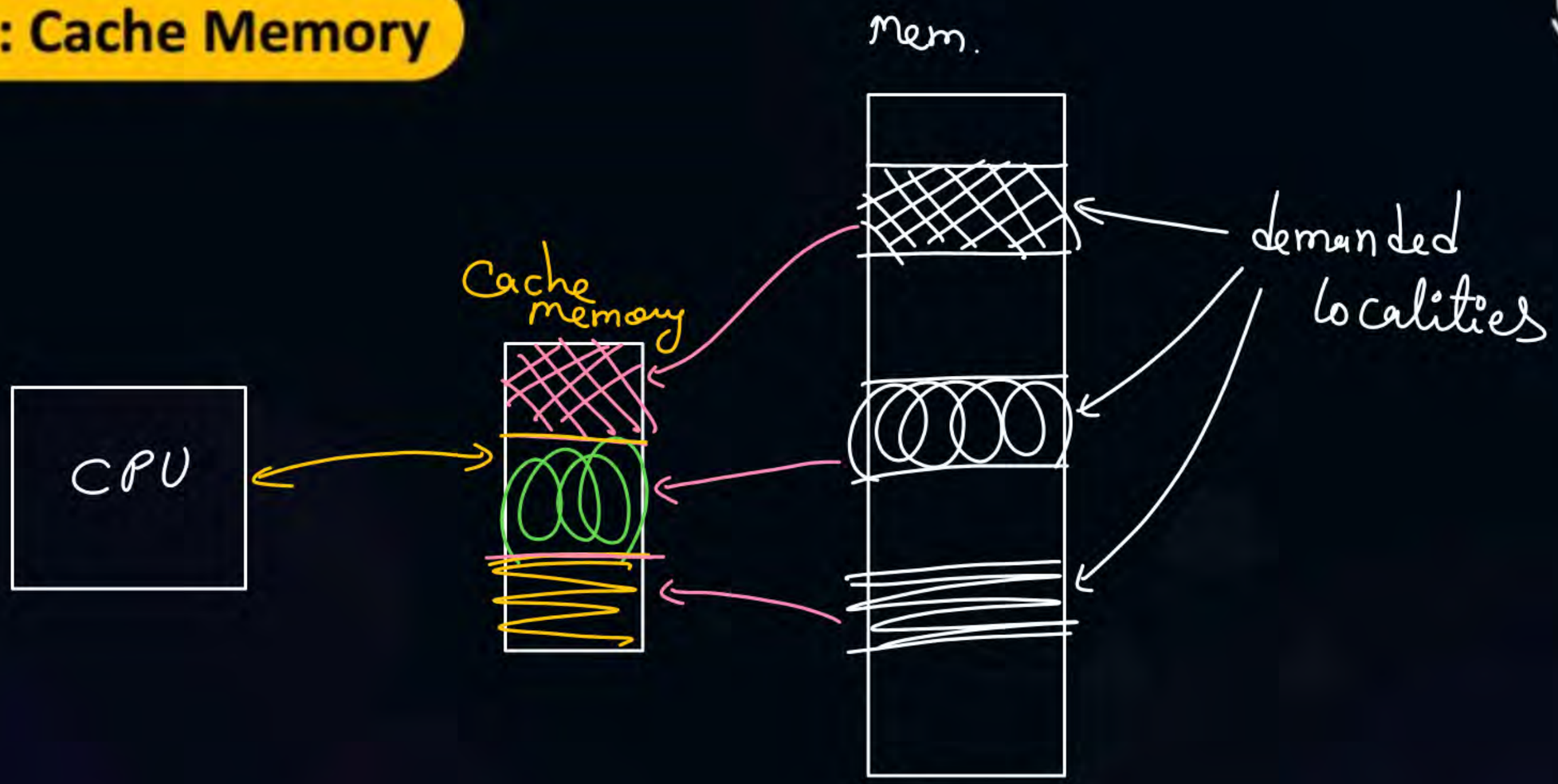


Topic : Locality of Reference

Types:

