CS546 Parallel and distributed Processing
Programming Assignment
Performance Evaluation

The assignment carries out Evaluation of the serial and Cuda-c/c++ version of Matrix Normalization algorithm.

We know that Parallel programs were capable of achieving more speedup, when compared with the speedup of the serial code hence supporting the fact that the parallel algorithms can provide great efficiency when compared to serial code.

The massive parallelism of GPUs provides ample of performance for certain algorithms in scientific computing. If we use 1 thread, CPU will definitely be very, very fast than GPU. But to benefit from GPU we create 1000 threads (>1), each inserting the value to memory of each value at its corresponding position. This might result in performance gain over CPU.

In this Assignment, I parallelized the cuda code to perform matrix normalization. And used shared memory to get high performance. By taking more threads for each block depending on the matrix size results better performance.

I've evaluated the elapsed time taken to run Matrix Normalization using CUDA and Serial code up to matrix size 8000 on Jarvis.

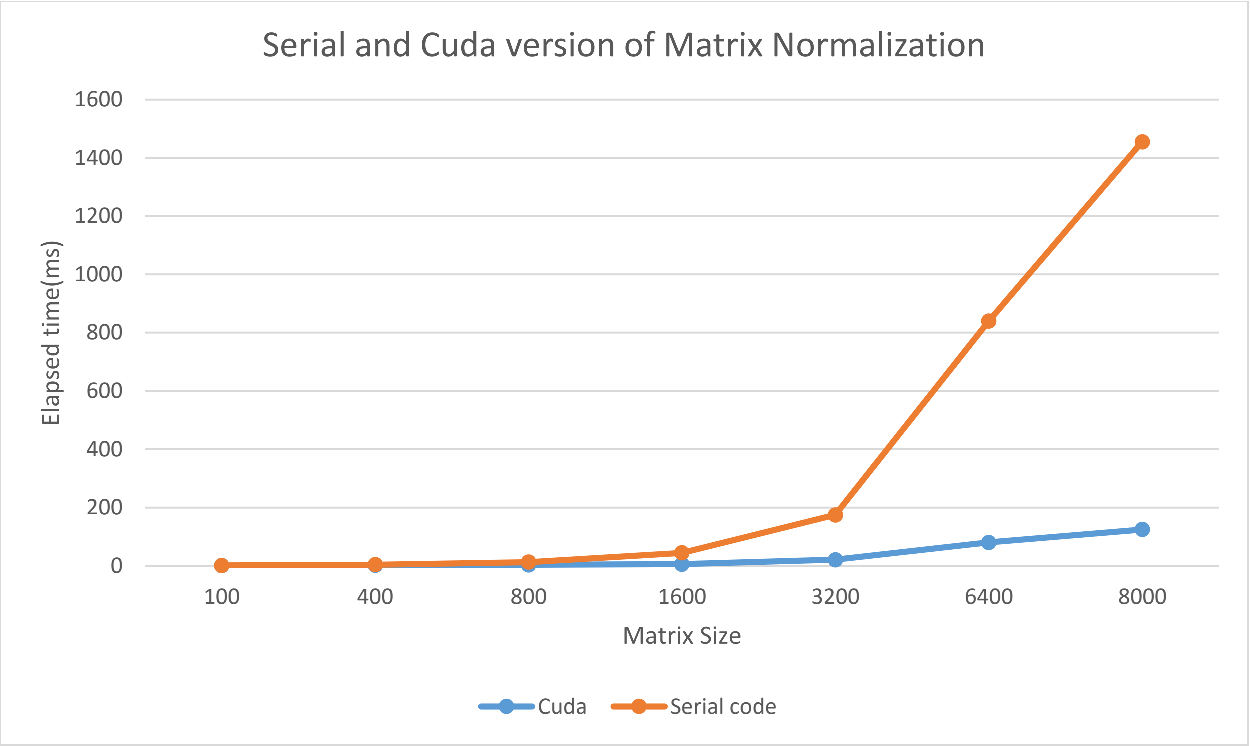
The below table and plot drawn based on elapsed time and Matrix size.

Note: Elapsed time in milliseconds

Matrix Size	100	400	800	1600	3200	6400	8000
Cuda	0.5	1.29	3.25	5.58	20.52	79.93	124.57
Serial	0.55	2.301	9.109	38.856	153.879	760.475	1330.83

Graph: X axis: Matrix Size

Y-axis: Elapsed time (ms)



Observation: From the above graph, we can observe that Cuda version is almost 10 times faster than the serial version of the matrix normalization. I parallelized the cuda code and used shared memory to get high performance. and based on the matrix size, I've taken number of blocks and number of threads in each block. we know that, Between 128 and 256 threads per block is a better choice and a good initial range for experimentation with different block sizes. I've calculated number of blocks and threads based on size of matrix. By increasing the matrix size, elapsed time taken for serial code is higher than cuda version code. The performance of cuda on gpu's is almost ten times faster than serial code.