REPORT
ECE 786 Programming Assignment-1
PART-A

The Kernel code:

```
__global__ void apply_quantum_gate(const float *input, float *output, const float
*matrix, int size, int qubit) {
   int in1 = blockIdx.x * blockDim.x + threadIdx.x;
   int mask = 1 << qubit;
   int in2 = in1 ^ mask;
   bool is_lower_half = (in1 & mask) == 0;
   float a, b, res1, res2;
   if (in1 < size && in2 < size && is_lower_half) {
      a = input[in1];
      b = input[in2];
      res1 = fmaf(matrix[0], a, matrix[1] * b);
      res2 = fmaf(matrix[2], a, matrix[3] * b);
      output[in1] = res1;
      output[in2] = res2;
   }
}
```

Explanation:
The function quantum_gate_transform is a CUDA kernel that applies a single-qubit quantum gate to a given state vector. The transformation follows the two-value update rule, where each thread processes two elements of the quantum state.

Host Code Execution:
The program reads the quantum gate matrix, state vector, and target qubit index from a file, then allocates memory on both the host and GPU. Data is transferred to the GPU using cudaMemcpy, and the quantum_gate_transform kernel is executed with an optimized grid configuration. Execution time is measured using CUDA events, results are copied back to the host, printed with precision, and finally, all allocated memory is freed.

Timing Report:

Input file	With cudaMalloc	With cudaMallocManaged
Input_qc4_q0	27.884 ms	27.445 ms
Input_qc7_q0	28.77 ms	26.631 ms
Input_qc7_q0	28.163 ms	28.655 ms

PART-B

Ran the GPGPU sim with input_qc4_q2.txt

1). What is the IPC of your program and how is this value calculated from the statistics?

-> gpu_sim_cycle = 5591
gpu_sim_insn = 5280
gpu_ipc = (gpu_sim_insn / gpu_sim_cycle) =(5280/ 5591) = 0.944

2). What is the data cache miss_rate and how is this value calculated from the statistics?

-> L1D_total_cache_accesses = 12
L1D_total_cache_misses = 7
L1D_total_cache_miss_rate = (L1D_total_cache_misses / L1D_total_cache_accesses)
= (7/ 12) = 0.5833