CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Cache Organization

Lecture No.- 11

By- Vishvadeep Gothi sir

Recap of Previous Lecture

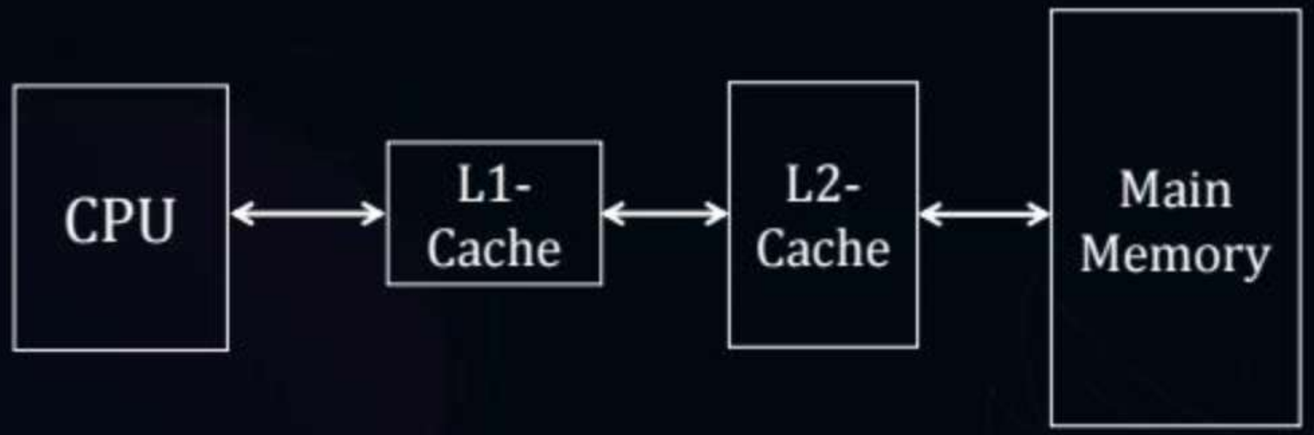
Topic Array Access with Cache

Topic Multilevel Cache

Topics to be Covered

- Topic: Multilevel Cache
- Topic: Cache Inclusion Policy
- Topic: Magnetic Disk

Topic : Multilevel Cache

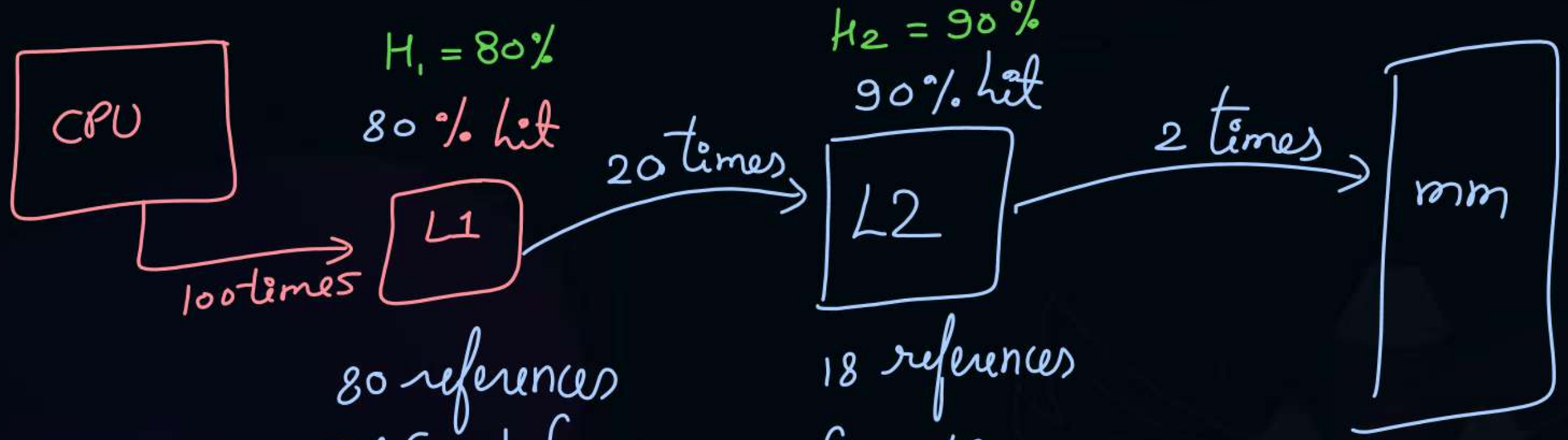


[NAT]

#Q. Consider a 3-level memory hierarchy with L1 cache, L2 cache and a main memory. The hit ratios of L1 is 90% and of L2 is 95%. The access times of L1, L2 and main memory are 15s, 60ns and 350ns respectively. The average memory access time is _____ns?

