1. Assume a reduction algorithm that finds the maximum of an array of 1024 integers.
   A. Write the sequential version of the program in C.
   B. Write a CUDA version of the program that does not take thread divergence into account.
   C. Update the version in B to take thread divergence into account.
   D. Update the program in C to make use of shared memory to reduce global memory bandwidth.

   Draw a bar graph that compares the execution time of each of the above 4 versions.

2. Repeat problem 1 with an array of 4096 elements.

3. What can we conclude from the results of problems 1 and 2 regarding the optimizations and the problem size?