Lecture 16 GPU Exercises

Q1:
Assume a hypothetical GPU with the following characteristics:

Clock rate 1.6 GHz, warp size 32 threads, Contains 16 SIMD processors, each containing 32 single-precision floating-point units, each instruction performs **one** single-precision floating-point operation.

What is the peak single-precision floating-point throughput for this GPU in GFLOP/sec, assuming that all memory latencies can be hidden?

Solution:

1.6 * 16 * 32 = 819.2 GFLOPS

Q2:
If a program contains 1024 threads, how many warps does it take to execute this program with 32-wide SIMD execution?

Solution:

1024 / 32 = 32