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

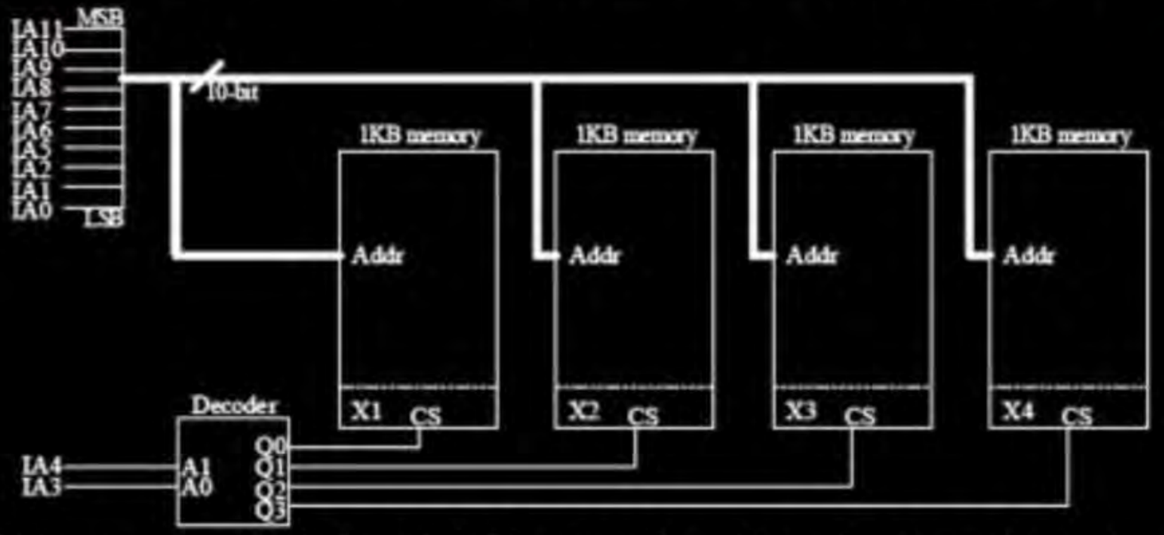
Topics to be Covered

- Topic: Locality of Reference
- Topic: Cache Memory
- Topic: Working of Cache

[MCQ]

#Q. A 4 kilobyte (KB) byte-addressable memory is realized using four 1 KB memory blocks. Two input address lines (IA4 and IA3) are connected to the chip select (CS) port of these memory blocks through a decoder as shown in the figure. The remaining ten input address lines from IA11-IA0 are connected to the address port of these blocks. The chip select (CS) is active high.

Vertical arrangement



IA_4	IA_3	
0	0	X1
0	1	X2
1	0	X3
1	1	X4

The input memory addresses (IA11-IA0), in decimal, for the starting locations (Addr=0) of each block (indicated as XI, X2, X3, X4 in the figure) are among the options given below. Which one of the following options is CORRECT?

A (0, 1,2,3)

B (0, 1024, 2048, 3072)

C (0, 8, 16, 24)

D (0, 0, 0, 0)

	IA_{11}	IA_{10}	-----	IA_4	IA_3	IA_2	IA_1	IA_0	
$X_1 \Rightarrow$	0	0	----	0	0	0	0	0	0
$X_2 \Rightarrow$	0	0	----	0	1	0	0	0	8
$X_3 \Rightarrow$	0	0	----	0	1	0	0	0	16
$X_4 \Rightarrow$	0	0	----	0	1	1	0	0	24

[NAT]

#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 _____?

