CS525: Homework 1.

Due Date: Thursday, 18th September, 2014

Make reasonable and well-stated assumptions as needed.

1. Consider a case of p processors hanging off a shared bus. The peak bus bandwidth is 40GB/s. Each of the processors performs a computation that requires access to 4 bytes for each floating point operation (FLOP). What is the peak FLOPs rating of such a system? What is the number of processors, p, for which this peak rate is achieved?

2. Consider now a case where each queued memory request has an overhead of 10 ns, i.e. it takes 10 ns to process. Furthermore, assume that each processor is making one four-byte memory request every 10 ns. As before, assume that the peak bus bandwidth is 40GB/s (or 10GW/s) and that each memory word request is used for one FLOP. Plot the performance (in terms of total floating point operations per second or FLOPS) of the system as a function of the number of processors p.

3. Repeat problem 1 above. In this case, assume that each processor has a local cache (on the processor side of the shared bus) and that the computation has a cache hit ratio of 90%. What is the peak performance of the system for the computation in problem 1? What is the number of processors for which this peak performance is achieved?

4. You are given an 8-port non-blocking switch (i.e., the switch has 8 ports and any port can be connected to any other port in a non-blocking fashion). You must use one of the ports for the local processor/ computer. You must also use one port for an I/O subsystem (a network disk). The other six ports are available to you. The switch is a gigabit ethernet switch, i.e., each wire is capable of carrying 1 Gb/s. How would you connect p processors using this switch to maximize the bisection bandwidth? What is the bisection bandwidth of your system (as a function of p)? What is the diameter of your system?

5. You are given the same switch as above (with the same constraints). In this case, you are looking to connect no more than 64 computers (as opposed to an arbitrarily large number of processors in problem 4). How would you connect the computers to maximize bisection bandwidth? What is the bisection bandwidth of your 64 computer system? What is the diameter of the system?

6. Consider a shared address space computer with a single shared directory. Each processor in this system has a local memory and remote memory. Each processor makes 90% of its accesses to the local memory and 10% of the accesses to the remote memory. Accesses to the remote memory must be handled through the directory, which is the single point of contention. Each remote access results in a lookup to the directory. The shared directory is capable of 10 billion lookups/second. Each processor generates one access every 10 ns. Of these, 90% are satisfied from the local memory. The computation performs 1 FLOP on each word. What is the peak performance of such a system?

7. Repeat problem 6 for the case when 95% of the accesses are satisfied from local memory.

8. Consider a distributed directory shared address space computer. Once again, each processor generates 1 memory access every 10 ns and 90% of these accesses are satisfied from the local memory. The remaining 10% of the accesses are satisfied from remote memory. However, access to remote memory takes 10 times the time it takes to access the local memory. Making reasonable assumptions, estimate the slowdown in the peak computation rate of each processor (i.e., what is the computation rate if there was only one processor; i.e., only local memory, versus the case where the same processor leaks 10% of the accesses to the remote memory)?

9. Assuming a non-congested network, a 10 word message takes 1100 ns and a 1000 word message takes 11000 ns. What are the startup and per-word transfer times on a cut-through network with negligible per-hop time?

10. Assuming that the network in problem 9 is a 3D mesh without wraparound links and that each processor on one side of the machine is sending a message to a processor on the other side of the machine (i.e., there is congestion across the bisection), what is the effective per-word time for these messages?