CS & IT ENGINEERING

COMPUTER ORGANIZATION AND ARCHITECTURE

Cache Organization

Lecture No.- 10

By- Vishvadeep Gothi sir

Recap of Previous Lecture

Topic Mapping Hardware

Topic Array Access With Cache

Topics to be Covered

- **Topic** Multilevel Cache
- **Topic** Dual Cache
- **Topic** Cache Inclusion Policies

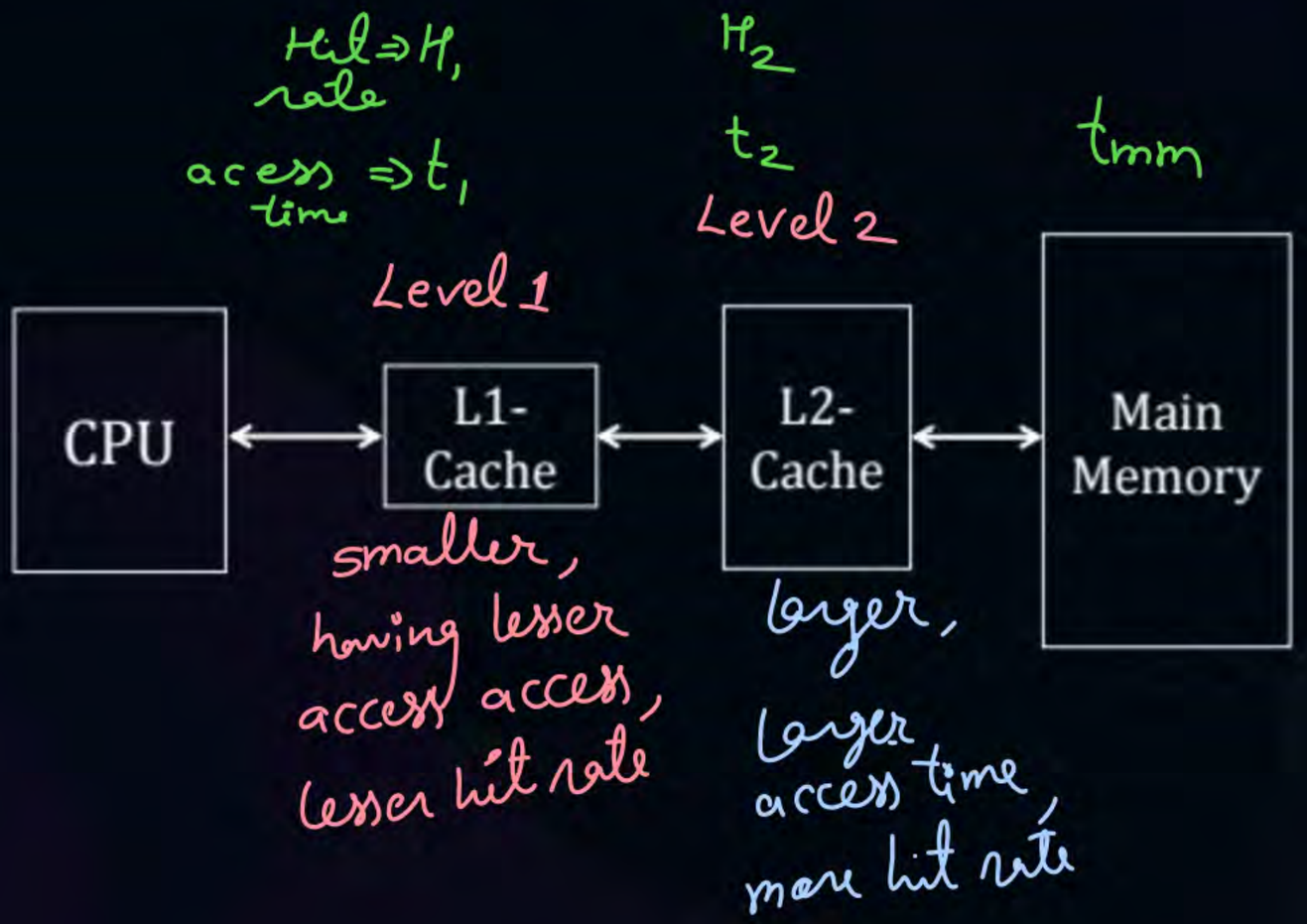
Topic : Goal of Using Cache

1. Minimize Access Time (avg.) ⇒ small cache
2. Maximize Hit Rate ⇒ Large cache

} ⇒ multilevel cache

3. Minimize Miss Penalty

Topic : Multilevel Cache



Topic : Average Access Time Multilevel Cache

Simultaneous :-

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

or

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

- $H_1 t_1$: Prob. of access of L1
- $(1-H_1)H_2\, t_2$: Prob. of access of L2
- $(1-H_1)(1-H_2)\, t_{mm}$: Prob. of access of mm.

