The Evolution of Computer Graphics

Tony Tamasi
SVP, Content & Technology, NVIDIA
Graphics

- Make great images
  - intricate shapes
  - complex optical effects
  - seamless motion

- Make them fast
  - invent clever techniques
  - use every trick imaginable
  - build monster hardware
What is a Graphics Processing Unit?

- Massively Parallel - 1000s of processors (today)
- Hardware managed parallelism - achievable performance
- Specialized processing - order of magnitude more efficient
- Latency Tolerant - throughput oriented
- Memory Bandwidth - saturate 100’s GB/sec
- Not cache dependent - ALU heavy architecture
Or, we could do it by hand...

Perspective study of a chalice
Paolo Uccello, circa 1450
What is reality?

- Reality is 80 million polygons.

Alvy Ray Smith
1999
Evolution of GPU Game Computing

Constrained
Fixed Function
Limited characters & animation
Simplistic AI
Indoors / basic environment
Multipass Rendering
Multi-Texture

Dynamic
Programmable Shading
Increased character count
IK skeletons / rigid bodies
Outdoors
Dynamic lighting
Programmable pixels

Programmable Graphics
Armies of characters
Complex physical simulations
Sophisticated AI
Procedural generation
Custom renderers / lighting
GPU Architecture Progression

1998
- 16-bit depth, Color, and textures

1999
- Multi-texture, 32b rendering

2000
- T&L, cube maps, Texture compression, Anisotropic filtering

2001
- Programmable vertex, 3D textures, shadow maps, multisampling

2002
- Early z-cull, Dual-monitor

2003
- Fragment programs, Color and depth compression

2004
- Flow control, FP textures, VTF

2005
- Transparency antialiasing

2006
- Unified shader, geometry shader, CUDA/C

2007
- Double Precision

Games:
- Far Cry
- Ballistics
- Test Drive 6
- Crysis
- Crysis
Graphics 1998
GPU 2003

• Dawn of programmable shading
• Vertex & Pixel Shaders
• More (8) pixels per clock
• Faster/Wider/Larger memory
• 32-bit and higher precision
GPU 2003

- **GeForce® FX 5800**
  - Shader Model 2
  - 4 ppc / 8tpc @ 500 Mhz
  - Single precision
  - AGP8x
  - 128b 16GB/sec memory system
  - 4 processors @ 500 MHz
  - Single threaded - massive SIMD width

- Programmed in OpenGL, DirectX or assembly
## GPU 2003

<table>
<thead>
<tr>
<th></th>
<th>Riva 128ZX 1998</th>
<th>GeForce FX5800 2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture Performance (tex/s)</td>
<td>0.1G</td>
<td>4 G</td>
</tr>
<tr>
<td>Antialiasing Performance (sam/s)</td>
<td>0.1G</td>
<td>2 G</td>
</tr>
<tr>
<td>Depth Performance (pix/s)</td>
<td>0.1G</td>
<td>4 G</td>
</tr>
<tr>
<td>Floating Point Performance (FLOPS)</td>
<td>n/a</td>
<td>12 G</td>
</tr>
<tr>
<td>Memory Bandwidth (B/s)</td>
<td>1.6G</td>
<td>16 G</td>
</tr>
</tbody>
</table>
The Graphics Pipeline

- Key abstraction of real-time graphics
- Hardware used to look like this
- Distinct chips/boards/units per stage
- Fixed data flow through pipeline

- Vertex Transform & Lighting
- Triangle Setup & Rasterization
- Texturing & Pixel Shading
- Depth Test & Blending
- Framebuffer
GPU Architecture 2003

Diagram showing the architecture with stages for Vertex Shader, Pixel Shader, Blend / Depth, and Memory, with Triangle Setup and Depth Cull at the top.
Graphics 2003
Performance Trends

- Texture
- Antialiasing
- Depth
- Floating Point
- Bandwidth

Graph showing performance trends from 1999 to 2007 for various NVIDIA GPUs:
- Riva TNT2
- GeForce GTS
- GeForce3
- GeForce4 Ti4600
- GeForce FX (5800)
- GeForce 6800 Ultra
- GeForce 7800 GTX
- GeForce 7900 GTX
- GeForce 8800 GTX
### GPU Processing CAGR 1998 - 2007

<table>
<thead>
<tr>
<th>Performance Category</th>
<th>Compound Annual Growth Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture Performance</td>
<td>1.9</td>
</tr>
<tr>
<td>Antialiasing Performance</td>
<td>2.2</td>
</tr>
<tr>
<td>Depth Performance</td>
<td>1.9</td>
</tr>
<tr>
<td>Floating Point Performance</td>
<td>2.3</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>1.6</td>
</tr>
</tbody>
</table>
# GPU 2008 Projection

