AMD’s Graphics Core Next (GCN) Architecture

Tony Chen
Jason Greaves
Overview

- Terascale
  - Limitation
- GCN architecture
  - Revisions
    - Features/Updates
- Mantle
Terascale

- ~2006 - 2011
- VLIW SIMD (Very Long Instruction Word Single Instruction Multiple Data)
- Meant to compete with Nvidia’s Tesla line
- Ranged from the Radeon HD 2000 - 6000 series graphics cards
- Found in some of the early APUs
- VLIW4
- One SIMD engine
  - 16 shader processor
    - 4 ALUs each
  - 64 ALUs per SIMD engine
The problem

- Terascale’s VLIW architecture
  - Very good for graphics instructions
  - Compiler optimized for dot product math
    - Cannot change the wavefront (instruction set) queue once it has been scheduled
  - High instruction level parallelism
  - Specialized hardware design
    - Double edge sword
    - Not good for general computing
    - Cannot continue until a dependency has been resolved
The compiler schedules the following wavefront queue:

Wavefront 'A'
Wavefront 'B'
Wavefront 'C'
Wavefront 'D'
Wavefront 'E'
Wavefront 'F'
Wavefront 'G'
Wavefront 'H'
Wavefront 'I'
Wavefront 'J'
Wavefront 'K'
Wavefront 'L'
Wavefront 'M'
Wavefront 'N'
Wavefront 'O'

It takes six clock cycles to handle all of the wavefronts in the schedule because some ALUs lie dormant while they wait for dependencies to be resolved.
GCN architecture

- Released early 2012 (announced late 2011)
- Select Radeon HD 7000s, 8000s, and Rx 200 series
- 28nm technology TSMC
- RISC SIMD (Reduced instruction set computing Single Instruction Multiple Data)
- Designed for general computing
GCN Compute Unit

- **Compute Unit (CU)**
  - 4 vector units
    - 16 ALUs each
  - Scheduler
    - Independent
    - Can avoid dependencies
  - 64 ALUs per CU
COMPUTE UNIT
Demonstration of efficient dependency handling

First Clock Cycle:
- **COMPUTE UNIT**
  - AAAAA
  - BBBB
  - DDDD
  - EE
  - Vector Unit
- **Hardware Scheduler**
  - XXXX
  - Vector Unit

Second Clock Cycle:
- **COMPUTE UNIT**
  - CCCC
  - EEEE
  - FFFF
  - GGGG
  - Vector Unit
- **Hardware Scheduler**
  - HHHH
  - Vector Unit

Third Clock Cycle:
- **COMPUTE UNIT**
  - JJJJ
  - KKKK
  - LLL
  - MMMM
  - Vector Unit
- **Hardware Scheduler**
  - NNNN
  - Vector Unit

Fourth Clock Cycle:
- **COMPUTE UNIT**
  - NNNN
  - OOOO
  - Vector Unit
- **Hardware Scheduler**
  - Vector Unit

Wavefront C is dependent on B, so the scheduler chooses wavefronts D and E for processing.

Wavefront E is complete, C can be processed. G is dependent on F, so H and I are processed instead.

With wavefront F completed, G can be processed. L must wait for K, so M takes its place.

Wavefronts L, N, and O are processed.

It takes four clock cycles to handle all of the wavefronts in the schedule, in 1/3 less time than VLIW4, even though a Compute Unit and SIMD Engine possess the same number of ALUs. In this example, all ALUs are able to work every clock cycle.
GCN Architecture

- HyperZ
- GCN 1.0
  - PowerTune
  - ZeroCore power
- GCN 1.1
- GCN 1.2
HyperZ

- November 2000
- developed by ATI (Array Technologies Inc)
- Processing technique for Z-buffering
  - Algorithm to determine which depth to use for a single pixel space
  - About ~50% of bandwidth used for Z-Buffer read/write
- Boost memory bandwidth / Improves efficiency
  - Compresses data (lossless format) to minimize bandwidth for reads and writes
  - Tags blocks of data instead of writing individual bits
  - Compares pixels before rendering to save bandwidth
- Updates over time (ex: HyperZ II, HyperZ HD)
GCN 1.0

- Supports 64 bit addressing
- Backwards compatible with 16 and 32 bit
- Unified virtual memory between CPU and GPU
  - Zero copy
    - Pointers are passed
AMD PowerTune

- Dec 15 2010
- Technology to dynamically scale the clock speed by software
  - Better power utilization
  - Runs cooler
  - Less noise from fan/cooling
- 3 states (High, Intermediate, Low)
ZeroCore Power

- 15W on idle
- Ability to shut off GPU functional units during “long idle”
  - 3W
  - Fan turns off
- Borrowed from mobile graphics technology
- Crossfire support (Multi-card)
GCN 1.1

- March 2013
- **AMD TrueAudio**
  - AMD claims audio effects can use up 14% of CPU
  - Dedicated DSP built into the GPU core
    - Better performance and power usage
  - Support more audio channels and objects
- **Revised PowerTune**
  - 8 states
  - Formula for state changing now considers the temperature
GCN 1.2

- September 2014 with the release of the Radeon R9 285
- Improved Tessellation Performance
- Lossless delta color compression
- Updated Instruction Set
  - 16 bit floating point and integer instructions
    - low power GPU compute
- New version of the Unified Video Decoder
  - 4K H.264 support (up to level 5.2 (4kp60))
  - better performance
- New version of the Video Codec Engine
  - ~40% faster than previous VCE
Video Encode Performance (1080p)
Time In Seconds - Lower Is Better

<table>
<thead>
<tr>
<th>Card</th>
<th>Time (Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R9 285</td>
<td>43</td>
</tr>
<tr>
<td>R9 280</td>
<td>73</td>
</tr>
<tr>
<td>R9 290</td>
<td>76</td>
</tr>
</tbody>
</table>

Video Decode Performance
Frames Per Second - Higher Is Better

<table>
<thead>
<tr>
<th>Card</th>
<th>1080p</th>
<th>4K</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTX 750 Ti</td>
<td>316</td>
<td>135</td>
</tr>
<tr>
<td>GTX 770</td>
<td>183</td>
<td>50</td>
</tr>
<tr>
<td>R9 280</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>R9 290</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>R9 285</td>
<td>331</td>
<td>152</td>
</tr>
</tbody>
</table>

http://www.anandtech.com/show/8460/ amd-radeon-r9-285-review/4
Partially Resident Textures

- Increasing physical size of textures in a game can have a negative impact
- Utilize Large texture files
  - Up to 32 TB large
- Small bits of textures streamed in as needed
Improved Anisotropic Filtering

- Algorithm improved
  - New filter kernel weights
  - Reduces shimmering artifacts
  - No performance cost
  - Fully angle invariant
Improved DirectX 11 Tessellation
TressFX

- Real time Hair physics
- Realistic Detail
- Graphics cards featuring GCN are well equipped to handle TressFX
Mantle

- Developed in 2013
- Low Level rendering API
- Developed in cooperation with DICE
- Alternative to Direct3D and OpenGL
- Works on GCN but not anything before it
Three Ingredients of Mantle

- Driver within AMD Catalyst software suite that lets applications speak directly to GCN architecture
- A GPU or APU enabled with the GCN architecture
- An application or game written to take advantage of Mantle
What’s Using Mantle?

- HD 7000, HD 8000, Rx 200
- Steamroller based “Kaveri” AMD Accelerated Processing Units
- All Jaguar-based APUs
- All Puma based APUs
In Conclusion

1. Better performance
2. Lower power consumption/Energy Efficiency
3. Better support for developers
4. Up to date
5. General computing
Questions?