Aaryan Bhagat 862468345

Q1

What we calculate and even find out later using gpgpu analysis is that 256 * 256 matrices multiplication has very much less load store operations as compared to 1024*64 matrices (almost 4 times less).

For 256*256 it will be 4 * 256 * 256 * 2 load operations

For 1024*64 it will be 64 * 1024 * 24 * 2 load operations

Q2

Each tile will multiply 4 such tiles from matrix B, so 4 memory access for one element.

Q3

Each element will be accessed 64 times as each row of matrix A will be multiplied with 64 columns of matrix B one by one.

Q4

For 128 * 128 matrices

Tile size 16

```
gpu_tot_sim_cycle = 27576
gpu_tot_ipc =     469.3705
gpgpu_n_load_insn  = 262144
gpgpu_n_store_insn = 16384
gpgpu_n_shmem_insn = 4456448
```

Tile size 8

```
gpgpu_n_load_insn  = 262144
gpgpu_n_store_insn = 16384
gpgpu_n_shmem_insn = 4456448
gpu_tot_sim_cycle = 27576
gpu_tot_ipc =      469.3705
```

Tile size 32

```
gpu_tot_sim_cycle = 27576
gpu_tot_ipc =      469.3705
gpgpu_n_load_insn  = 262144
gpgpu_n_store_insn = 16384
gpgpu_n_shmem_insn = 4456448
```