

Name: Santhosh Rishi Keshireni

Roll No: 2021CS10564

(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]

→ We need to find the % of times, data is found in main memory and not in either cache.

$$\% = \% \text{ of miss in cache-1} \times \% \text{ of miss in cache-2}$$

$$= \frac{3}{100} \times \frac{5}{100}$$

$$= \frac{0.15}{100} \quad \boxed{= 0.15\%}$$

Name: Santhosh Rishi Reshineni

Roll No: 2021CS10564

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]

→ let us consider the access time during a hit to be x cycles. The processor with lesser avg CPI has ~~best~~ better performance.

$$CPI_A = x + \frac{2}{100} \times 9 + \frac{10}{100} \times \frac{35}{100} \times 9 + \frac{35}{100} \times x$$

$$CPI_B = x + \frac{3}{100} \times 9 + \frac{3}{100} \times \frac{35}{100} \times 9 + \frac{35}{100} \times x$$

$$CPI_A = \cancel{1.35}x + 0.495$$

$$CPI_B = \cancel{1.35}x + 0.3645$$

since $CPI_B < CPI_A$, processor B has better performance

Name: Santhosh Rishi Deshineni

Roll No: 2021CS10567

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	_____	8	1	_____	21	8	3
2.	32	2,048	_____	_____	128	23	7	2
3.	32	1,024	2	8	64	_____	_____	1
4.	32	1024	_____	2	16	23	4	_____

Table 1: Cache organization

1.	32	2048	8	1	256	21	8	3
2.	32	2048	4	4	128	23	7	2
3.	32	1024	2	8	64	25	6	1
4.	32	1024	32	2	16	23	4	5

$$B = 2^b \quad m = t + s + b$$

$$S = 2^s \quad C = S \times E \times B$$

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]

→ DRAM has to be refreshed 100 times/msec or 10^5 times/sec. So $10^5 \times (64 \times 10^{-9})$ time every memory cycle.

time taken in 1 memory cycle.
 $= 10^5 \times (64 \times 10^{-9}) \times 100 \times 10^{-9}$ sec

$$\text{percentage} = \frac{10^5 \times (64 \times 10^{-9}) \times 100 \times 10^{-9}}{(64 \times 10^{-9})} \times 100$$

$$= 1\%$$

Name: Santhosh Rishij Deshinemi

Roll No: 2021CS10564

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	P1 write X	mem responds no wrbk	M(X)	I	I
P2: Write X	hit miss	no request P2 write X	no one P1 responds no wrbk	M(X) I	I M(X)	I
P3: Read X	miss	P3 read X	P2 responds wrbk occurs	S(X) I	S(X)	S(X)
P1: Read X	miss	P1 read X	mem responds no wrbk	S(X)	S(X)	S(X)
P3: Write X	hit	P3 write X	no one responds no wrbk	M(X) I	I	I M(X)
P3: Read Y	miss	P3 read Y	mem responds X wrbk	I	I	S(Y)
P2: Write Y	miss	P2 write Y	mem responds no wrbk	I	M(Y)	I

Table 2: Snoop based Cache Coherence Table

Name: Santhosh Rishi Peshineri

Roll No: 2021CS10054

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) $\text{sizeof}(\text{int}) = 4$. (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?

A. Note that the size of the cache is 512 bytes and the size of the array is $128 \times 2 \times 4$ bytes = 1024 bytes. This means that the cache can fully fit one row of the array. However, $x[0][i]$ & $x[1][i]$ map to the same set because of this. Therefore, we always have a miss as they visit each other.

\therefore miss rate = 1

B. In this case, they map to different sets. Therefore, a miss occurs after each block in memory that is one every 4 reads.

\therefore miss rate = $\frac{1}{4}$

Name: Santhosh Rishi Keshireni

Roll No: 2021CS10087

C. The cache now has 16 sets, 2 lines per set and 16 byte blocks (4 ints fit in a block). In other words, the first 4 ints map to 0th set, then 1st set until ⁽⁵⁰⁻⁶³⁾ 64 ints mapping to 15th set. This then repeats for next 64 ints in a row-major order. Although, both 1st & 2nd row map to same set, 2 lines allow both to fit in cache. Thus, misses only at new occurrence.

$$\boxed{\text{miss rate} = \frac{1}{4}}$$

(Replacements will occur in cache after $i=63$)

~~E~~

D. No, a larger cache size will not reduce the miss rate. A larger cache helps when there are conflict misses as in case A. The misses in C occur only at first occurrence of a block and not subsequently. Thus, increasing or even doubling the cache will still keep miss rate at $\frac{1}{4}$.

E. Yes, a larger block size will reduce the miss rate. A larger block implies that we can go longer without a new memory block. (more iterations). Since, we have 2 lines and we are working with 2 ints, the miss rate will be: $\frac{1}{(\text{no. of ints in 1 block})}$. Thus, increasing block size helps here.

Name: Santhosh Rishi Deshineni

Roll No: 2021CSE0564

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) The cache has 4096 sets and 8-byte lines or 8-byte blocks that is 2 structs per block. It can fit 8192 structs. One write miss will bring in 2 structs and cause 7 write hits later in the loops.
 \therefore % of writes which hit = $\frac{7}{8} = 87.5\%$

Note that the structs in a column map to different sets.

Name: Sonthush Rishi Raghini

Roll No: 2021CS10054

(B) Code corresponds to iterating over each element ~~from~~ ⁱⁿ order of over array in row-major order. There will continuously be 1 write miss followed by 7 write hits. That is the 1st character write will bring in the next 7 characters.
 \therefore write hit % = $\frac{7}{8} = 87.5\%$

(C) Code corresponds to iterating over array in row-major order in chunks of 4-bytes. 1 write miss will be followed by 1 hit (as a block contains 8-bytes) and this repeats
 \therefore write hit % = $\frac{1}{2} = 50\%$

Name: Santhosh Rishi Reddineni

Roll No: 2021CS10564

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) $0xC14C2EF4$

$$P = -1 \times 2^4 \quad Q = + (1 + F) \times 2^{-1}$$
$$P \times Q = - (1 + F) \times 2^3$$

(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 _____ nanosecs. [2 marks]

$$\rightarrow (100 + 3 - 1) \times 20 \text{ nsec}$$
$$= 2040 \text{ nsec}$$

\rightarrow 2040 nsec

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

\rightarrow ~~True~~, ~~True~~ false sharing occurs to ensure that there is only 1 word in every line. If we initially already have 1 word/block then false sharing isn't needed.

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

1) when it is in modified state and ~~B~~ is being evicted
(since it has the only valid copy)

2) when it is in modified state and another processor wants to read something from this block, B.

Name: Santhosh Rishi Keshivani

Roll No: 2021CS10884

- (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 _____ [3 marks]

500 bytes

no. of bits in instruction
 $= 4 + 6 + 6 + 6 + 12 = 34$ bits

5 bytes (byte-aligned)

100 instructions \rightarrow 500 bytes

- (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 _____ GHz, ignoring delays in the pipeline registers. [3 marks]

4 GHz

$\frac{1}{3 \times 10^9} \rightarrow \tau_2$
 $\tau_1 \rightarrow \tau_3$

\Rightarrow

$\frac{\tau_2}{\tau_3} = \frac{1}{1 \times 3 \times 10^9}$

$\tau_1 = \frac{1}{4 \times 10^9} \Rightarrow 4 \text{ GHz}$

- (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

- (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