Midterm Exam

ECE 741 – Advanced Computer Architecture, Spring 2009

Instructor: Onur Mutlu

TAs: Michael Papamichael, Theodoros Strigkos, Evangelos Vlachos

February 25, 2009

NAME: _____

Problem	Points	Score
1	40	
2	20	
3	15	
4	20	
5	25	
6	20	
7 (bonus)	15	
Total	**140+15**	

- This is a closed book midterm. You are allowed to have only two letter-sized cheat sheets.
- No electronic devices may be used.
- This exam lasts 1 hour 50 minutes.
- If you make a mess, clearly indicate your final answer.
- For questions requiring brief answers, please provide brief answers. Do not write an essay. You can be penalized for verbosity.
- Please show your work when needed. We cannot give you partial credit if you do not clearly show how you arrive at a numerical answer.
- **Please write your name on *every* sheet.**

Name: _____

Problem 1 (Short answers – 40 points)

i. (3 points)
A cache has the block size equal to the word length. What property of program behavior, which usually contributes to higher performance if we use a cache, does not help the performance if we use THIS cache?

ii. (3 points)
Pipelining increases the performance of a processor if the pipeline can be kept full with useful instructions. Two reasons that often prevent the pipeline from staying full with useful instructions are (in two words each):

iii. (3 points)
The reference bit (sometimes called "access" bit) in a PTE (Page Table Entry) is used for what purpose?

The similar function is performed by what bit or bits in a cache's tag store entry?

iv. (3 points)
The fundamental distinction between interrupts and exceptions is that

interrupts are caused by_____

and exceptions are caused by _____.

Interrupts are handled mostly when convenient. Why?

Why are interrupts not always handled when convenient? Give an example.

Name: _____

v. (3 points)

A microprocessor manufacturer decides to advertise its newest chip based only on the metric IPC (Instructions per cycle). Is this a good metric? Why or why not? (Use less than 20 words)

If you were the chief architect for another company and were asked to design a chip to compete based solely on this metric, what important design decision would you make (in less than 10 words)?

vi. (3 points)

p% of a program is perfectly parallelizable. We wish to run it on a multiprocessor. Assume we have an unlimited number of processing elements. If the maximum speedup achievable on this program is 100, what is p?

vii. (5 points)

Name two techniques that eliminate the need for hardware-based interlocking in a pipelined processor:

viii. (6 points)

Remember that the history buffer is a structure that enables precise exceptions in a pipelined machine. Briefly describe how the history buffer works during the execution of a register-to-register ADD instruction (ADD RD ← RS1 + RS2) by completing the following sentences:

When the add instruction is decoded _____

When the add instruction completes execution _____

When the add instruction is the oldest completed instruction in the pipeline and

i) An exception has occurred during its execution,

_____

ii) No exception has occurred during its execution,

_____

Name: _____

Assume we would like to use the exact same solution (history buffer) for executing a store instruction to memory. Why is this difficult to do?

Very briefly describe one solution to handling store instructions in conjunction with a history buffer. (Use less than 30 words)

ix. (5 points)
Suppose you are designing a small, 16KB, 2-way set-associative L1 and a large, 32MB, 32-way set-associative L3 cache for the next processor your company will build. Which one of the following design decisions would you make and why? Justify your choice.

Access L1 tag store and data store: in parallel OR series (circle one and explain)

Access L3 tag store and data store in parallel OR series (circle one and explain)

Name: _____

x. (6 points)

Suppose you are designing a computer from scratch and that your company's budget allows a very small amount of memory bandwidth. Which of the following characteristics would you choose in the ISA and the microarchitecture, and why? Explain briefly.

Variable length instructions or fixed length instructions?

Complex instructions or simple instructions?

A large L2 cache or a small L2 cache? (L2 is the last-level cache)

An aggressive prefetcher or a conservative prefetcher?

Large cache blocks vs. small cache blocks?

Name: _____

Problem 2 (20 points)

You designed a microprocessor. It came back from the fab with an error: one of the bits is stuck. We call the bit a stuck-at-0 fault if the bit is always 0 (i.e., you cannot store a 1 in it). We call the bit a stuck-at-1 fault if the bit is always 1 (you cannot store a 0 in it).

Consider each of the structures below independently. Assume the structure contains a stuck at 0 or 1 fault. Does the fault affect the correctness of the chip? Does the fault affect performance? Explain. (Note: For each structure, consider separately stuck-at-0 and stuck-at-1 faults.) No error detection and correction mechanisms are present.

A bit in the register scoreboard:

The dirty bit in one of the tag store entries of the L2 cache:

The LRU bit for one of the sets of a 2-way set associative cache:

A bit in the "predicted next instruction address" supplied by the branch predictor:

A bit in the stride value stored in a stride prefetcher:

Name: _____

Problem 3 (15 points)

Your job is to evaluate the potential performance of two processors, each implementing a different ISA. The evaluation is based on its performance on a particular benchmark. On the processor implementing ISA A, the best compiled code for this benchmark performs at the rate of 10 IPC. That processor has a 500 MHz clock. On the processor implementing ISA B, the best compiled code for this benchmark performs at the rate of 2 IPC. That processor has a 600 MHz clock.

