CS2323 Computer Architecture 2019

Homework 1

- Your submission should be named as RollNumber_CA_HW1.pdf. For example, if your roll number is cs16mtech11075, then your submission should be cs16mtech11075_CA_HW1.pdf. Except pdf, no other format is acceptable. **10 marks will be deducted for not following these instructions or if you submit a zipped file.**
- If you submit hand-written solution after scanning, make sure all the text is legible.
- The reasoning for obtaining the answer should be clearly shown to obtain full marks. At the same time, be concise.
- There are 6 bonus marks for writing your solution in Latex. You need to include the following line at the beginning of your solution (i.e., in your *.tex file) to get the bonus:
  This document is generated by \LaTeX
- Late Submission Penalty: 20% for each late day (including weekend)

===

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.

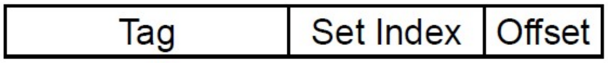
Q2. (1 mark) Write reason why on increasing the associativity, there is only marginal decrease in the miss rate (max 2 sentence).

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

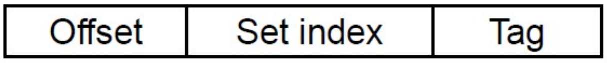
Sequence1: 0, 63, 1, 62, 2, 61, 3, 60, 4, 59, 5, 58, 6, 57, 7, 56, 8, 55, 9, 54, 10, 53, 11, 52

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

Address subdivision method used by Cache 1



Address subdivision method used by Cache 2



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?

(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

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?

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
ii. P2 has write miss
iii. P0 has write miss
iv. P3 has read miss
v. P3 has write miss
vi. P2 has read miss

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.

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.

(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.

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.

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).

Q10.

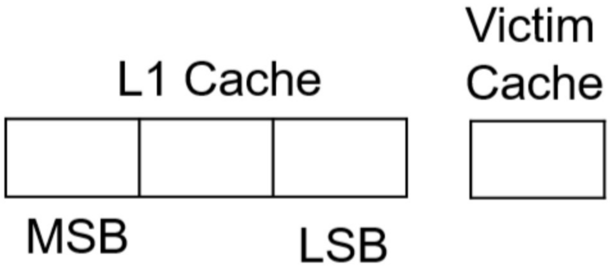
(a) (3 mark) We have a small 3-entry, 3-way cache and the block size is 1B. Consider an access stream with addresses

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).

(b) (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



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.