CS2323 Computer Architecture 2019
Homework 1

==

Q1. (2 marks) To see how much fraction of energy of L1 cache and L2 cache come from dynamic or leakage energy, consider this data.

From processor core, 10^6 accesses come to L1 cache. Total execution time of the program is 1000 ns.

L1 cache:
Leakage 0.2W
Dynamic access energy for each access 0.217 nJ
Hit-rate = 95%

L2 cache:
Leakage 6.9 W
Dynamic access energy for each access 1.47 nJ
Hit rate = 100%

Find (DynamicEnergy*100)/TotalEnergy for L1 cache and L2 cache.

L1 Cache
Dynamic Energy = 10^6 * 0.217nJ = 217000nJ (some students multiplied 0.95 to it, which is not fully correct, but we are not deducting marks).
Leakage Energy = 0.2W * 1000ns
Dynamic Energy Fraction = 217000nJ/217200nJ = 99.90%

L2 Cache
Accesses to L2 cache: 0.05 * 10^6
Dynamic Energy = 0.05 * 10^6 * 1.47 nJ
Leakage Energy = 6.9W * 1000 ns
Dynamic Energy Fraction = 73500nJ/80400nJ = 91.4179%

Only first digit after decimal will be checked.

Q2. (1 mark) Write reason why on increasing the associativity, there is only marginal decrease in the miss rate (max 2 sentence).

On increasing associativity, only conflict misses are reduced. Capacity and Compulsory misses are unchanged. Hence, after certain level of associativity, there is only marginal decrease in miss rate.

Q3. (10 marks) Assume that a processor uses 8-bit address space. Assume that the address pattern that accesses the cache is:

Sequence1: 0, 63, 1, 62, 2, 61, 3, 60, 4, 59, 5, 58, 6, 57, 7, 56, 8, 55, 9, 54, 10, 53, 11, 52

Assume two different caches use the following two different address subdivision methods (figure is not drawn to scale).

Both the caches are direct-mapped, with a block size of 4 and have 8 sets each. In other words, their architectures are identical, except that they use different subdivision methods.

(a) Compute the tag and set for each address for both subdivision methods (hint: you can write a small C program to do that). You need not show this in your submission. For each address, show whether it leads to a hit or a miss and finally, what is the hit ratio (hits/accesses) for each cache?

Address		Method 1			Method 2		
Decimal	Binary	set	tag	Result	set	tag	Result
0	0	0	0	Miss	0	0	Miss
63	111111	7	1	Miss	7	7	Miss
1	1	0	0	Hit	0	1	Miss
62	111110	7	1	Hit	7	6	Miss
2	10	0	0	Hit	0	2	Miss
61	111101	7	1	Hit	7	5	Miss
3	11	0	0	Hit	0	3	Miss
60	111100	7	1	Hit	7	4	Miss
4	100	1	0	Miss	0	4	Miss
59	111011	6	1	Miss	7	3	Miss
5	101	1	0	Hit	0	5	Miss
58	111010	6	1	Hit	7	2	Miss
6	110	1	0	Hit	0	6	Miss
57	111001	6	1	Hit	7	1	Miss
7	111	1	0	Hit	0	7	Miss

56	111000	6	1	Hit	7	0	Miss
8	1000	2	0	Miss	1	0	Miss
55	110111	5	1	Miss	6	7	Miss
9	1001	2	0	Hit	1	1	Miss
54	110110	5	1	Hit	6	6	Miss
10	1010	2	0	Hit	1	2	Miss
53	110101	5	1	Hit	6	5	Miss
11	1011	2	0	Hit	1	3	Miss
52	110100	5	1	Hit	6	4	Miss

Hit rate for method 1 = 18/24 = 75%, hit rate for method 2 = 0%
(0.5 mark each for finding hit rate for method 1 and 2. 2 mark for finding all hit/miss decisions correctly for each of the methods. Partial marking can be done. Total marks = 5)

(b) Repeat (a) but with the following sequence:
Sequence2: 0, 64, 128, 192, 1, 65, 129, 193, 11, 75, 139, 203, 9, 137, 201, 73

Address		Method 1			Method 2		
Decimal	Binary	set	tag	Result	set	tag	Result
0	00000000	0	0	Miss	0	0	Miss
64	01000000	0	2	Miss	0	0	Hit
128	10000000	0	4	Miss	0	0	Hit
192	11000000	0	6	Miss	0	0	Hit
1	00000001	0	0	Miss	0	1	Miss
65	01000001	0	2	Miss	0	1	Hit

