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 1

The transfer times for host to device and device to host were very similar, but when the problem set gets larger, we can see Host to Device transfer time is faster. This change can specifically be seen in the 5000x5000 example in the chart below.

Figure 1 - Transfer Timer vs. Matrix Size

## 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 2 – CPU Core vs GPU Core

## Part 3

Increasing block widths for the GPU kernel has shown to decrease the computation time. We can additionally see that the number of memory accesses decrease as the block width increases, this is due to the blocks sharing cache.

Figure 3 - 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:

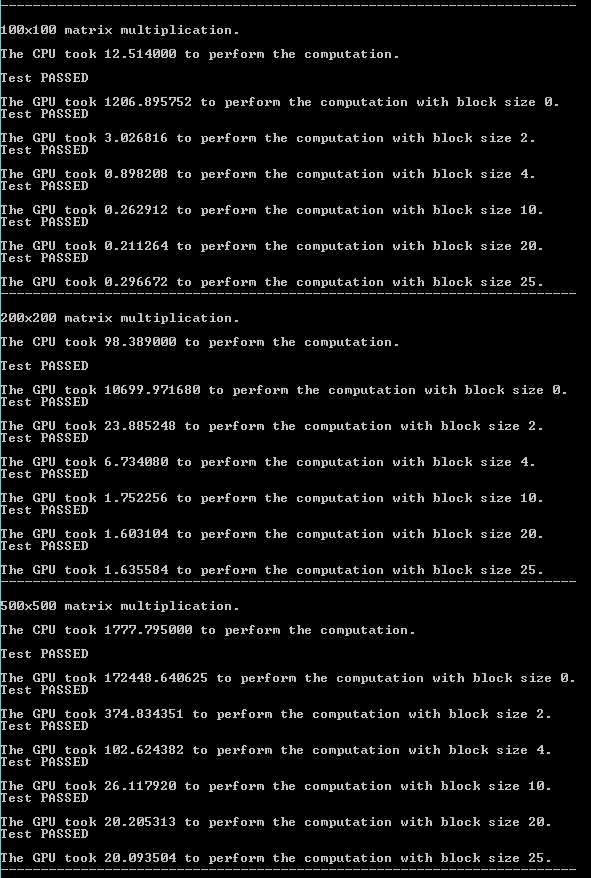


Figure 4 - Matrix Multiplication Results #1

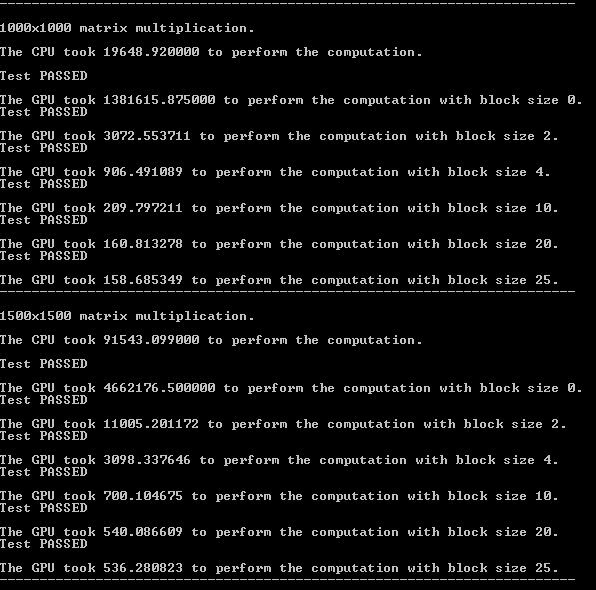


Figure 5 - Matrix Multiplication Results #2

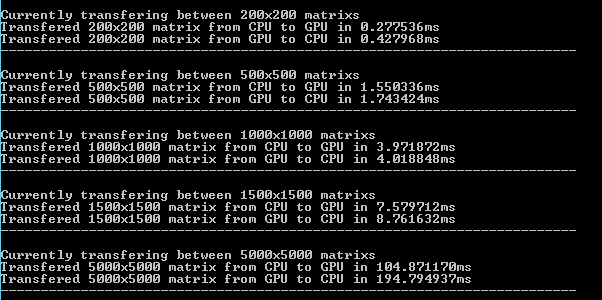


Figure 6 - Transfer Times

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