CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Cache Organization

Lecture No.- 02

By- Vishvadeep Gothi sir

Recap of Previous Lecture

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

Topics to be Covered

Topic Cache Write

Topic Write Through & Write Back

Topic Write Allocate & No Write Allocate

[NAT] GATE-2022

#Q. A cache memory that has a hit rate of 0.8 has an access latency 10 ns and miss penalty 100 ns. Optimization is done on the cache to reduce the miss rate. However, the optimization results in an increase of cache access latency to 15 ns, whereas the miss penalty is not affected. The minimum hit rate (rounded off to two decimal places) needed after the optimization such that it should not increase the average memory access time is 0.85 ?

Sim.

old.

$$t_{avg} = 0.8 * 10 + 0.2 * 100$$
$$= 28 \text{ ns}$$

new:-

$$28 = H * 15 + (1-H) 100$$
$$28 = 15H + 100 - 100H$$
$$85H = 72$$
$$H = \frac{72}{85} = 0.8470$$
$$= 0.85$$

Hier.

old

$$t_{avg} = 10 + 0.2 * 100$$
$$= 30 \text{ ns}$$

new

$$30 = 15 + (1-H) 100$$
$$15 = 100 - 100H$$
$$H = 0.85$$

[NAT] GATE-2016

#Q. A file system uses an in-memory cache to cache disk blocks. The miss rate of the cache is shown in the figure. The latency to read a block from the cache is 1 ms and to read a block from the disk is 10 ms. Assume that the cost of checking whether a block exists in the cache is negligible. Available cache sizes are in multiples of 10 MB.

simultaneous

The smallest cache size required to ensure an average read latency of less than 6 ms is 30 MB?

