

Name: Jaivardhan Singh

Roll No: 2021CS10074

(COL 216) Computer Architecture

May 1, 2023

Major Exam

Duration: 120 minutes

(60 marks)

Beware: Be concise in your writing. You can use rough sheets for calculations. But you cannot submit any additional sheet for grading on Gradescope. So make sure you are certain when you write something (after rough work, or use a dark pencil). If you cheat, you will surely get an F in this course.

1. Consider a processor with a 16 Kbyte unified L1 cache. The miss rate for this cache is 3% and the hit time is 2 clock cycles. The processor also has an 8 Mbyte, on-chip L2 cache. 95% of the time, data requests to the L2 cache are found. If data is not found in the L2 cache, a request is made to a 4 Gbyte main memory. The time to service a memory request is 100,000 clock cycles. On average, it takes 3.5 clock cycles to process a memory request. How often is data found only in main memory, and not in either of the two caches? [3 marks]

Miss rate of L1 cache of 3%

Local miss rate of L2 cache = 5%

Data is found in memory and in neither of the two caches $3\% \times 5\%$ of the time

= 0.15% (global miss rate)

Name: Jaivardhan Singh

Roll No: 2021CS10078

2. Each instruction fetch means a reference to the instruction cache and 35% of all instructions reference data memory. Processor A has two 8 Kbyte, L1 caches - one for data and one for instructions. Computer B has a single, unified 16 Kbyte L1 cache that holds both instructions and data. For A, the average miss rate in the L1 instruction cache is 2%, the average miss rate in the L1 data cache is 10%, and the miss penalty for both data and instruction caches is 9 clock cycles. For B, the average miss rate is 3% for the cache as a whole, and the miss penalty is again 9 clock cycles. Which processor has better performance? [3 marks]

I is number of instructions.

Memory stalls in Processor A

$$\begin{aligned} &= \cancel{100\%} I \times 2\% \times 9 + I \times 35\% \times 10\% \times 9 \\ &= I \times 5.5\% \times 9 \end{aligned}$$

Memory stalls in processor B

$$\begin{aligned} &= I \times 3\% \times 9 + I \times 35\% \times 3\% \times 9 \\ &= I \times 1.35 \times 3\% \times 9 = I \times 4.05\% \times 9 \end{aligned}$$

∴ Processor B has better performance (lower number of memory stalls per instruction).

Note: This analysis does not consider the structural hazards ~~caused by a shared cache.~~ ~~caused~~ caused by a shared cache. i.e both memory and IF stages cannot take place simultaneously.

Name: Jaivardhan Singh

Roll No: 2021CS10079

3. The following table gives the parameters for a number of different caches. Your task is to fill in the missing fields in the table. Recall that m is the number of physical address bits, C is the cache size (number of data bytes), B is the block size in bytes, E is the associativity, S is the number of cache sets, t is the number of tag bits, s is the number of set index bits, and b is the number of block offset bits. [4 marks]

Cache	m	C	B	E	S	t	s	b
1.	32	<u>2048</u>	8	1	<u>256</u>	21	8	3
2.	32	2,048	<u>4</u>	<u>4</u>	128	23	7	2
3.	32	1,024	2	8	64	<u>25</u>	<u>6</u>	1
4.	32	1024	<u>32</u>	2	16	23	4	<u>5</u>

256×8

Table 1: Cache organization

$$S = 2^s \quad S_1 = 2^8 = 256 \quad C_1 = S_1 \times B \times E = 2048$$

$$B = 2^b \quad B_2 = 2^2 = 4 \quad C_2 = S_2 \times B_2 \times E_2 \rightarrow E_2 = 4$$

$$S_3 = \log_2 64 = 6 \quad t + s + b = 32 \rightarrow t_2 = 25$$

$$B_4 = \frac{1024}{2 \times 16} = 32 \quad b_4 = \log_2 32 = 5$$

4. A dynamic RAM has a memory cycle time of 64 nsec. It has to be refreshed 100 times per msec and each refresh takes 100 nsec. What percentage of the memory cycle time is used for refreshing? [3 marks]

$$\text{time for refreshing every nanosecond} = 100 \times 100 \text{ ns}$$

$$= 10 \mu\text{s}$$

$$\text{percentage of time spent refreshing} = \frac{10 \mu\text{s}}{1 \text{ ms}}$$

