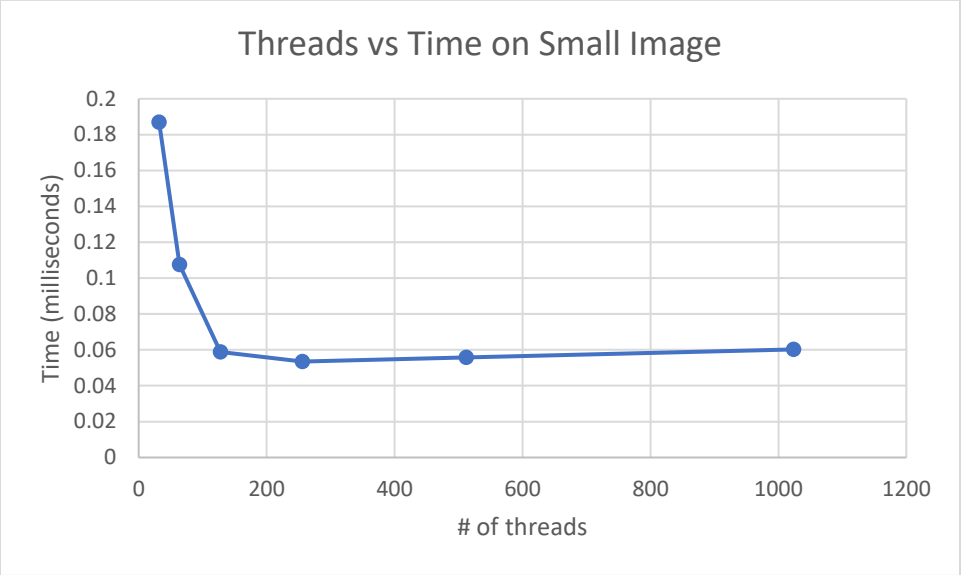
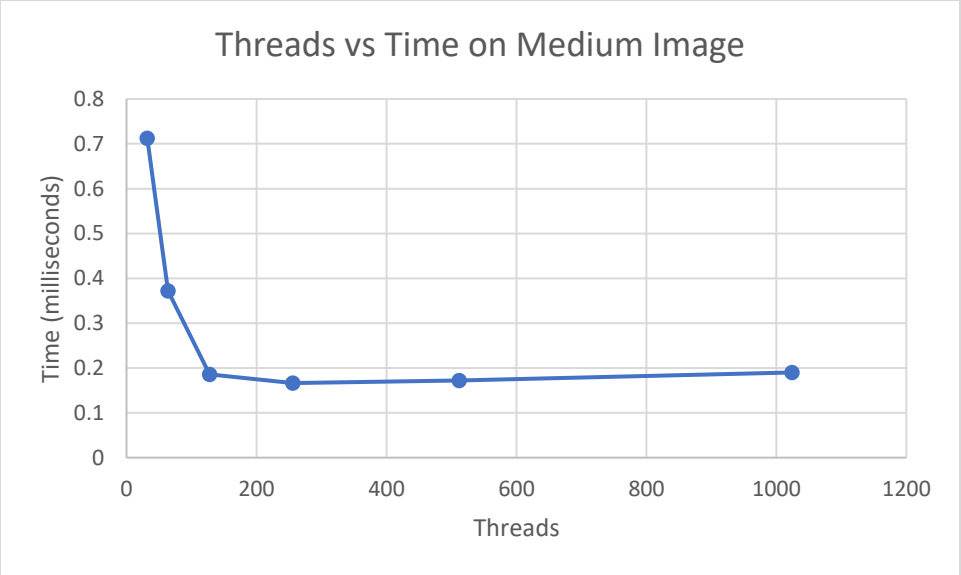
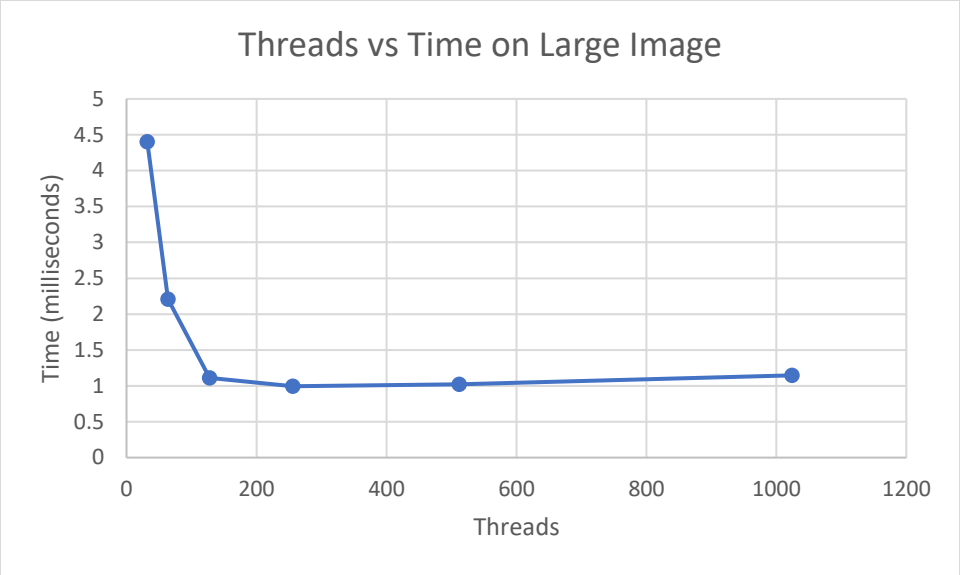
Hart Russell

CS-605

Homework 4







As we can see by the data, 256 threads seem to be the sweet spot for fastest run time, regardless of image size. We see greatly increasing returns up until 128 threads, but performance is similar after. Block size is a function of image size divided by thread size because thread size should be a multiple of warp size[1]. In this way, we also ensure that the kernel is called the right number of times for each pixel in the image.

An interesting note: I ran this same program on my GTX 1060 using Visual Studio and the latest version of the CUDA toolkit. The run time for every thread count was nearly identical from 32->1024. I'm not sure whether this is because the Tesla P100s in tuckoo have a much bigger memory bus width (4096 bits in the P100 vs 192 bits in the GTX 1060). This could cause the data to be throttled to/from the GPU. Or perhaps there are some big optimizations in the newest version of CUDA (running 10.1 on my Windows PC vs 8.0 on tuckoo).

[1] https://stackoverflow.com/questions/9985912/how-do-i-choose-grid-and-block-dimensions-for-cuda-kernels/9986071#9986071