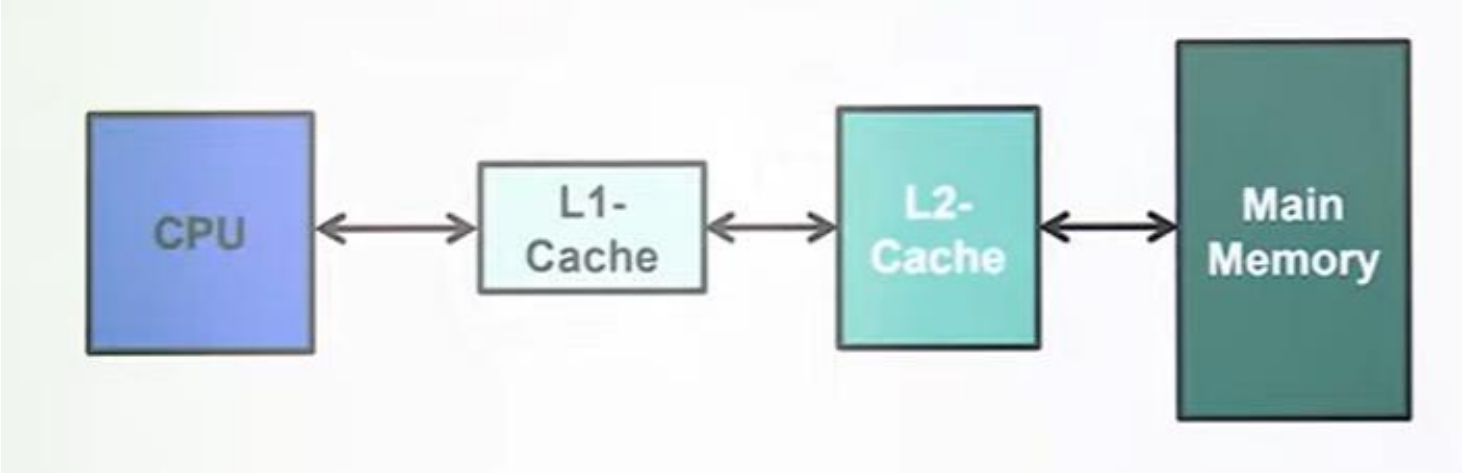
Why multilevel cache?

Sridhar,Associare Professor,CSE BIT Mesra

- 1)Minimize access time(use small cache-L1)
- 2)Maximize hit rate(use large cache-L2)
- 3)Minimize miss penalty(use multibank memory,use non-blocking cache)



AMAT

1. Simultaneous Access

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

2. Hierarchical Access

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

or

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

Cache Inclusion Policy

- Inclusion policy(content of L1 is also present in L2)

Inclusion

all requests are for read.

Scenario	Operation
Hit in L1	Read content from L1.
Miss in L1 & Hit in L2	Copy the block from L2 to L1. There will not be any role of L2 for evicted block from L1
Miss in L1 & L2 Both	First the block is copied to L2, then from there the block is copied to L1.

Evicted blocks from L1 and L2 will not have any role of either of them.

Exclusion

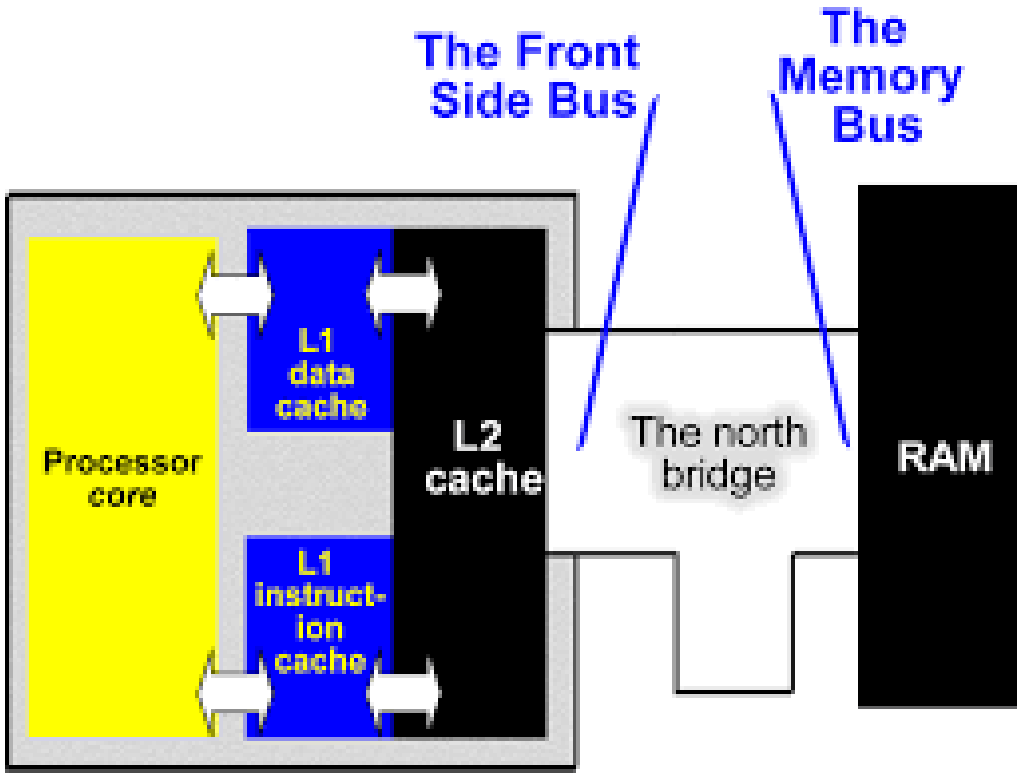
- Content of L1 is not present in L2 also
- L2 is filled with only replaced(victim)blocks of L1.Hence L2 is a victim cache

all requests are read.

Exclusion

Scenario	Operation
Hit in L1	Read content from L1
Miss in L1 & Hit in L2	Move the block from L2 to L1. Evicted block of L1 is moved to L2.
Miss in L1 & L2 Both	Copy the block from mm to L1. Evicted block of L1 is moved to L2.

Dual cache at L1



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

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

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

Ans-1)25ns,2)17.5ns,3)20.5ns

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.

Average memory access time for only instructions access

Average memory access time for only data access

Average memory access time

HW

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 _____?