Dustin Cook

CSC 410, Christer Karlsson

Programming Assignment 2

10/29/2018

# Program Description

This program multiplies a matrix by a vector using cuda. There are two source files, mull\_1.cu and mull\_2.cu. Both files have their own main, the main difference between them is the cuda kernel and how blocks/threads are determined. In mull\_1.cu, the kernel first performs the multiplications using a linear addressing method threadIdx.x + blockDim.x \* blockIdx.x. This is the same for mull\_2.cu. The two kernels differ on how they sum up the rows. Mull\_2.cu uses a single thread per row to sum up the result. Mull\_2.cu reduces the rows first by storing the rows locally and then uses a logarithmic method to sum up the rows. Mull\_2.cu proves to be sufficiently faster with nvcprof showing a speed increase of as much as 78%. Mull\_2.cu can be considered my final result, and mull\_1.cu is just there as a bad example now.

# Libraries Uses

* Cuda (Thrust for simplifying memory allocation of vectors)
* STD iostream, math.h, random, chrono (for random numbers and user input)

# Program Structure

This program has a simple structure. It first allocates the memory and initializes the vectors for both the host and device. It then passes it to the kernel, and the kernel stores the result in a third vector.