CS/ECE 752: Advanced Computer Architecture I

Prof. David A. Wood

Final Exam
December 18, 2008
Approximate Weight: 20%

CLOSED BOOK
TWO SHEETS OF NOTES

NAME: _____

DO NOT OPEN THE EXAM UNTIL TOLD TO DO SO!

Read over the entire exam before beginning. Verify that your exam includes all 9 pages. Budget your time according to the weight of the questions, and your ability to answer them. Limit your answers to the space provided, if possible. If not, write on the BACK OF THE SAME SHEET. Use the back of the sheet for scratch work. WRITE YOUR NAME ON EACH SHEET.

Problem	Possible Points	Points
Problem 1	10	
Problem 2	15	
Problem 3	30	
Problem 4	15	
Problem 5	15	
Problem 6	15	
Total	**100**	

NAME: _____

Problem 1: (10 points)

Circle either the word True or the word False for the questions below.

True or False: Compiler optimizations such as loop unrolling and if-conversion can improve program performance on both statically scheduled processors and dynamically scheduled processors.

True or False: Loop unrolling is the preferred optimization if a loop has a loop-carried dependence.

True or False: Blocking (aka tiling) is a compiler optimization that primarily tries to increase a program's spatial locality.

True or False: The Sun Niagara processor uses the ICOUNT policy to select the which thread should run next.

True or False: Wider interfaces to DRAM are the prefered way to achieve higher bandwidth in modern systems.

True or False: Using open DRAM pages always reduces latency, by acting as a small cache.

True or False: Stream buffers are one way to implement non-binding prefetching.

True or False: Chaining in a vector processor is basically the same as bypassing in a scalar pipeline.

True or False: TLB Reach can be increased either by increasing the page size or by increasing the number of TLB entries, or both.

True or False: Albonesi described a technique to dynamically disable some of the sets in a cache to reduce power dissipation at the expense of performance.

The following questions **WILL NOT** impact your grade, but I'm interested in your answers to help improve the course next time I teach it.

True or False: I wish CS752 would include more information on multiprocessors.

True or False: I wish CS752 would focus more on reliability and power management.

True or False: I wish CS752 would include more systems architecture.

True or False: I wish we had spent more time talking about the papers we read.

True or False: I would like having the Powerpoint slides as a reference, but wish the lecture was more of a discussion, using the chalkboard, not slides, to illustrate key points.

Problem 2: (15 points)

The IA-64 instruction set has many features to help the compiler to a better job of scheduling (i.e., optimizing) the code to eliminate stalls. Compare and contrast the support for *speculative loads* and *advanced loads*. Describe the new instructions and any *architectural* state needed to support them. Give examples of what problems these new loads address and how they help.

NAME: _____

Problem 3: (30 points)

Consider a memory hierarchy with the following components and properties:

- An L1 data cache with 64 KByte capacity, 64 byte blocks, direct-mapped placement, a write-through policy, and physical tags. An L1 hit stalls the in-order pipeline for 2 cycles (load-use delay).
- A two-entry victim cache (i.e., victim buffer) between the L1 and L2 with fully-associative placement, LRU replacement, and a "swap-on-hit" policy. A reference that misses in the L1 cache but hits in the victim cache stall a total of 4 cycles.
- An L2 unified cache with 2 Mbyte capacity, 128 byte blocks, 4-way set-associative, LRU replacement policy, a writeback policy, and physical tags. A reference that misses in both the L1 cache and victim cache but hits in the L2 cache stall a total of 15 cycles.
- A main memory system with 4 Gbyte capacity. A reference that misses in all caches and is satisfied by the main memory stall a total of 250 cycles.
- Physical addresses are 32 bits and the smallest addressable unit is one byte.
- Virtual addresses are 32 bits and pages are 8Kbytes.
- Address translation is performed in parallel with the L1 cache via a a 64 entry, fully-associative TLB.

Part A: (10 points)

How many bits are required to implement each of the components described above? Be sure to identify and include the state needed to implement the various policies. Complete the table below. Show your work in the space below the table, using additional pages if necessary.

Table 1:

Cache	Bits per Block	Blocks per set	Bits per set	Total bits in cache
L1 Cache				
Victim Cache				
L2 Cache				

NAME: _____

Part B: (20 points)

Consider only the L1 data cache and victim cache and assume that all entries in both caches are initially invalid. For the memory reference stream below, determine which references hit or miss at each level (accessed from top to bottom). Assume that the victim cache is only accessed on L1 misses and all accesses that miss in the victim cache will hit in the L2 cache. Assume byte addressing and all accesses are 32-bit loads.

Table 2:

Physical Address	L1 Index	Hits in?	Victim Cache Contents MRU, LRU
0x00000000	0	L2	invalid, invalid
0x00010010			
0x00020020			
0x00040040			
0x00000000			
0x00010010			
0x00020020			
0x00040040			
0xEEE00000			
0x00000000			
0x00010010			
0x00020020			
0x00040040			

Complete the tables to the left. In Table 2, indicate in which cache each reference hits. In Table 3, indicate the total number of hits, misses, and stall cycles for each cache.

Hint: Calculate the L1 index bits for each access in the middle column. Then track which entries are added to or removed from the victim cache.

Table 3:

Cache	Number of Hits	Number of Misses	Hit Latency	Total Stall Cycles
L1 Cache			3	
Victim Cache			5	
L2 Cache			15	
All				

Problem 4: (15 points)

Part A: (4 points)

Explain what *virtual address synonyms* (also called *aliases*) are and how they can arise.

