Microprocessor & Computer Architecture (µpCA)

UE19CS252

Dr. D. C. Kiran

Department of
Computer Science and Engineering

Microprocessor & Computer Architecture (µpCA)

Unit 4: Cache Optimization

Dr. D. C. Kiran
Department of Computer Science and Engineering

Microprocessor & Computer Architecture (µpCA)

Syllabus

~~**Unit 1: Basic Processor Architecture and Design**~~

~~**Unit 2: Pipelined Processor and Design**~~

~~**Unit 3: Memory**~~

Unit 4: Input/Output Device Design

~~3 C~~

~~Introduction to Cache Optimization~~

~~Recuse Miss Rate~~

Reduce Miss Penalty

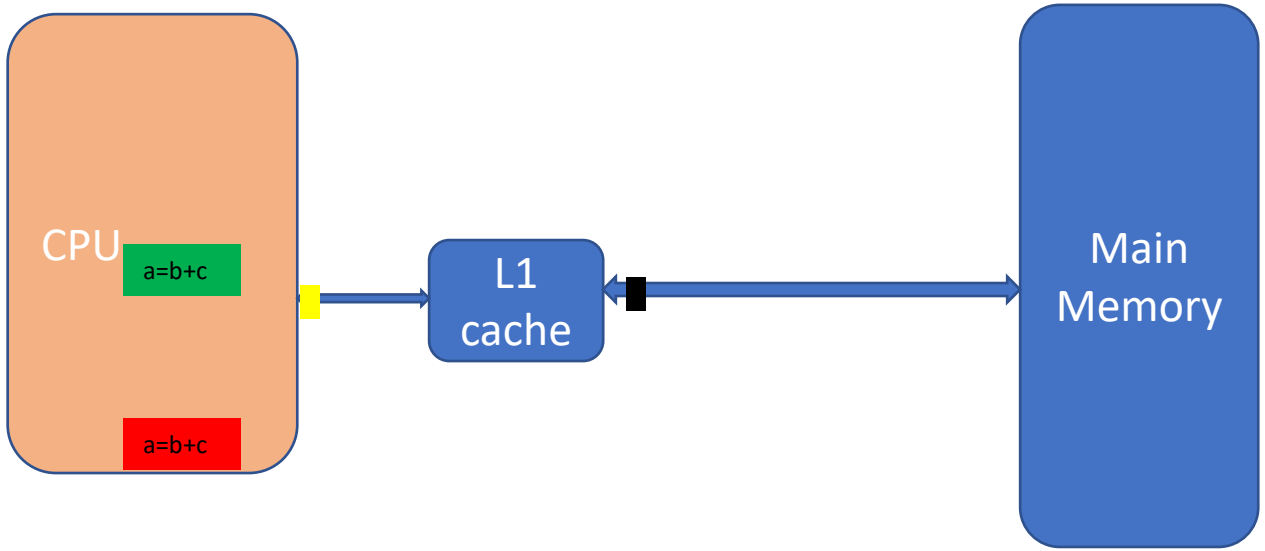
4th Optimization: Multilevel Caches to Reduce Miss Penalty

5th Optimization: Giving priority to Read misses over Write misses to reduce miss penalty

Unit 5: Advanced Architecture

What is the Problem?

