Agenda

• What is a GPU?
• GPU evolution!!
• GPU costs
• GPU pipeline
• Fermi architecture
• Kepler architecture
• Pascal architecture
• Turing architecture
• AMD Radeon R9
• AMD vs Nvidia
A graphics processing unit is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.
What is it a GPU?

HOW DO GRAPHICS CARDS WORK?
GPU Evolution

Moore’s Law at work

1997
RIVA 128
3M transistors

2000
GeForce 256
23M transistors

2005
GeForce 3
60M transistors

2010
GeForce FX
125M transistors

GeForce 8800
681M transistors

GeForce GTX 580
3B transistors

image:reference
Evolution of Nvidia GeForce 1999 - 2018
First GPU

World’s First GPU

Nvidia GeForce 256

Modern GPU

GeForce GTX 1050

IN 2018
CUDA

CUDA is a parallel computing platform and programming model developed by NVIDIA for general computing on graphical processing units (GPUs). With CUDA, developers are able to dramatically speed up computing applications by harnessing the power of GPUs.
Graphics Processing

Input geometry + attributes
Memory
Geometry
Textures
Buffers
Frame buffers
Vertex processing
Vertex
Primitive processing
Primitive
Rasterization
Fragment processing
Fragment
Pixel processing
Pixel
GPU pipeline

3D THEORY
What is a Graphics Pipeline?
Why GPUs so fast?
Why GPUs so fast?
GPU has more transistors.
simplified Nvidia GPU architecture

MP = Multi Processor
SM = Shared Memory
SFU = Special Functions Unit
IU = Instruction Unit
SP = Streaming processor
reference
Nvidia Fermi GPU Architecture
## Specifications

<table>
<thead>
<tr>
<th>GPU</th>
<th>GT200 (Tesla)</th>
<th>GF110 (Fermi)</th>
<th>GK104 (Kepler)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transistors</td>
<td>1.4 billion</td>
<td>3.0 billion</td>
<td>3.54 billion</td>
</tr>
<tr>
<td>CUDA Cores</td>
<td>240</td>
<td>512</td>
<td>1536</td>
</tr>
<tr>
<td>Graphics Core Clock</td>
<td>648MHz</td>
<td>772MHz</td>
<td>1006MHz</td>
</tr>
<tr>
<td>Shader Core Clock</td>
<td>1476MHz</td>
<td>1544MHz</td>
<td>n/a</td>
</tr>
<tr>
<td>GFLOPs</td>
<td>1063</td>
<td>1581</td>
<td>3090</td>
</tr>
<tr>
<td>Texture Units</td>
<td>80</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>Texel fill-rate</td>
<td>51.8 Gigatexels/sec</td>
<td>49.4 Gigatexels/sec</td>
<td>128.8 Gigatexels/sec</td>
</tr>
<tr>
<td>Memory Clock</td>
<td>2484 MHz</td>
<td>4008 MHz</td>
<td>6008MHz</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>159 GB/sec</td>
<td>192.4 GB/sec</td>
<td>192.26 GB/sec</td>
</tr>
<tr>
<td>Max # of Active Displays</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>TDP</td>
<td>183W</td>
<td>244W</td>
<td>195W</td>
</tr>
</tbody>
</table>

reference for table
Pascal architecture

Nvidia Pascal

The GTX 1070, the second commercially available card to use the Pascal architecture.

<table>
<thead>
<tr>
<th>Transistors</th>
<th>14 nm and 16 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td></td>
</tr>
<tr>
<td>Predecessor</td>
<td>Maxwell</td>
</tr>
<tr>
<td>Successor</td>
<td>Volta (workstation/datacenter) Turing (general)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison table of some Kepler, Maxwell, and Pascal chips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedicated texture cache per SM</td>
</tr>
<tr>
<td>GK104</td>
</tr>
<tr>
<td>N/A</td>
</tr>
<tr>
<td>Texture (graphics or compute) or read-only data (compute only) cache per SM</td>
</tr>
<tr>
<td>N/A</td>
</tr>
<tr>
<td>Programmer-selectable shared memory/L1 partitions per SM</td>
</tr>
<tr>
<td>48 Kib shared memory + 16 Kib L1 cache (default)[26]</td>
</tr>
<tr>
<td>32 Kib shared memory + 32 Kib L1 cache[24]</td>
</tr>
<tr>
<td>16 Kib shared memory + 48 Kib L1 cache[24]</td>
</tr>
<tr>
<td>Unified L1 cache/texture cache per SM</td>
</tr>
<tr>
<td>Dedicated shared memory per SM</td>
</tr>
<tr>
<td>N/A</td>
</tr>
<tr>
<td>L2 cache per chip</td>
</tr>
</tbody>
</table>

