High Bandwidth Memory for Graphics Applications
Contents

• Differences in Requirements: System Memory vs. Graphics Memory
• Timeline of Graphics Memory Standards
  • GDDR2
  • GDDR3
  • GDDR4
  • GDDR5 SGRAM
• Problems with GDDR
• Solution - Introduction to HBM
• Performance comparisons with GDDR5
• Benchmarks
• Hybrid Memory Cube
Differences in Requirements

System Memory
- Optimized for low latency
- Short burst vector loads
- Equal read/write latency ratio
- Very general solutions and designs

Graphics Memory
- Optimized for high bandwidth
- Long burst vector loads
- Low read/write latency ratio
- Designs can be very haphazard
Brief History of Graphics Memory Types

- Ancient History: VRAM, WRAM, MDRAM, SGRAM
- Bridge to modern times: GDDR2
- The first modern standard: GDDR4
- Rapidly outclassed: GDDR4
- Current state: GDDR5
GDDR2

- First implemented with Nvidia GeForce FX 5800 (2003)
- Midway point between DDR and ‘true’ DDR2
- Stepping stone towards DDR-based graphics memory
- Second-generation GDDR2 based on DDR2
GDDR3

- Designed by ATI Technologies, first used by Nvidia GeForce FX 5700 (2004)
- Based off of the same technological base as DDR2
- Lower heat and power consumption
- Uses internal terminators and a 32-bit bus
GDDR4

- Based on DDR3, designed by Samsung from 2005-2007
- Introduced Data Bus Inversion (DBI)
- Doubled prefetch size to 8n
- Used on ATI Radeon 2xxx and 3xxx, never became commercially viable
GDDR5 SGRAM

- Based on DDR3 SDRAM memory
- Inherits benefits of GDDR4
- First used in AMD Radeon HD 4870 video cards (2008)
- Current generation of graphics memory
- Uses three clock lines, 128/256 bit data bus
- Uses 20nm die
Problems

• Memory bandwidth has become a bottleneck
• Increasing bandwidth with GDDR5 would be too power-inefficient
• Edge contacts limit placement of memory
PCB area occupied by ASIC + Memory (R9 290X)
The Solution: High Bandwidth Memory

- Designed by AMD from 2008-2013
- Stack memory chips on top of each other
- Increases bus size, memory-area density
- Route buses through interposer rather than substrate
Comparisons and Dividends

<table>
<thead>
<tr>
<th></th>
<th>GDDR5</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 Bits Per Chip</td>
<td>Bus Width</td>
<td>1024+ Bits Per Chip</td>
</tr>
<tr>
<td>1750MHz</td>
<td>Clock Speed</td>
<td>500MHz</td>
</tr>
<tr>
<td>28GB/s</td>
<td>Bandwidth</td>
<td>&gt;100GB/s</td>
</tr>
<tr>
<td>1.5V</td>
<td>Voltage</td>
<td>1.3V</td>
</tr>
</tbody>
</table>
Benchmarks

**HBM**
- R9 Fury, Fury X (4GB HBM, 4096-bit bus)

**GDDR5**
- Nvidia GTX 980 (4GB GDDR5, 256-bit bus)
- Nvidia GTX 980TI (6GB GDDR5, 384-bit bus)
- AMD R9 390x (8GB GDDR5, 512-bit bus)
Hybrid Memory Cube?

- In development by Intel, Micron, Samsung, Xilinx, Altera, Microsoft...
- First specification defined in 2013
- Stacks RAM directly on top of processors
- Uses significantly less energy and space for significantly higher performance
- Does not compete with HBM
Citations

• http://www.hardwaresecrets.com/ddr-vs-gddr-memories/
• https://en.wikipedia.org/wiki/DDR2_SDRAM#Relation_to_GDDR_memory
• https://en.wikipedia.org/wiki/GDDR
• http://www.hwstation.net/img/news/allegati/Qimonda_GDDR5_whitepaper.pdf
• http://www.pcworld.com/article/2922599/amd-talks-up-high-bandwidth-memory-that-will-power-its-next-gpus-pokes-nvidia-too.html
• http://www.hybridmemorycube.org/technology.html