Programming Assignment 1
Shawn Salekin
ECE786

Part A
======

1. The kernel code takes the thread id and use it to calculate the pair of elements in the quantum state that should be computed together. It checks for 1 or 0 at the "t-th" bit and calculates the matrix once it finds that. However, when it figures out that pairing criteria does not match, that thread does not do anything. So there is some optimization that can be done

2. cudaMemAlloc is around 40 us and cudaMemAllocManaged is around 27 us on few of the inputs I have
tested. I am unsure why that is. It is supposed to be the other way - perhaps my usage of cudaMemAllocManaged is
not correct.

Part B
======
3. I have used the sample with that has the smallest quantum state representation.
With performance simulator, these are the stats returned by GPGPU Sim

gpgpu_simulation_time = 0 days, 0 hrs, 0 min, 3 sec (3 sec)
gpgpu_simulation_rate = 1752 (inst/sec)
gpgpu_simulation_rate = 1856 (cycle/sec)

That brings our IPC to 0.9439. (divide 1752/1856)

Data Cache Miss rate is calculated by the following:
   L1D_total_cache_accesses = 12
   L1D_total_cache_misses = 8
(ang GPGPU returns the number)
   L1D_total_cache_miss_rate = 0.6667