AMAT = Hit time + Miss rate x Miss penalty

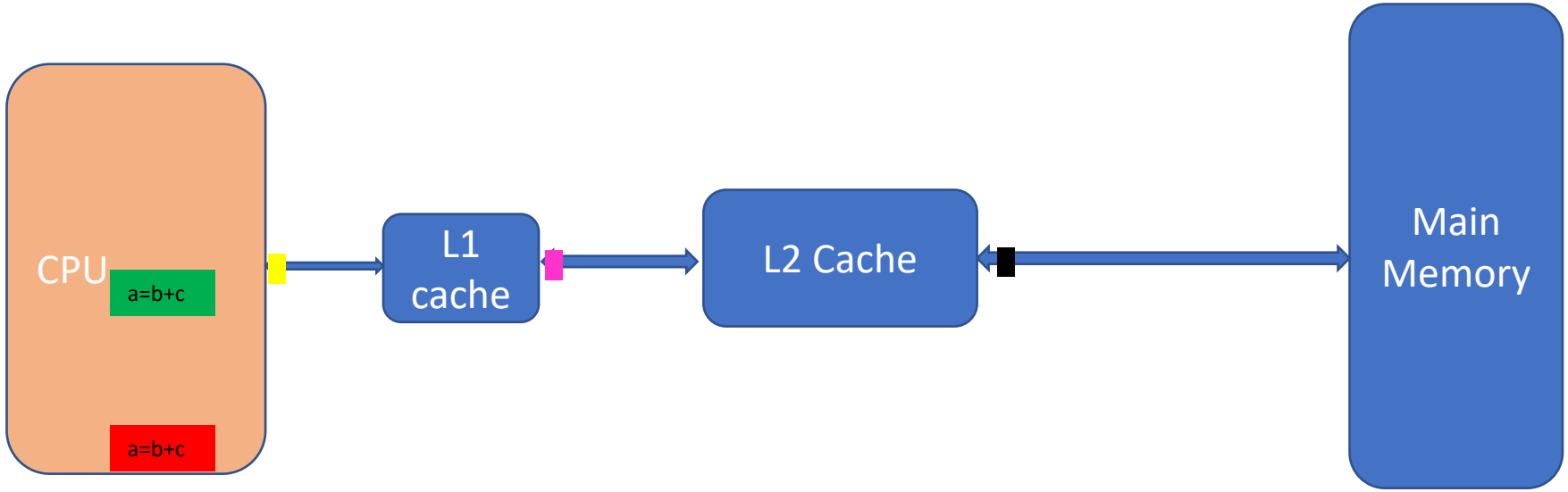


- **Reducing Miss penalty is also important similar to Reducing miss rate.**

- The performance gap between processor & memory raises a question:
  - –Should I make the cache faster to keep the pace with the speed of the processor? **Or**
  - –Make the cache larger to overcome the widening gap between the processor & the main memory?

What is the Problem?

- Answer for these questions is to do both.
  - **–Adding another level of cache between memory & original cache simplifies the decision.**
- First level cache can be small enough to match the clock cycle time of the processor.
- Second level cache be can be large enough to capture many accesses that would go to main memory, thus reducing miss penalty.



AMAT? For two Level Cache

•The original formula is:

$$\textbf{AMAT} = \textbf{Hit Time}_{L1} + \textbf{Miss Rate}_{L1} \times \textbf{Miss Penalty}_{L1}$$

Miss in Level 1 Cache, leads to access data from Level 2 cache

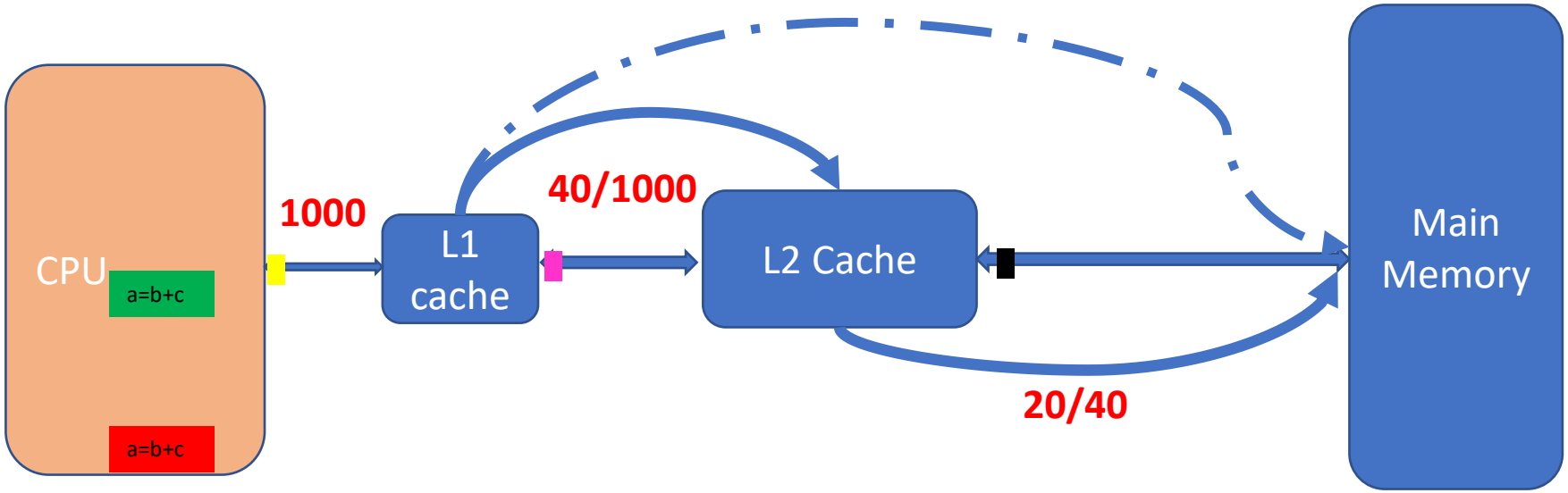
$$\textbf{Miss Penalty}_{L1} = \textbf{Hit Time}_{L2} + \textbf{Miss Rate}_{L2} \times \textbf{Miss Penalty}_{L2}$$

