Quiz 0430

1. Briefly describe the relationship between cuda thread, block, grid。Are GPUs just SIMD vector multiprocessors ?

2. Assume a hypothetical GPU with the following characteristics:
    - Clock rate 1.5 GHz
    - Contains 16 SIMD processors, each containing 16 single-precision floating-point units
    - Has 100 GB/sec off-chip memory bandwidth

Without considering memory bandwidth, what is the peak single-precision floating-point throughput for this GPU in GFLOP/sec(GFLOP Giga single-precision floating-point), assuming that all memory latencies can be hidden? Is this throughput sustainable given the memory bandwidth limitation?