What is the performance in MIPS (millions of instructions per sec) of the processor implementing ISA A?

What is the performance in MIPS (millions of instructions per sec) of the processor implementing ISA B?

Which is the higher performance processor? A B Don't know
Explain.

Name: _____

Problem 4 (20 points)

As you remember from the lecture, SPEC numbers are measures of the time it takes to execute certain representative benchmark programs. A higher number means the execution time of the corresponding benchmark(s) is smaller. Some have argued that this gives unfair advantage to processors that are designed using a faster clock, and have suggested that the SPEC numbers should be normalized with respect to the clock frequency, since faster clocks mean shorter execution time and therefore better SPEC numbers. Is this suggestion a good or a bad idea? Explain.

If you are told that your design will be evaluated on the basis of its SPEC/MHz number, what major design decision would you make?

Name: _____

Problem 5 (25 points)

We have a byte-addressable toy computer that has a physical address space of 512 bytes. The computer uses a simple, one-level virtual memory system. The page table is always in physical memory. The page size is specified as 8 bytes and the virtual address space is 2 KB.

Part A.

i. (1 point)
How many bits of each virtual address is the virtual page number?

ii. (1 point)
How many bits of each physical address is the physical frame number?

We would like to add a 128-byte *write-through* cache to enhance the performance of this computer. However, we would like the cache access and address translation to be performed simultaneously. In other words, we would like to index our cache using a virtual address, but do the tag comparison using the physical addresses (virtually-indexed physically-tagged). The cache we would like to add is direct-mapped, and has a block size of 2 bytes. The replacement policy is LRU. Answer the following questions:

iii. (1 point)
How many bits of a virtual address are used to determine which byte in a block is accessed?

iv. (2 point)
How many bits of a virtual address are used to index into the cache? Which bits exactly?

v. (1 point)
How many bits of the virtual page number are used to index into the cache?

vi. (5 points)
What is the size of the tag store in bits? Show your work.

Name: _____

Part B.

Suppose we have two processes sharing our toy computer. These processes share some portion of the physical memory. Some of the virtual page-physical frame mappings of each process are given below:

PROCESS 0	
Virtual Page	Physical Frame
Page 0	Frame 0
Page 3	Frame 7
Page 7	Frame 1
Page 15	Frame 3

PROCESS 1	
Virtual Page	Physical Frame
Page 0	Frame 4
Page 1	Frame 5
Page 7	Frame 3
Page 11	Frame 2

vii. (2 points)
Give a complete physical address whose data can exist in two different locations in the cache.

viii. (3 points)
Give the indexes of those two different locations in the cache.

ix. (5 points)
We do not want the same physical address stored in two different locations in the 128-byte cache. We can prevent this by increasing the associativity of our virtually-indexed physically-tagged cache. What is the minimum associativity required?

x. (4 points)
Assume we would like to use a direct-mapped cache. Describe a solution that ensures that the same physical address is never stored in two different locations in the 128-byte cache.

Name: _____

Problem 6 (20 points)

A processor has an 8-bit physical address space and a physically addressed cache. Memory is byte addressable. The cache uses perfect LRU replacement.

The processor supplies the following sequence of addresses to the cache. The cache is initially empty. The hit/miss outcome of each access is shown.

Address	Outcome
0	Miss
2	Hit
4	Miss
128	Miss
0	Hit
128	Hit
64	Miss
4	Hit
0	Miss
32	Miss
64	Hit

Your job: Determine the block size, associativity, and size of the cache. Note: It is not necessary to give an explanation for every step, but you should show sufficient work for us to know that you know what you are doing.

BONUS Problem 7 (15 points)

Suppose you have designed the next fancy hardware prefetcher for your system. You analyze its behavior and find the following:

i) The prefetcher successfully prefetches block A into the cache before it is required by a load instruction. The prefetched block evicts a never-to-be-used block from the cache, so it does not cause cache pollution. Furthermore, you find that the prefetch request does not waste bus bandwidth needed by some other request.

ii) The prefetcher successfully prefetches block B into the cache before it is required by a load instruction. The prefetched block evicts a never-to-be-used block from the cache, so it does not cause cache pollution. Furthermore, you find that the prefetch request does not waste bus bandwidth needed by some other request.

Upon further analysis, you find that the prefetching of block A actually reduced execution time of the program whereas prefetching of block B did *not* reduce execution time significantly. Describe why this could happen. Draw two execution timelines, one with and one without the prefetcher, to illustrate the concept.

Name: _____

EXTRA PAGES

Name: _____

EXTRA PAGES