- Hit Time_{L2}= Item found in the Level 2 Cache
- Miss Rate_{L2}= How frequently element not found in Level 2 Cache
- Miss Penalty_{L2} = Extra time spent to bring item from RAM.

AMAT? For two Level Cache

$$AMAT = Hit\ Time_{L1} + Miss\ Rate_{L1} \times [Hit\ Time_{L2} + Miss\ Rate_{L2} \times Miss\ Penalty_{L2}]$$

4th Optimization: Multilevel Caches to Reduce Miss Penalty



Suppose that in 1000 memory references there are 40 misses in the first level cache and 20 misses in the second –level cache.

What is Miss Rate of Level 1 Cache? **40/1000**

What is Miss Rate of Level 2 Cache? **20/1000** or **20/40**

4th Optimization: Multilevel Caches to Reduce Miss Penalty

Local miss rate:

•The number of misses in the cache divided by the total number of memory accesses to this cache.

Ex: For first level cache it is, **Miss Rate$_{L1}$ (40/1000)**

For second level cache it is, **Miss Rate$_{L2}$ (20/40)**

Global miss rate:

•The number of misses in the cache divided by the total number of memory accesses generated by the processor.

Ex: Global miss rate for level1 cache is still **Miss Rate$_{L1}$ (40/1000)**

but, for level2 cache it is : **Miss Rate$_{L1}$ x Miss Rate$_{L2}$**

(40/1000) x (20/40)= (20/1000)

Microprocessor & Computer Architecture (µpCA)

4th Optimization: Multilevel Caches to Reduce Miss Penalty

Suppose that in 1000 memory references there are 40 misses in the first level cache and 20 misses in the second –level cache. ***What are the various miss rates?***

Assume the miss penalty from the L2 cache to memory is 200 clock cycles, the hit time of the L2 cache is 10 clock cycles, hit time for L1 cache is 1 clock cycle.

What is the average memory access time ?

Solution:

The Miss rate [global & local] for the 1st level cache is 40/1000= 4%
The local miss rate for 2nd Level cache is 20/40 = 50%
The Global miss rate of 2nd Level cache is 20/1000=2%

AMAT = Hit Time$_{L1}$ + Miss Rate$_{L1}$ x [Hit Time$_{L2}$ + Miss Rate$_{L2}$ x Miss Penalty$_{L2}$]

AMAT =1 + 4% x [10 +50% x 200]

=1+0.04 x[10+0.5x200]

=1+0.04 x 110

=5.4

Average Memory Stalls Per Instruction

Average Memory Stalls Per instruction = (AMAT - Hit time$_{L1}$) x Average # of memory references per instruction

or

Average memory stalls per instruction = Misses per instruction $_{L1}$ x Hit time$_{L2}$ + Misses per instruction $_{L2}$ x Miss Penalty $_{L2}$

or

Average memory stalls per instruction = (Miss Rate $_{L1}$ x Hit time$_{L2}$ + Miss Rate $_{L2}$ x Miss Penalty $_{L2}$) *Memory Reference per Instruction

or

Average memory stalls per instruction = (Miss Rate $_{L1}$ x Miss Penalty$_{L1}$ + Miss Rate $_{L2}$ x Miss Penalty $_{L2}$)* Memory Reference per Instruction

