ECE786: PROGRAMMING ASSIGNMENT ONE

PART A : CUDA CODE

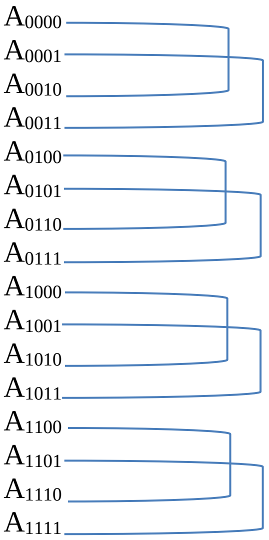
ALGORITHM

1. Problem Statement Analysis

The problem statement computes matrix multiplication of a one-bit quantum gate (U) on the nth bit of an N qubit quantum state (A).

Example
U: Quantum Gate → 2 x 1
N: Number of states in the qubit
A: N state qubit → 2^N x 1
n: bit on which the quantum gate is applied.

For a given value of n, the quantum gate is applied on a pair of qubits of the quantum state that only differ in their nth bit.
Example: N = 4, n = 0



2. *Thread Organization*

The proposed solution uses a one-dimensional thread organization across gird as well as the thread blocks. This implies each gird will have a 1D array of thread blocks and each corresponding thread block will have a 1D array of threads.

3. *Kernel Function*

Each thread will perform matrix multiplication on one set of input pairs. The global thread id will be mapped to its corresponding pairs.

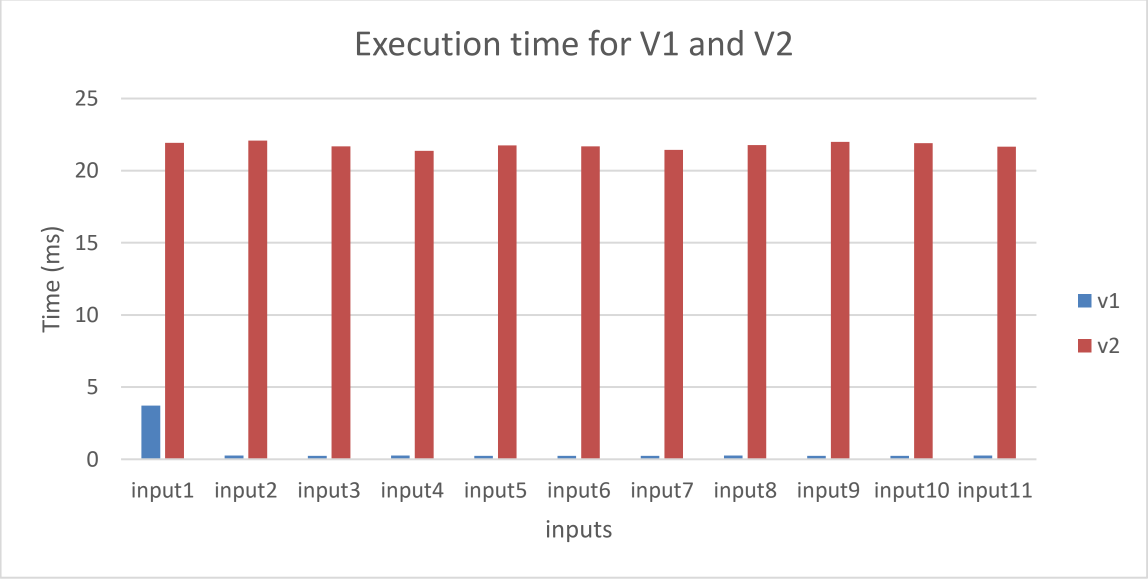
$$index_1 = (tid \ \% \ 2^n) + (2^{n+1} * (tid/2^n))$$
$$index_2 = index_1 + 2^n$$

tid: global thread id
n: position of bit on which gate is applied.

TIMING DIFFERENCE

- ⇒ The two versions of the cuda code differ in the way the memory is managed between the host and the device. *cudaMallocManaged()* uses unified memory. It acts as an abstraction for programmers, underneath the host and device each access their own memory. *cudaMalloc()* requires the programmer to explicitly handle memory separately on the host and device.
- ⇒ Use of *cudaMallocManaged()* significantly reduces the prototyping time for cuda code development. However, analysing the runtime for both versions of the code *cudaMalloc()* followed by *cudaMemCpy()* is considerably faster. The comparison is depicted in the figure below.

