EECS 470 *Final Exam*

Fall 2021

Name: ____________________KEY____________________ unique name: ________________

Sign the honor code:

I have neither given nor received aid on this exam nor observed anyone else doing so.

____

Scores:

NOTES:

- Open book and Open notes
- Calculators are allowed, but no PDAs, Portables, Cell phones, etc.
- Don't spend too much time on any one problem.
- You have about 120 minutes for the exam.
- There are ***10*** pages including this one.
- Do not write on the back of any pages. Anything you want graded must be on the page of the question.
- **Be sure to show work and explain what you've done when asked to do so.**
- **The last page has two "answer areas" for the last question. Clearly mark which one you want graded or we will grade the first one.**

1. Pick and hope [12 points, -2 per wrong/blank, minimum 0]

Circle the best answer.

1) In the MESI protocol taught in class, which of the following transitions might happen due to a transaction by another processor?

M → E   S → E   I → S   [✓ M → S]   S → M

2) You would expect a 1024-byte direct-mapped cache with a block size of 64 bytes to get about what hit rate on a memory access with a stack distance of 2?

0%   80%   [✓ 88%]   92%   96%   100%

3) In the R10K scheme, if you have a ROB size of 16, RS size of 8, and ARF size of 32, what is the maximum number of PRF entries you would expect to have?

40   [✓ 48]   50   56   70   80

4) With what type of cache type can you be certain you won't need dirty bits?

write-back   write-allocate   2-way associative   [✓ write-through]   skew

5) Say you have a 4KB, 2-way associative write-back cache with 32-byte blocks and a 40-bit address space. How many bits would you need to index this cache?

14   12   10   8   [✓ 6]   4

6) Say you have a pipelined multiplier which is on the critical path of your processor. If you increase the number of pipeline stages you would expect which of the following?

- **The clock period and the CPI would go up.**
- **The clock period would go up and the CPI would go down.**
- **[✓ The clock period would go down and the CPI would go up.]**
- **The clock period and the CPI would go down.**

7) In the P6 scheme taught in class, the RAT points to a ________ entry.

PRF   ARF   RS   [✓ ROB]   CDB

8) Say you had an ISA where every instruction was predicated where any GPR can be used as a predicate. How many *more* instruction encodings would be needed for an add instruction (Rd=Ra+Rb) than in a non-predicated ISA? Assume there are 32 GPRs in both ISAs.

[✓ $2^{20}-2^{15}$] 2^{15} 2^{20} $2^{20}+2^{15}$ 2^{5} 2^{35}

2. It’s looking MESI out there [13 points, -1 per wrong box]

Consider a case of having 2 processors using a snoopy MESI protocol where the memories can snarf data. Both have a 2-line direct-mapped cache with ***each line consisting of 16 bytes***. The caches begin with all lines marked as invalid. Fill in the following tables indicating:

- If the processor gets a hit or a miss in its cache.
- What bus transaction(s) (if any) the processor performs (BRL, BWL, BRIL, BIL)
- If a HIT or HITM (or nothing) occurs on the bus during snoop.
- For misses only, indicate if the miss is compulsory, capacity, conflict, or coherence. A coherence miss is one where there would have been a hit, had some other processor not interfered.

Finally, indicate the state of the processor after all of these memory operations have completed. The operations occur in the order shown. **[15 points, -0.5 per wrong or blank, minimum of 0]**

Processor	**Address**	**Read/Write**	**Cache Hit/Miss**	**Bus transaction(s)**	**HIT/ HITM**	**“4C” miss type (if any)**
1	0x00	Write	Miss	BRIL		Compulsory
1	0x14	Read	Miss	BRL		Compulsory
2	0x10	Read	Miss	BRL	HIT	Compulsory
1	0x06	Write	Hit			
2	0x19	Write	Miss	BIL	HIT	Coherence
1	0x10	Write	Miss	BRIL	HITM	Coherence
1	0x50	Write	Miss	BRIL/BWL		Compulsory
1	0x30	Read	Miss	BRL/BWL		Compulsory
2	0x00	Read	Miss	BRL	HITM	Compulsory
1	0x50	Read	Miss	BRL		Conflict

Final state:

	Proc 1	
	Address	**State**
Set 0	0x00	S
Set 1	0x50	E

	Proc 2	
	Address	**State**
Set 0	0x00	S
Set 1		I

3. Caching in [8 points]

Consider the following C-code segment:

```
          char A[4096];            // each element is 1 byte
          for(j=0;j<100000;j++)
               for(i=0;i<Y;i=i+X)
                    A[i]=A[i]+1;
```

Assume that only accesses to the array A go to the data cache (the other values are in registers). For this code, what would be the expected *hit-rate* for the various values of X and Y if the data cache were 1 KB with 32-byte lines that was direct-mapped?

