CS/ECE 752: Advanced Computer Architecture II

Prof. David A. Wood

Final Exam
December 19, 1998
Approximate Weight: 35%

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 8 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	15	
Problem 2	10	
Problem 3	20	
Problem 4	10	
Problem 5	15	
Problem 6	15	
Total	**85**	

NAME: _____

Problem 1: (15 points)

Consider a processor with the following caches for a microprocessor with a 64-bit virtual address and 43-bit physical address.

- L1 instruction cache: 32 Kbytes, 64 byte blocks, 4-way set associative, NMRU replacement, and a "next fetch index" to predict which line in the cache should be fetched next.
- L1 data cache: 64 Kbytes, 256 byte blocks, 64 byte subblocks, 2-way set associative, LRU replacement, writeback.

Part A: (3 points)

How many index bits are needed to access each cache?

Part B: (8 points)

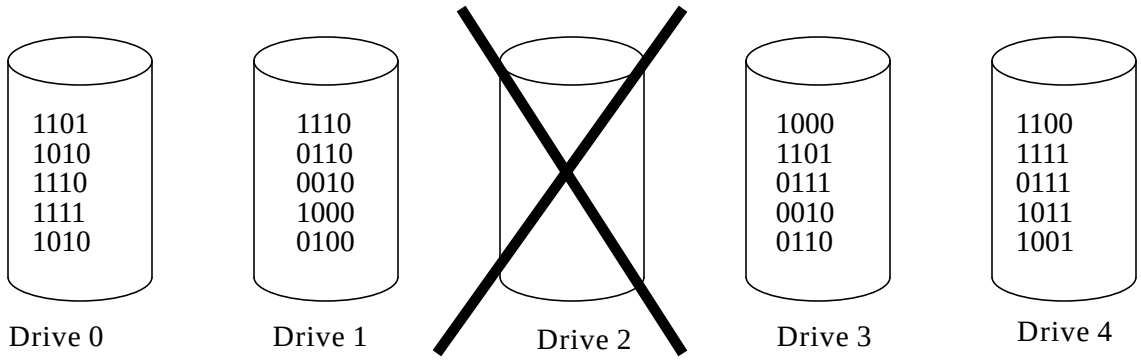
What is the *minimum* number of bits needed to implement each cache assuming virtual address tags?

Part C: (4 points)

What is the *minimum* number of bits needed to implement each cache assuming physical address tags?

NAME: _____

Problem 2: (10 points)



Consider the RAID storage system above. Each sector contains four bits and each drive contains five sectors. The *i*th stripe includes the *i*th sector on each disk, for a total of five stripes. In this example, drive 2 has just failed, and all data on the drive has been destroyed.

Part A: (5 points)

Assume the storage system is using RAID Level 3, where all the parity information is stored on Drive 4. Reconstruct the contents of Drive 2. *Show the algorithm and your work.*

Part B: (5 points)

Assume the storage system is using RAID Level 5, where the parity information is distributed so that the parity of stripe *i* is on drive *i*. Reconstruct the contents of Drive 2. *Show the algorithm and your work.*

NAME: _____

Problem 3: (20 points)

There are numerous techniques for improving the performance of caches. Some reduce the frequency of misses, some reduce the miss penalty, some reduce hit times, and so on. Many optimizations involve trade-offs, making one performance factor worse in return for reducing another. In some cases, the so-called optimization can actually hurt performance. For the following optimizations, briefly summarize the trade-offs involved and how it helps or can hurt performance. Be as specific as possible, for example, if an optimation reduces conflict misses, say so, rather than just say "it reduces misses".

Increasing block size

+ Decreases compulsory and capacity misses by "prefetching" data

+ Increases capacity misses by decreasing number of unique blocks in cache

- Increases memory traffic

- May increase miss penalty

Software controlled prefetching

+ May decrease stalls due to demand misses

- Introduces explicit instruction overhead which may be higher than the benefit

- Requires software support

Victim caches

+ Decreases conflict misses (and some capicity misses)

- May increase miss penalty

- Extra hardware

Skewed associative caches

+ Decreases conflict misses by hashing data

- More complex replacement protocol

Writeback buffers

+ Decreases read miss penalty by keeping writeback off the critical path

- A little extra hardware

Lockup-free (non-blocking) caches

+ Allows multiple outstanding cache misses, allowing misses to overlap

+ Increases memory system bandwidth

- Requires dynamically scheduled processor to fully exploit

- Complex hardware

NAME: _____

Problem 4: (10 points)

The (hypothetical) Phoenix P3000 processor has 64K byte instruction and data caches and is projected to run at 2 Gigahertz. Because of the very high frequency, the designers do not feel it is feasible to perform address translation before they access the cache. You have been asked to figure out how to deal with synonyms.

Part A: (5 points)

In the baseline design, the architects propose to eliminate synonyms by choosing the page size and associativity so that all synonyms map to the same set. Show the equation that relates page size to associativity. If the page size is 8K bytes, what is the minimum possible associativity with this approach?

Part B: (5 points)

Your circuit designers feel that the associativity required by the baseline approach is too high to meet the cycle time goals. What other approaches would you consider? List at least two other alternatives and at least one pro and one con for each.

NAME: _____

Problem 5: (15 points)

The (hypothetical) Phoenix P3000 processor is projected to run at 2 Gigahertz and issue up to sixteen instructions per cycle. The level two cache is unified, two megabytes, and has 128 byte blocks. Despite the large on-chip caches, the architects project that the P3000 will require at least 10 GB/s of main memory bandwidth. Ignore ECC.

Part A: (5 points)

Consider building the memory system out of standard 4-bit wide RAS/CAS DRAMs with a 60 ns access time. What is the minimum number of chips that would be needed to achieve the projected bandwidth? If the chips are 64 megabit chips, what is the *minimum* memory size? Show your work.

Part B: (5 points)

Consider building the memory system for the P3000 using Direct Rambus (16 bit channel, 400 Mhz clock). Assuming 64-megabit chips, what is the *minimum* number of RDRAM chips and minimum memory size for the P3000? Show your work.

Part C: (5 points)

Explain why a single RDRAM chip is capable of providing (approximately) the full bandwidth of a Direct Rambus channel.

NAME: _____

Problem 6: (15 points)

Non-binding prefetch and speculative loads (ala EPIC) are two techniques for tolerating long memory latencies.

Part A: (5 points)

Briefly describe each of these techniques.

Part B: (10 points)

Discuss the pros and cons of these techniques. Are these schemes best suited to statically or dynamically scheduled processors, or does it matter? Are these best for tolerating L1 cache misses that hit in the L2 cache, or L2 cache misses? How do these schemes affect the rest of the pipeline design.