Part B: (4 points)

Explain what problem(s) virtual address synonyms can cause in a cache memory system.

Part C: (7 points)

Can virtual address synonyms cause problems in the L1 data cache described in Problem 3? If so, explain why and what can be done to prevent or mitigate the problems.

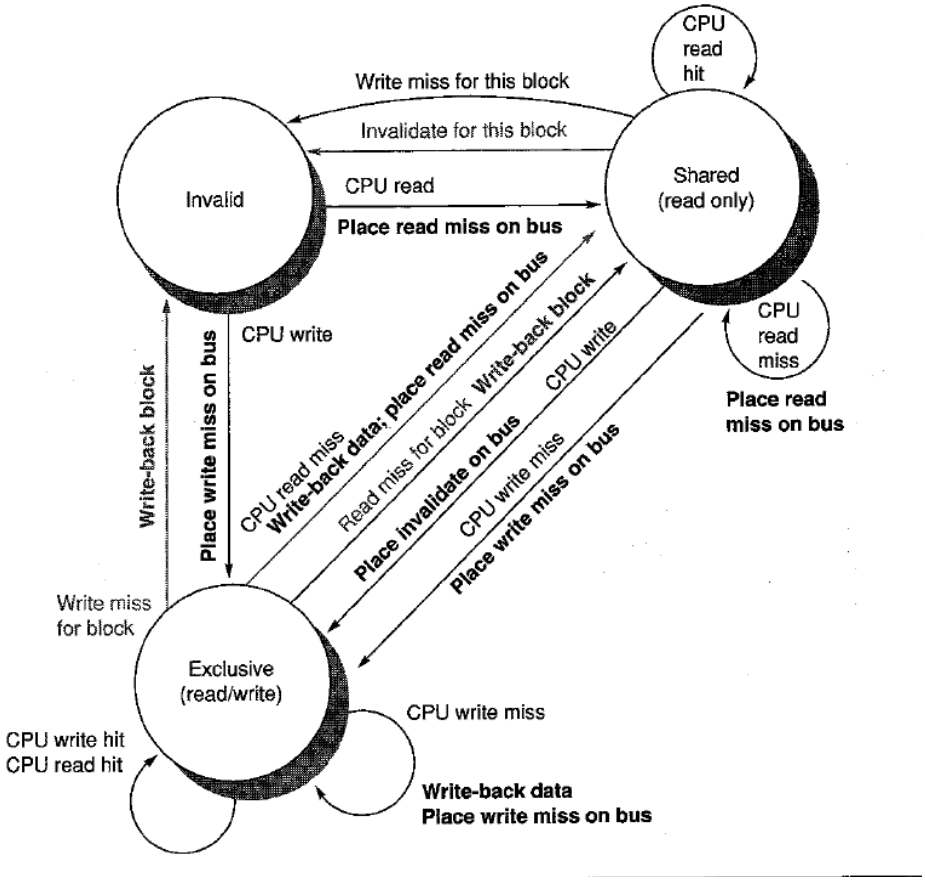
NAME: _____

Problem 5: (15 points)

Consider a multiprocessor system that uses broadcast snooping cache coherence. Four processors, P1, P2, P3, and P4 perform the following sequence of loads and stores to/from lines A and B. Assume A and B do not conflict in the data caches. Assume the protocol described in the book (reprinted on the next page), which uses the three states: Invalid (I), Shared (S), and Exclusive (E). The table below shows the state of the memory system as time flows down. The cache blocks are represented with the following notation: *address:(state, data)*. For example, *A:(I,0)* means that cache block A is in state I with data value 0. A data value of x means the value is unknown or undefined. Complete the table below, updating the cache and memory states in response to the sequence of loads and stores. Indicate actions taken by the cache and memory controllers: hits, requests to get a block shared or exclusive, and responses to requests. You may use arrows (as shown) to indicate that the state has not changed in that cycle.

P1	P2	P3	P4	MEM
A:(I,x) B:(I,x)	A:(I,x) B:(I,x)	A:(I,x) B:(I,x)	A:(I,x) B:(I,x)	A:0 B:0
load A miss, get A shared A:(S,0) B:(I,x)	↓	↓	↓	respond with A A:0 B:0
↓	**load B** miss, get B shared A:(I,x) B:(S,0)	↓	↓	respond with B A:0 B:0
	load A			
	store B = 1			
		load A		
load B				
			store A = 3	
load A				
			load B	

NAME: _____


CPU read hit
Write miss for this block
Invalidate for this block
Invalid
CPU read
Place read miss on bus
Shared (read only)
CPU write
Write-back block
Place write miss on bus
CPU read miss
Write-back data; place read miss on bus
Read miss for block
Write-back block
Place invalidate on bus
CPU write
CPU write miss
Place write miss on bus
CPU read miss
Place read miss on bus
Write miss for block
Exclusive (read/write)
CPU write miss
CPU write hit
CPU read hit
Write-back data
Place write miss on bus

Problem 6: (15 points)

The 3C's model classifies misses as compulsory, capacity, or conflict. Different cache design features attempt to reduce one or more of these types of misses. Complete the table below, indicating whether each feature tends to increase (+), tends to decrease (-), or largely leaves unchanged (0) the *frequency of a particular type of cache miss in a single, unified cache*.

Cache Design Feature	Compulsory Misses	Capacity Misses	Conflict Misses
Increasing cache size			
Increasing associativity			
Using skewed associativity rather than regular associativity			
Increasing block size			
Breaking the block into multiple subblocks			
Using critical-word first transfers			
Using non-binding software prefetch into the cache			