[-2 per wrong/blank box, min 0]

	X=2	X=4	X=64
Y=2048	15/16	7/8	0
Y=1025	511/513	╳	15/17

4. ISA vs Microarchitecture [6 Points]

Indicate whether each of the following design choices in a processor is typically a feature of the ISA or of the microarchitecture. For each write either "ISA" or "MA". -1.5 points per wrong or blank answer.

___MA_ A 64-bit wide data bus to memory

___MA_ Gshare branch prediction

___ISA_ Predicated instruction execution

___ISA_ A floating-point unit that uses wide floating-point values for additional accuracy Initials

___ISA_ An additional set of user-visible registers

___ISA_ 32-bit instructions

___MA_ 64KB L1 cache

___MA_ 5-stage pipeline

5. Copypasta processors [14 points]

HobTech, a small startup in Ann Arbor, has asked you to design a multicore system with different processor sizes (asymmetric multicore). Your design is to consist of one large processor with the rest of the die are being taken up by small cores.

You have three cores available for use. These numbers include caches and all uncore components needed for the cores to function correctly.

Core	Area	Performance	Dynamic Power	Static power
A	$20mm^2$	7 GIPS	50 Watts	10 Watts
B	$10mm^2$	4 GIPS	25 Watts	5 Watts
C	$2mm^2$	1 GIPS	1 Watt	1 Watt

You have a total of $30mm^2$ area on your chip.

a) Say HobTech wants to optimize performance for a program that is 90% perfectly parallel and 10% perfectly serial. How should you allocate your selection of cores? That is, how many of each would result in the best performance? To be clear, when running the serial part, nothing else can run at the same time and when running the parallel part, all processors can be utilized. Justify your answer **[6]**

Say we have 9 GI in parallel and 1 GI in series
A & 5C: 1GI / 7 GIPS + 9GI / 12GIPS = .89s
B & 10C: 1GI / 4GIPS + 9GI/ 14GIPS = .89s
A & B: 1GI / 7GIPS + 9GI / 11GIPS = .96s
Tie between A & 5C and B & 10C

b) Say HobTech has limited you two configurations:

- **Configuration X:** 1 A core and the rest C cores
- **Configuration Y:** 1 B core and the rest C cores.

If the workload we run has a fraction S of the work be perfectly serial and (1-S) be perfectly parallel, for what values of S will Configuration X use less energy to accomplish the same work? Show your work. You should assume that when a core is unused we can't shut it down, but we can insure no transistors are switching. Clearly state and justify any assumptions you are making. **[8]**

Say we have 1GI instruction total
We want to minimize the J/GI = watts / GIPS. Serial execution will always be the static power of C cores plus the static and dynamic power of A/B. For parallel, you can choose to shut off as many cores as desired.
X - Parallel: 70W / 12GIPS (all cores on) or 20W / 5GIPS (A off). 20W / 5GIPS smaller
Y - Parallel: 50W / 14GIPS (all cores on) or 25W / 10GIPS (B off). 25W / 10GIPS smaller
E_X = 65W * s / 7GIPS + 20W * (1-s) / 5 GIPS = 5.29 * s + 4
E_Y = 40W * s / 4GIPS + 25W * (1-s) / 10 GIPS = 7.5 * s + 2.5
s = .677
E_X < E_Y when s > .677 (see last page for more).

6. Short answer—numbers and letters [12 points]

You must briefly show your work to receive credit.

a) Given a 40-bit virtual memory system with a 32-bit physical memory and a physically-addressed L1 cache that is 64KB, where each line has a 25-bit tag, what is the minimum degree of associativity of the cache? **[4]**

32-25-2 = 5 index bits
(64kB/4B) / 2^5 = 2^9 way associative

b) Say your out-of-order processor has 32 RoB entries, 8 RS entries, and the ISA supports 16 architected registers. How many bits would you need for the RRAT (just the pointers in the table, not valid bits, the free list or anything else)?

i. If you are using the P6 scheme taught in class? **[2]**

32 = 2^5. 5 bits * 16 reg = 80 bits

ii. If you are using the R10K scheme taught in class? **[2]**

PRF = 48 = 2^6. 6 bits * 16 = 96 bits

c) Say you have an unified LSQ with non-speculative load-to-store forwarding. Each instructions reads/writes 4 bytes. Which of the following loads are able to issue (either from a store forward or to the memory)? **[4]**