129	10000001	0	4	Miss	0	1	Hit
193	11000001	0	6	Miss	0	1	Hit
11	00001011	2	0	Miss	1	3	Miss
75	01001011	2	2	Miss	1	3	Hit
139	10001011	2	4	Miss	1	3	Hit
203	11001011	2	6	Miss	1	3	Hit
9	00001001	2	0	Miss	1	1	Miss
137	10001001	2	4	Miss	1	1	Hit
201	11001001	2	6	Miss	1	1	Hit
73	01001001	2	2	Miss	1	1	Hit

Hit rate for method 1 = 0%, hit rate for method 2 = 12/16 = 75%
(0.5 mark each for finding hit rate for method 1 and 2. 2 mark for finding all hit/miss decisions correctly for each of the methods. Partial marking can be done. Total marks = 5)

Q4. (4 marks) Consider two processors (P1 and P2) which run the same instruction set architecture (ISA). The frequency of P1 and P2 are 2.2GHz and 1.6GHz, respectively.
In this ISA, there are four classes of instructions A, B, C, and D. The CPI of each of these classes are given in the following table.

	A	B	C	D
P1	2	2	4	4
P2	2	1	2	3

There is a program which has 10^6 instructions divided into classes as follows: 30% class A, 20% class B, 35% class C, and 15% class D. Which processor is faster for this program?

Processor 1:

Total number of cycles for a given program is 10^6(0.3* 2+0.2* 2+0.35* 4 + 0.15* 4) = 3* 10^6.

Total execution time = 3*10^6 sec/2.2*10^9 = 1.36msec [1.5 marks]

For Processor P2

Total number of cycles = 1.95*10^6

Total execution time = 1.95*10^6/1.6 * 10^9 = 1.21875 msec [1.5 marks]

Hence, Processor P2 is 1.12 times faster than P1. [1 marks. No need to write how much faster.]

Q5. (3 marks) Assume that a system has 4 processors (P=4). Assume that directory-based coherence protocol is used. Show the state of (P+1) bit directory for a cache block after each of these operations to that block.

i. P1 has read miss 0 1 0 0 0 (exclusive bit is the last)
ii. P2 has write miss 0 0 1 0 1
iii. P0 has write miss 1 0 0 0 1
iv. P3 has read miss 1 0 0 1 0
v. P3 has write miss 0 0 0 1 1
vi. P2 has read miss 0 0 1 1 0

Q6. (2 marks) Two applications are running on a processor which has shared L2 cache.
For application1: L2 cache misses with 2 and 6 ways (of last level cache) are 4000 and 3600, respectively.
For application2: L2 cache misses with 2 and 6 ways (of last level cache) are 2040 and 1600, respectively
Assume that in between 2 and 6 ways, number of misses scale linearly (i.e., use linear interpolation).
Assume the cache has 8 ways, then which application should get how many ways for minimizing the total number of misses. An application needs to get at least two ways.
Answer: Application1 should get 2 ways and Application2 should get 6 ways [2 marks]. There are different ways to solve it .

Q7.
(a) (1 marks) Three applications P, Q, R have a transactions rate of 44 per minute, 77 per minute and 91 per minute respectively. They run one after another. If each of them make 600 transactions, find the correct average value of transactions per minute. Also write which mean would you use to get the average.
Harmonic mean,
Answer = 3/(1/44+1/77+1/91) = 64.23

(b) (3 marks) We execute 70 instructions in 45 cycles, then 80 instructions in 35 cycles and then 90 instructions in 40 cycles. Show the average computed using both weighted AM and weighted HM. Show the weights used clearly and your computations.
In the question, it was not clear whether we are asking for IPC or CPI. Hence, we will consider any of those answers as correct, as long as student has done its computations correctly. Here, we are showing solution for IPC:
IPC1=70/45
IPC2=80/35
IPC3=90/40
Total number of cycles 45 + 35 + 40 = 120
Total number of instructions 70 + 80 + 90 = 240
Average IPC = 2.

