ECE 4270: Computer Architecture, Spring 2023
Lab 5: CUDA Programming

Scope

In this lab assignment, you will develop a CUDA that takes adds 2 matrices each of m x n. In the previous introductory lab session, we leveraged the elementary CUDA principles to find the square of two numbers. Employing the same understanding, write a CUDA program to add two matrices of m x n dimension.

Your CUDA program should accept as input 2 matrices each of m x n dimension and return the sum of the two matrices.

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad B = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$$

$$A + B = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

Problem 1: Explain how writing CUDA kernel code to have thread blocks waiting for the execution of other thread blocks in the same grid to complete can lead to problems, even if the dependencies are acyclic.

Problem 2: A friend wants your help to write CUDA code that performs a reduction for each field (A, B, C, and D, all integers) in an array of structures. The friend tried reducing A, then B, then C, then D, but got poor performance. Explain why and suggest a simple fix to solve the problem.

Grading Rubric

Code: CUDA program, and solutions to the problems (75)
Report: 25 points

Code and CUDA programs (75 points):

In order to get full credit for the code, your CUDA program should be able to take as an input 2 matrices and return the sum of the matrices. Add a readme file detailing how your program has to run as well as the input used in testing your program.

Lab report (25 points):
Your report should give details about the work distribution within the group (who did what), milestones in your work and your implementation decisions (why did you choose the way you did it, and/or how did you do that).