Reference

Average Memory Stalls Per Instruction

Suppose that in 1000 memory references there are 40 misses in the first level cache and 20 misses in the second –level cache. ***What are the various miss rates?***

Assume the miss penalty from the L2 cache to memory is 200 clock cycles, the hit time of the L2 cache is 10 clock cycles, hit time for L1 cache is 1 clock cycle and **there are 1.5 memory references per instruction.**

What is the Average Stall Cycles Per Instruction?

Solution:

$$\text{\# of misses per instruction} = \frac{\#\ of\ memory\ references}{\#\ of\ memory\ references\ per\ instructions} = \frac{1000}{1.5} = 667 \text{ ins.}$$

Thus, for L1 cache ,

 # misses for 40 memory accesses for 1000 instructions = 40 x 1.5 = 60 misses and for 20 misses for L2 cache it is 1.5 x 20 = 30 misses.

Average Memory Stalls Per Instruction

Average memory stalls per instruction = Misses per instruction $_{L1}$ x Hit time$_{L2}$ + Misses per instruction $_{L2}$ x Miss Penalty $_{L2}$

= (60/1000) x 10 + (30/1000) x 200

= 0.06 x 10 + 0.03 x200 = 0.06 + 6

= 6.6 clock cycles.

or

Average memory stalls per instruction = (AMAT - Hit time$_{L1}$) x Memory references per instruction

= (5.4 - 1.0) x 1.5 = 6.6 clock cycles.

or

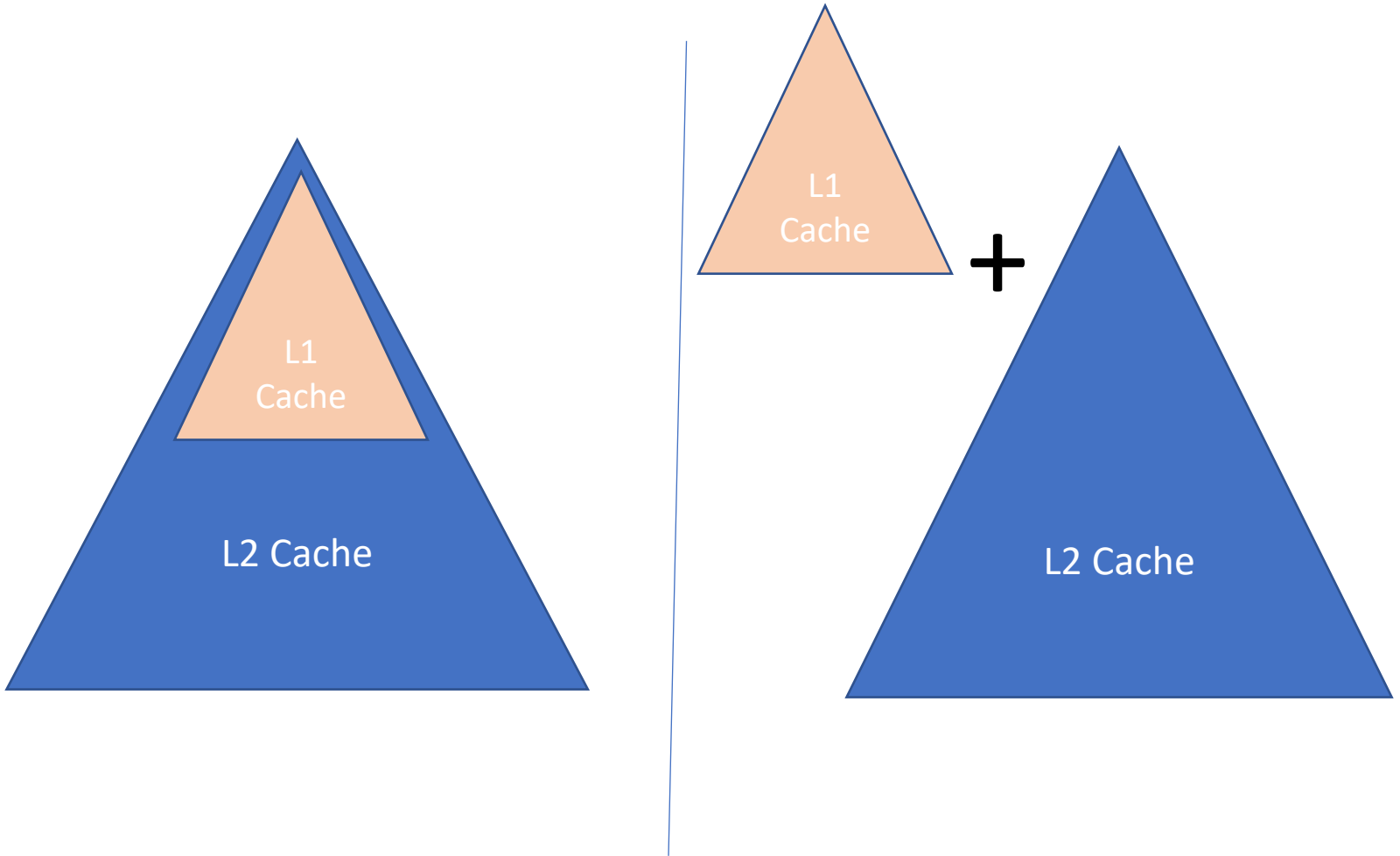
Average memory stalls per instruction = (Miss Rate $_{L1}$ x Hit time$_{L2}$ + Miss Rate $_{L2}$ x Miss Penalty $_{L2}$)* Memory Reference per Instruction

