Name: Rohit Pujar Nagraj
USC ID: 4664 4939 18

# Programming Homework 5 - Report

## 2.1 Naive Matrix Multiplication on CUDA

```
❯ ./p1
time is 12374345.000000 ns
Value of C[451][451] = 2048
```

Kernel Execution Time: 12,374,345 ns

## 2.2 Block Matrix Multiplication on CUDA

```
❯ ./p2
time is 3590915.000000 ns
Value of C[451][451] = 2048
```

Kernel Execution Time: 3,590,915 ns

**Observations:** The block matrix multiplication is around 4 times faster than the naive implementation. This is because CUDA GPUs have around 32kB of shared memory, which is around 16 times faster than access bandwidth (source) compared to global memory. In the block matrix multiplication, we explicitly pull a block into the shared memory. Each data point is accessed n times while calculating C. And, we save some time during every access, leading to the observed speedup.