GDDR Memory Enabling AI and High Performance Compute

Wolfgang Spirkl, Fellow at Micron Technology

GTC, S9968, 20-March-2019
Agenda

- The Demand for faster Memory and storage
- Competing Compute/Memory Solutions
- GDDR6 for AI applications and more
- Micron GDDR6 AI demonstration
Accelerated Data Cycle

Driven by Increasing Data Value

- Creates continuous need to capture, process, move & store data
- Generates ever-increasing demand for memory & fast storage
- AI is amplifying the Accelerated Cycle
AI Landscape for Memory & Storage

Data Center
System/Infrastructure:
- Cloud
- On Prem.

Workloads:
- Model Development
- Batch Training
- Batch Inference

Memory/Storage:
- HBM
- GDDR6
- DDR4
- P-Mem
- 3D TLC
- 3D QLC

Smart Edge/Intelligent Endpoint
System/Infrastructure:
- Edge Compute
- Smart Access Point
- Autonomous Vehicle/Robot
- Mobile Device
- Smart IoT/Sensor

Workloads:
- Online Training
- On-demand Inference
- Local Inference
- Data Collection

Memory/Storage:
- HBM
- GDDR6
- DDR4
- LP4X
- P-Mem
- 3D TLC
- 3D QLC

GTC 2019, Micron GDDR6
AI Workloads Unleash the Need For More Memory & Storage

Significant Growth Across Private, Public & Hybrid Cloud

Source: Micron

GTC 2019, Micron GDDR6
- The Demand for faster Memory and storage
- Competing Compute/Memory Solutions
- GDDR6 for AI applications and more
- Micron GDDR6 AI demonstration
AI Acceleration is Driving Demand for Memory Bandwidth

- AI accelerators increase compute performance
  - GPU, TPU, etc..

- Accelerated applications are more likely to be memory bound [3]

- Micron supports next gen technologies
  - GDDR6, HBM2E

---

[3] Forrester report on memory and storage impact on AI
Memory Options

**DDR LPDDR**
- Low power
- Mainstream
- Cost efficient

**GDDR**
- High performance
- High bandwidth
- Cost efficient

**HBM**
- Maximum bandwidth
- Complex
- Costly

---

Data Center

Smart Edge

Intelligent Endpoint
The Demand for faster Memory and storage
Competing Compute/Memory Solutions
GDDR6 for AI applications and more
Micron GDDR6 AI demonstration
## GDDR5/GDDR6 Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>GDDR5</th>
<th>GDDR6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>512Mb – 8Gb</td>
<td>8Gb – 32Gb</td>
</tr>
<tr>
<td>VDD, VDDQ</td>
<td>1.5V + 1.35V</td>
<td>1.35V</td>
</tr>
<tr>
<td>VPP</td>
<td>N/A</td>
<td>1.8V</td>
</tr>
<tr>
<td>Package</td>
<td>BGA-170 14mm x 12mm 0.8mm ball pitch</td>
<td>BGA-180 14mm x 12mm 0.75mm ball pitch</td>
</tr>
<tr>
<td>Signaling</td>
<td>POD15 / POD135</td>
<td>POD135</td>
</tr>
<tr>
<td>Data rate</td>
<td>≤8 Gbps</td>
<td>≤16 Gbps</td>
</tr>
<tr>
<td>I/O Width</td>
<td>x32/x16</td>
<td>2-ch x16/x8</td>
</tr>
<tr>
<td>Access Granularity</td>
<td>32B</td>
<td>2-ch 32B each or 1-ch 64B w/ PC mode</td>
</tr>
<tr>
<td>I/O Count</td>
<td>61</td>
<td>62 / 74</td>
</tr>
<tr>
<td>ABI, DBI</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CRC</td>
<td>CRC-8 (BL8)</td>
<td>2x CRC-8 (BL16); compressed 2x CRC-8 (BL8)</td>
</tr>
<tr>
<td>RDQS Mode</td>
<td>✓ (BL8)</td>
<td>✓ (BL16)</td>
</tr>
<tr>
<td>ODT</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$V_{REFC}$</td>
<td>external</td>
<td>external / internal</td>
</tr>
<tr>
<td>$V_{REFD}$</td>
<td>ext. / int.</td>
<td>internal</td>
</tr>
<tr>
<td>Temp Sensor</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