Average Using Weighted A.M:
The weights are with cycles, i.e 45/120, 35/120 and 40/120. [1 mark]
Weighted A.M of IPC will come out to be 2 (=(70+80+90)/120) [0.5 mark]

Average using weighted H.M:
The weights are with instructions i.e 70/240, 80/240 and 90/240. [1 mark]
Weighted H.M of IPC will come out to be 2 (=240/(45+35+40) [0.5 mark]

Q8. (2 marks) An application spends 29% of time in initialization, 39% of time in vision-processing function and remaining time in signal-processing function.
In System0, all the tasks are run on a single-core CPU.
System1 has an signal-processing accelerator and a vision-processing accelerator which give a speedup of 12X and 7X, respectively over the single-core CPU execution.
Find the speedup of system1 over system0 assuming that both the accelerators are used on system1.
In System0, initialization will take 29units, vision processing will take 39 units and signal-processing will take 32 units.
In System1, initialization will take 29units, vision processing will take 39/7 units and signal-processing will take 32/12 units.
Speedup = 100/37.23 = 2.685
Q9. (5 marks) Consider a processor that runs at 3 GHz and 1 Volt. The processor is capable of executing safely at voltages between 0.8 V to 1.2 V. Voltage and frequency follow a linear relationship (i.e., if voltage doubles, frequency doubles as well). When running a given CPU-bound program, the processor consumes 150 W, of which 40 W is leakage. The program takes 40 seconds to execute. Compute the following values (and also show at what

frequency/voltage they are obtained): (i) The smallest time it takes to execute the program (1 mark). (ii) The lowest power to execute the program (2 mark). (iii) The lowest energy to execute the program (2 mark).

(i) For smallest time, we should increase the frequency to highest, which is 3*1.2 GHz. At this, voltage = 1.2V

minimum execution time is = 40/1.2 = 33.33 seconds

(ii) Lowest power occurs at lowest voltage 0.8V.

Static power = 32W, Dynamic power = 56.32W. total power = 88.32W

(iii) Lowest energy occurs at lowest voltage 0.8V. We don't need to solve any quadratic equation because the dynamic power depends on cubic of voltage, so a decrease in dynamic power will weigh over increase in execution time.

Total energy = 88.32W * 40/0.8 = 4416 J

Q10.

(a) (3 mark) We have a small 3-entry, 3-way cache and the block size is 1B. Consider an access stream with addresses

P, Q, R, S, P, Q, R, S, P, Q, R, S

Show whether each of the access is a hit or a miss with (a) LRU (least recently used) replacement policy and (b) MRU (most recently used) replacement policy. (you don't need to show the state of the cache. Just show the hit/miss decision for each access and total number of misses).

(a) With LRU, all accesses will be miss, since reuse distance is higher than associativity => 12 misses.

(b) With MRU policy, 6 misses will happen as follows:

M, M, M, M, H, H, M, H, H, M, H, H

Lesson: in this case, MRU policy is better because it keeps at least some blocks in cache that will be used in future.

(c) (3 mark) Now consider that we use LRU policy and add a victim cache which has just one block. When a block is replaced from L1 cache, it is put in victim cache. If there was an element in victim cache already, it is simply discarded. An element hitting in victim cache is swapped with the LRU element of the L1 cache.

Show the state of cache and victim cache after each access in following format. Also show whether the access led to hit or miss. Access sequence is same as above: P, Q, R, S, P, Q, R, S, P, Q, R, S

After first 4 compulsory misses, no miss will be there, because total size of L1+victim cache is sufficient to hold the whole working set:

Solution diagram is shown below

Q11 [5 marks] Consider a processor with base CPI of 3.

Case 1: The processor has only one level of cache. It has I-cache miss rate of 1% and D-cache miss rate of 3%. Find CPI.

Case 2: The processor has two levels of cache. L1 I-cache miss rate is 1% and L1 D-cache miss rate is 3%. Unified L2 cache has access time of 6ns. Assume that all L1 I-cache misses are hit in L2 cache. For accesses to L2 cache coming due to misses in L1 D-cache, the local miss rate of L2 cache is 4%. Find CPI.

For both cases, main memory access latency (i.e., miss penalty) is 60ns. Loads are 20% and stores are 10% of total instructions. Clock frequency is 2 GHz.

Case 1:

I cache miss overhead = 120 * 0.01 = 1.2

D cache miss overhead = 0.3 * 120 * 0.03 = 1.08

Effective CPI = 3 + 1.08 + 1.2 = 5.28 cycles

Case 2:

I cache miss overhead = 12 * 0.01 = 0.12

D cache miss overhead = 0.3 * (0.03 * 12 + 0.03 * 0.04 * 120) = 0.1512

Effective CPI = 3 + 0.1512 + 0.12 = 3.2712 cycles