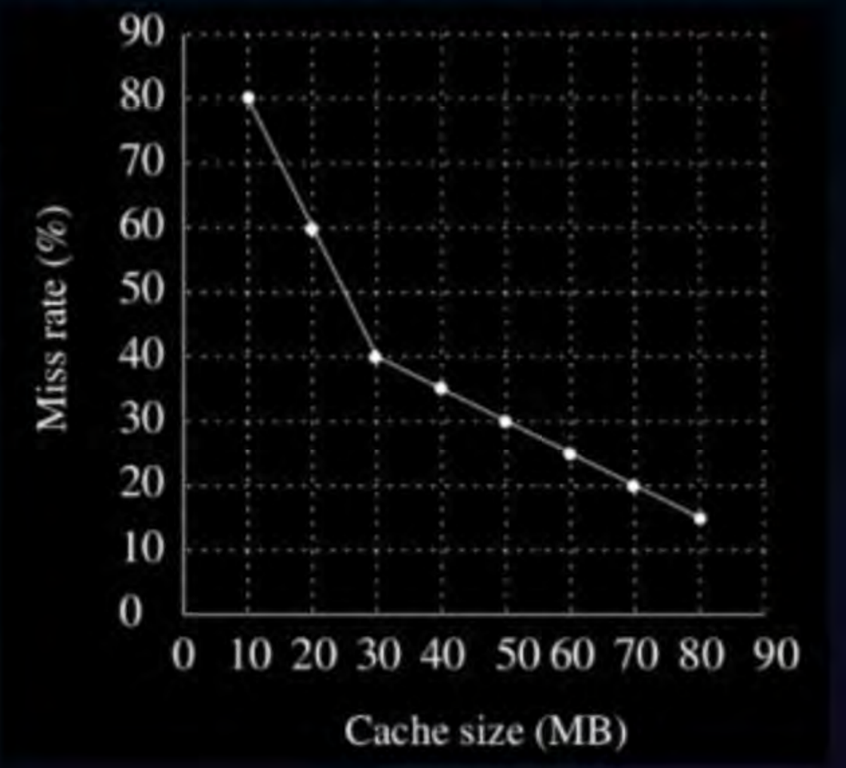
$6 \geq H * 1 + (1-H)\ 10$

$6 \geq H + 10 - 10H$

$H \geq \frac{4}{9}$

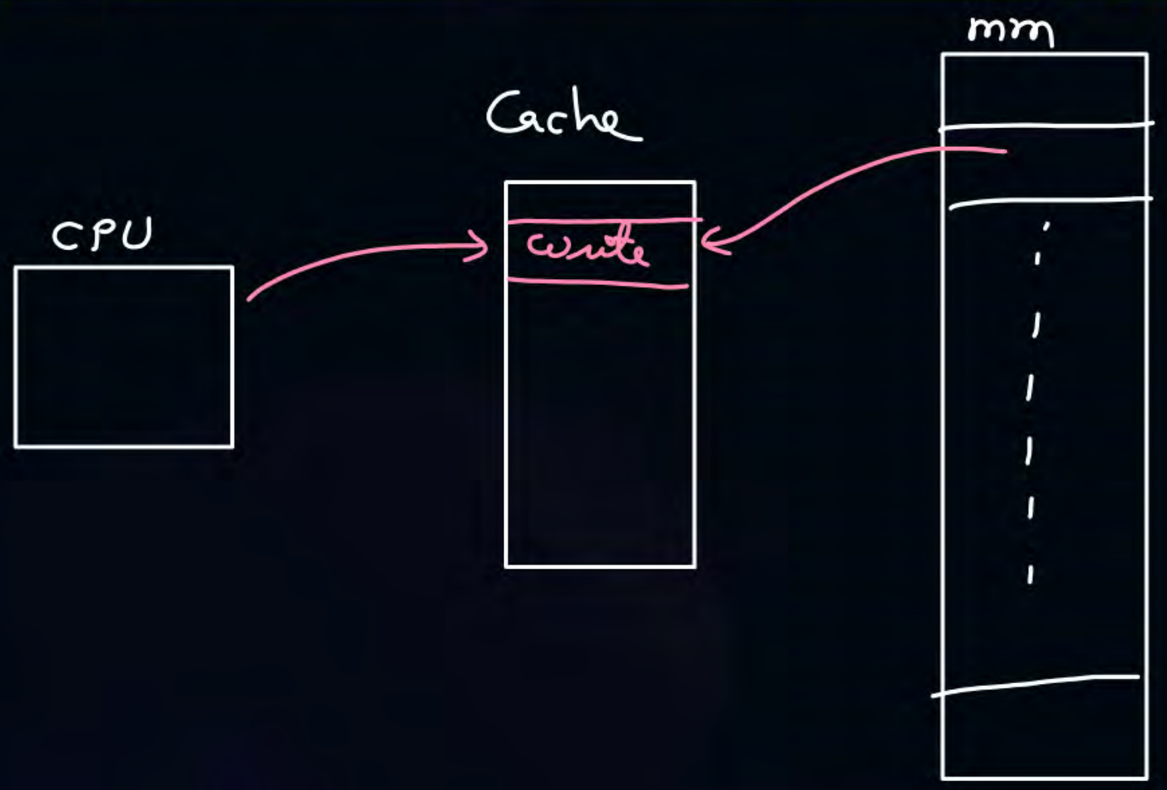
$H \geq 0.44$

$(1-H) \leq 0.56$



$$H \propto \frac{1}{T_{avg}}$$

Topic : Cache Write or Write Propagation

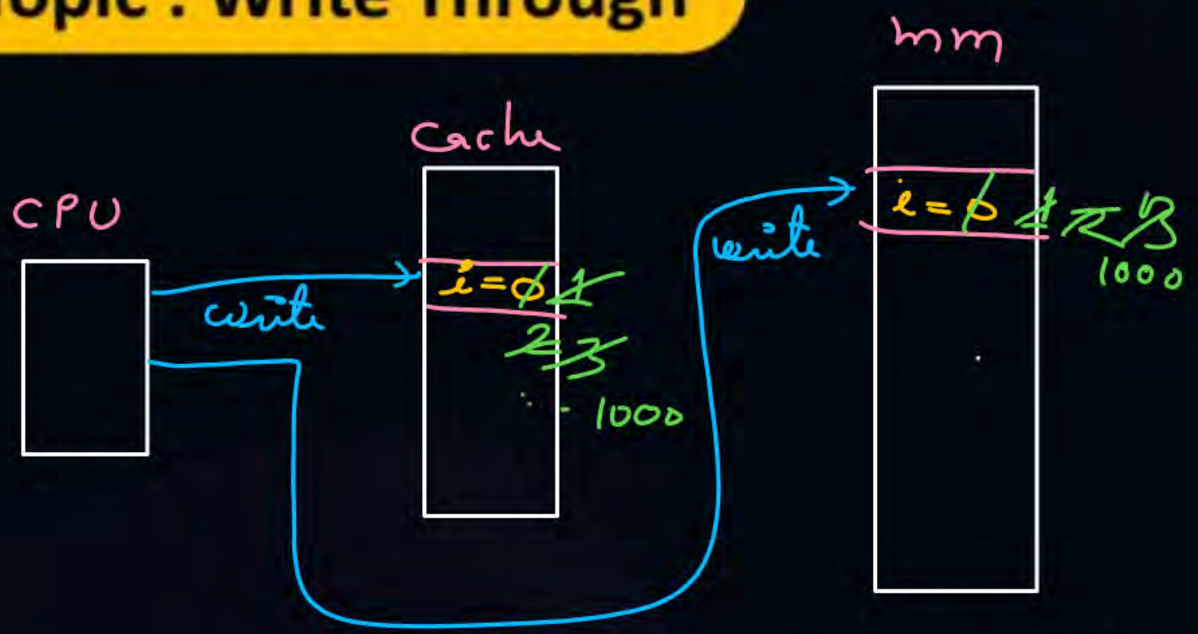


Topic : Cache Write or Write Propagation

1. Write Through
2. Write Back

Topic : Write Through

CPU performs write in cach and mm both simultaneously.



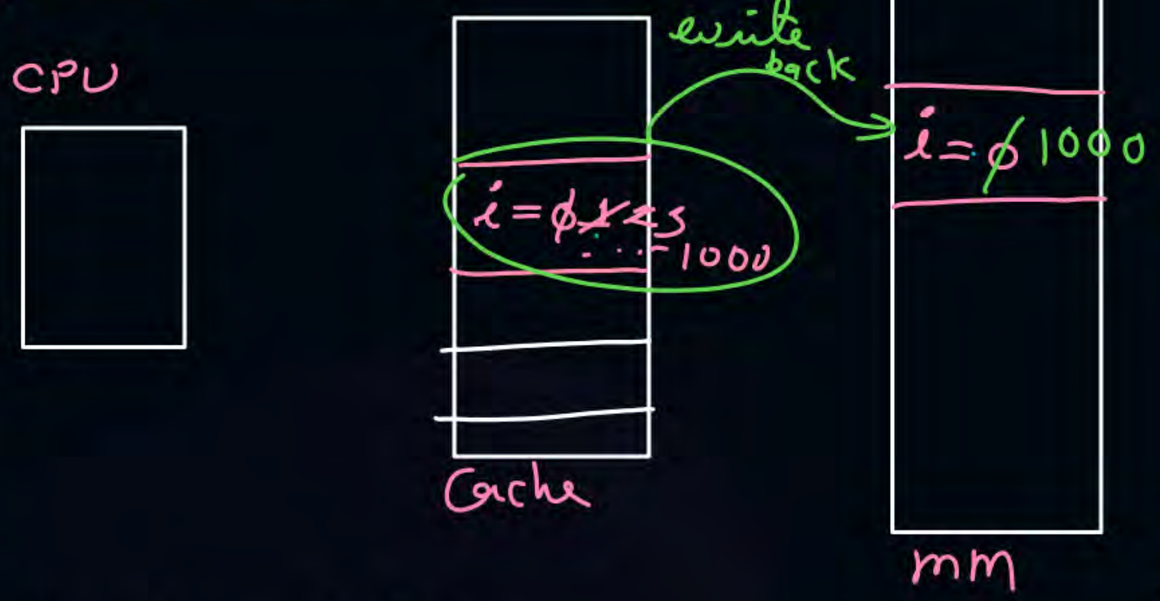

```
for (i=0; i<1000; i++)
{
}
```

while replacement, a block is replaced always without writing it back to mm

☺ ⇒ consistency in cache & mm content

☹ ⇒ Time consuming (irrespective of hit or miss in cache, CPU accesses mm for write operation)

Topic : Write Back



CPU writes only in cache and when a block is replaced from cache then write it back to mm.

↓

write back of only modified (dirty) blocks are done.

Non-modified blocks are directly replaced.

→ To save time.

