COL380 REVIEW QUESTIONS

(MPI and Cuda)

1. What is a communicator in MPI? What are different ways of creating communicators?
2. What is a blocking MPI Send – what does it block for?
3. What does a blocking MPI Receive block for? Explain how to use a non-blocking MPI Receive.
4. What is MPI call matching? What matches what?
5. Do the number of elements in a matching MPI Send and Recv need to also be the same?
6. Can an MPI_Barrier be replaced with MPI_Bcast (with 0 elements)?
7. What is RDMA, and what are the possible advantages and disadvantages of RDMA?
8. In what ways does MPI help reduce the number of data copies?
9. Create a datatype that allows efficient block-wise scatter of a dense 2D matrix. (Assume an NxN matrix is divided into blocks of N/PxN/P elements, and each block is sent to one of P^2 processes.
10. Consider a block-wise format of two sparse matrices A and B stored in a file (as in asignment 4). Provide an efficient algorithm for P processes to read A and B, perform AxB, and write the result in a file. (You may read from or write to an arbitrary offset of a file. However, you must still avoid races.)
11. Provide an efficient Send/Recv-based implementation of MPI_Reduce. Analyze the cost.
12. Explain the Cuda kernel launch arguments (the ones contained in <<< >>>)?
13. What is the notion of a block of threads in Cuda? In what ways can two threads within the same block and two threads in different blocks interact? (Interact means synchronize or communicate.)
14. What are Cuda streams? How can multiple grids be created in the same stream? How do two threads in such two grids interact?
15. How do threads in two different streams interact?
16. What are device memory, managed memory, shared memory, and consant memory (in Cuda)?
17. What are the ways in which a Cuda program is able to write data to or read data from shared memory?
18. Give examples of efficient and inefficient memory IO from shared memory and device memory (respectively).
19. Are both blocking and non-blocking Kernel launch posible? Explain.
20. What are Kernel events? How do you use them?
21. Explain SIMD architecture. Explain SIMT programming model.
22. What is latency hiding, and what are some ways to hide latency in Cuda?