COSE222: Computer Architecture
Assignment #4

Due: December 15, 2020 (Tuesday) 11:59pm on Blackboard

Solutions (Total score: 121)

Please answer for the questions. Write your student ID and name on the top of the document. Submit your homework with "**PDF**" format only. (You can easily generate the pdf files from Microsoft Word or HWP. You can also handwrite your answers to scan the handwritten documents with "**PDF**" format. You may use the document capture applications such as "Office Lens" for scanning your documents with your smartphones.)

The answer rules:
(1) You can write answers in both Korean and English.
(2) Please make your final answer numbers have two decimal places.
(3) Performance of A is improved by *NN* % compared to performance B if PerfA / PerfB = 1.*NN*.

1. The following code is written in C, where elements within the same **row** are stored contiguously. Assume each element of arrays is a 64-bit integer. The data type of all variables is a 64-bit integer also. Cache blocks are allocated on write miss. (*Hint: in this code, there are 5 variables: a, b, i, j, and sum*)

```
int sum;
for (i = 0; i < 1024; i++)
    sum = 0;
    for (j = 0; j < 1024; j++)
        sum += a[i][j];
    b[i] = sum;
```

(a) Which variable references exhibit temporal locality? [2]

(b) Which variable references exhibit spatial locality? [2]

(c) Let's focus on the data transfers for arrays `a` and `b`. How many `ld` and `sd` instructions are issued while executing this code? [4]

Array a:

Array b:

(d) Let's focus on the cache miss rates for arrays `a` and `b`. Let us assume the cache block size is 32 bytes and elements of arrays `a` and `b` are aligned within a cache block. Calculate the compulsory (cold) miss rate for arrays `a` and `b` respectively. (Hint: assume that the cache size is infinite). [8]

Cod miss rate of A:

Cold miss rate of B:

(e) In this problem let's focus on the miss rates of arrays a and b. Assume the cache block size is 32 bytes and the cache has 128 blocks. The cache is a fully associative cache. Calculate the capacity miss rates for arrays a and b. [8]

Capacity miss rate of A:

Capacity miss rate of B:

2. Below is a list of 64-bit memory address references, given as **word addresses**. (1 word = 64-bits)

`0xFD, 0xBA, 0x2C, 0xB5, 0x0E, 0xBE, 0x58, 0xBF, 0x02, 0x2B, 0xB4, 0x03`

(a) Let us assume a direct-mapped cache has 16 blocks and a single block includes 2 word. What is the size of this cache? [2]

(b) For each of these references, identify the binary word address, the tag, and the index given a direct-mapped cache with 16 one-word blocks. Also list whether each reference is a hit or miss, assuming the cache is initially empty. [6]

Word address	Binary address	Tag	Index	Hit/Miss
0xFD				
0xBA				
0x2C				
0xB5				
0x0E				
0xBE				
0x58				
0xBF				
0x02				
0x2B				
0xB4				
0x03				

(c) Calculate the miss rate (in percentage) of the above cache. [2]

(d) Assume that the hit time of the above cache is one cycle and it takes 8 cycles to read **one** word from the memory (i.e. this value decides the miss penalty of the cache). Calculate the average memory access time (AMAT) of the cache for the given reference stream. [2]

(e) Assume that the direct-mapped cache has the same size of the data memory, but the size of a single cache block is increased to **4 words**. For each of the above references, identify the binary word address, the tag, and the index of the cache. Also list if each reference is a hit or a miss, assuming the cache is initially empty. [6]

Word address	Binary address	Tag	Index	Hit/Miss
0xFD				
0xBA				
0x2C				
0xB5				
0x0E				
0xBE				
0x58				
0xBF				
0x02				
0x2B				
0xB4				
0x03				

(f) Calculate the miss rate (in percentage) of the above cache. [2]

(g) Assume that the hit time of the above cache is one cycle and it takes 8 cycles to read **one** word from the memory (this configuration is the same to the above problem d). Calculate the average memory access time (AMAT) of the cache for the given reference stream. [2]

3. For a direct-mapped cache design with a 64-bit address, the following bits of the address are used to access the cache. (1 word = 64-bits)

Tag	Index	Offset
63-10	9-5	4-0

(a) What is the cache block size (in words)? [4]

(b) How may blocks does the cache have? [4]

(c) What is the ratio between total bits required for such a cache implementation over the data storage bits? Let us assume each cache block includes 1-bit "valid" field. [8]

Beginning from power on, the following **byte-addressed** cache references are recorded.

