# Final Exam Topics and Question Types

The final exam will cover my Chapter 7 lecture notes (excluding pipelining) and my notes on GPUs, GPU programming and CUDA. You are expected to know:

- The Grand Challenge problems.
- Definitions of the different types of parallel computers.
- Speedup, Amdahl's Law, the Amdahl effect, scaling, and load-balancing.
- The different types of multiprocessors, e.g., UMA, NUMA, and SMP and how they differ;
- The reduction algorithm for summing a sequence of numbers on a shared memory multiprocessor (page 8 of my notes);
- The reduction algorithm for summing a sequence of numbers on a message-passing multiprocessor (pp 10 -11 of my notes);
- Hardware multithreading: what it is, the difference between instruction-level and thread-level parallelism, the three different types of multithreading (fine-grained, coarse-grained, and SMT), and the differences among these methods;
- Flynn's taxonomy: what it is and the differences between SIMD and MIMD architectures in particular;
- Network performance metrics: latency, throughput, and delays;
- The difference between blocking and non-blocking networks and examples of each;
- Network topology metrics: diameter, bisection width, bisection bandwidth, and total bandwidth;
- The definitions of crossbar networks, buses, and multistage networks;
- The definitions and properties of meshes, rings, buses, fully-connected networks, butterfly and omega networks, including their diameters, bisection widths, maximum edges per node and maximum edge lengths, as a function of the network parameters (such as its dimension or number of nodes.)
- The purpose of a GPU and the ways in which a GPU differs from a CPU;
- The basic organization of the NVidia GPU family (organization in terms of streaming multiprocessors, streaming processors (cores), memory hierarchy (register file, shared memory, device memory), special purpose subprocessors (SFU, etc);
- How to read a CUDA program and identify the different parts, such as block and grid dimensions, thread IDs, block IDs, and what is executed synchronously, what is executed on the host, and what is executed on the device;
- The basic terminology of GPU programming including grids, thread blocks, threads, and warps.

Some questions may be short code fragments to be analyzed. Some questions may contain representations of network topologies, about which various questions will be asked. Some may ask for a short algorithm or piece of an algorithm to be completed. Some questions will be short answer, some multiple choice, and some true-false. The format will be similar to those of the first and second exams.