Ans = 59 to 60

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

$2^{10} = 1K$

$= 16 * 50\ \mu sec$

$= 800\ \mu sec$

$= 0.8\ msec$

% of time for R/W = $\frac{2 - 0.8}{2} * 100\%$

$= 60\%$

$2^{10} = 1024$

$2^4 * 1024 * 50\,ns$

$= 819200\,ns$

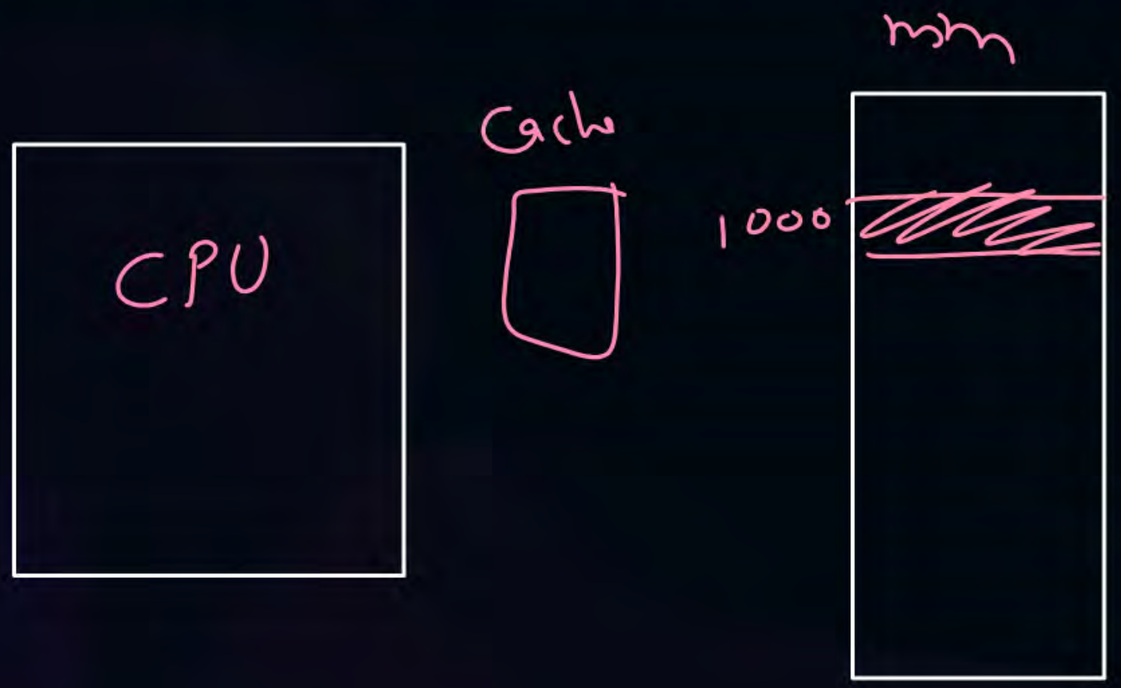
$= 0.81920\ ms$

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

$\cong 59\%$

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.

