Tezan Sahu
170100035
tezansahu@iitb.ac.in

Homework 3
ME766: High Performance Scientific Computing

Due Date: 08/05/21

Problem Statement

Create two matrices, A and B, each of size $(N \times N)$. Initialise the matrices to random floating point numbers (choose A & B to be the same as HW2). Write CUDA (or OpenCL) code for computing $C = AB$.
Vary the size of the problem from $N = 100...10000$ & analyse the running time of the codes.

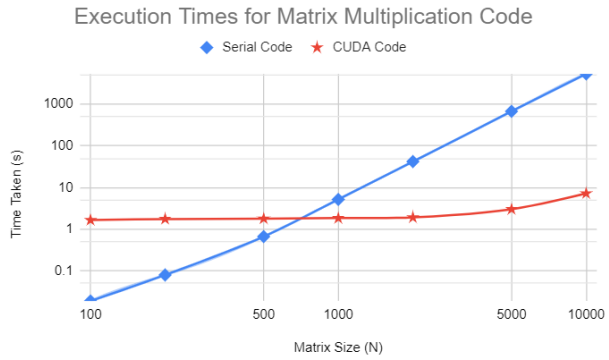
CUDA Implementation

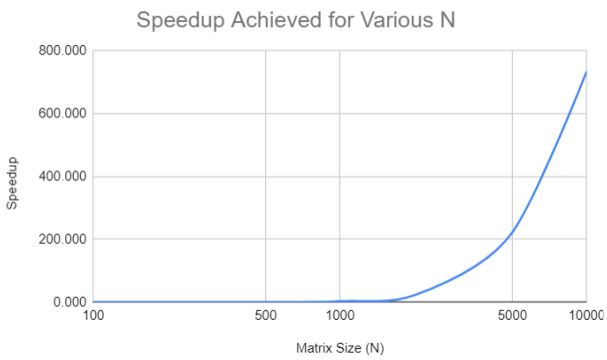
- The CUDA code has been written in the `matmul.cu` file
- Matrices A & B are initialized using seeded random numbers, same as in HW2
- For matrix size $N \times N$, we initialize a kernel of size `<<N, N>>`, i.e., containing N blocks, each with N threads.
- The `blockIdx.x` is used to index rows, while `threadIdx.x` is used to index columns of the matrices inside the kernel
- Each thread j in each block i computes the value corresponding to $C[i, j]$ of the product matrix

Results & Discussion

The execution times of the CUDA version of matrix multiplication as compared to the corresponding serial versions is mentioned in the table below. The respective speedups are also indicated.

N	Execution Time (s)		Speedup
	Serial Code	CUDA Code	
100	0.019	1.662	0.011
200	0.08	1.757	0.046
500	0.666	1.793	0.371
1000	5.188	1.848	2.807
2000	41.911	1.922	21.806
5000	672.882	3.024	222.514
10000	5314.329	7.228	735.242





- The running time for serial matrix multiplication code varies approximately $O(N^3)$
- For smaller values of N (typically < 700), the serial code takes less time for execution compared to CUDA code, mostly due to the extra overheads of copying data between host & GPU being much more than the actual computation costs
- For larger values of N ($>= 1000$), it is evident that the GPU provides immense benefits through parallelization, as speedups rise exponentially.

Machine Configuration

OS & CPU Specifications

Operating System	Ubuntu 18.04 *[ParamSanganak]*
Architecture	x86_64
CPU op-mode(s)	32-bit, 64-bit
Byte Order	Little Endian
CPU(s)	40
Thread(s) per core	1
Core(s) per socket	20
Socket(s)	2
Model name	Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz
CPU MHz	999.908

GPU Specifications

Model	Tesla V100-SXM2-16GB
Bus Type	PCIe
DMA Size	47 bits