`0x000, 0x004, 0x010, 0x084, 0x0E8, 0x0A0, 0x400, 0x01E, 0x08C, 0xC1C, 0x0B4, 0x884`

(d) For each reference, complete the following table. "Replace" represents which bytes replaced if any. [6]

Address	Tag	Index	Offset	Hit/Miss	Replace
0x000					
0x004					
0x010					
0x084					
0x0E8					
0x0A0					
0x400					
0x01E					
0x08C					
0xC1C					
0x0B4					
0x884					

(e) What is the hit ratio? [2]

(f) List the final state of the cache, with each valid entry represented as a record of <index, tag, data>. For example, [5]

```
<0, 3, Mem[0xC00]-Mem[0xC1F]>
```

4. Cache access time is usually proportional to the capacity of cache. Assume that main memory accesses take 50 ns and that 36% of all instructions access data memory. The following table shows data for L1 cache attached to each of two processors, P1 and P2.

Processor	L1 size	L1 miss rate	L1 hit time
P1	2 KB	8%	0.5 ns
P2	4 KB	4%	0.8 ns

(a) Assuming that the L1 hit time determines the cycle time for P1 and P2, what are their respective clock rates? [2]

(b) What is the Average Memory Access Time (AMAT) for P1 and P2 (in cycles)? [4]

(c) Assuming a base CPI is 1.0 without any memory stalls, what is the total CPI for P1 and P2? Which processor is faster? When we say a "base CPI of 1.0", we mean instructions complete in one cycle, unless either the instruction access or the data access causes a cache miss. [4]

For the next problems, we will consider the addition of an L2 cache to P1 (to presumably make up for its limited L1 cache capacity). Use the L1 cache capacities and hit times from the previous table when solving these problems. The L2 miss rate indicated is its local miss rate, namely the L2 miss counts divided by the total L2 access counts.

L2 size	L2 miss rate	L2 hit time
1 MB	80%	5.0 ns

(d) What is the AMAT for P1 with the addition of an L2 caches? Is the AMAT better or worse with the L2 cache? [4]

(e) Assuming a base CPI of 1.0 without any memory stalls, what is the total CPI for P1 with the addition of an L2 cache? [4]

5. Consider the following program and cache behaviors.

Data reads per 1000 instructions	Data writes per 1000 instructions	Instruction cache miss rate	Data cache miss rate	Block size (bytes)
250	100	0.3%	2.0%	64

(a) Suppose a CPU with a write-through, write-allocate cache achieves a CPI of 2. Assume that the size of a single write request generated by one data write is 8 bytes (64-bit) for write-through policy. What are the read and write bandwidths (measured by bytes per cycle) between RAM and the cache? (Assume each miss generates a request for one block.) [8]

(b) For a write-back, write-allocate cache, assuming 30% of replaced data cache blocks are dirty. Note that a whole evicted block is written to memory with the write-back cache if the evicted block is dirty. What are the read and write bandwidths needed for a CPI of 2? [8]

6. Let us assume that the size of virtual address is 48-bits and the size of physical memory is 4 GB. Word size is 64-bits and page size is 4 KB. All addresses are byte-addressed.

(a) What is the maximum size of the virtual memory supported by this system? [2]

(b) What is the size of physical address? [2]

(c) Let us assume the TLB has 512 entries and TLB is two-way set associative. Which virtual address bits are used to index the TLB? Which virtual address bits are used as tag of the table? [8]