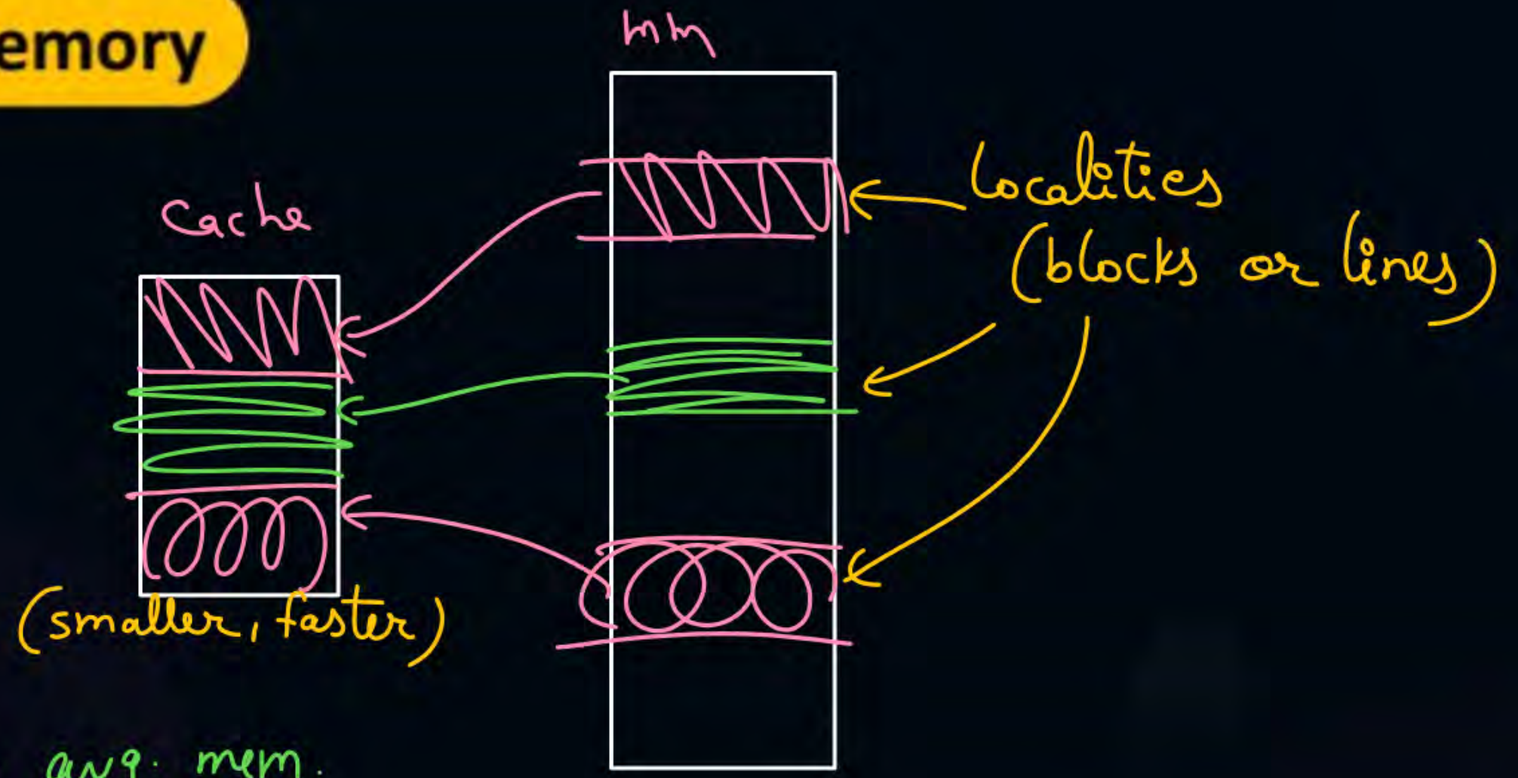


Topic : Locality of Reference

Types:

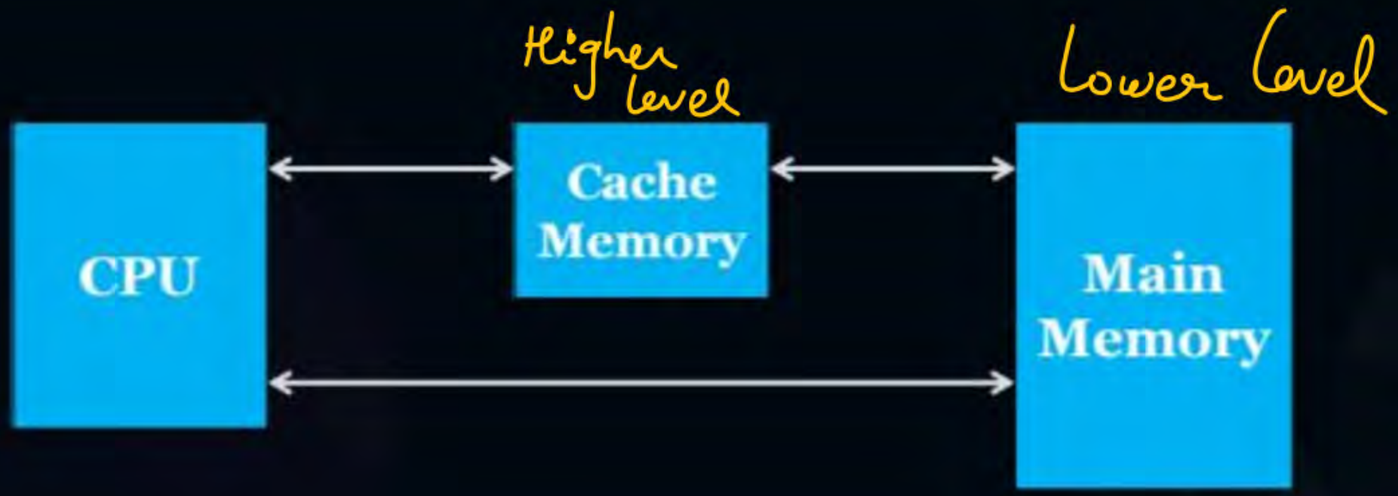
1. Spatial (according to space)
2. Temporal (according to time)