$$= \frac{10 \times 10^{-6}}{10^{-3}}$$

$$= 1\%$$

Name: Jaivardhan Singh

Roll No: 2021C40074

5. Consider a symmetric shared-memory multiprocessor (3 processors sharing a bus) implementing a snooping cache coherence protocol (MSI). For each of the events below, explain the coherence protocol steps (does the cache flag a hit/miss, what request is placed on the bus, who responds, is a writeback required, etc.) and mention the eventual state of the data block in the caches of each of the 3 processors. Assume that X and Y are not in any of the caches at the start of the sequence, the caches are direct-mapped, and blocks X and Y map to the same set in each cache (X and Y cannot co-exist in a cache at any time). [7 marks]

Request	Cache hit/miss	Request on bus	Who responds/ Write Back happens?	Cache 1 state	Cache 2 state	Cache 3 state
P1: Write X	miss	write miss	memory. No write back	M(X)	I	I
P2: Write X	miss	write miss	P1 responds. No write back	I	M(X)	I
P3: Read X	miss	read miss	P2 responds write back	I	S(X)	S(X)
P1: Read X	miss	read miss	Memory responds No write back	S(X)	S(X)	S(X)
P3: Write X	hit	Invalidate	No write back	I	I	M(X)
P3: Read Y	miss	read miss	memory responds write back	I	I	S(Y)
P2: Write Y	miss	read miss write	memory responds no write back	I	M(Y)	I

Table 2: Snoop based Cache Coherence Table

Name: Jaivardhan Singh

Roll No: 2021CS10071

6. You are given the following code to analyze:

```
1 int x[2][128];
2 int i; int sum = 0;
3
4 for (i = 0; i < 128; i++) {
5     sum += x[0][i] * x[1][i];
6 }
```

Assume we execute this under the following conditions: (a) $\text{sizeof(int)} = 4$. (b) Array x begins at memory address $0x0$ and is stored in row-major order. (c) In each case below, the cache is initially empty. (d) The only memory accesses are to the entries of the array x . All other variables are stored in registers. Given these assumptions, estimate the miss rates for the following cases: [10 marks]

A. Case 1: Assume the cache is 512 bytes, direct-mapped, with 16-byte cache blocks. What is the miss rate?

B. Case 2: What is the miss rate if we double the cache size to 1,024 bytes?

C. Case 3: Now assume the cache is 512 bytes, two-way set associative using an LRU replacement policy, with 16-byte cache blocks. What is the cache miss rate?

D. For case 3, will a larger cache size help to reduce the miss rate? Why or why not?

E. For case 3, will a larger block size help to reduce the miss rate? Why or why not?

Read requests = 2 per loop \times no. of loops = 512

Address of $x[0][i] = 4i$, address of $x[1][i] = 4 \times 128 + 4i$
Block no. = $i // 16$, Block no. of $x[1][i] = 32 + i // 16$

A) $512 / 16 = 32$ blocks. Both $x[0][i]$ and $x[1][i]$ map to the same index. assuming reads follow the @ pattern $\left. \begin{array}{l} \text{read } x[0][i] \\ \text{read } x[1][i] \end{array} \right\}$ the two arrays

keep replacing each other in cache and: all reads are misses.
 \therefore miss rate = 100%. $x[0][i]$ brought to cache then $x[0][i+1]$ due to conflict then $x[0][i+2]$ & so on

~~B) There are 64 blocks in cache. They still map to the same block.~~
~~miss rate = 100%~~

C) Each block can hold 4 integers. There are 16 sets. Both $x[0][i]$ & $x[1][i]$ can co-exist due to set assoc. $x[0][4n]$ is a miss followed by 3 hits $x[0][4n+1] \dots 4n+2, \dots 4n+3$. Same for $x[1][4n]$.
 \therefore miss rate = 25%.

D) A larger cache size won't reduce miss rate as we only need a block once. After it has been replaced we don't make any requests from that block i.e. there are no capacity misses.

E) A larger block size will reduce miss rate by increasing the number of hits following a miss. By increasing block size we reduce the number of compulsory misses by reducing the number of blocks we need.

Name: Jaivardhan SinghRoll No: 2021CS10674B) 1024 bytes \rightarrow 64 indices

$$x[0][i] \rightarrow \text{cache block} = \left[(4i // 16) \right] \% 64$$

$$x[1][i] \rightarrow \text{cache block} = \left[32 + (4i // 16) \right] \% 64$$

$x[0][i]$ and $x[1][i]$ always map to different cache blocks.