Input	execution time (ms)(cudaMalloc)	execution time (ms)(cudaMallocManaged)
input1	3.724288	21.921791
input2	0.25088	22.083584
input3	0.248832	21.683201
input4	0.252896	21.372929
input5	0.242688	21.745665
input6	0.248832	21.696512
input7	0.24576	21.438463
input8	0.258048	21.768192
input9	0.247808	21.992449
input10	0.243712	21.916672
input11	0.26624	21.659649

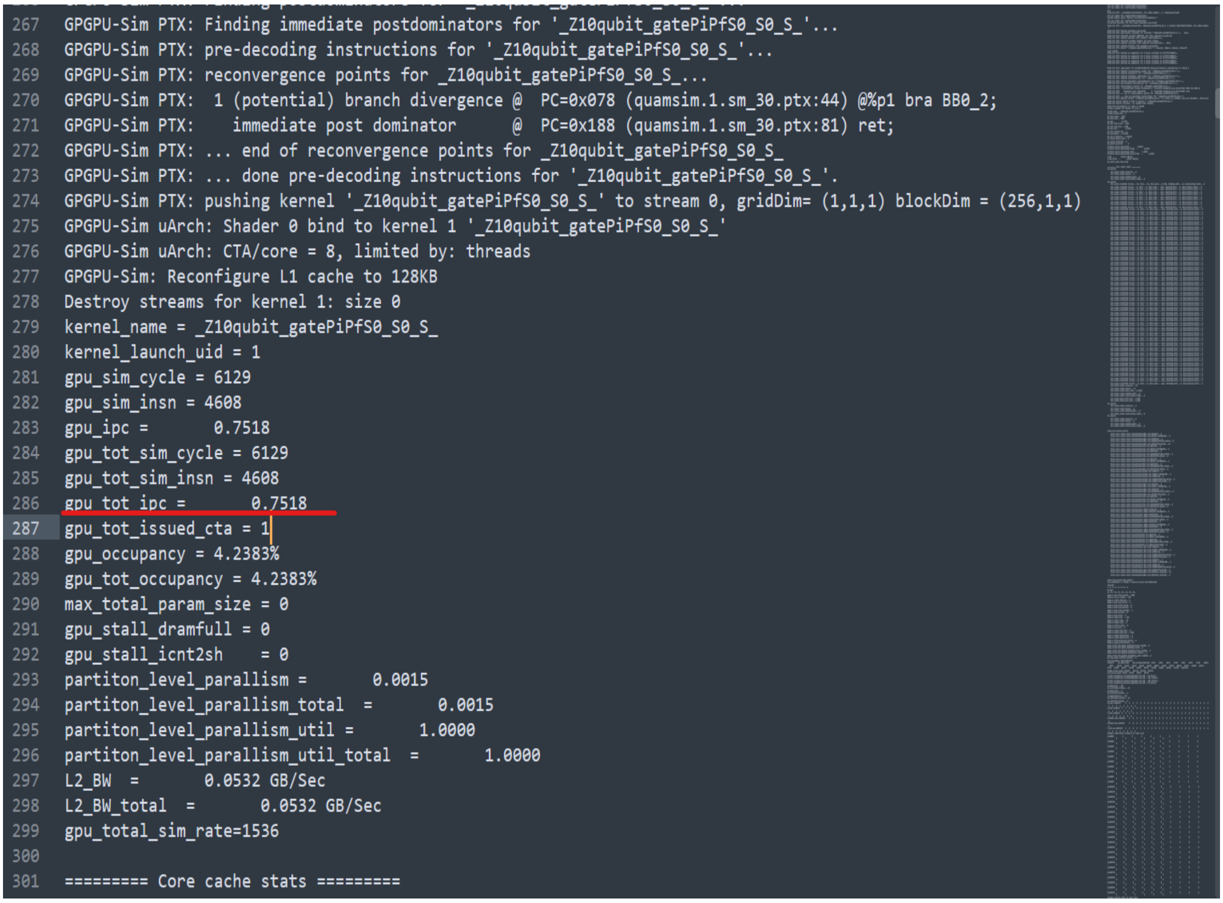


PART B: SIMULATION ON GPGPUSIM

WHAT IS THE IPC OF YOUR PROGRAM AND HOW IS THIS VALUE CALCULATED FROM THE STATISTICS?