Topic : Cache Memory



- use of cache reduces avg. 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, H) $= \frac{\text{no. of hits}}{\text{total mem. references}}$

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

Miss penalty :-

CPU's time spent during cache miss.

- Reading content from mm
- Bringing missed block from mm to cache.

Topic : Average Memory Access Time

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

Topic : Types of Cache Accesses

Simultaneous Access: (Parallel access)

Request for cache and main-memory are generated simultaneously

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

t_{cm} = cm access time

t_{mm} = mm —||—

Topic : Types of Cache Accesses

Hierarchical Access: (serial access)

Only cache is accessed first

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

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

$$= t_{cm} + (1-H)\, t_{mm} \quad \text{.....(3)}$$

Topic : When to Use Which Formula

if t_{cm}, t_{mm} given

No

which means hit time & given miss time

⇓

Use general formula eq^{n} (1)

yes

if words 'level' or 'hierarchy' given

default for PSUs

↓

No

Use simultaneous access

(Cache search time or lookup time is negligible)

eq (2)

yes

Hierarchical access eq (3)

default for GATE, Test series

[NAT]

#Q. Assume that for a certain processor, a read request takes 200 nanoseconds on a cache miss and 25 nanoseconds on a cache hit. Suppose while running a program, it was observed that 60% of the processor's read requests result in a cache hit. The average read access time in nanoseconds is_____?

(General formula)

$= 0.6 * 25 + 0.4 * 200$

$= 15 + 80$

$= 95$ ns

[NAT]

#Q. In a two-level hierarchy, if the top level has an access time of 15 ns and the bottom level has an access time of 80 ns, what is the hit rate on the top level required to give an average access time of 20ns?

$t_{cm} = 15\,ns$

$t_{mm} = 80\,ns$

$t_{avg} = 20\,ns$

Hierarchical

$$20 = 15 + (1-H)\,80$$

$$5 = 80 - 80H$$

$$H = \frac{75}{80}$$

$$= 0.9375 \quad \text{Ans.}$$

[NAT]

#Q. In previous question if hit rate of the top-level memory is 100%, then the access time of bottom level memory will be 80 ns?

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

$$= t_{cm} + 0 * (t_{cm} + t_{mm})$$

$$= t_{cm}$$

[NAT]

#Q. MM access time = 20ns

Cache access time = 5ns

Tavg = 4ns

H = ?

Hierarchical access

Invalid Quest^n

because $t_{avg} < t_{cm}$

↓

not possible

[NAT]

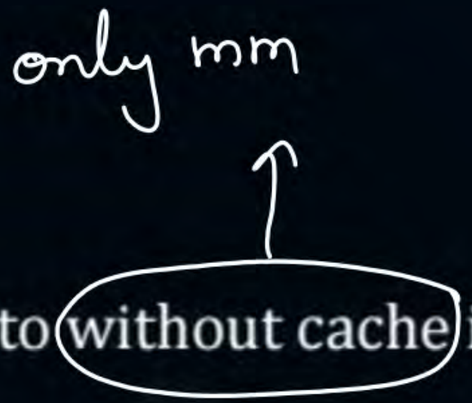
#Q. MM access time = 200ns

Cache access time = 15ns

H = 90%

Hierarchical access

Speed up of memory system with cache as compared to without cache is?



