Computer Architecture

Fall, 2024
Week 14
2024.12.02

組別：_____ 簽名：_____

[group 1]

1. Qustion:

Compare the following cache types: direct-mapped cache, N-way set-associative cache, and fully associative cache.

1. Which cache type has the longest data access time, and why?
2. Which cache types encounter data replacement issues? Please suggest a replacement policy for each. Fill in the answer in the given table.

Cache Type	Replacement Issues	Replacement Policy
Direct-Mapped Cache		
N-Way Set-Associative		
Fully Associative Cache		

Ans:

1. Fully Associative Cache, because the data can be in any cache line. The system must search the entire cache (using associative lookup) to find the data.

2.

Cache Type	Replacement Issues	Replacement Policy
Direct-Mapped Cache	No	Direct replacement
N-Way Set-Associative	Yes	LRU, Random, or FIFO
Fully Associative Cache	Yes	LRU, Random, or FIFO

[group 2]

2. Question:

Please calculate the effective CPI with the given conditions.

CPU base CPI = 3, clock rate = 5GHz

L1 miss rate to main memory or L2 = 1.5%

L2 miss rate to main memory = 0.01%

access time to main memory = 80ns

access time to L2 = 10ns

a. Use only one cache(L1)

b. Use two caches(L1、L2)

Ans:

a. miss penalty = 80/0.2 = 400

effective CPI = 3 + 0.015*400 = 3+6=9

b. miss penalty of L1 to access L2 = 10/0.2 = 50

miss penalty of L2 to access main memory = 80/0.2 = 400

effective CPI = 3 + 0.015*50 + 0.0001*400 = 3+0.75+0.04 = 3.79

[group 6]

3. Question:

Assume that main memory accesses take 70 ns. The following table

shows data for L1 caches attached to each of two processors, P1 and P2.

What is the Average Memory Access Time for P1 and P2 (in cycles)?

	L1 Size	L1 Miss Rate	L1 Hit Time
P1	2 KiB	8.0%	0.66 ns
P2	4 KiB	6.0%	0.90 ns

Hit time = 1 cycle

Ans:

For P1 all memory accesses require at least one cycle (to access L1). 8% of memory accesses additionally require a 70 ns access to main memory. This is 70/0.66 = 106.06 cycles. However, we can't divide cycles; therefore, we must round up to 107 cycles. Thus, the Average Memory Access time is 1 + 0.08*107 = 9.56 cycles

For P2, a main memory access takes 70 ns. This is 70/0.9 = 77.78 cycles. Because we can't divide cycles, we must round up to 78 cycles. Thus the Average Memory Access time is 1 + 0.06*78 = 5.68 cycles

[group 7]

4. Question:

Select the correct answer, explain the wrong answer: (T or F)

(1) Increate miss ratio can improve cache performance

(2) I-cache is the instruction cache and D cache is the disk cache

(3) CPU with 1ns clock, hit time = 1 cycle, miss penalty = 30, I-cache miss rate = 5%. AMAT = 2

(4) The miss penalty for a four-word-wide bank of DRAMs is 17. With 1 memory bus clock to send the address, 15 memory bus clocks to each FRAM access initiated and 1

memory bus clock to send a word of data. Consider the case that cycle time overlaps with transfer time.

(5) For a fixed size cache, increasing associativity expands tag, shrinks index

Ans:

(1) F decrease miss ratio

(2) F D cache is data cache

(3) F AMAT = 2.5 (1*1 + 30 * 0.5 = 2.5)

(4) F, because the cycle time and transfer time overlaps, therefore, 1 + 15 = 16 (don't need to transfer)

(5) T

[group 8]

5. Question:

Which of the following statements are true?

a. Virtual memory "block" is called a page.

b. When CPU performance increases, the miss penalty becomes more significant.

c. In a fully associative cache, each memory block is placed in a specified cache location.

d. The main reason for adding an L2 cache between the L1 cache and main memory is to reduce the average hit time.