Slot	**LSQ Addresses**
A(head)	Store 0x20
B	**Load 0x40**
C	**Load 0x20**
D	**Load ???**
E	Store ???
F	Store 0x34
G	**Load ???**
H	**Load 0x50**
I(tail)	**Load 0x34**

B,C,I
____
Place answer on line above

7. Short answer—words [12 points]

You must answer each question in 20 words or less.

a) Loads and Stores. Answer each question in 20 words or less.

i) Why does a complier have to be careful about hoisting a load above a store? **[3]**

If there is a store->load address conflict then the load could have the incorrect value if hoisted.

ii) How does a dynamic out-of-order processor deal with the same issue? **[3]**

LSQ only sends loads to memory when there are no stores in front of them that they're dependent on.

b) A RAS helps predict where returns will branch to. Why do function return statements require a special structure but function calls do not? **[3]**

Function calls are handled by the branch predictor instead, and since returns are dependent on the initial jump location, they cannot be handled by just a branch predictor.

c) What is the main advantage of a virtually addressed cache over a physically addressed one? **[3]**

Since you don't need to translate the virtual address to a physical one before the cache lookup in a virtually addressed cache, the latency of a cache lookup will be lower.

8. I have no memory of that. [8 points]

You are back to working for HobTech, and they want you to design a processor using a memory technology that supports *very* low memory bandwidth. Which of the following options would you likely choose to deal with this limitation? Very briefly justify each answer.

a) RISC or CISC ISA? **[2]**

CISC because it is usually more code dense.

b) Large cache blocks or small cache blocks? **[2]**

Small cache blocks, so you reduce evictions and only retrieve the memory you need.

c) Store-to-Load forwarding or no load-to-store forwarding? **[2]**

Yes, as the LSQ will then reduce the number of loads eventually sent to memory.

d) Prefetching or no prefetching **[2]**

No, while it helps performance, it will potentially fetch useless lines such as in the case of a branch mis predict which will result in more fetches being sent to memory.

Also correct: Yes, because if the prefetcher is given the lowest priority, it can utilize the bandwidth when it would have otherwise gone unused.

9. It’s All about the Pentiums [15 points]

Consider the following tables that represent the state of a processor that implements what we have called the P6 scheme:

RAT

Arch Reg. #	ROB# (-- if in ARF)
0	--
1	~~4~~ 9
2	~~5~~ --
3	~~2~~ --
4	--
5	~~--~~ 8

ROB

Buffer Number	PC	Done with EX?	Dest. Arch Reg #	Value	
0	~~20~~	~~Y~~	~~1~~	~~12~~	← Tail
1	~~24~~	~~N~~	~~1~~	~~--~~	
2	~~28~~	~~Y~~	~~3~~	~~-1~~	
3	~~32~~	~~Y~~	~~--~~	~~--~~	
4	~~36~~	~~Y~~	~~1~~	~~14~~	
5	~~40~~	~~N~~	~~2~~	~~--~~	
6	~~100~~	~~Y~~	~~3~~	~~0~~	
7	104	N	1	--	← Head
8	108	N	5	--	
9	112	N	1	--	

RS

RS#	Op type	Op1 ready?	Op1 RoB/value	Op2 ready?	Op2 RoB/value	Dest ROB
0	~~*~~	~~Y~~	~~12~~	~~Y~~	~~-1~~	~~1~~
1	~~+~~	~~N~~	~~1~~	~~Y~~	~~12~~	~~2~~
2	~~+~~	~~Y~~	~~-1~~	~~Y~~	~~-1~~	~~5~~
3	+	Y	-1	N	7	8
4	*	Y	3	N	7	9

ARF	**Reg#**	**0**	**1**	**2**	**3**	**4**	**5**
	Value	1	~~2~~ -12	3	~~4~~ 0	6	-1

The instruction at PC 32 is a branch that has been predicted not-taken, but it is actually taken. The destination of the branch is PC 100, where the following code resides:

```
R3=R3+R0              // A
R1=R1+R3              // B
R5=R5+R1              // C
R1=R2*R1              // D
```

Show the state of the above tables if instruction A has retired, instruction B has been issued but has not finished execution, while C and D have progressed as far along as possible. ***Be sure to label the head and tail of the ROB***. Please place instruction A in slot 6 of the ROB. When other arbitrary decisions need to be made, you are to just make them. **[15 points]**

Notes:

On problem 6 if you don’t think to turn off the main processor when doing stuff in parallel (for energy savings) you get a different answer: 0.76 exactly.