Practice Questions for CSE531 (Spring 2023)

1. The following statements are true or false? Justify your answer with a two-line statement:

   (a) Locality of reference is only critical for parallel computers and not for serial processors with deep memory (cache) hierarchies.

   (b) Data locality is critical only for message passing machines and not for shared address space machines.

   (c) Conventional microprocessors increasingly rely on parallelism within the processor for speed.

   (d) Best parallel algorithms can only be generated from the best sequential (serial) algorithm.

2.

   (a) Consider a problem of adding 256 numbers using 32 threads. Each thread adds 8 numbers. The 32 partial sums are then added to the global sum one after the other. Assuming single cycle addition and no communication cost, what is the maximum speedup and efficiency of this code.

   (b) Consider the general problem of adding N numbers on p processors. What is the most efficient way of performing this computation? What is the parallel runtime of your formulation?

3. Consider the "sparse" matrix illustrated in Figure 1. The matrix has 8 rows. Consider a matrix with identical structure with N rows. Write a message-passing pseudo-code for multiplying this matrix with a vector. What is the runtime of this code (assuming time tc for a single multiply-add and using the (ts,tw) model of communication). Hint: Ordering is important in this example.

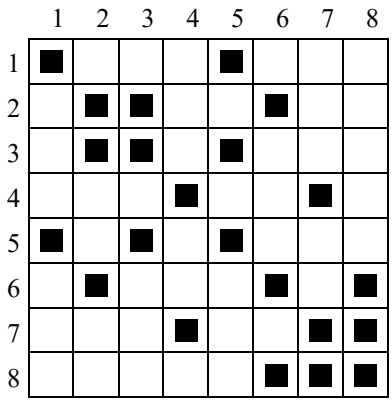

Figure 1: Sparsity structure of the matrix.

4. Define two different ways of performing "all-to-all personalized communication" in a message-passing machine? When would you prefer one over the other? Explain.

5. Compare the CUDA and OpenCL programming paradigms from the programmer productivity, performance, and supported library availability perspectives.

6. Describe a parallel formulation of matrix-vector multiplication in which the matrix is 1D block-partitioned along the columns and the vector is equally partitioned all the processors. Is the parallel runtime of your algorithm the *same* as the row-wise 1D block-partitioned version discussed in the class? Write the overhead function (To) of this formulation.

7. Show an embedding of a p-node three-dimensional mesh into a p-node hypercube. Does your embedding work for any value of p? Explain.