Abhi Dankar

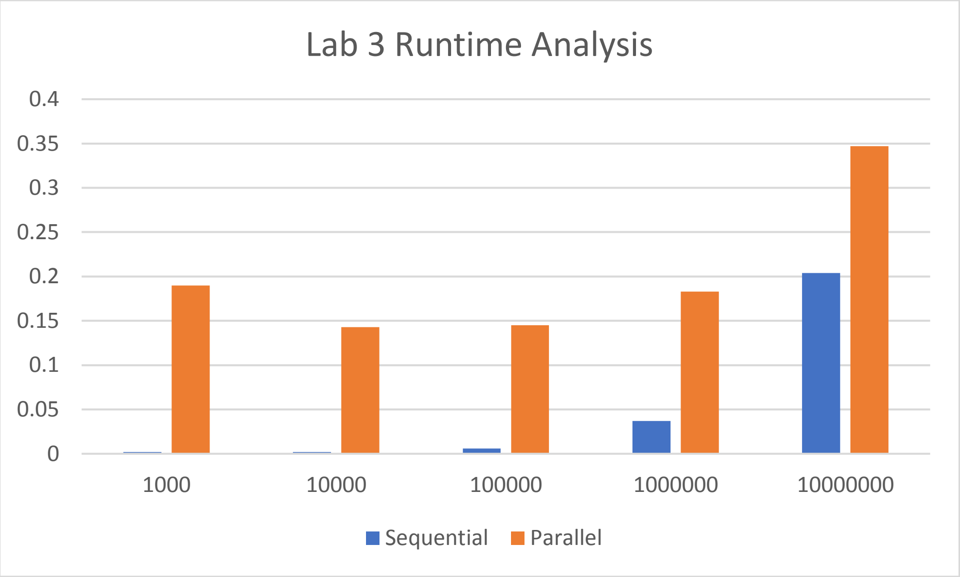
Professor Zahran

May 3, 2019

Lab 3 Report

I ran my code on cuda1 and compiled with **nvcc -o maxgpu maxgpu.cu**.

For the threads per block, I chose the max number allowed of 1024. This led to less reconciliation between the blocks, as within blocks we can work with shared memory. As such, the dimensions of the block are 32x32.



The parallel version performs significantly worse than the sequential, especially with smaller problem sizes. The parallel code gets better (in comparison to the sequential) with bigger problem sizes since more data offers more opportunity for speedup via parallelization. However, the overhead is never compensated for with the speedup in any case. The overhead caused by copying back and forth between the device and host as well as invoking the kernel are the primary causes of the slowdown.