Topic : Probability of Access

assumption
Total 100 times CPU refers mem.



80 references fulfilled from L1

18 references from L2

80% references fulfilled from L1
H_1

18% references
$(1-H_1) * H_2$

2% references
$(1-H_1) * (1-H_2)$

	Probability of access			
L1	0.8	(80%) $\Leftarrow$	H_1	$\Leftarrow P_1$
L2	0.18	(18%) $\Leftarrow$	$(1-H_1)H_2$	$\Leftarrow P_2$
MM	0.02	(2%) $\Leftarrow$	$(1-H_1)(1-H_2)$	$\Leftarrow P_m$

Hierarchical access:-

$$T_{avg} = H_1 * t_1 + (1-H_1)\left[H_2 * (t_1+t_2) + (1-H_2)(t_1+t_2+t_{mm})\right]$$

$$= H_1 t_1 + (1-H_1) H_2 (t_1+t_2) + (1-H_1)(1-H_2)(t_1+t_2+t_{mm})$$

- $H_1 t_1$ → Hit in L1
- $(1-H_1) H_2 (t_1+t_2)$ → miss in L1 but hit in L2
- $(1-H_1)(1-H_2)(t_1+t_2+t_{mm})$ → Miss in L1 & L2

or

$$= P_1 t_1 + P_2 (t_1+t_2) + P_m (t_1+t_2+t_{mm})$$

Simultaneous access :-

$$T_{avg} = P_1 * t_1 + P_2 * t_2 + P_m * t_{mm}$$

[NAT]

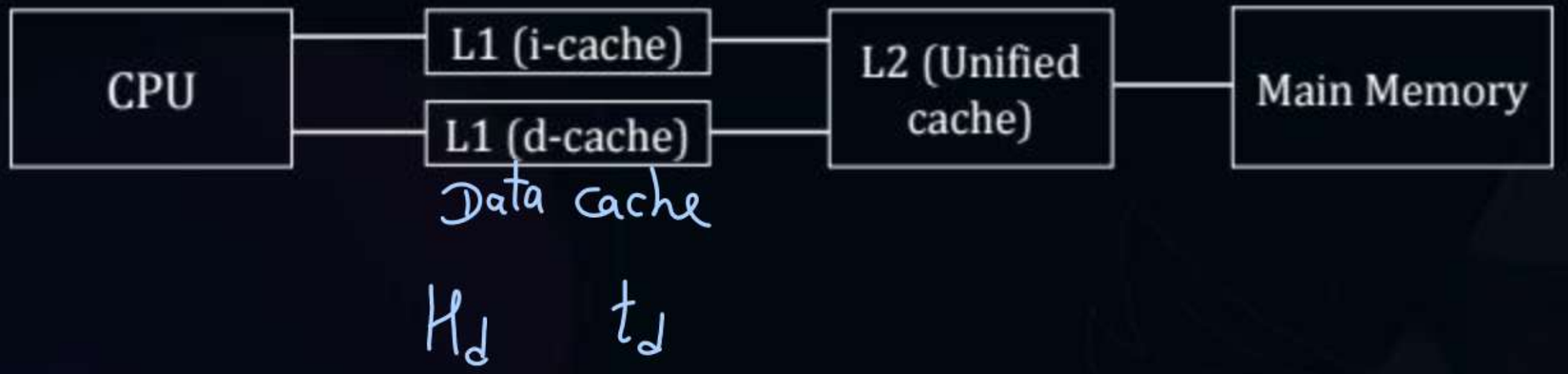
$(1-H_1)(1-H_2)$ H_1 $(1-H_1)H_2$

#Q. Consider a 3-level memory hierarchy with L1 cache, L2 cache and a main memory. The probability of access of L1 is 95%, of L2 is 4.5% and of main memory is 0.5%. The access times of L1, L2 and main memory are 10ns, 50ns and 400ns respectively. The average memory access time is _____ns?