⇒ For input1.txt provided with the assignment the IPC was 0.7518. This value is provided as a part of the complete simulation as depicted in the screenshot below.

⇒ It can be easily verified that the IPC is computed using the given formula: *gpu_sim_insn / gpu_sim_cycle*



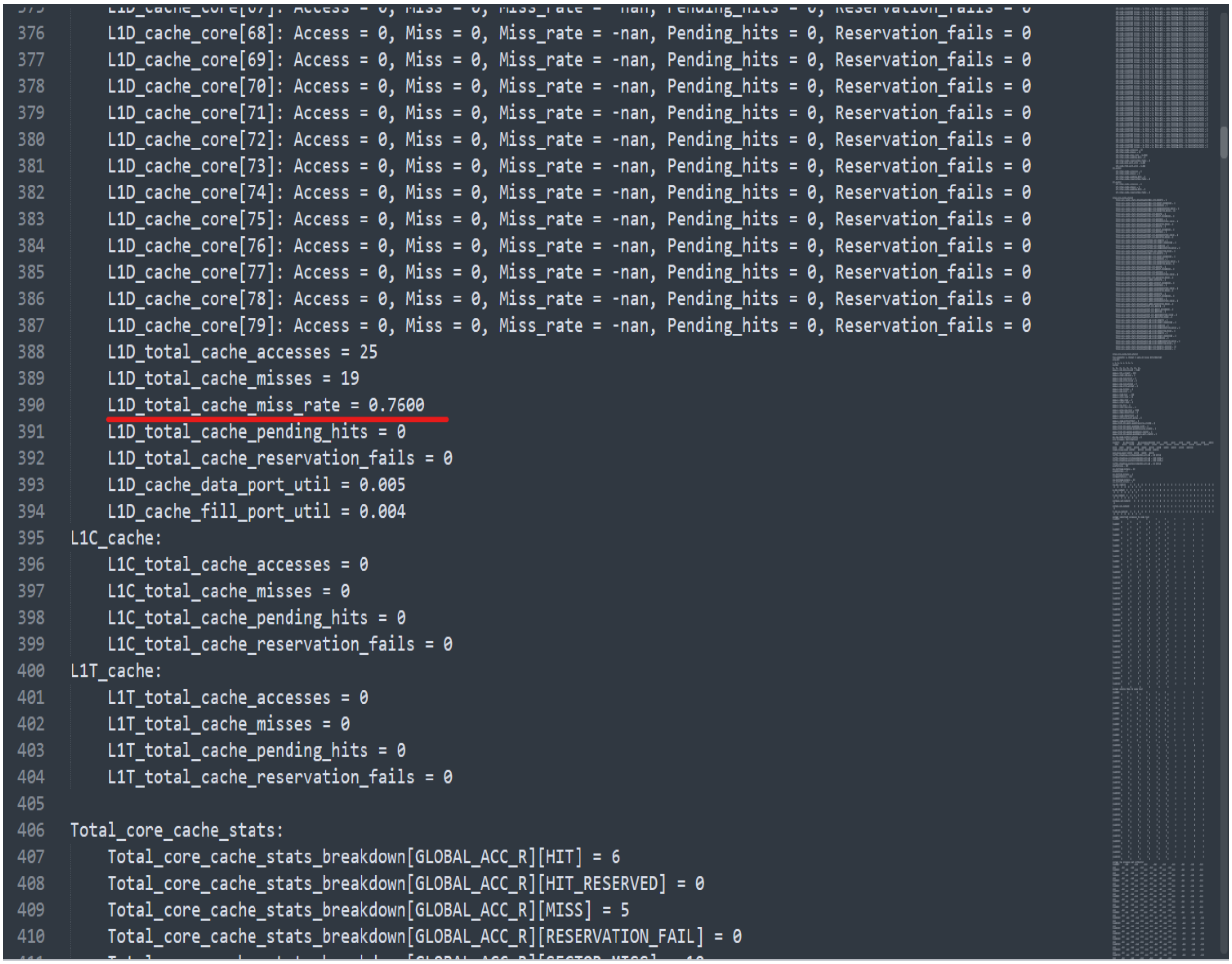
```
GPGPU-Sim PTX: Finding immediate postdominators for '_Z10qubit_gatePiPfS0_S0_S_'...
GPGPU-Sim PTX: pre-decoding instructions for '_Z10qubit_gatePiPfS0_S0_S_'...
GPGPU-Sim PTX: reconvergence points for _Z10qubit_gatePiPfS0_S0_S_...
GPGPU-Sim PTX:  1 (potential) branch divergence @  PC=0x078 (quamsim.1.sm_30.ptx:44) @%p1 bra BB0_2;
GPGPU-Sim PTX:    immediate post dominator      @  PC=0x188 (quamsim.1.sm_30.ptx:81) ret;
GPGPU-Sim PTX: ... end of reconvergence points for _Z10qubit_gatePiPfS0_S0_S_
GPGPU-Sim PTX: ... done pre-decoding instructions for '_Z10qubit_gatePiPfS0_S0_S_'.
GPGPU-Sim PTX: pushing kernel '_Z10qubit_gatePiPfS0_S0_S_' to stream 0, gridDim= (1,1,1) blockDim = (256,1,1)
GPGPU-Sim uArch: Shader 0 bind to kernel 1 '_Z10qubit_gatePiPfS0_S0_S_'
GPGPU-Sim uArch: CTA/core = 8, limited by: threads
GPGPU-Sim: Reconfigure L1 cache to 128KB
Destroy streams for kernel 1: size 0
kernel_name = _Z10qubit_gatePiPfS0_S0_S_
kernel_launch_uid = 1
gpu_sim_cycle = 6129
gpu_sim_insn = 4608
gpu_ipc =       0.7518
gpu_tot_sim_cycle = 6129
gpu_tot_sim_insn = 4608
gpu_tot_ipc =       0.7518
gpu_tot_issued_cta = 1
gpu_occupancy = 4.2383%
gpu_tot_occupancy = 4.2383%
max_total_param_size = 0
gpu_stall_dramfull = 0
gpu_stall_icnt2sh    = 0
partiton_level_parallism =       0.0015
partiton_level_parallism_total  =       0.0015
partiton_level_parallism_util =       1.0000
partiton_level_parallism_util_total  =       1.0000
L2_BW  =       0.0532 GB/Sec
L2_BW_total  =       0.0532 GB/Sec
gpu_total_sim_rate=1536

========= Core cache stats =========
```

