Overview

- Kepler Architecture
- Streaming Multiprocessor (SMX)
- Improvements over Fermi Architecture
- Hyper-Q Technology
- Dynamic Parallelism
- Nvidia GRID
- Kepler Applications
Nvidia Kepler Microarchitecture

- 192 single-precision CUDA cores
- 28 nm manufacturing process
- 15 SMX processors
- Six 64-bit memory controllers
Kepler SMX Processor

- 192 Single-precision
- 64 Double-precision
- 32 Load/Store
- 32 Special Function
- 16 Texture
- 65536 32-bit Registers
- 4 Warps
Nvidia GPU Cache Hierarchy

- L1/L2 Unified Cache
- 64 KB Configurable Shared Memory and L1 Cache
- 48 KB Read-Only Data Cache
- 1536 KB L2 Cache
- Content protected using SECDED EEC
Kepler Quad-Warp Scheduler

- 1 Warp = 32 parallel threads
- Each SMX can handle 4 warps and 8 instruction dispatch units
- Each warp allows two independent instructions per cycle
Kepler vs Fermi (Last Gen)

- Most cores operate on GPU core clock instead of shader clock
  - 3x increase in performance per watt at cost of larger area
- 5 additional atomic operations
  - Increases performance of parallel programming
- Replaced texture binding table with bindless texture states
  - Allows for unlimited simultaneous texture access
- Shuffle instruction added to ISA
  - Shares data between threads in the same warp

![Performance Comparison Chart]

![Kepler vs Fermi Diagram]

Kepler: Fast & Efficient

**SM Fermi**

**SMX Kepler**

3x Perf / Watt

32 cores

192 cores
Hyper-Q Technology

- 32 simultaneous hardware-managed connections/queues
- Vastly increases GPU utilization by keeping many more cores busy
- Removes false dependencies that caused bottlenecks
Dynamic Parallelism

- Removes dependence on CPU for launching new kernels
- Allows GPU to generate work for itself
- Parallel/recursive/nested loop heavy code can run solely on GPU
- Frees up CPU to work on other tasks
Kepler and Cloud Computing: GRID

- First true GPU Virtualization solution
- Comes in GPU or standalone modules
- Contains up to 4 Kepler based cards per unit
- Boasts extreme energy efficiency
NVIDIA VGX Accelerated Cloud

Hypervisor
- GPU Hypervisor
- Virtual GPU Manager
- Resource Manager

Virtual Machine
- Guest OS
  - Apps
  - Virtual Desktop
    - Delivery Protocol
    - User Selectable Machine Image (USM)
    - Remote Display

KEPLER
- GPU MMU
- Per-VM Dedicated Channels

State
Graphic Commands
Kepler Applications: Jetson TK1

Tegra K1 SOC
- NVIDIA Kepler GPU with 192 CUDA cores
- NVIDIA 4-Plus-1™ quad-core ARM® Cortex-A15 CPU
16 GB eMMC
1 GB memory
Gigabit Ethernet
USB 3.0
SD/MMC
miniPCIe
HDMI 1.4
SATA
Line out/Mic in
RS232 serial port
Expansion ports for additional display, GPIOs, and high-bandwidth camera interface
Jetson TK1 (Cont)

Tegra K1

- **GPU**
  - Kepler GPU (192 CUDA Cores)
  - Open GL 4.4, Open GL E33.0, DX11, CUDA 6

- **CPU**
  - Quad Core Cortex A15 “r3”
  - With 5th Battery-Saver Core; 2MB L2 cache

- **Camera**
  - Dual High Performance ISP
  - 1.2 Gigapixel throughput, 100MP sensor

- **Power**
  - Lower Power
  - 28HPM, Battery Saver Core

- **Display**
  - 4K panel, 4K HDMI
  - DSI, eDP, LVDS, High Speed HDMI 1.4a
Questions?

IS THIS REAL LIFE?!