$t_{avg} = 15 + 0.1 * 200$

$= 35\ ns$

$$\text{speed up} = \frac{200\ ns}{35\ ns}$$

$$= \underline{5.71}\ \text{Ans.}$$

[NAT]

no cache => only mm

#Q. A computer system contains a cache. Uncached memory access takes 7 times longer than access to cache. If cache has a hit ratio 0.9. The ratio of cached memory access time to uncached memory access time is?

with cache

$t_{cm} = x$

$t_{mm} = 7x$

$$= \frac{t_{avg}}{t_{mm}}$$

$$= \frac{0.9 * x + 0.1 * 7x}{7x}$$

$$= \frac{1.6}{7} = 0.22857$$

[NAT]

#Q. Consider a system with a CPU which runs on 2MHz clock rate. The system contains cache in its memory hierarchy. The cache is hierarchically accessed and take 5 cycles for an access. The main memory access takes 75 cycles. The hit rate of cache needed to get a break even for using the cache is _____%?

$T_{avg} = t_{mm}$

$$75 = 5 + (1-H)\,75$$

$$70 = 75 - 75H$$

$$H = \frac{5}{75}$$

$$= 0.067 \text{ or } 6.67\%$$

Topic : T_{avg} When Locality of Reference is Used

replace t_{mm} by t_{bt}

t_{bt} = block transfer time from mm to cache

Sim :- $t_{avg} = H * t_{cm} + (1-H)\, t_{bt}$

Hier :- $t_{avg} = t_{cm} + (1-H)\, t_{bt}$

[NAT]

#Q. In a two-level hierarchy, the top level has an access time of 10 ns and hit rate of 90%. If the block transfer from main memory to cache takes 500ns in case of miss then average memory access time is 60?

$= 10 + 0.1 * 500$

$= 60$ ns

[NAT]

#Q. Byte transfer time from mm to cache = 50ns

block size = 16 bytes

Block transfer time from mm to cache = 800 ns?

16 * 50ns = 800ns

[NAT]

#Q. Byte transfer time from mm to cache = 50ns

block size = 16 words $= 16 * 4B = 64\ B$

1 Word = 4 bytes

Block transfer time from mm to cache = 3200 ns?

$t_{bt} = 64 * 50\ ns$

$= 3200\ ns$

[NAT]

#Q. Word transfer time from mm to cache = 50ns

block size = 16 words

Block transfer time from mm to cache = 800 ns?

$t_{bt} = 16 * 50 = 800ns$

[NAT]

#Q. Word transfer time from mm to cache = 50ns

block size = 64 bytes $= \frac{64B}{4B} = 16$ words

1 Word = 4 bytes

Block transfer time from mm to cache = 800 ns?

$t_{bt} = 16 * 50$

$= 800ns$

[NAT]

#Q. First byte transfer time from mm to cache = 10ns

Remaining each byte transfer time from mm to cache = 2ns

block size = 32 bytes

Block transfer time from mm to cache = 72 ns?

$$t_{bt} = 10ns + (31 * 2ns)$$
$$= 72ns$$

[NAT]

#Q. First word transfer time from mm to cache = 10ns

Remaining each word transfer time from mm to cache = 2ns

block size = 128 bytes = $\frac{128\ B}{2B}$ = 64 words

Word size = 2 bytes

Block transfer time from mm to cache = 136 ns?

$t_{bt} = 10ns + (63 * 2ns)$

$= 136\ ns$

[NAT] GATE-20 Ans :- 13.3 to 13.5

#Q. A direct mapped cache memory of 1 MB has a block size of 256 bytes. The cache has an access time of 3 ns and a hit rate of 94%. During a cache miss, it takes 20ns to bring the first word of a block from the main memory, while each subsequent word takes 5 ns. The word size is 64 bits. The average memory access time in ns (round off to 1 decimal place) is _____ ?

8B

$t_{bt} = 20 + (31 * 5)\,ns = 175\,ns$

block size $= \frac{256\,B}{8\,B} = 32$ words

sim.

$t_{avg} = 0.94 * 3 + 0.06 * 175\,ns = 13.3\,ns$

Hier.

$= 3 + 0.06 * 175\,ns = 13.5\,ns$

2 mins Summary

Topic Locality of Reference

Topic Cache Memory

Topic Working of Cache

Happy Learning

THANK - YOU