Ans:

A, B

C) In fully associative cache, each memory block can be placed in any cache location.

D) The main reason is to reduce the miss rate.

[group 9]

6. Question:

Give an example scenario where cache memory plays a critical role in reducing access time.

Ans:

Executing program frequently access value in the memory like loop through an array

Without cache: the CPU would need to fetch element from the main memory every time, and each time need a lot of time(lower level cost more time to access)

With cache memory:

- **First Access:** When the CPU accesses the first element, the cache misses, and the data is fetched from the main memory and stored in the cache.
- **Subsequent Accesses:** As the CPU continues looping through the array, if the cache is well-design satified "spatial locality", the next element might also be fetched with the previous several element, so CPU can access directly from the cache with much higher speed.

[group 10]

7. Question:

Assume we have a two level cache machine. Given the following information, please calculate the CPI considering miss penalties. (Ideal CPI=2;CPU clock rate=2GHz)
(1)Data transfer time table

from	to	time cost
L1	processor	1ns
L2	L1	5ns
main memory	L2	10ns

(2)Data access time table

storage	time cost
L1	2ns
L2	12ns
main memory	120ns

(3)Miss and Hit rate table

	percentage
L1 Hit	97.5%
L1 Miss & L2 Hit	2%
L1 Miss & L2 Miss	0.5%

(4)It costs 0ns to place or replace a block in any storage.

Ans:

L1 miss but L2 hit:

12ns(access L2) + 5ns(L2 transfer to L1) + 2ns(access L1) + 1ns(L1 to processor) = 20ns

20ns/0.5ns = 40 cycles

L1 and L2 both miss:

120ns(access memory) + 10ns(memory to L2) + 12ns(access L2) + 5ns(L2 transfer to L1) + 2ns(access L1) + 1ns(L1 to processor) = 150ns

150ns/0.5ns = 300cycles

CPI = 2 + 0.02*40 + 0.005*300 = 2 + 0.8 + 1.5 = 4.3(ans)

[group 11]

8. Question:

Given:

- CPU base CPI = 1
- Clock rate = 2GHz
- Miss rate/instruction = 3%

- Main memory access time = 80ns

Calculate the effective CPI if the system uses only a primary cache.

Ans:

1.Calculate the miss penalty in cycles:

Clock Cycle Time = 1/2GHz = 0.5 {ns}

Miss Penalty = 80 ns / 0.5 ns = 160 {cycles}

2.Calculate the Effective CPI:

Effective CPI = 1 + 0.03 \times 160 = 1 + 4.8 = 5.8

Effective CPI = 5.8.

[group 12]

9. Question:

For a direct-mapped cache design with a 32-bit address, the following bits of the address are used to access the cache.

Tag	Index	Offset
31–10	9–5	4–0

(1) What is the cache block size (in words)?

(2) How many entries does the cache have?

(3) What is the ratio between total bits required for such a cache implementation over the data storage bits?

Ans:

(1) 8 words

(2) 32 entries

(3) (22+1+8*32)*32 : (32*8)*32 bits= 279:256 約等於 1.09

[group 13]

10. Question:

2. Which of the following statements about page faults are correct?

(多選題)

A. A page fault means that the requested page is not currently in memory.

B. Hardware alone can detect and resolve page faults efficiently without software intervention.

C. Page faults can incur a huge miss penalty, often taking millions of cycles to process.

D. Page hits occur when the required page is not resident in memory.

Ans:

A, C

[group 14]

11. Question:

Assume It costs:

1 memory bus clock to send the address

10 memory bus clocks for each DRAM access initiated

2 memory bus clocks to send a word

When sending 10 words, How many time of Miss penalty we can reduce if we use “five-bank, one-word-wide bus of DRAMs” instead of “one-word-wide bank of DRAMs”?

Ans:

(1+10*10+10*2)/(1+10*10/5+10*2)=121/41=2.95