Contents

● Background on GDDR
● The Memory Bottleneck
● HBM Overview
● HBM Schematic Improvements
● Benchmarks
● HMC / 3DXpoint Comparison
● HBM2 Standard
● GPU Comparisons
Background Info - GPU and GDDR

- Specialized hardware for rendering/encoding/decoding images and video
- Designed for highly parallel and computationally intensive operations
- Typically produced as standalone cards
- Does not use normal DDRx RAM
- GDDR has a higher bandwidth, wider bus, can request and receive data in the same cycle
The Problem
The Memory Problem

- The Von Neumann Memory Bottleneck
- Processor speeds have overtaken memory access speeds
- GDDR5 is rising in power consumption.
- Large footprint of GDDR5 chips expands form factor.
The Solution
HBM Overview

- Developed by AMD and SK Hynix
- JEDEC industry standard in October 2013
- Multiple DRAM dies stacked in a single package connected by TSV and Microbumps
- Connected to the processor unit directly via Interposer Layer
- Two fully independent channels between each stack and chip
HBM Overview Continued

Improvements over Standard GDDR RAM

- Very High Bandwidth
- Lower Effective Clock Speed
- Smaller Package
- Lower Power Consumption
- Shorter Interconnect Wires
- Individual Banks Can Be Refreshed
HBM vs. GDDR5 Form Factor
Space Savings of Cache Memory

1GB GDDR5

1GB HBM
94% less surface area

Areal, to scale
GDDR5 vs HBM1

- Bus Width: 32 bit
- Clock Speed: 1750MHz
- Transfer Rate per pin: 7GB/s
- Bandwidth: 28GB/s per chip
- Bandwidth per Watt: 10.5GB/W
- Operating Voltage: 1.5V
- Area: 24mm x 28mm

- Bus Width: 1024 bit
- Clock Speed: 500MHz
- Transfer Rate per pin: 1GB/s
- Bandwidth: 128GB/s per chip
- Bandwidth per Watt: 35GB/W
- Operating Voltage: 1.3V
- Area: 5mm x 7mm
Benchmark of GDDR5 vs HBM (GTX 980ti vs R9 Fury X)

Stock 980ti and Fury X perform the same at 1080p/1440p.

Stock Fury X outperforms the stock 980ti at 4k.
### Other 3D RAM Solutions: HBM vs HMC vs 3D XPoint

<table>
<thead>
<tr>
<th>Type</th>
<th>HBM</th>
<th>HMC</th>
<th>3D XPoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developer</td>
<td>AMD, SK Hynix, Samsung</td>
<td>Arm, Micron, IBM, Samsung</td>
<td>Micron and Intel</td>
</tr>
<tr>
<td>Applications</td>
<td>VRAM</td>
<td>Multi-Core Servers, DRAM</td>
<td>Mass Storage, DRAM, Hybrid</td>
</tr>
<tr>
<td>Max Bandwidth</td>
<td>256 GB/s</td>
<td>480 GB/s</td>
<td>N/A (Application Based)</td>
</tr>
<tr>
<td>Other</td>
<td>JEDEC Standard, Per Bank Refresh</td>
<td>Also Uses TSV and Microbumps Stack 4-8 Memory Cells Not JEDEC Standard</td>
<td>Non-Volatile, High Read/Write Endurance</td>
</tr>
</tbody>
</table>
The Future
HBM Gen 2

- Finalized by JEDEC in January 2016
- Improvements over Gen 1
  - 8 Dies per stack
  - 2Gb/s per pin
  - 256 GB/s bandwidth
  - 8GB per package
- Already on market
  - NVIDIA Tesla P100
- Very important for high bandwidth applications such as VR and networking
## GPU Memory Math

<table>
<thead>
<tr>
<th></th>
<th>AMD Radeon R9 290X</th>
<th>NVIDIA GeForce GTX 980 Ti</th>
<th>AMD Radeon R9 Fury X</th>
<th>Samsung’s 4-Stack HBM2 based on 8 Gb DRAMs</th>
<th>Theoretical GDDR5X 256-bit sub-system</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Capacity</strong></td>
<td>4 GB</td>
<td>6 GB</td>
<td>4 GB</td>
<td>16 GB</td>
<td>8 GB</td>
</tr>
<tr>
<td><strong>Bandwidth Per Pin</strong></td>
<td>5 Gb/s</td>
<td>7 Gb/s</td>
<td>1 Gb/s</td>
<td>2 Gb/s</td>
<td>10 Gb/s</td>
</tr>
<tr>
<td><strong>Number of Chips/Stacks</strong></td>
<td>16</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td><strong>Bandwidth Per Chip/Stack</strong></td>
<td>20 GB/s</td>
<td>28 GB/s</td>
<td>128 GB/s</td>
<td>256 GB/s</td>
<td>40 GB/s</td>
</tr>
<tr>
<td><strong>Effective Bus Width</strong></td>
<td>512-bit</td>
<td>384-bit</td>
<td>4096-bit</td>
<td>4096-bit</td>
<td>256-bit</td>
</tr>
<tr>
<td><strong>Total Bandwidth</strong></td>
<td>320 GB/s</td>
<td>336 GB/s</td>
<td>512 GB/s</td>
<td>1 TB/s</td>
<td>320 GB/s</td>
</tr>
<tr>
<td><strong>Estimated DRAM Power Consumption</strong></td>
<td>30W</td>
<td>31.5W</td>
<td>14.6W</td>
<td>n/a</td>
<td>20W</td>
</tr>
</tbody>
</table>
Questions?
Sources

- Image Source: http://images.anandtech.com/doci/9390/HBM_7_Interposer.png