Hierarchical :- (default)

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

or

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

or

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

[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?

$$t_{avg} = 15 + 0.1\,(60 + 0.05 * 350)$$

$$= 22.75\ ns$$

Topic : Probability of Access

ex:- $H_1 = 80\%$
$H_2 = 90\%$

	Probability that CPU gets Content	
from L1	H_1	$= 0.8$
from L2	$(1-H_1)H_2$	$= 0.2 * 0.9 = 0.18$
from mm	$(1-H_1)(1-H_2)$	$= 0.2 * 0.1 = 0.02$

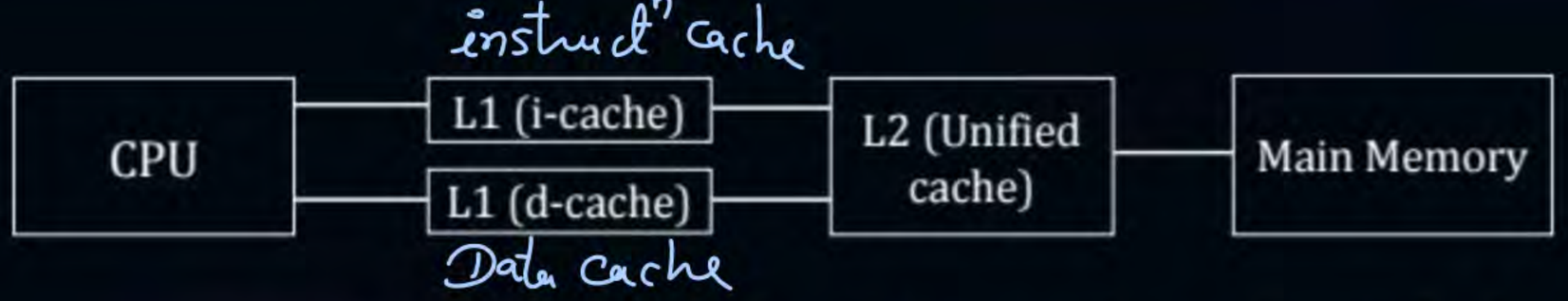
[NAT]

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

$$t_{avg} = 0.95 * 10 + 0.045 * (10+50) + 0.005 * (10+50+400)$$
$$= 14.5 \text{ ns}$$

Topic : Dual Cache



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

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

$$t_{avg} = \%\ of\ inst^n\ access * t_{avg\ inst^n} + \%\ of\ data\ access * t_{avg\ data}$$

[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 $\Rightarrow 25\ ns$
2. Average memory access time for only data access $\Rightarrow 17.5\ ns$
3. Average memory access time $\Rightarrow 20.5\ ns$

$$t_{avg\ inst^n} = 10 + 0.2\left(50 + 0.05 * 500\right) = 25\ ns$$

$$t_{avg\ data} = 10 + 0.1\left(50 + 0.05 * 500\right) = 17.5\ ns$$

$$t_{avg} = 0.4 * 25\ ns + 0.6 * 17.5$$

$$= 20.5\ ns$$

[NAT] GATE–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 _____?

$$t_{avg\,i} = 2 + 0.2\left[8 + 0.1 * 90\right] = 5.4\,ns$$

$$t_{avg\,d} = 2 + 0.1\left[8 + 0.1 * 90\right] = 3.7\,ns$$

$$t_{avg} = (0.6 * 5.4) + (0.4 * 3.7)$$

$$= 4.72\,ns$$

[NAT]

#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 8.88
2. Performance gain (speed up) of perfect cache as compared to cache with stalls

extra cycles

mem. operand access

cache with 0 miss

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

extra cycles (stalls) for inst^n access = $1 * 0.02 * 200 = 4$ cycles

extra cycles (stalls) for data access = $0.36 * 0.04 * 200 = 2.88$ cycles

with stalls CPI = $2 + 4 + 2.88 = 8.88$

Topic : Cache Inclusion Policy

if a mm block is present in L1 then it should be present in L2 also or not.

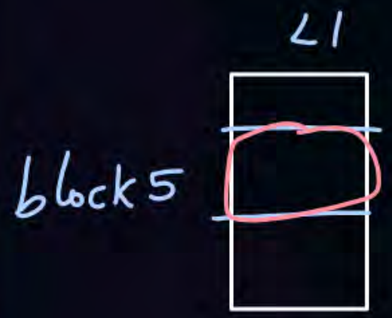
Topic : Inclusion Policy

All the content of L1 must be present in L2 also.