= ((40/1000*10)+(20/1000*200)) *1.5

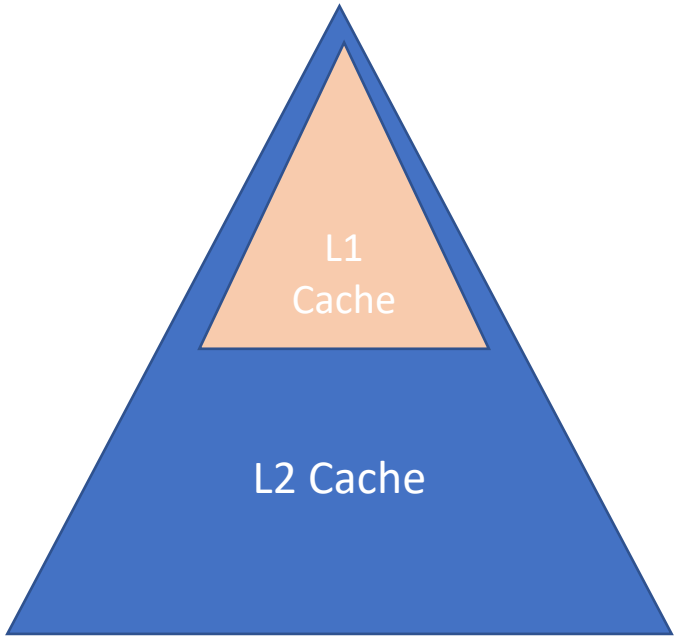
= 6.6 Clock Cycles

Why Global Miss Rate of 2nd Level Cache is Important?

Which one is Correct?



Why Global Miss Rate of 2nd Level Cache is Important?



First perspective :

Global cache miss rate is very similar to the single cache miss rate of the second level cache.

Second Perspective:

- Local cache miss rate is not a good measure of secondary caches.
- It is a function of the miss rate of the first level cache.
- Can vary by changing the first – level cache.

Note: Global cache miss rate should be used when evaluating second level caches.

Parameters for choosing 2nd Level Cache

- Miss Rate of 2nd level cache is function of 1st Level Cache.
- Speed of 2nd Level Cache will affect the Miss Penalty of 1st Level Cache.
- Design of 2nd Level cache should compensate all the deficiency in designing of 1st Cache.
- Thus, the strategies or Policy used in 2nd Level cache need not be same as design of 1st Level Cache.

