High Performance Computing

Assignment 7

1. **10 marks.** Consider a machine with 16 Gflops/s peak and stream bandwidth of 16 GB/s.
   (a) What is the maximum attainable performance if your code has flops/byte ratio of 0.5?
   (b) Same question for flops/byte = 1.
   (c) Same question for flops/byte = 2.
   (d) Plot the maximum attainable performance (units of flops/s) as a function of code flops/byte.

2. **10 marks.** Consider a communication network with latency α and bandwidth $1/\beta$. Suppose you want to perform a reduce-scatter collective operation between p processes. This operation is a reduction of length n (e.g., the result of the sum is n items) followed by scattering each n/p part to the p processors. Write pseudocode for an algorithm that implements this operation using point-to-point communication, i.e., using only (nonblocking) sends and receives. Write the model for the communication time in terms of α, β, n, and p.

3. **10 marks.** Draw a fat tree network with full bisection bandwidth for 16 nodes using only 8-port switches. Explain why your network has full bisection bandwidth.

4. The SUMMA algorithm is a well-known algorithm for performing distributed matrix multiplication, $C = AB$.
   (a) **4 marks.** Write pseudocode for this algorithm, assuming A and B are $n \times n$ matrices, using p^2 processors on a $p \times p$ processor grid.
   (b) **6 marks.** Suppose that both A and B are symmetric matrices. Write pseudocode to explain how the SUMMA algorithm can be optimized for this case. Note that the product of two symmetric matrices is not necessarily symmetric.

5. **10 marks.** What are three similarities and three differences between Intel Xeon Phi coprocessors and GPUs?