NVIDIA’S TEGRA K1 SYSTEM-ON-CHIP

Michael Ditty, Tegra Architecture
Co-authors: John Montrym, Craig Wittenbrink
Tegra K1

**GPU**
- Kepler GPU (192 CUDA Cores)
  - OpenGL 4.4, OpenGL ES3.1+AEP, DX12, CUDA 6

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

**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
Overview

- Kepler into Mobile
- Tegra ISP
- Power Management
- Mobile Enablement
- Demo Intro
A Major Discontinuity in Mobile Graphics

ES2.0, DX9
Programmable Pixel Shaders

ES3.1+AEP, OGL4.4, DX12
Tessellation, Compute Shaders, ASTC, GPGPU
Mobile Roadmap Meets GeForce

MOBILE ARCHITECTURE

Tegra 3

Tegra 4

Fermi

GEFORCE ARCHITECTURE

Kepler

Tegra K1

Maxwell

Advancements
<table>
<thead>
<tr>
<th>Metric</th>
<th>Tegra 4</th>
<th>Tegra K1</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32 ops</td>
<td>48</td>
<td>384</td>
<td>Per clock</td>
</tr>
<tr>
<td>Z-only Primitives</td>
<td>0.1</td>
<td>1</td>
<td>Per clock</td>
</tr>
<tr>
<td>Zcull</td>
<td>-</td>
<td>256</td>
<td>Pixels/clk</td>
</tr>
<tr>
<td>Raster</td>
<td>8</td>
<td>64</td>
<td>Samples/clk</td>
</tr>
<tr>
<td>Texture</td>
<td>4</td>
<td>8</td>
<td>Bilinear filters/clk</td>
</tr>
<tr>
<td>ZROP</td>
<td>8</td>
<td>64</td>
<td>Samples/clk</td>
</tr>
<tr>
<td>L2 size</td>
<td>32</td>
<td>128</td>
<td>KBytes</td>
</tr>
</tbody>
</table>
Tegra K1 / Kepler Graphics Core Architecture

- 192 CUDA cores
- Unified Memory Cache
- Dedicated Accelerators
  - Geom / Tessellation
  - Z Cull
  - Z / Color ROP
Power Efficiency

- Clock and power gating
  - Multi-level Clock Gating
  - Power Gating
  - Rail Gating
- Architectural power improvements
  - Interconnect and Data Paths architected for mobile
  - Shader Bypass
  - GPU L2 Cache and Compression
- Work reduction
  - Aggressive Culling Of Z, Stencil, Attribute Fetch
  - Early Z
Dual Next Gen ISP

Performance
- 1.2Gp total pixel throughput
- 600Mp each ISP
- 4096 simultaneous focus points
- 14 bits input
- 100Mp camera support

Interoperability
- Reconfigurable ISP fabric
- Full GPGPU interoperability
- Memory or Isochronous sourcing
Tegra K1 Computational Photography Architecture

GPU + ISP + CPU

Frame/Image Bus

Kernels CPU
- K0
- K1
- ... Kn

Kernels GPU
- K0
- K1
- ... Kn

ISP-A
- ATO
- LS
- NR
- LAC0
- H1
- FB
- AT1
- DS
- FX

ISP-B
- ATO
- LS
- NR
- LAC0
- H1
- FB
- AT1
- DS
- FX

State Bus

CSI

V1-Mux

S0
- S1
- ... Sn

GPU + ISP + CPU

CPU Kernels

GPU Kernels

ISP Kernels
GPU Power Management

GPU Idle State Transitions

- Active
- Power gating
- Rail gating

Power Usage vs. Time

Idle Transition
Multi-core Gaming

- CPU
- Kepler GPU
- Video Decode
- ISP
- ISP
- Video Encode
- Audio
- Display
- 64bit DRAM
Multi-core gaming power management

- Balance power & performance across cores and power rails.
- Clocking policies must look at more than active time.
- Power optimization must be done globally, not locally to each unit.
Multi-core video processing

“Live” Local Tone Mapping

Kepler GPGPU Processing
30fps

Original

LTM
Multi-core video processing

- CPU
- Kepler GPU
- Video Decode
- Video Encode
- ISP
- ISP
- Display
- Audio
- 64-bit DRAM
Multi-core video processing

Utilize burst performance for latency reduction
Tegra K1 Benchmarks

Performance relative to Fastest Competitor

- GFXBench 3.0 Manhattan
- GFXBench 3.0 Trex-HD
- AndEBench-Pro

- Shield Tablet
- Competitor X
- Shield Portable
Mobile Compute

- CUDA
- VisionWorks Toolkit
- Renderscript

Tango Tablet

Automotive Computer Vision
Tegra K1 Compute Benchmark

Compubench RS (Geometric Mean)

Performance relative to Fastest Competitor

- Shield Tablet
- Competitor X
Consumer Devices

Shield Tablet

Acer Chromebook 13

Xiaomi MiPad
NVIDIA Dabbler
Improving the user experience with Tegra K1

- Watercolor
  - GPGPU simulates realistic water
- Oil painting
  - 3D modeling enables realistic lighting
- Low Pen-to-Ink latency
  - Optimized GPU rendering paths to reduce latency.
Conclusion

- New capabilities in mobile
  - Compute, OpenGL 4.4, Advanced Imaging Pipeline

- Great performance
  - Over 2x the performance of current mobile devices

- Enabling new platforms and ecosystems
We would like to thank the GPU & Tegra teams across NVIDIA who collaborated to make this chip possible.