2013/Autumn Computer Architecture PhD Qualifying Exam

Student ID _____ Name _____

1. [25pts] Branch Prediction:
   A. Is branch prediction more important in single-issue in-order core or 4-way superscalar processor? Why?
   B. Which is more difficult to predict – indirect branches or direct branches? Why?
   C. Assume you have a 5-stage pipeline processor but the details of the processor microarchitecture are not known. To minimize any branch prediction penalty, you suggest a fine-grained multithread architecture. What is the minimum number of thread needed to minimize the impact of branch penalty? Why?
   D. Current GPUs have thousands of threads and can execute multiple threads simultaneously –e.g., NVIDIA GPUs execute 32 threads simultaneously in units of warps and 32 threads execute the same instruction. If the instruction is a branch, 32 threads can have different behaviors as some threads have branch-taken and others have branch-not-taken. Thus, what impact does branch on such parallel architectures? Discuss how important is branch prediction in such parallel architectures.

2. [15pts] Coherence miss
   A. What is a coherence miss?
   B. How does the "increasing number of bytes per line" impact the amount of coherence miss? Increase, decrease, or no impact? Please explain. Assume all other cache parameters remain the same.
   C. How does the "increasing the number of ways" impact the amount of coherence miss? Increase, decrease, or no impact? Please explain. Assume all other cache parameters remain the same.

3. [20pts] Scalable Processor
   You are responsible for designing a scalable processor. Currently, the processor is a 4-issue out-of-order processor with an instruction window size of 32. However, you decide to modify the instruction window size to 512. What is the performance impact of such modification? Describe the microarchitectural challenges required in modifying the processor to support a system of window size 512.

4. [20pts] In designing future multicore processors, one design decision is whether to use the on-chip cache as a shared or a private cache. Assume each processor has a local, private L1 cache and a slice of L2, which can be used as a shared L2 cache, or a private L2. You need to determine whether shared or a private cache organization is better. If the following constraints are provided, which cache organization would you prefer – shared or private? Why?
   A. The on-chip network that interconnects the cores together can be designed with a new technology (such as optics) such that it provides near zero-cycle latency and very high bandwidth.
   B. Applications with high degree of temporal locality but with a small workings set.

5. [20pts] Processor Architecture:
   A. In trying to increase the performance and throughput of a processor, both superscalar processor and VLIW processors have been implemented. What are the similarities and differences between the two different types of processors, in terms of how they improve the performance?
   B. Power consumption is a critical aspect of future microprocessor design and has led to different companies building multicore processors. If power consumption was your main concern, which processor (superscalar or VLIW) would you use as the building block of your multicore processor? Why?