$$T_{avg} = 0.95 * 10 + 0.045 * (10+50) + 0.005 * (10+50+400)$$

$$= \underline{\underline{14.5}} \text{ ns} \quad \text{Ans.}$$

Topic : Dual Cache



H_i t_i

inst^n cache

Data cache

H_d t_d

$$T_{avg\ inst^n} = H_i * t_i + (1-H_i)\left[H_2 * (t_i + t_2) + (1-H_2)(t_i + t_2 + t_{mm})\right]$$

or

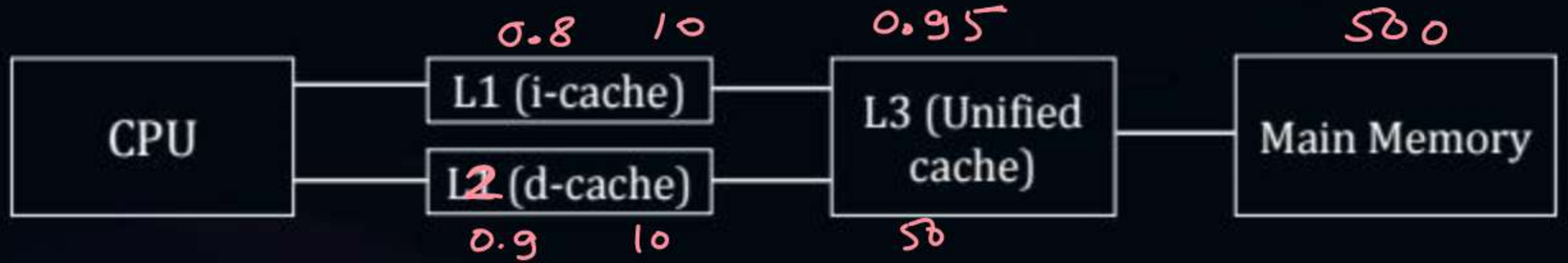
$$= t_i + (1-H_i)\left[t_2 + (1-H_2)\, t_{mm}\right]$$

$$T_{avg\ data} = t_d + (1-H_d)\left[t_2 + (1-H_2)\, t_{mm}\right]$$

$$T_{avg} = \%\ of\ inst^n\ reference * t_{avg\ inst^n} + \%\ of\ data\ references * t_{avg\ data}$$

[NAT]

#Q. The multilevel memory hierarchy is given.



The hit ratio of L1, L2, L3 and main memory are 0.8, 0.9, 0.95 and 1.0 respectively. The access times of respective memories are 10ns, 10ns, 50ns and 500ns. Among total memory references 60% of them are for data.

1. Average memory access time for only instructions access ⇒ 25 ns
2. Average memory access time for only data access ⇒ 17.5 ns
3. Average memory access time ⇒ 20.5 ns

$$T_{avg\ inst^n} = 10 + 0.2\left[50 + 0.05 * 500\right]$$

$$= 25\ ns$$

$$T_{avg\ data} = 10 + 0.1\left[50 + 0.05 * 500\right]$$

$$= 17.5\ ns$$

$$T_{avg} = 0.4 * 25 + 0.6 * 17.5$$

$$= 20.5\ ns$$

[NAT]

GATE- PYQ (2017)

Ans = 4.72

#Q. The read access times and the hit ratios for different caches in a memory hierarchy are as given below:

Cache	Read access time (in nanoseconds)	Hit Ratio
I-cache	2	0.8
D=cache	2	0.9
L2-cache	8	0.9

The read access time of main memory in 90nanoseconds. Assume that the caches use the referred-word-first read policy and the write-back policy. Assume that all the caches are direct mapped caches. Assume that the dirty bit is always 0 for all the blocks in the caches. In execution of a program, 60% of memory reads are for instruction fetch and 40% are for memory operand fetch. The average read access time in nanoseconds (up to 2 decimal places) is _____?

$$T_{avg\ inst^n} = 2 + 0.2\left[8 + 0.1 * 90\right] = 5.4\ ns$$

$$T_{avg\ data} = 2 + 0.1\left[8 + 0.1 * 90\right] = 3.7\ ns$$

$$T_{avg} = 0.6 * 5.4 + 0.4 * 3.7$$

$$= \underline{4.72}\ ns \quad \underline{Ans}$$

[NAT]

extra cycle

mem. data access

#Q. Consider a program execution which has 36% instructions for load and store. The CPI without memory stalls is 2. The program experiences 2% miss for instruction cache and 4% miss for data cache. The cache miss penalty is 200 cycles.