☺ Time Saving

☹ Inconsistency

Topic : T_{avg} in Write Through Cache

default

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

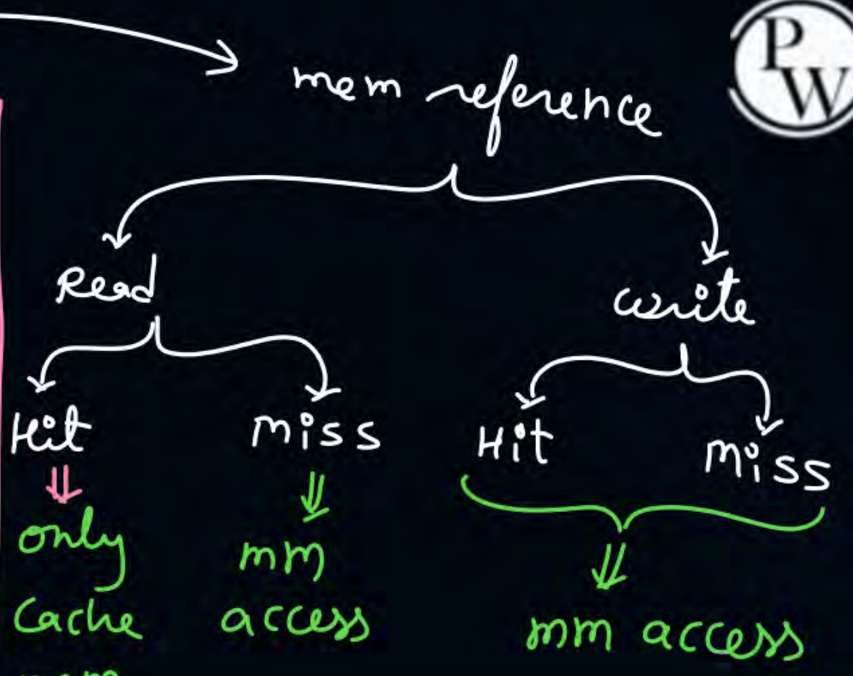
or

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

$$T_{avg\ write} = \max(t_{cm}, t_{mm}) = t_{mm}$$

$$T_{avg} = \%\ of\ read * t_{avg\ read} + \%\ of\ write * t_{avg\ write}$$

Effective hit rate = % of read * hit ratio



[MCQ]

#Q. A system has a write through cache with access time of 100ns and hit ratio of 90%. The main memory access time is 1000ns. The 70% of memory references are for read operations.

1. Average memory access time for read operations only $= 0.9 * 100 + 0.1 * 1000 = 190\ ns$
2. Average memory access time for write operations only $\Rightarrow 1000\ ns$
3. Average memory access time for read-write operations both $= 0.7 * 190 + 0.3 * 1000 = 433\ ns$
4. Effective Hit ratio $= 0.7 * 0.9 = 0.63$

Topic : T_{avg} in Write Back Cache

→ assume fraction dirty replaced blocks ⇒ d

Sim.

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

Hier.

$$= H * t_{cm} + (1-H) \left[t_{cm} + t_{bt} + d * t_{bt} \right]$$

Topic : Write Allocate vs No Write Allocate

Write Allocate: → used with write back cache

The block is loaded from main memory to cache, on a write miss

No Write Allocate: → used with write through

The block is modified in the main memory and not loaded into the cache

write through with no write allocate

Read

Hit

CPU reads required content from cache

Miss

CPU reads required content from mm. Then CPU brings missed block from mm to cache.

If any block is replaced from cache then replace it directly. (no write back needed)

write

Hit

CPU performs write in cache and main memory simultaneously

Miss

CPU performs write in mm and does not bring missed block into cache

write back with write allocate

Read

Hit

CPU reads required content from cm

miss

CPU reads required content from mm. Then CPU brings missed block from mm to cache. If any block is replaced from cache then write it back to mm if it is dirty

write

Hit

CPU performs write in cache. and marks this block as dirty.

miss

CPU brings missed block from mm to cache then performs write in cache. If any block is replaced from cache then write back it to mm only if it is dirty.

[NAT] GATE – 2014

Ans = 1.68

#Q. The memory access time is 1 nanosecond for a read operation with a hit in cache, 5 nanoseconds for a read operation with a miss in cache, 2 nanoseconds for a write operation with a hit in cache and 10 nanoseconds for a write operation with a miss in cache. Execution of a sequence of instructions involves 100 instruction fetch operations; 60 memory operand read operations and 40 memory operand write operations. The cache hit-ratio is 0.9. The average memory access time (in nanoseconds) in executing the sequence of instructions is?