WHAT IS THE DATA CACHE MISS_RATE AND HOW IS THIS VALUE CALCULATED FROM THE STATISTICS?

⇒ For input1.txt, the data cache miss rate is 0.76. This value is provided by the L1D_cache metrics as shown in the below screenshot.

⇒ This is computed by the following formula: *LD1_total_cache_misses/ LD_1 total_cache_accesses.*



```
    L1D_cache_core[67]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[68]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[69]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[70]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[71]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[72]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[73]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[74]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[75]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[76]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[77]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[78]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_cache_core[79]: Access = 0, Miss = 0, Miss_rate = -nan, Pending_hits = 0, Reservation_fails = 0
    L1D_total_cache_accesses = 25
    L1D_total_cache_misses = 19
    L1D_total_cache_miss_rate = 0.7600
    L1D_total_cache_pending_hits = 0
    L1D_total_cache_reservation_fails = 0
    L1D_cache_data_port_util = 0.005
    L1D_cache_fill_port_util = 0.004
L1C_cache:
    L1C_total_cache_accesses = 0
    L1C_total_cache_misses = 0
    L1C_total_cache_pending_hits = 0
    L1C_total_cache_reservation_fails = 0
L1T_cache:
    L1T_total_cache_accesses = 0
    L1T_total_cache_misses = 0
    L1T_total_cache_pending_hits = 0
    L1T_total_cache_reservation_fails = 0

Total_core_cache_stats:
    Total_core_cache_stats_breakdown[GLOBAL_ACC_R][HIT] = 6
    Total_core_cache_stats_breakdown[GLOBAL_ACC_R][HIT_RESERVED] = 0
    Total_core_cache_stats_breakdown[GLOBAL_ACC_R][MISS] = 5
    Total_core_cache_stats_breakdown[GLOBAL_ACC_R][RESERVATION_FAIL] = 0
    Total_core_cache_stats_breakdown[GLOBAL_ACC_R][SECTOR_MISS] = 19
```