2016 Computer Architecture PhD Qualifying Exam

Student ID _____ Name _____

Please keep the answers short.

1. [10pts] Give at least 2 reasons why a processor manufactures do not build a superscalar processor that can fetch and execute 10 instructions in parallel.

2. [15 pts] 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] 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?

4. [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? Also, please discuss – how important is branch prediction in such parallel architectures?

5. [15pts] Explain register dependencies and memory dependencies. In particular, in what ways are they similar and in what ways are they different?

6. [15pts] Suppose the width of physical address of a system is 28 bits (the maximum physical address space = 2^28B). Calculate the size of storage (in bits) for the tags for the following cache configurations. Each tag needs extra 2 bits for valid, modified bits. Ignore extra bits for LRU replacement for set-associative caches. Please show work.
A. 64KB direct-mapped cache, block size = 128B
B. 256KB 4-way associative cache, block size = 32B
C. 4KB fully associative cache, block size = 64B