1. CPI with memory stalls $\Rightarrow 8.88$
2. Performance gain (speed up) of perfect cache as compared to cache with stalls

$$= \frac{8.88}{2} = 4.44$$

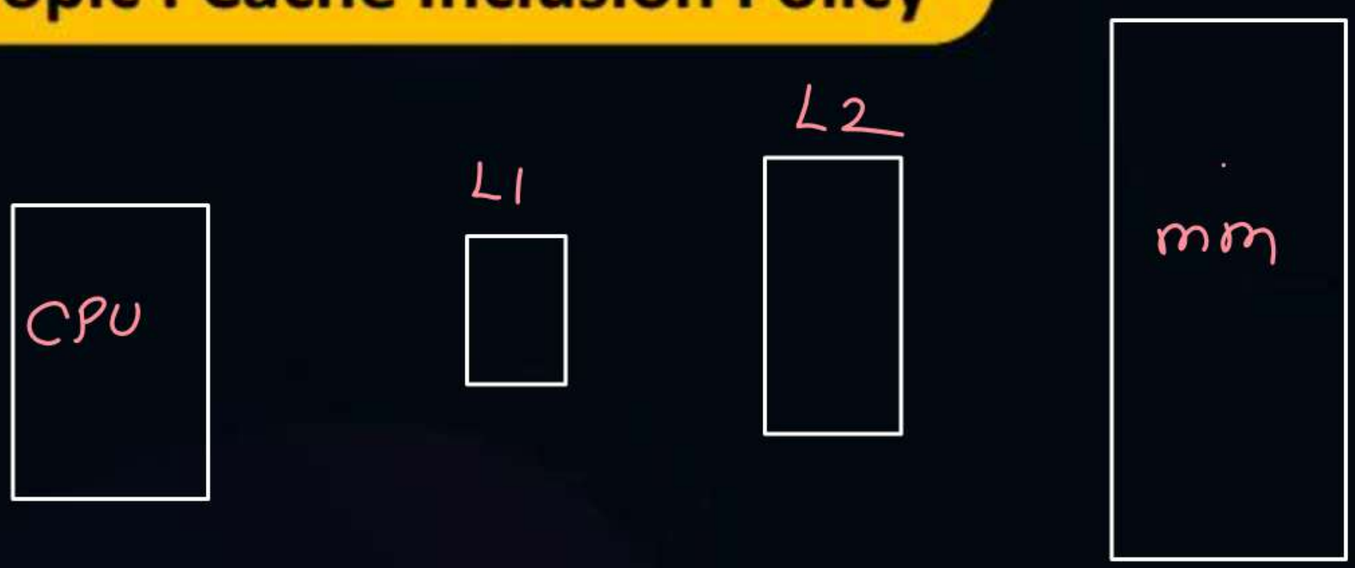
extra cycles for $inst^{ns}$ = $1 * 0.02 * 200 = 4$ cycles

extra cycles for data = $0.36 * 0.04 * 200 = 2.88$ cycles

CPI with cache miss included = $2 + 4 + 2.88$

$= 8.88$ cycles

Topic : Cache Inclusion Policy



Block which are present in L1 are also present in L2 or not?

Value inclusion policy :-

with inclusion of blocks, the value of blocks also must be same.

Topic : Inclusion Policy

all the blocks which are present in L1 must be present in L2 also.

⇓

L2 is all inclusive of L1.

Topic : Inclusion Policy

for read access

1. Hit in L1 ⇒ CPU reads content from L1

2. Miss in L1 & Hit in L2 ⇒ CPU reads content from L2
   Copy missed block from L2 to L1. If any block evicted from L1, then for that there is no role of L2.

3. Miss in L1 & Miss in L2 ⇒ CPU reads content from mm
   Copy missed block from mm to L2; then from L2 to L1.
   If any block evicted from L1 then there is no any role of L2.

Topic : Exclusion Policy

blocks which are present in L1 are not present in L2

L2 stores blocks which are replaced (evicted) from L1; hence L2 is called as "victim cache".

Topic : Exclusion Policy

1. Hit in L1 CPU reads content from L1

2. Miss in L1 & Hit in L2 CPU reads content from L2.
   Move (cut & paste) the missed block from L2 to L1. If any block is evicted from L1, then it is moved to L2.

3. Miss in L1 & Miss in L2 CPU reads content from mm.
   Copy the missed block from mm to L1. If any block is evicted from L1, then it is moved to L2.

2 mins Summary

Topic Array Access with Cache

Topic Multilevel Cache

Topic Magnetic Disk

Happy Learning

THANK - YOU