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture Performance (tex/s)</td>
<td>1.9</td>
<td>4 G</td>
</tr>
<tr>
<td>Antialiasing Performance (sam/s)</td>
<td>2.2</td>
<td>2 G</td>
</tr>
<tr>
<td>Depth Performance (pix/s)</td>
<td>1.9</td>
<td>4 G</td>
</tr>
<tr>
<td>Floating Point Performance (FLOPS)</td>
<td>2.3</td>
<td>12 G</td>
</tr>
<tr>
<td>Memory Bandwidth (B/s)</td>
<td>1.6</td>
<td>16 G</td>
</tr>
</tbody>
</table>

Within 10%
Era of Visual Computing

Fixed-Function Pipelines
“3D Accelerators”

Programmable Shaders
DX8 – DX9 – DX10

Programmable Graphics
CUDA – DX11 – Future

RIVA 128
GeForce 3
GeForce 6
G80 CUDA
GTX200
Next Gen

1997 August
2001
2005
2007
2008
2009
GPU 2008

• Dawn of fully programmable generation
• Unified architectures
• New graphics functionality – geometry shading
• Programmable in C
The Graphics Pipeline

- Remains a useful abstraction
- Hardware used to look like this
Modern GPU’s: Unified Architecture

Vertex shaders, pixel shaders, etc. become *threads* running different programs on flexible cores
Why unify?

Heavy Geometry
Workload Perf = 4

Heavy Pixel
Workload Perf = 8
Why unify?

Unified Shader

- **Vertex Workload**
- **Pixel Workload**

Heavy Geometry
Workload Perf = 11

Heavy Pixel
Workload Perf = 11

© 2008 NVIDIA Corporation.
GPU 2008

• GeForce® GTX 200
  – Shader Model 4
  – Double precision
  – PCI-Express Gen2
  – 512b 140GB/sec memory system
  – 32ppc / 80 tpc @ 600 Mhz
  – 240 processors @ 1.5 GHz
  – Many threaded - scalar

  – Programmed in OpenGL, DirectX, CUDA™ C
GPU Processing

<table>
<thead>
<tr>
<th>Year</th>
<th>CPU GFLOPS</th>
<th>GPU GFLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2003</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2004</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2005</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2006</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2008</td>
<td>0</td>
<td>1200</td>
</tr>
</tbody>
</table>

© 2008 NVIDIA Corporation.
State of the art 2008

Grid - Codemasters

Far Cry 2 - Ubisoft

Call of Duty 5 - Activision

Bionic Commando - Capcom
GPU 2013

2003
Fragment programs,
Color and depth compression

2004
Flow control,
FP textures, VTF

2005
Transparency, antialiasing

2006
Unified shader,
geometry shader, CUDA/C

2008
Double precision
GPU 2013

- Arbitrary dataflow
- General purpose programming model
- Special purpose hardware
- Hardware managed threading and pipelining
- Freely intermingle “graphics” & computation
Performance Trends

- Texture
- Antialiasing
- Depth
- Floating Point
- Bandwidth

Graph showing performance trends for various GPU models from 1999 to 2007.
# GPU 2013

<table>
<thead>
<tr>
<th></th>
<th>CAGR 1998 - 2007</th>
<th>GeForce GTX 200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture Performance (tex/s)</td>
<td>1.9</td>
<td>48 G</td>
</tr>
<tr>
<td>Antialiasing Performance (sam/s)</td>
<td>2.2</td>
<td>153 G</td>
</tr>
<tr>
<td>Depth Performance (pix/s)</td>
<td>1.9</td>
<td>38 G</td>
</tr>
<tr>
<td>Floating Point Performance (FLOPS)</td>
<td>2.3</td>
<td>1080 G</td>
</tr>
<tr>
<td>Memory Bandwidth (B/s)</td>
<td>1.6</td>
<td>141 G</td>
</tr>
</tbody>
</table>
Indirect Lighting  
Soft Shadows  
Fracture  
Subsurface Scatter  
Turbulence  
Participating Media  
Detailed Characters  
Ambient Occlusion  
Simulations  
Rich Environments  
Fluids
The Future of Graphics Processing

- Programmable and Specialized Processing
- Monster performance and power efficiency
- Graphics and Arbitrary C/C++ Programming
- Global Illumination and Rasterization
- Rendering and Simulation

...Evolution
The Evolution of Computer Graphics