# CS398 Report

## Data Presentation for block size of 16

Fig 1. Speed tests for each size of image.

From the graph above, we can see that design 1 is faster than design 2.

## Answered Questions

For the section below we are only considering an image size of 512x512.

### How many multiplication/addition operations are being performed in your convolution kernel? explain.

Number of masks = 8,

Number of additions per mask = 9,

Number of multiplications per mask = 9,

Number of pixels = W\*H\*C = 512\*512\*3 = 786,432

Number of total operations = 8\*2\*9\*W\*H\*C = 113,246,208

### How many global memory reads are being performed by your kernel? explain.

Design 1

Number of pixels = W\*H\*C = 512\*512\*3 = 786,432

Tile width = 16, Block width = 18

Number of global reads = Number of pixels \* Block width \* Block width / Tile width \*Tile width = 995,328

Design 2

Number of pixels = W\*H\*C = 512\*512\*3 = 786,432

Tile width = 14, Block width = 16

Number of global reads = Number of pixels \* Block width \* Block width / Tile width \*Tile width = 1,027,176 (rounded up).

### How many global memory writes are being performed by your kernel? explain.

Number of global writes = Number of pixels = W\*H\*C = 512\*512\*3 = 786,432.

For every thread, only one value is written to the final output. Therefore the number of global writes is equal to the number of pixels

### What is the minimum, maximum, and average number of real operations that a thread will perform? Real operations are those that directly contribute to the final output value.

9 additions, 9 multiplications, 1 division.

### How much time is spent as an overhead cost for using the GPU for computation? Consider all code executed within your host function with the exception of the kernel itself, as overhead. How does the overhead scale with the size of the input?

### What do you think happens as you increase the mask size (say to 1024) while you set the block dimensions to 16x16? What do you end up spending most of your time doing? Does that put other constraints on the way you’d write your algorithm (think of the shared/constant memory size)?

### Do you have to have a separate output memory buffer? Put it in another way, why can’t you perform the convolution in place?

### What is the best parameter for block size?

Fig 2. Speed up for different block size for each design.

The best parameter for block size as seen in above table is 16 for both design 1 and design 2.