After $x[0][4i]$ and $x[1][4i]$ are read the next 6 reads of $x[0/1][4i+1/2/3]$ will all be hits as they won't replace each other in cache \rightarrow 75% hit rate, ~~25%~~ 25% miss rate

Name: Jaivardhan Singh

Roll No: 2021CS10071

7. You are writing a new 3D game. You are currently working on a function to blank the screen buffer before drawing the next frame. The screen you are working with is a 640×480 array of pixels. The machine you are working on has a 32 KB direct-mapped cache with 8-byte lines. The C structures you are using are as follows:

```
1 struct pixel{
2     char r;
3     char g;
4     char b;
5     char a;
6 };
7
8 struct pixel buffer[480][640];
9 int i, j;
10 char *cptr;
11 int *iptr;
12 }
```

Assume the following: (a) $\text{sizeof(char)} = 1$ and $\text{sizeof(int)} = 4$ (b) buffer begins at memory address 0. The cache is initially empty. (c) The only memory accesses are to the entries of the array buffer. Variables i, j, cptr, and iptr are stored in registers.

(A) What percentage of writes in the following code will hit in the cache? [4 marks]

```
1 for (j = 639; j >= 0; j--) {
2     for (i = 479; i >= 0; i--){
3         buffer[i][j].r = 0;
4         buffer[i][j].g = 0;
5         buffer[i][j].b = 0;
6         buffer[i][j].a = 0;
7     }
8 }
```

(B) What percentage of writes in the following code will hit in the cache? [3 marks]

```
1 char *cptr = (char *) buffer;
2 for (; cptr < (((char *) buffer) + 640 * 480 * 4); cptr++)
3     *cptr = 0;
```

(C) What percentage of writes in the following code will hit in the cache? [3 marks]

```
1 int *iptr = (int *)buffer;
2 for (; iptr < ((int *)buffer + 640*480); iptr++)
3     *iptr = 0;
4 }
```

A) Each pixel takes 4 bytes. Two consecutive pixels can be stored in ~~the~~ a cache block at once pixel[x][y] + pixel[x][y+1]. However we traverse the array by column so two consecutive iterations of the inner loop write to pixels in different blocks, by the time we write to its neighbouring pixel (after the outer loop completes an iteration) the block will already be replaced in the cache. So Every pixel generates a cold miss when its first char is written and then 3 hits as the other 3 bytes are already in the cache.
75% hit rate. * continued on the other side.

B) When cptr points to address 8i the entire block from 8i to (i+7) is brought in the cache. The next 7 writes are to these bytes. So 1 cold miss leads to 7 hits. $\frac{7}{8}$ hit rate
 $= 87.5\%$

Name: Jaivardhan Singh

Roll No: 2021C110074

c) When the integer at address i is brought in the cache, the integer at address $i+4$ also comes with it. \therefore A miss at i is followed by a hit at $i+4$. \therefore Hit rate = $\frac{1}{2} = 50\%$

A). $\frac{32 \text{ KB}}{8 \text{ B}} = 4096 \text{ blocks.}$

Address of buffer $[i][j] = 4 \times (j + 1024 \times i)$

Block of buffer $[i][j] = 256i + (4j \% 8)$

Index in cache of buffer $[i][j] = [256i + (4j \% 8)] \% 4096$

\therefore buffer $[i][j]$ is replaced by buffer $[i+64][j]$ and then by buffer $[i+128][j]$ ~~and then~~. These will then be replaced by entries from the next iteration of the outer loop. ~~At~~ ~~no point do we get~~

a write hit from a block fetched in the previous iteration of the outer loop. i.e. buffer $[i+128][j]$

& buffer $[i][2j+1]$ as it would have been overwritten by either $i+64$ or $i-64$

depending on the value of i .