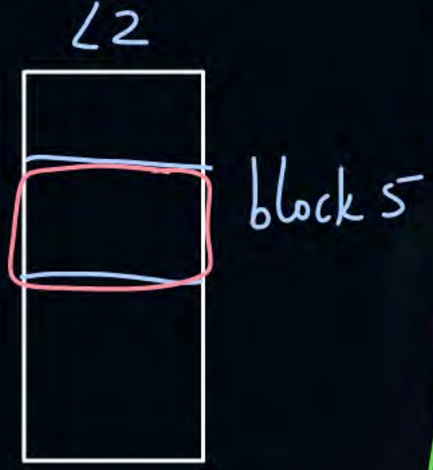
or

L2 is inclusive of L1

Size of L1 $\leq$ (Size of L2)



Values in block in L1 and L2 may be different.



If values also same then policy is called as value inclusion policy

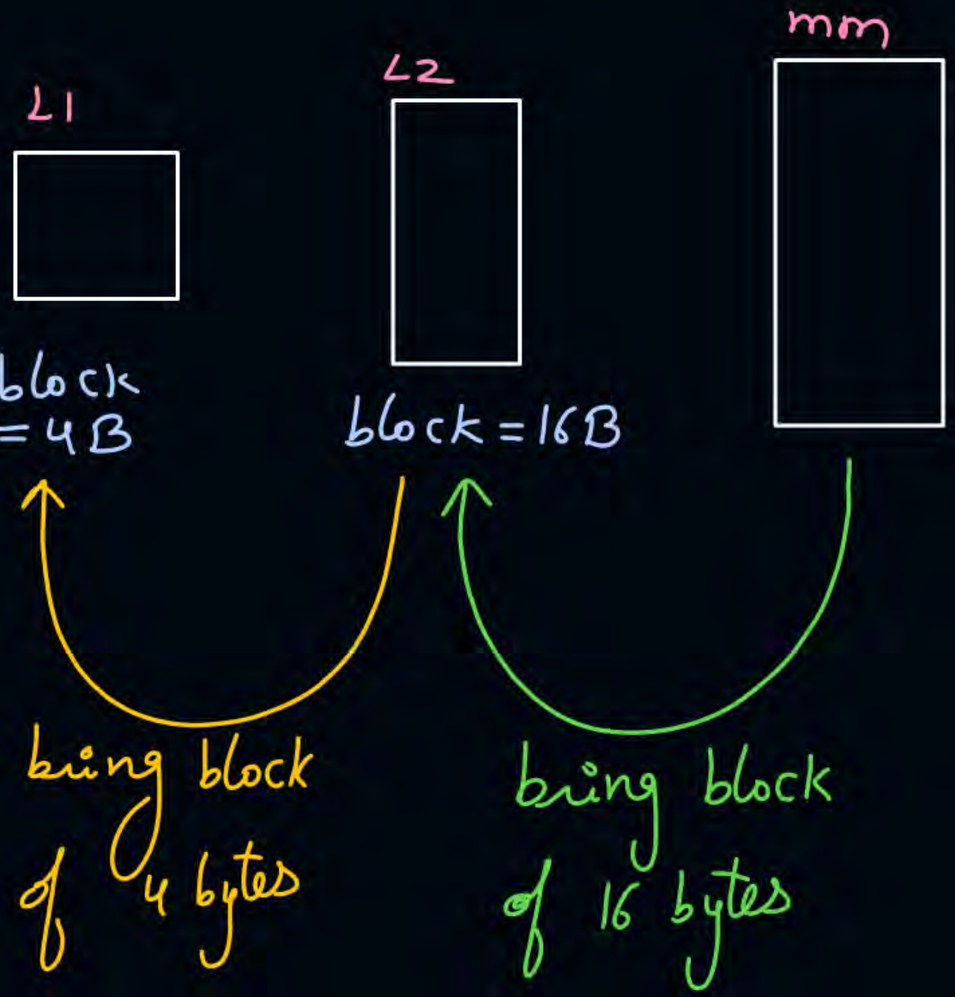
Topic : Inclusion Policy for read access

1. Hit in L1
   Read content from L1

2. Miss in L1 & Hit in L2
   Reads content from L2. Copies missed block from L2 to L1.
   If a block is evicted from L1 then there is no any role of L2.

3. Miss in L1 & Miss in L2
   Reads content from mm. Copies missed block from mm to L2, then from L2 to L1. If a block is evicted from L1 then there is no any role of L2.

Inclusion :-


L1
L2
mm
assume
block
= 4B
block = 16B
bring block
of 4 bytes
bring block
of 16 bytes

Topic : Exclusion Policy

Content of L1 is not present in L2

or

L2 is not inclusive of L1.

L2 is victim cache here, as it always holds only removed blocks of L1.

Size of L1 & Size of L2 are independent.

Topic : Exclusion Policy for read access

1. Hit in L1 CPU reads content from L1

2. Miss in L1 & Hit in L2 reads content from L2.
   Move (remove from L2) the missed block from L2 to L1.
   If a block is evicted from L1, then move it to L2.

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

2 mins Summary

Topic Multilevel Cache

Topic Dual Cache

Topic Cache Inclusion Policies

Happy Learning

THANK - YOU