2015/Spring Computer Architecture PhD Qualifying Exam

Student ID _____ Name _____

1. [15pts] Pipelining (State the reason for the answer shortly)

   A. [3] What is the CPI (cycle-per-instruction) of a single-cycle processor, which can execute an instruction every cycle?
   B. [3] What is the speedup of the ideally pipelined processor compared to the single-cycle design? Explain the reason for your answer using the execution time decomposition (Execution time = instruction count * CPI * clock cycle time). Suppose the number of pipeline stages is N.
   C. [3] Explain two reasons that the clock cycle time of a pipelined processor can be longer than the ideally pipelined processor.
   D. [6] Explain any possible reasons that the CPI of a pipelined processor can be longer than the ideally pipelined processor.

2. [15pts] Suppose you are comparing two processors with different cache designs. Both processors have two-level caches (L1 and L2 caches). Assume a perfect instruction L1 cache. The cache configurations of the two processors are as follows:

   Processor A: L1 cache: 16KB, 1 cycle hit latency, L2 cache: 512KB, 10 cycle hit latency,
   Processor B: L1 cache: 64KB, 3 cycle hit latency, L2 cache: 2MB, 12 cycle hit latency

   For both processors, external memory access latency is 100 cycles. Suppose you will run only two applications with the following characteristics

   Application 1 running on processor A: L1 miss rate = 20%, L2 miss rate = 50%
   Application 2 running on processor A: L1 miss rate = 2%, L2 miss rate = 10%
   Application 1 running on processor B: L1 miss rate = 10%, L2 miss rate = 20%
   Application 2 running on processor B: L1 miss rate = 1%, L2 miss rate= 5%

   Assume that you use the two applications equally. (Each application is used for 50% of the total system run time) Which processor will you choose? Explain the reason with quantitative analysis.

3. [15pts] The ISA for a processor requires maximum 2 input operands and 1 output operand per instruction and has 32 architectural registers. The superscalar out-of-order processor is based on a unified reservation and ROB. (ROB: L entries, reservation station: R entries, issue width: W).

   A. [5pts] When does a write operation occur on the register file and the ROB?

   B. [5pts] How many read and write ports are necessary for the ROB to support the maximum issue/commit rate? (explain how the ports are used)

   C. [5pts] Explain how WAR and WAW dependencies are eliminated with the reservation station and ROB.

4. [15pts] Memory dependency

   A. [5pts] Executing load and store instructions involves address calculation and memory access (cache access) parts. In out-of-order execution processors, certain re-orderings of load and store executions can lead to an incorrect program execution. In the following sequence of store and load (in the program order), describe the condition that the execution of two instructions can lead to an incorrect out-of-order execution. (The condition must include the ordering of address calculation and memory access parts of two instructions.)

   A: store r1, 10(r2)
   B: load r5, 150(r6)

   B. [10pts] To issue load instructions as early as possible, out-of-order processors can issue load speculatively, even if there is a chance of dependence violation. Explain the necessary operations for executing loads and stores to support such speculative load execution.

5. [15pts] This problem assumes the following system configurations

   - 32-bit architecture, which uses 4GB address space for each process
   - Page size: 4KB page, the size of each page table entry : 4B

   A. [5pts] What is the page table size for each process for one-level page table (flat page table)?

   B. [5pts] What are the minimum and maximum page table sizes for a process, if the system uses two-level page tables? (For the minimum case, a process uses the memory which fits in one page.)

   C. [5pts] To hide the access latency for TLBs, TLBs may be accessed in parallel with accesses to L1 caches. To support such parallel accesses to TLBs and L1 caches, the organization of L1 caches may be restricted to meet certain conditions. If the maximum associativity is limited to 8 ways, what is the largest cache capacity for such L1 caches? (a physical address must be mapped to only one set in the cache)

6. [25pts] Multi-processors

   C. [8pts] Discuss the pros and cons of updated-based coherence protocols compared to invalidation-based protocols. You may explain why invalidation-based protocols are dominantly used in commercial multi-processors, and also discuss on what conditions updated-based protocols can excel invalidation-based ones.

   D. [8pts] In multi-processors with invalidation-based coherence protocols, a cache block can be invalidated by another core, which is about to update the cache block. In such a system, is it possible that a cache can hold a correct value for a word (4B), even if the cache block containing the word has recently been invalided?

   E. [9pts] Discuss the pros and cons of shared L2 caches, compare to private L2 caches. Assume that the total cache capacity on a chip does not change. Explain why the current commercial multi-cores commonly have shared L3 caches, instead of shared L2 caches.