### Package Configuration

- **GDDR5**:
  - 1 channel x 32 bits
- **GDDR6**:
  - 2 channels x 16 bits

GDDR6: x16/x8 on dual channel equals GDDR5 x32/x16

### Dual channel organization
- Maintains fine granularity (32 bytes per column access)
- In spite of doubled prefetch size

### New features for high data rates:
- Optimized signal ball-out for low-effort PCB design
- Per lane DFE and $V_{REFD}$
- Transmitter equalization

---

GTC 2019, Micron GDDR6
Calculating GDDR Bandwidth

- Bandwidth = number of bits/s between GPU and memory
- Memory bus is like traffic lanes
  - More lanes, the greater the flow
  - Higher lane speed, the greater the flow

<table>
<thead>
<tr>
<th>Memory Bandwidth is</th>
<th>GDDR6 Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of memory components ( \times ) number of lanes per component ( \times ) Data rate per lane (Gbps)</td>
<td>8 ( \times ) 32 ( \times ) 16</td>
</tr>
<tr>
<td>Memory Bandwidth (GB/s)</td>
<td>512</td>
</tr>
</tbody>
</table>
# GDDR Bandwidth / Memory Bus

## Technology Specifications

<table>
<thead>
<tr>
<th>Technology</th>
<th>Speed (Gbps)</th>
<th># of comp.</th>
<th># of lanes</th>
<th>Memory bus (bit)</th>
<th>Bandwidth (GB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBM2</td>
<td>2</td>
<td>4</td>
<td>1024</td>
<td>4096</td>
<td>1024</td>
</tr>
<tr>
<td>GDDR6</td>
<td>16</td>
<td>12</td>
<td>32</td>
<td>384</td>
<td>768</td>
</tr>
<tr>
<td>GDDR6</td>
<td>14</td>
<td>12</td>
<td>32</td>
<td>384</td>
<td>672</td>
</tr>
<tr>
<td>GDDR5X</td>
<td>11</td>
<td>12</td>
<td>32</td>
<td>384</td>
<td>528</td>
</tr>
<tr>
<td>GDDR5</td>
<td>7</td>
<td>12</td>
<td>32</td>
<td>384</td>
<td>336</td>
</tr>
<tr>
<td>GDDR6</td>
<td>14</td>
<td>8</td>
<td>32</td>
<td>256</td>
<td>448</td>
</tr>
<tr>
<td>GDDR5X</td>
<td>11</td>
<td>8</td>
<td>32</td>
<td>256</td>
<td>352</td>
</tr>
<tr>
<td>GDDR5</td>
<td>7</td>
<td>8</td>
<td>32</td>
<td>256</td>
<td>224</td>
</tr>
</tbody>
</table>

---

GTC 2019, Micron GDDR6
GDDR6 20 Gbps Data Eye
Measured performance beyond the specification

Figure 15: Measured 20Gb/s data eye based on a PRBS6 pattern
Speech Recognition Craves Memory Bandwidth

- “Deep Speech” recognition application
  - Mozilla’s tensorflow implementation

- Hardware
  - NVIDIA RTX 2080 Ti
  - 11GB GDDR6
    - 384 bit bus @14Gb/s/pin, 672GB/s

- Experiment setup
  - Adjust GDDR6 clock rate
  - Measure speech recognition inference rate:
    - $\text{Inference rate} = \frac{\text{Audio file duration}}{\text{Inference time}}$

Sensitivity = $\frac{\Delta IR/IR}{\Delta mBW/mBW} \approx 33\%$
AI Demonstrates the Need for Memory Speed
SPEECH RECOGNITION DEMO – Micron Booth #1713

Memory Sensitivity

14.74x
Inference Rate

27.70%
Memory Sensitivity

Memory Data Rate: 14 Gbps/pin

Hardware Activity

System Monitor

8 Core 2.9GHz
CPU 29.5%

Micron 32GB DDR4
Memory 11%

Micron 500GB MX500
Disk 0%

NVIDIA 2080Ti
GPU Core 28%

Micron 11GB GDDR6
GPU Memory 29%
Conclusions

- AI Landscape demands higher performance memory to feed the compute needs
- Micron delivers a broad range of memory solutions for AI applications from data center to cloud to edge to endpoint devices
- GDDR6 high performance memory optimized for applications beyond graphics

Experience Micron speech recognition AI with GDDR6 in our booth 1713!