Note: The loop actually goes from higher i to lower
but that does not change the answer. ~~only~~

$$128 + 2^1 - 127 = 4$$

$$128 - 127 = -1$$

$$\therefore 2^2 \times 2^3$$

$$1.01$$

Name: Jaivardhan Singh

Roll No: 2021CS10074

8. Choose the correct answer or write short answers to the following questions. [20 marks]

(a) Consider the IEEE-754 single precision floating point numbers $P = 0xC1800000$ and $Q = 0x3F5C2EF4$. Which one of the following corresponds to the product of these numbers represented in the IEEE-754 single precision format? [3 marks]

- (a) $0x404C2EF4$ (b) $0x405C2EF4$ (c) $0xC15C2EF4$ (d) $0xC14C2EF4$

Ans \rightarrow (c)

$$0xC1800000 = -2^4 (128 + 2^1 - 127)$$

\therefore we simply increment exponent of $0x3F5C2EF4$ by 4 sign = negative.

Binary representation of $0x3F5C2EF4$:
 $01001100001111101011100010111011110100$
 0000×5
 110000010101

(b) Consider a 3-stage pipelined processor having a delay of 10 nanosecs, 20 nanosecs, and 14 nanosecs for the first, second, and the third stages, respectively. Assume that there is no other delay and the processor does not suffer from any pipeline hazards. Also assume that one instruction is fetched every cycle. The total execution time for executing 100 instructions on this processor is 2040 nanosecs. [2 marks]

Cycle time = longest stage = 20 ns.

No. of cycles = $100 + 2$ initial cycles = 102 cycles.
 time = $102 \times 20 = 2040$ ns.

(c) "False sharing occurs only if a cache block contains multiple words" - True or False? Why? [2 marks]

Assuming word level addressing.
 True. False sharing occurs when we modify a certain word in memory which invalidates all other copies of that block. When another processor tries to read another word from the same block it generates a cache miss even though the value of that word in the invalidated block hadn't been changed by the write. This cannot occur if only one word is in the block as all writes would be to that block.

(d) In a MSI coherence protocol, when is a cache controller forced to write back a block, B? [3 marks]

A block will be written back in two cases.

- 1) The cache has that block in M state and receives a read miss for that block on the bus. It will have to switch that block to shared state, send it on the bus and write back to memory.
- 2) The cache has that block in M state and the processor generates a read/write miss for that cache index. i.e. the block has to be replaced in the cache. In this case block B is written back to memory & the new block is put in its place in the cache.

Name: Jaivardhan Singh

Roll No: 2021CS10079

- (e) Consider a processor with 64 registers and an instruction set of size twelve. Each instruction has five distinct fields, namely, opcode, two source register identifiers, one destination register identifier, and twelve-bit immediate value. Each instruction must be stored in memory in a byte-aligned fashion. If a program has 100 instructions, the amount of memory (in bytes) consumed by the program text is 500. [3 marks]

one register number takes ~~6~~ 18 bits. \therefore = 18 register bits (6x3)
12 instructions \therefore we need atleast 4 opcode bits &
12 bit immediate value.

total instruction size = $18 + 12 + 4 = 34$ bits.
Since the instructions are stored in a byte aligned fashion they take memory in multiples of 8. Each 34 bits instruction takes 40 bits = 5 bytes. 100 instr = $100 \times 5 = 500$ B

- (f) Consider a 3 GHz processor with a three-stage pipeline and stage latencies τ_1 , τ_2 and τ_3 such that $\tau_1 = 3\tau_2/4 = 2\tau_3$. If the longest pipeline stage is split into two pipeline stages of equal latency, the new frequency is 4 GHz, ignoring delays in the pipeline registers. [3 marks]

τ_2 is the longest stage. $\tau_1 = \frac{3\tau_2}{4}$ $\tau_3 = \frac{2\tau_2}{8}$

After splitting

$$\tau_1 = \frac{3\tau_2}{4} \quad \tau_2' = \frac{\tau_2}{2} \quad \tau_2'' = \frac{\tau_2}{2} \quad \tau_3 = \frac{2\tau_2}{8}$$

new longest stage = $\tau_1 = \frac{3\tau_2}{4}$.

$$\text{New freq} = \frac{1}{\tau_1} = \frac{4}{3} \frac{1}{\tau_2} = \frac{4}{3} \times \text{old freq} = 4 \text{ GHz}$$

- (g) Tick all that apply. Concepts taught in class to improve program performance are: [2 marks]

(a) pipelining (b) branch prediction (c) pipeline stalls (d) caching (e) cache coherence

a) b) d)

- (h) Tick all that apply. Concepts taught in class to maintain program correctness are: [2 marks]

(a) pipelining (b) branch prediction (c) pipeline stalls (d) caching (e) cache coherence

c) e)