CS223 : Computer Architecture & Organization

Lecture 23 [01.04.2022]

Advanced Cache Optimizations-Tutorials

Dr. John Jose

Associate Professor

Department of Computer Science & Engineering

Indian Institute of Technology Guwahati, Assam.

Multilevel Caches

Assume a 2-level cache system with the following specifications. L1 Hit Time = 1 cycle, L1 Miss Rate = 2.5%, L2 Hit Time = 6 cycles, L2 Miss Rate = 17% (% L1 misses that miss), L2 Miss Penalty = 120 cycles. Compute the average memory access time.

AMAT = ht1+mr1xMP
= ht1+mr1x(ht2+mr2xMP)
= 1 + 0.025 x (6 +0.17 x 120)= 1.66 CC

Optimization

A cache has access time (hit latency) of 10 ns and miss rate of 5%. An optimization was made to reduce the miss rate to 3% but the hit latency was increased to 15 ns. Under what condition this change will result in better performance (Lower AMAT)?

Optimization

A cache has access time (hit latency) of 10 ns and miss rate of 5%. An optimization was made to reduce the miss rate to 3% but the hit latency was increased to 15 ns. Under what condition this change will result in better performance (Lower AMAT)?

AMAT 1 = HT1 + MR1 x MP HT1 = 10ns; MR1=0.05

AMAT 2 = HT2 + MR2 x MP HT2 = 15ns; MR2=0.03

AMAT2<AMAT1

15 + 0.03x MP < 10 +0.05xMP

5 <0.02MP → MP> 250 ns

Optimization

A cache has hit rate of 95%, block size of 128B, cache hit latency of 5ns. Main memory takes 50 ns to return first word (32 bits) of a block and 10 ns for each subsequent word.

(a) What is the miss latency of the cache?

(b) If doubling the cache block size reduces the miss rate to 3%, does it reduces AMAT?

Optimization

hit rate of 95%, 28B blocks, cache hit latency of 5ns. Main memory takes 50 ns to return first word (32 bits) of a block and 10 ns for each subsequent word. (a) What is the miss latency of the cache?
(b) If doubling the cache block size reduces the miss rate to 3%, does it reduces AMAT?

Hr= 0.95; BS= 128B; Ht =5 ns ; 1word= 4B (32 bits)

words/ block = 128B/4B = 32

(A) MP = 50 + (31x10) = 360 ns

AMAT1 = 5 + 0.05 x 360 = 23 ns

(B) # words/ block = 256B/4B = 64 ;

MP = 50 + 63 x10 =680 ns

AMAT2 = 5 + 0.03 x 680 = 25.4 ns

Doubling block size will not reduce AMAT

Cache Mapping

A 16KB direct mapped 256B block unified cache is attached to a 16MB main memory system. The word length as well as instruction length of the processor is 16 bits. Consider a program that consists of a main routine M which in turn calls a subroutine S. M consists of 12 instruction words which are loaded in the main memory from the address 0x4230FA onwards. The last five instructions of M is a loop that is iterated 10 times. The second instruction in the loop is a call to subroutine S. S consists of 4 instruction words loaded in the main memory from the address 0x70F168. The last instruction of S is a subroutine return back to M. The only two data words that are used by M and S are at addresses 0x748074 and 0x846064. Assume the caches are initially empty. Ignore OS level interruption and subsequent cache impact on context switching. List out the block numbers (in decimal) in the cache that are non-empty after the execution of the program.

Cache Mapping by OS

M

1	0x 4230FA
2	0x 4230FC
3	0x 4230FE
4	0x 423100
5	0x 423102
6	0x 423104
7	0x 423106
8	0x 423108
9	0x 42310A
10	0x 42310C
11	0x 42310E
12	0x 423110

10

Tag	Index	Offset
10	6	8

1	0x 748074

1	0x 846064

S

1	0x 707168
2	0x 70716A
3	0x 70716C
4	0x 70716E

Cache Mapping by OS

M

1	0x 4230FA (48)
2	0x 4230FC (48)
3	0x 4230FE (48)
4	0x 423100 (49)
5	0x 423102 (49)
6	0x 423104 (49)
7	0x 423106 (49)
8	0x 423108 (49)
9	0x 42310A (49)
10	0x 42310C (49)
11	0x 42310E (49)
12	0x 423110 (49)

10

Tag	Index	Offset
10	6	8

1	0x 748074 (0)

1	0x 846064 (16)

S

1	0x 707168 (49)
2	0x 70716A (49)
3	0x 70716C (49)
4	0x 70716E (49)

Cache Mapping by OS

Find the number of cache misses occurred during the execution of the program.

M1

M4

S1, M10, S1, M10, ….. (10 TIMES) = 22 MISSES

How many cache block evictions happened during the execution of the program?

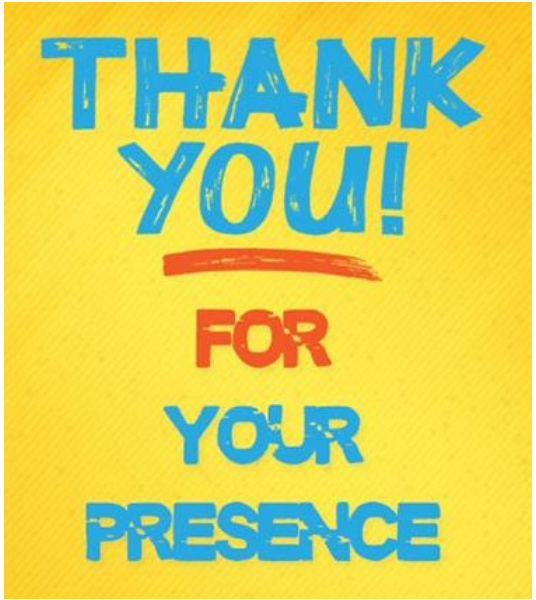
20 EVICTIONS

List out the block numbers (in decimal) in the cache that are non-empty after the execution of the program.

ALL BLOCKS EXCEPT 0,16,48,49

Reference

- **Computer Architecture-A Quantitative Approach** (5th edition), John L. Hennessy, David A. Patterson, Morgan Kaufman.
- Chapter 2: Memory Hierarchy Design
  - Section 2.1: Introduction
  - Section 2.2: Ten Advanced Cache Optimizations of Cache Performance
- **NPTEL Video Links:**
  - **https://tinyurl.com/yf94dnby**
  - **https://tinyurl.com/yfgmfsmm**



johnjose@iitg.ac.in
http://www.iitg.ac.in/johnjose/