Parameters for choosing 2nd Level Cache

Example Given the data below, what is the impact of second-level cache associativity on its miss penalty?

- Hit time$_{L2}$ for direct mapped = 10 clock cycles.
- Two-way set associativity increases hit time by 0.1 clock cycle to 10.1 clock cycles.
- Local miss rate$_{L2}$ for direct mapped = 25%.
- Local miss rate$_{L2}$ for two-way set associative = 20%.

Answer For a direct-mapped second-level cache, the first-level cache miss penalty is

$$\text{Miss penalty}_{\text{1-way L2}} = 10 + 25\% \times 200 = 60.0 \text{ clock cycles}$$

Adding the cost of associativity increases the hit cost only 0.1 clock cycle, making the new first-level cache miss penalty:

$$\text{Miss penalty}_{\text{2-way L2}} = 10.1 + 20\% \times 200 = 50.1 \text{ clock cycles}$$

Microprocessor & Computer Architecture (µpCA)

5th Optimization: Giving priority to Read Misses over Write

Case 1: Write Back Policy

Scenario

- P is referred by the Processor.
- P Should replace A.
- A is Dirty

Known Solution

- Write A back in Memory
- Replace P by A

Cache

Valid Bit	Dirty Bit	TAG	Data
1	1	10100	A=11
1	0	10100	B=5
1	0	10100	C=6

Main Memory

10100…..1010..0000	A= 0
10100…..1010..0001	B=5
10100…..1010..0010	C=6
00010…..1010..0000	P=100

Optimized Solution (Give Priority for CPU Reference / Read)

- Place A on to buffer to save time.
- Replace P by A, to provide quick access .
- Write A into memory parallelly when Processor is using P.

Microprocessor & Computer Architecture (µpCA)

5th Optimization: Giving priority to Read Misses over Write

Case 2: Write Through

Scenario

- **P is referred by the Processor.**
- **P Should replace A.**

Cache

Valid Bit	TAG	Data
1	10100	A=11
1	10100	B=5
1	10100	C=6

		A=11

Write Buffer

Main Memory

10100…..1010..0000	**A= 0**
10100…..1010..0001	B=5
10100…..1010..0010	C=6
00010…..1010..0000	P=100

Microprocessor & Computer Architecture (µpCA)

5th Optimization: Giving priority to Read Misses over Write

Case 2: Write Through

Scenario

- **P is referred by the Processor.**
- **P Should replace A.**

Known Solution

- **Place A on the Write Buffer**
- **Replace P by A**

What may go wrong?

If A is referred again!

Cache

Valid Bit	TAG	Data
1	00010	P=100
1	10100	B=5
1	10100	C=6

		A=11

Write Buffer

Main Memory

10100…..1010..0000	A= 0
10100…..1010..0001	B=5
10100…..1010..0010	C=6
00010…..1010..0000	P=100

Solution1: Wait till A is written back in Memory

Solution2: (Give Priority for CPU Reference / Read)

Search for A in buffer, if found compare A in Memory,

i. if both are same, Replace.

ii. Otherwise, read from Buffer

Consider the following code sequence.

Ex: STR R3, [R0,512]
LDR R1, [R0.1024])
LDR R2, [R0,512] Assume Direct mapped:

Write –through cache that maps 512 and 1024 to the same block. Four word write buffer that is not checked on a read miss. Will the value in R2 always be equal to the value in R3? o R2!

