ELEC 374 | Digital Systems Engineering

Machine Problem# 3

April 3rd , 2020

Daniyal Maniar | 20064993

Professor: Ahmad Afsahi, P.Eng

Teaching Assistant: Amir Hossein Sojoodi

# Results

## Part 2

The time for the CPU to compute the matrix multiplication was exponentially faster than the GPU computation with a kernel based on one element per thread count. These results show that CPU core is much faster than a GPU core for computations. This is due to the fact that a GPU consists of many weaker cores that are able to work in parallel versus a CPU which has a limited number of powerful cores that compute algorithms sequentially. The results of this experiment can be seen in the chart below:

Figure 1 – CPU Core vs GPU Core

## Part 3

Increasing block widths for the GPU kernel has shown to decrease the computation time. Even though the processing time was reduced from increasing the block width,

Changing block widths in the GPU used to compute the matrix multiplication generally decreased the time to compute the multiplication. As can be seen in the chart however the time to compute the matrix multiplication began increased when the block width exceeded a size of 16. The increase in computation time can be attributed to the increase in local memory swaps the large sized blocks needed to perform. The local memory within blocks is quite limited and an increase in block size for matrix multiplications leads to larger proportions of the input matrices needing to be stored within the local memory. When the block size increases passed a limit the local memory starts to require many more local memory swaps than for smaller block sizes.

Figure 2 - Block width comparison

1. How many times is each element of each input matrix loaded during the execution of the kernel?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Block Width 1 | Block Width 2 | Block Width 4 | Block Width 10 | Block Width 20 | Block Width 25 |
| Times Loaded | 2Nx2N | NxN | (N/2)x(N/2) | (N/5)x(N/5) | (N/20)x(N/20) | (N/25)x(N/25) |

1. What is the floating-point computation to memory-access ratio in each thread?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Block Width 1 | Block Width 2 | Block Width 4 | Block Width 10 | Block Width 20 | Block Width 25 |
| Computations | 2Nx2N | 4Nx4N | 8Nx8N | 20Nx20N | 40Nx40N | 50Nx50N |
| Ratio of memory-access to floating-point computation | 1:1 | 1:4x4 | 1:16x16 | 1:80x80 | 1:800x800 | 1:1250x1250 |

The script output for each of the matrix experiment are provided below.

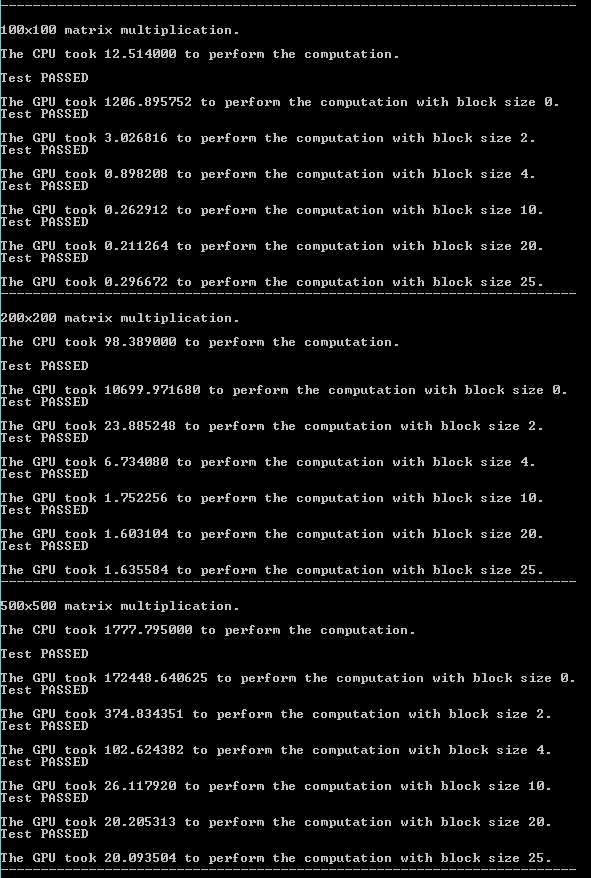


Figure 3 - Matrix Multiplication Results #1

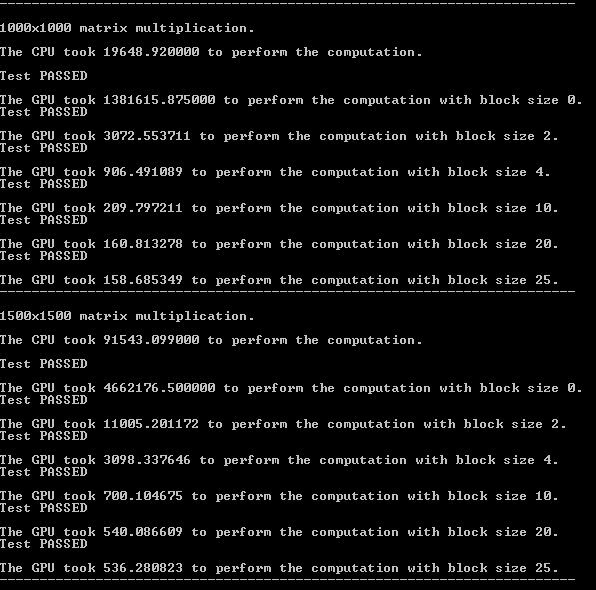


Figure 4 - Matrix Multiplication Results #2

The code for the problem is located in “.\question3\question3\kernel.cu”.