Reference
GP100 (Tesla P100)
Pascal Streaming Multiprocessor

Instruction Cache

Instruction Buffer

Warp Scheduler

Dispatch Unit

Dispatch Unit

Register File (32,768 x 32-bit)

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Core

Core

Core

Core

DP Unit

LP/ST

SFU

Texture / L1 Cache

Tex

Tex

Tex

64KB Shared Memory
## Compute Capabilities: GK110 vs GM200 vs GP100

<table>
<thead>
<tr>
<th>GPU</th>
<th>Kepler GK110</th>
<th>Maxwell GM200</th>
<th>Pascal GP100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Capability</td>
<td>3.5</td>
<td>5.2</td>
<td>6.0</td>
</tr>
<tr>
<td>Threads / Warp</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Max Warps / Multiprocessor</td>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>Max Threads / Multiprocessor</td>
<td>2048</td>
<td>2048</td>
<td>2048</td>
</tr>
<tr>
<td>Max Thread Blocks / Multiprocessor</td>
<td>16</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Max 32-bit Registers / SM</td>
<td>65536</td>
<td>65536</td>
<td>65536</td>
</tr>
<tr>
<td>Max Registers / Block</td>
<td>65536</td>
<td>32768</td>
<td>65536</td>
</tr>
<tr>
<td>Max Registers / Thread</td>
<td>255</td>
<td>255</td>
<td>255</td>
</tr>
<tr>
<td>Max Thread Block Size</td>
<td>1024</td>
<td>1024</td>
<td>1024</td>
</tr>
<tr>
<td>Shared Memory Size / SM</td>
<td>16 KB/32 KB/48 KB</td>
<td>96 KB</td>
<td>64 KB</td>
</tr>
</tbody>
</table>
Cross-section Illustrating GP100 adjacent HBM2 stacks
Cross-section Photomicrograph of a P100 HBM2 stack and GP100 GPU
PASCAL DYNAMIC LOAD BALANCING

STATIC PARTITIONING

DYNAMIC BALANCING

GPU Resource Allocation

Time

Graphics
Compute
Idle
Nvidia Turing GPU Architecture
Nvidia Turing TU102 GPU
## NVIDIA Pascal GP102 vs Turing TU102

<table>
<thead>
<tr>
<th>GPU Features</th>
<th>GTX 1080Ti</th>
<th>RTX 2080 Ti</th>
<th>Quadro P6000</th>
<th>Quadro RTX 6000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>Pascal</td>
<td>Turing</td>
<td>Pascal</td>
<td>Turing</td>
</tr>
<tr>
<td>GPCs</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>TPCs</td>
<td>28</td>
<td>34</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>SMs</td>
<td>28</td>
<td>68</td>
<td>30</td>
<td>72</td>
</tr>
<tr>
<td>CUDA Cores / SM</td>
<td>128</td>
<td>64</td>
<td>128</td>
<td>64</td>
</tr>
<tr>
<td>CUDA Cores / GPU</td>
<td>3584</td>
<td>4352</td>
<td>3840</td>
<td>4608</td>
</tr>
<tr>
<td>Tensor Cores / SM</td>
<td>NA</td>
<td>8</td>
<td>NA</td>
<td>8</td>
</tr>
<tr>
<td>Tensor Cores / GPU</td>
<td>NA</td>
<td>544</td>
<td>NA</td>
<td>576</td>
</tr>
<tr>
<td>RT Cores</td>
<td>NA</td>
<td>68</td>
<td>NA</td>
<td>72</td>
</tr>
<tr>
<td>GPU Base Clock MHz</td>
<td>1480 / 1480</td>
<td>1350 / 1350</td>
<td>1506</td>
<td>1455</td>
</tr>
</tbody>
</table>

(Reference / Founders Edition)
Turing Streaming Multiprocessor (SM)
New shared memory architecture

- 2x L1 Bandwidth
- Lower L1 Hit Latency
- Up to 2.7x L1 Capacity
- 2x L2 Capacity

Reference
Turing Shading Performance

50% improved performance per core

Relative Shader Performance

Example shader: YRMark, Sniper Elite 4, Deus Ex, SoW, 3DMark, RoTR, AoS
44 compute units (2816 IEEE 2008-compliant shaders)

