E0 243 Computer Architecture Assignment 02
Part B

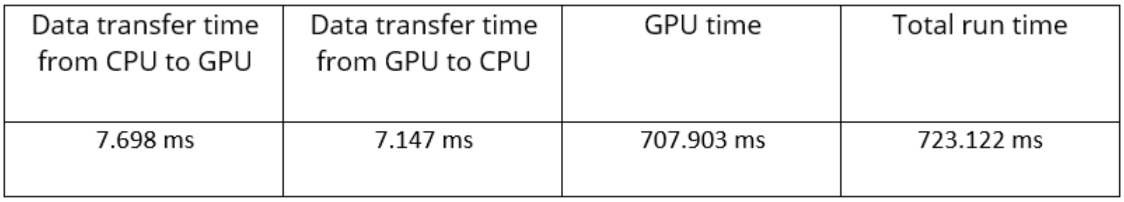
Kawin M
(kawinm@iisc.ac.in)

November 25, 2021

Used **GPGPU-Sim** for running the CUDA program.

Configuration: SM7 TITAN V

Performance



Data transfer time from CPU to GPU	Data transfer time from GPU to CPU	GPU time	Total run time
7.698 ms	7.147 ms	707.903 ms	723.122 ms

Data transfer took 2.05% of total runtime and GPU run time took 97.9% of total runtime.

Approach

Each thread in the Kernel call computes the value of one cell of the output matrix.

For the input N = 128, a kernel with 256 threads per block and 32 blocks per grid is created and called.

L1 Cache

L1 Cache Miss Rate = 0.9017

L2 Cache Miss Rate - 0.0149

The L1 Cache Hit Rate is improved when number of threads executed per block is increased, because of increase in common memory accesses among threads. For example, for input N = 128 and when a kernel with 2048 threads per block and 4 block is called, the L1 Cache Miss rate came down to be around 26.11%.

GDDR Bandwidth = 82.17 GB/s