ECE 6360 Parallel and Heterogeneous Computing

Homework 6: CUDA Threads and Kernels

A kernel is launched as `kernelFunc<<<100, 256>>>(...)`

How many threads run in total, how many warps does each block contain (assume a warp size of 32), and how many warps are there in the whole grid?

List the four built-in variables provide by CUDA inside of a __global__ function. Which of them are the same for every thread in a block?

Assume you are writing a kernel to control the color in a 600x400 output image.

The device has a maximum number of threads per block of 1024. If you design the grid to have square blocks while maximizing the number of threads per block, how many warps will be generated?

How many of those warps will experience control divergence?

We want to add two vectors A and B, each with N=22000 components. The device supports 1024 threads per block.

What are the configuration parameters required to minimize the number of blocks used? How many threads will be unused?

Write a line of code to calculate the grid index inside of the kernel:

What are the block and thread identifiers associated with element 15000?