- Device Flat (Generic) Addressing support
- Masked Quad Sum of Absolute Difference (MQSAD) with 32b Accumulation and Saturation
- Precision improvement for native LOG/EXP ops to 1ULP
# GCN Architecture Efficiency

<table>
<thead>
<tr>
<th></th>
<th>AMD Radeon™ HD 7970 GHz Edition</th>
<th>AMD Radeon™ R9 290X</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometry Processing</td>
<td>2.1 billion primitives/sec</td>
<td>4 billion primitives/sec</td>
<td>1.9x</td>
</tr>
<tr>
<td>Compute</td>
<td>4.3 TFLOPS</td>
<td>5.6 TFLOPS</td>
<td>1.3x</td>
</tr>
<tr>
<td>Texture fill rate</td>
<td>134.4 Gtexels/sec</td>
<td>176 Gtexels/sec</td>
<td>1.3x</td>
</tr>
<tr>
<td>Pixel fill rate</td>
<td>33.6 Gpixels/sec</td>
<td>64 Gpixels/sec</td>
<td>1.9x</td>
</tr>
<tr>
<td>Peak Bandwidth</td>
<td>264 GB/sec</td>
<td>320 GB/sec</td>
<td>1.2x</td>
</tr>
<tr>
<td>Die area</td>
<td>352 mm²</td>
<td>438 mm²</td>
<td>1.24x</td>
</tr>
<tr>
<td>Peak GFLOPS/mm²</td>
<td>12.2</td>
<td>12.8</td>
<td>1.05x</td>
</tr>
</tbody>
</table>
Asynchronous Compute

- Up to 8 Asynchronous Compute Engines (ACE)
  - Independent scheduling and work item dispatch for efficient multi-tasking
  - Operate in parallel with graphics command processor
  - Each can manage up to 8 queues
  - L2 cache and GDS access
  - Fast context switching

- Dual DMA engines
  - Can saturate PCIe 3.0 x16 bus bandwidth (16 GB/sec bidirectional)
specifications

RENDER BACK ENDS
MASSIVE PIXEL FILL RATE

- Upto16 Render Back Ends
  - 64 - 64b pixels per cycle
  - 256 Depth Test (Z) / stencil Ops

1 MB L2 R/W CACHE

- Upto16 64 KB L2 cache partition
  - Additional 33% last level cache capacity
  - Additional 33% internal fabric bandwidth
  - Up to 1TB/s L2/L1 bandwidth
MEMORY INTERFACE

- 512-bit GDDR5 at 5.0 Gbps
  - 320 GB/sec memory bandwidth

- High density interface design
  - 512-bit memory interface occupies ~20% less area than the 384-bit memory interface utilized on the AMD Radeon™ HD 7970 ASIC
  - Over 20% more total bandwidth
  - 50% increase in bandwidth per mm²
## PRODUCT SPECIFICATIONS

**AMD RADEON™ R9 290 SERIES**

<table>
<thead>
<tr>
<th>Feature</th>
<th>R9 290X</th>
<th>R9 290</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stream Processors</td>
<td>2,816</td>
<td>2,560</td>
</tr>
<tr>
<td>Engine Clock</td>
<td>Up to 1 GHz</td>
<td>Up to 947 MHz</td>
</tr>
<tr>
<td>Compute Performance</td>
<td>5.6 TFLOPS</td>
<td>4.9 TFLOPS</td>
</tr>
<tr>
<td>Memory Configuration</td>
<td>4GB GDDR5 / 512-bit</td>
<td>4GB GDDR5 / 512-bit</td>
</tr>
<tr>
<td>Memory Speed</td>
<td>5.0 Gbps</td>
<td>5.0 Gbps</td>
</tr>
<tr>
<td>Power Connectors</td>
<td>1 x 6-pin, 1 x 8-pin</td>
<td>1 x 6-pin, 1 x 8-pin</td>
</tr>
<tr>
<td>PCI-E Standard</td>
<td>PCI-E 3.0</td>
<td>PCI-E 3.0</td>
</tr>
<tr>
<td>AMD TrueAudio Technology</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>API Support</td>
<td>DirectX 11.2, OpenGL 4.3, Mantle</td>
<td>DirectX 11.2, OpenGL 4.3, Mantle</td>
</tr>
</tbody>
</table>
WHO HAS THE BEST OVERCLOCKING YIELDS?

POLARIS OR PASCAL
Costs !!!

Really?
Conclusion

We have several GPU architectures from different companies. Which is the best option? This depends on your PC and budget!!!

For the last 10 years GPUs graphically evaluated and we now see this magic.
Thank you