Winter ’12 CIS 429/529 Final Review

You may bring one page of notes, front and back.

You may bring a calculator but shouldn’t need one.

Questions will be in short-answer format with partial credit for partial answers.

Topics:

- All midterm topics
- Vector architectures (DLP): vector registers and instructions, vector-length register, strip mining
- SIMD Extensions (DLP): FP-unit utilization, limitations
- GPU architectures (DLP): CUDA threads, warps, blocks, grids, scheduling
- Snoopy coherence protocols (TLP): benefits/drawbacks, MSI states, coherence misses
- Directory coherence protocols (TLP): benefits/drawbacks, MSU states, hybrid protocols
- Warehouse-scale computers (RLP): MapReduce, basic organization, operational costs

Sample questions:

1. [15] Assume that you have an array of 1024 single-precision floating point numbers which need be multiplied by 3 as part of some function:

   a. (5) Assuming you have access to a vector architecture with vector registers which each hold 64 single-precision floats, explain how you would implement the above operation and estimate the number of cycles required to complete the operation. Assume that you have only one FP unit (i.e., one lane), that the unit is fully pipelined, that the unit does not need to complete one instruction before starting the next, and that the unit completes one operation per cycle.

   b. (5) Assuming that you have access to a processor with SIMD extensions capable of executing eight single-precision floats in parallel on a FPU unit, explain how you would implement the above operation and estimate the number of cycles required to complete the operation. Assume that you only have one FP unit and that the unit completes one instruction per ten cycles.

   c. (5) Assuming that you have access to a GPU with 16 multithreaded processors, each with 32 single-precision stream processors, explain how you would implement the above operation and estimate the number of cycles required to complete the operation. Discuss your implementation in GPU terms such as grids, blocks, warps, and threads. Assume that each stream processor completes one instruction per cycle.

2. [5] Briefly explain why most coherence protocols invalidate cache entries rather than providing the cache with an up-to-date version of the data.

3. [5] List three contributing factors to the cost of operating a WSC.