Notes on GPUS

- What does a CPU do? Small computations (add, move) very fast, typically 1-3 GHz.
- Moore's Law "the number of transistors on integrated circuits doubles approximately every 18 months"
- CPUs are often multi-core (2, 4, 8 common today) – what does each core do?
- Memory access is much slower than CPU computation - latency
    - Register 0.5 ns "bite in your mouth"
    - RAM Memory 100 ns "open fridge, make sandwich"
    - (USB stick) Flash memory 100,000 ns
    - Hard disk 1,000,000 ns (.001 seconds) "drive to store, buy ingredients"
    - IP Packet 10,000,000 (0.01 seconds)
- So often the CPU "stalls" while it waits for information to be moved from some kind of memory to a register, or vice versa.
- Typical CPU solution: cache
    - L1 1 ns "in hand"
    - L2 3 ns "on counter"
    - L3 10 ns "in lunchbox"
- Typical GPU solution: many more threads, each running the same code, but with different data.
- No cache or much less cache leaves space on chip for more memory for threads.
- What lends itself to SIMD – single instruction, multiple data; data parallel
- Nvidia Geforce GTX 780 2304 cores! About $680 (May, 2016)
- CUDA: way to program the GPU for many purposes including non-graphics