	Read	Write
hit	1 ns	2 ns
miss	5 ns	10 ns

$T_{avg\ read} = 0.9 * 1 + 0.1 * 5ns = 1.4ns$

$T_{avg\ write} = 0.9 * 2 + 0.1 * 10 = 2.8\ ns$

Total read = 100 + 60 = 160

Total write = 40

% of read = $\frac{160}{160+40}$ = 0.8 or 80%

% of write = $\frac{40}{160+40}$ = 0.2 or 20%

$T_{avg} = 0.8 * 1.4 + 0.2 * 2.8$

$= 1.68$ ns

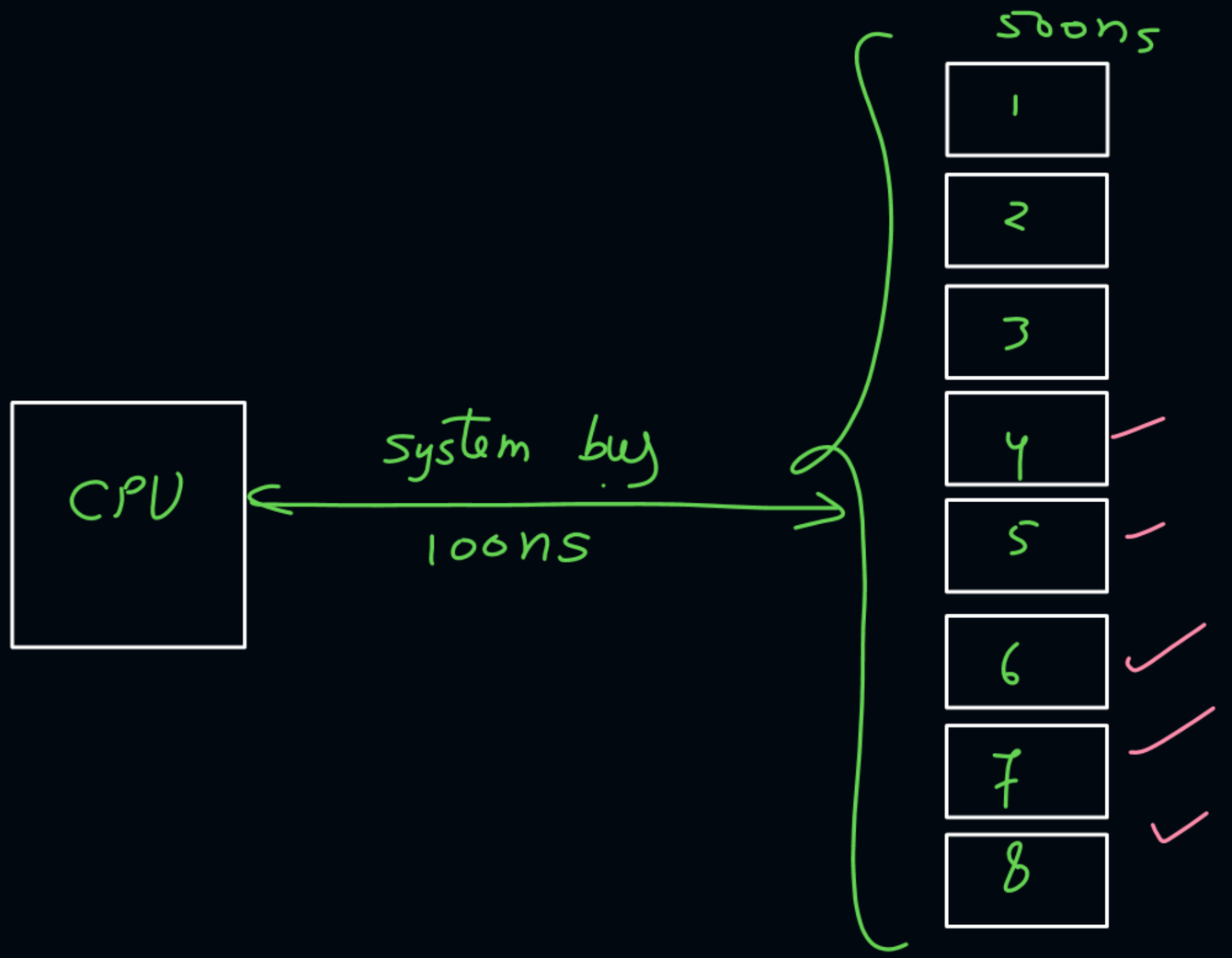
[NAT] GATE-14

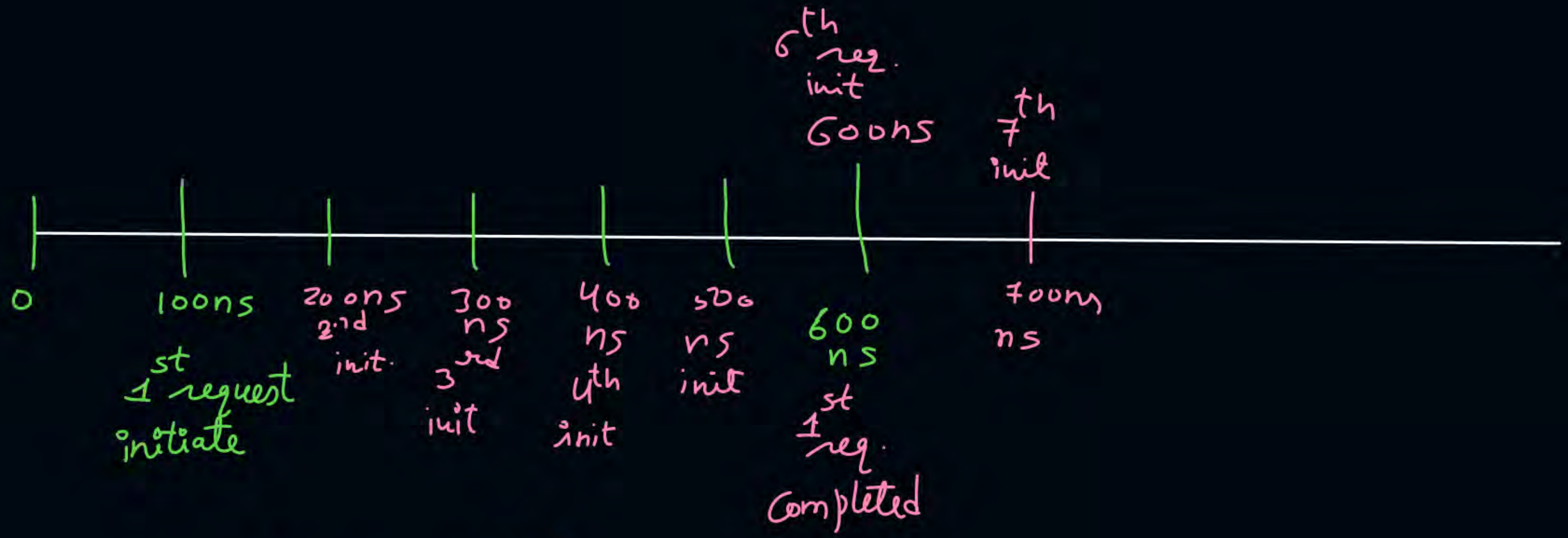
#Q. Consider a main memory system that consists of 8 memory modules attached to the system bus, which is one word wide. When a write request is made, the bus is occupied for 100 nanoseconds (ns) by the data, address, and control signals. During the same 100 ns, and for 500 ns thereafter, the addressed memory module executes one cycle accepting and storing the data. The (internal) operation of different memory modules may overlap in time, but only one request can be on the bus at any time. The maximum number of stores (of one word each) that can be initiated in 1 millisecond is 10000 ?

$$= \frac{1ms}{100ns}$$

$$= \frac{10^{-3}}{100 * 10^{-9}}$$

$$= 10000$$


500ns
1
2
3
4
5
6
7
8
CPU
system bus
100ns



CPU can initiate
one write request per 100 ns.

H.W.

#Q. Size of data sent to main memory from CPU:

1. For write hit, when a write through cache is used?
2. For write miss, when a write through cache is used?
3. For write hit, when a write back cache is used?
4. For write miss, when a write back cache is used?

H.W.

#Q. Size of data sent from main memory to cache:

1. For write hit, when a write through cache is used?
2. For write miss, when a write through cache is used?
3. For write hit, when a write back cache is used?
4. For write miss, when a write back cache is used?

[Question]

H.W.

#Q. Consider a computer with the following features:

- 90% of all memory accesses are found in the cache (hit ratio = 0.9)
- The block size is 2 words and the whole block is read on any miss
- The CPU sends references to the cache at the rate of 10^7 words per second
- 25% of the above references are writes (writes = 25%, reads = 75%)
- The bus can support 10^7 words per second, read or writes (total bus bandwidth =10^7)
- The bus reads or writes a single word at a time
- Assume at any one time, 30% of the block frames in the cache have been modified

Calculate the percentage of the bus bandwidth used on the average when:

1. Cache is write through with no write allocate
2. Cache is write back with write allocate

2 mins Summary

Topic Cache Write

Topic Write Through & Write Back

Topic Write Allocate & No Write Allocate

Happy Learning

THANK - YOU