- GPU history? http://searchvirtualdesktop.techtarget.com/definition/GPU-graphics-processing-unit
    - Offload processing from CPU, as more graphical programs get written
- GPU Programming research paper: http://compsci.hunter.cuny.edu/~sweiss/course_materials/csci360/lecture_notes/gpus.pdf
    - 1999 - Nvidia, first GPU
    - Video graphics array controller (VGA) traditional
    - 3D functions:
        - triangulation
        - rasterization
        - texture mapping and shading
- GPU Computing: http://lorenabarba.com/gpuatbu/Program_files/Cruz_gpuComputing09.pdf
    - Massively parallel
    - Hundreds of cores
    - Thousands of threads
    - Cheap
    - Highly available
    - Programmable: CUDA (Nvidia's Compute Unified Device Architecture)
        - 2006
        - Compiling and toolkit for programming NVIDIA GPUs
        - API extends C programming language
        - Abstraction from hardware
    - What is HPC?
    - Composed of processor cores, texture, ROP, Setup raster, Frame buffer, and Thread scheduler
    - Depends on non-parallel part: Amdahl's law
    - OpenCL - industry standard **learn opencl?**
    - Architecture is layered (Application, C + extension, CUDA)
    - Kernel is simple C
    - Memory
        - Global mem (4gb)

    - Shared mem (16kb)
    - Registers (16kb)
- Latency
    - Global: 400-600 cycles
    - Shared mem: fast
    - Register: fast
- Purpose
    - Global: IO for grid
    - Shared: thread collaboration
    - Register: thread space