1. Spatial (space) ⇒ if CPU refers nearby addresses
2. Temporal (time) ⇒ if CPU refers same add. again

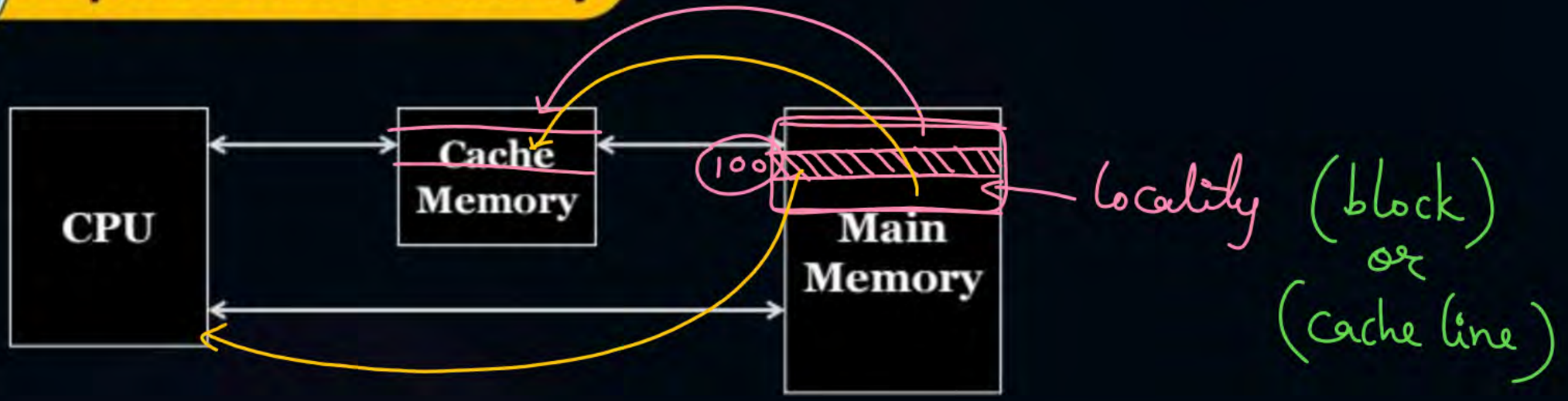
Topic : Cache Memory



Based on locality of reference, CPU's current demanded localities are kept into a smaller and faster memory called as cache.

$\Rightarrow$ Use of cache reduces CPU's mem. access time.

Topic : Cache Memory



Topic : Working of Cache Memory

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

$$H \text{ or } h = \frac{\text{no. of hits in cache}}{\text{total mem. references}}$$

$$\text{miss ratio} = (1 - h)$$

when there is a miss for read in cache then the block (which contains the missed content) is copied from mm to cache.

Topic : Average Memory Access Time

$$T_{avg} = H * \text{mem access time for hit} + (1-H) * \text{mem. access time for miss} \quad \text{.....} (1)$$

---

ex:- CPU refers mem => 100 times
hits => 80 times
miss => 20 times

CPU's mem access time when hit = 10 ns
—— " —— miss = 120 ns

$$H = \frac{80}{100} = 0.8 \text{ or } 80\%$$

$$T_{avg} = 0.8 * 10 + 0.2 * 120 \text{ ns}$$
$$= 32 \text{ ns}$$

Total mem access time for all hits = 80 * 10ns = 800 ns

" " " miss = 20 * 120ns = 2400 ns

Total mem. access time = 3200ns

avg mem access time = $\frac{3200\text{ns}}{100}$

= 32.

$\frac{800 + 2400}{100}$

$= \frac{80 * 10 + 20 * 120}{100} = \frac{80 * 10}{100} + \frac{20 * 120}{100}$

Topic : Types of Cache Accesses

Simultaneous Access: (Parallel)

Request for cache and main-memory are generated simultaneously

t_{cm} = cache mem. access time

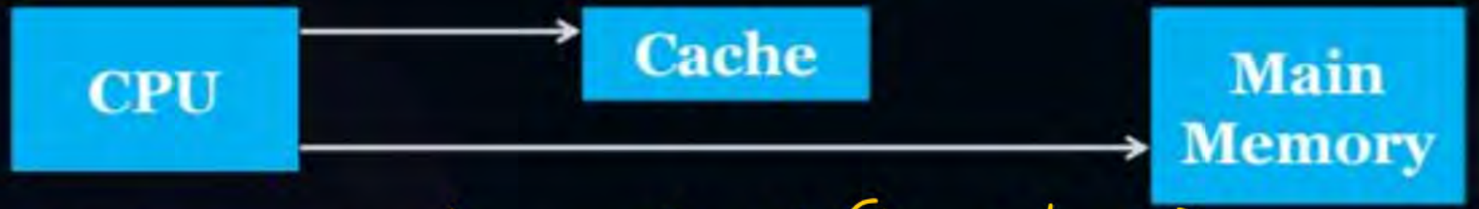
t_{mm} = m.m. —11—

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

Topic : Types of Cache Accesses

Hierarchical Access: (serial)

Only cache is accessed first



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

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

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

Topic : When to Use Which Formula

t_{cm}, t_{mm} given

No → Use general formula eqⁿ (1)

Yes → Level or hierarchy given

No (default for PSU) → Use simultaneous

⇓

Cache search time is negligible

Cache lookup —11—

Yes → Use hierarchical (default for GATE)

2 mins Summary

Topic Locality of Reference

Topic Cache Memory

Happy Learning

THANK - YOU