EN 600.320/420 Parallel Programming

GPU Architecture
Pascal Architecture
Pascal Architecture
Pascal Architecture

- **CORE** = single-precision unit
- **DP** = double-precision
  - Half the flops as single-precision
- **SFU** = special functional unit
  - Fused multiply and add
- **LD/ST** = load/store

- 16-bit support
  - For machine learning
Pascal Die Shot
I7 die for comparison

- More cache real estate
- Fewer cores
GPU Architecture

- Fundamental unit is the CUDA processor
  - Integer arithmetic logic unit ALU
  - Double-precision floating point FPU
  - Fused multiply-add instruction Fully pipelined

- CUDA processors are grouped into stream multiprocessors (SM)
- SMs run in SIMD lockstep
Each stream-multiprocessor has 64 CUDA processors
Each GPU can have 60 multiprocessors

→ GPU has up to 3840 CUDA cores (GP 100)
What’s new?

- How has the architecture evolved to be suitable for generally purpose computing
Double Precision

- Until Fermi, double precision was 4 cycles per instruction
- Why?
Memory Changes

- Caching
  - L1 per CUDA core
  - L2 per SM
- Error Correcting (ECC) memory
- Why weren’t these needed for graphics cards
Scheduling

- Concurrent program execution
- Lightweight context switching
Unfied Memory

- Single address space for GPU+CPU
  - Managed/virtual memory for CUDA programs
Multi-GPU (interconnect)

- 5x performance of PCI/e
- Solve bigger problems
nVidia GPU compute evolution

- Fermi: a general purpose compute platform
  - That runs graphics pipelines too
- Kepler/Maxwell – incremental, but good
- Pascal – multi-GPU (HPC) and 16-bit (ML)

Has moved beyond its history as a graphics pipeline