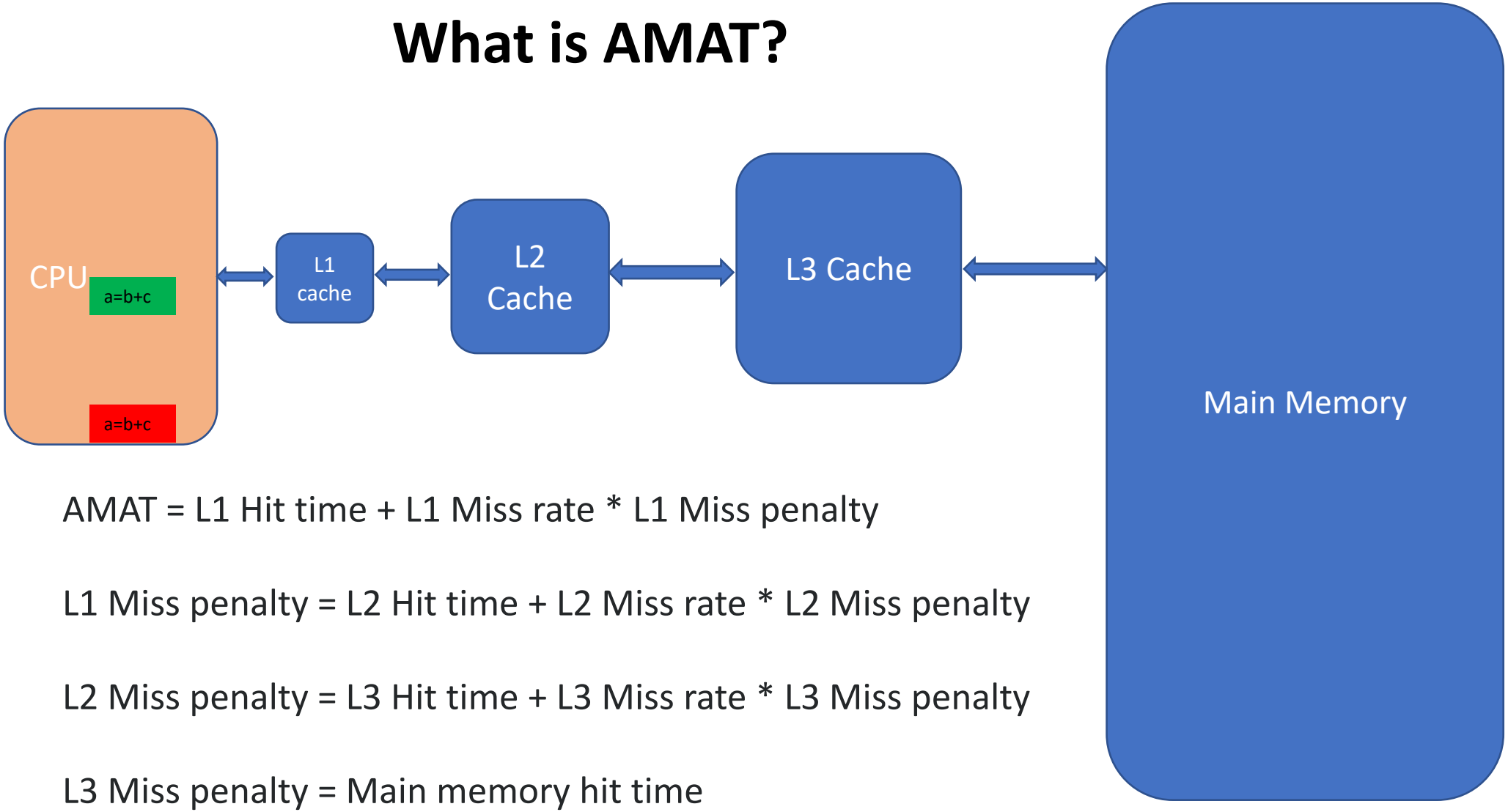
PES UNIVERSITY ONLINE

Ans:

- This is a read-after-write data hazard in memory.
- The data in R3 are placed into the write buffer after the STR.
- The following LDR instruction uses the same cache index and is therefore a miss.
- The second LDR instruction, tries to put the value in location 512 into the register R2.
- This also results in a miss.
- If the write buffer hasn't completed writing to location 512 in memory,
- The read of location 512 will put the old, wrong value into the cache block and then into R2.
- Without proper precautions, R3 would not be equal to R2!

Think About it?

What is AMAT?



AMAT = L1 Hit time + L1 Miss rate * L1 Miss penalty

L1 Miss penalty = L2 Hit time + L2 Miss rate * L2 Miss penalty

L2 Miss penalty = L3 Hit time + L3 Miss rate * L3 Miss penalty

L3 Miss penalty = Main memory hit time

Improving Hit Time

THANK YOU

Dr. D. C. Kiran
Department of Computer Science and Engineering

dckiran@pes.edu

9829935135