Expanding the Boundaries of the AI Revolution:

An In-depth Study of High Bandwidth Memory

Changyong Ahn & Nayoung Lee | March 2019
Outline

- SK hynix Overview
- Memory challenges of Deep Learning
- HBM Overview
- HBM Deep Dive
- Future HBM solution
Memory challenges of Deep Learning
Machine Learning/Deep Learning Use Cases

- Security
- Data analysis
- Financial Service
- Law

Machine Learning

- Autonomous Driving
- Gaming
- Medical Diagnostics
- Smart home
Memory Challenges of Deep Learning

Deep Neural Network Fundamental Concepts

Deep Neural Network

\[ Y_j = \text{activation} \left( \sum_{i=1}^{3} W_{ij} x_i \right) \]

Source: Standford

<table>
<thead>
<tr>
<th>Year</th>
<th>CNN</th>
<th># of layers</th>
<th># of Parameters</th>
<th>Memory size (MB)</th>
<th>Top5 Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>LeNet</td>
<td>8</td>
<td>60K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>AlexNet</td>
<td>7</td>
<td>60 million</td>
<td>240</td>
<td>15.3%</td>
</tr>
<tr>
<td>2014</td>
<td>GoogleNet</td>
<td>19</td>
<td>4 million</td>
<td></td>
<td>6.67%</td>
</tr>
<tr>
<td>2014</td>
<td>VGG Net</td>
<td>16</td>
<td>138 million</td>
<td>574</td>
<td>7.3%</td>
</tr>
<tr>
<td>2015</td>
<td>ResNet</td>
<td>50/152</td>
<td></td>
<td>519</td>
<td>3.6%</td>
</tr>
</tbody>
</table>
Memory Solution for ML/DL Systems

Memory Sub system hierarchy change

- **<10ns**
  - LLC
- **50ns**
  - IPM
- **100ns – 1us**
  - DRAM
  - Storage Class Memory
- **50-100us**
  - Fast Storage (SSD)
- **~10ms**
  - HDD

### “In Package Memory”

<table>
<thead>
<tr>
<th>TSV Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interposer</td>
</tr>
<tr>
<td>Substrate(&quot;PCB&quot;)</td>
</tr>
</tbody>
</table>

### Target Market/Price

<table>
<thead>
<tr>
<th>Conventional DRAM</th>
<th>IPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad &amp; Cheap</td>
<td>Specific &amp; high Premium</td>
</tr>
</tbody>
</table>

### Standardization

<table>
<thead>
<tr>
<th>Conventional DRAM</th>
<th>IPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>JEDEC</td>
<td>Semi Custom</td>
</tr>
</tbody>
</table>

### Qualification Period

<table>
<thead>
<tr>
<th>Conventional DRAM</th>
<th>IPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively short</td>
<td>Relatively long</td>
</tr>
</tbody>
</table>

### Key factors

<table>
<thead>
<tr>
<th>Conventional DRAM</th>
<th>IPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price Competitiveness</td>
<td>Reliability / Performance</td>
</tr>
</tbody>
</table>

* Source: SK hynix

1) In-Package Memory 2) SCM("Storage Class Memory") : 3DXP, PCRAM
HBM Overview
HBM, What’s the difference?

GDDR/DDR/LPDDR

- FBGA

HBM

- KGSD

HBM in 2.5D SiP

- Directly soldered on PCB or used as a DI MM
HBM Advantages

**More Bandwidth**

**High Power Efficiency**

**Small Form Factor**

---

<table>
<thead>
<tr>
<th></th>
<th>DDR4</th>
<th>LPDDR4(X)</th>
<th>GDDR6</th>
<th>HBM2 (JEDEC)</th>
<th>HBM2E (TBD)</th>
<th>HBM3 (TBD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data rate</td>
<td>3200Mbps (up to 4266 Mbps)</td>
<td>14Gbps (up to 16Gbps)</td>
<td>2.4Gbps</td>
<td>2.8Gbps</td>
<td>&gt;3.2Gbps (TBD)</td>
<td></td>
</tr>
<tr>
<td>Pin count</td>
<td>x4/x8/x16</td>
<td>x16/ch (2ch per die)</td>
<td>x16/x32</td>
<td>x1024</td>
<td>x1024</td>
<td>x1024</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>5.4GB/s</td>
<td>12.8(17)GB/s</td>
<td>56GB/s</td>
<td>307GB/s</td>
<td>358GB/s</td>
<td>&gt;500GB/s</td>
</tr>
<tr>
<td>Density (per package)</td>
<td>4Gb/8Gb</td>
<td>8Gb/16Gb/24Gb/32Gb</td>
<td>8Gb/16Gb</td>
<td>4GB/8GB</td>
<td>8GB/16GB</td>
<td>8GB/16GB/24GB (TBD)</td>
</tr>
</tbody>
</table>

**To Achieve 1TB Bandwidth …..**

40ea of DDR4-3200 Module

160ea of DDR4-3200

4ea HBM2 in a single 50mm x 50mm Sip

Note: Advil is a registered trademark
HBM Deep Dive
HBM Architecture

HBM2 core die supports 4 pseudo channels or 2 channels
Each channel consists of 2 Pseudo Channels. Only BL4 is supported

<table>
<thead>
<tr>
<th>Items</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Stack</td>
<td>4/8(Core) + 1(Base)</td>
</tr>
<tr>
<td>Ch./Slice</td>
<td>2</td>
</tr>
<tr>
<td>Total Ch. for KGSD</td>
<td>8/16 (8ch based operation)</td>
</tr>
<tr>
<td>IO/Ch.</td>
<td>128</td>
</tr>
<tr>
<td>Total IO/KGSD</td>
<td>1024 (=128 x 8)</td>
</tr>
<tr>
<td>Address/CMD</td>
<td>Dual CMD</td>
</tr>
<tr>
<td>Data Rate</td>
<td>DDR</td>
</tr>
</tbody>
</table>
Next-Gen. System Architecture Leveraging HBM

HBM and 2.5D SiP integration unlock new system architecture

HPC & Server
(B/W & Capacity)

Network & Graphics
(B/W)

Client-DT & NB
(B/W & Cost)
HBM Test Flow

General DRAM Test Flow

Wafer
- DRAM Die
  - WFBI
  - Hot/Cold Test
  - Repair
- Package Process

PKG
- TDBI
- Hot/Cold Test
- Speed Test

HBM Test Flow

Wafer
- DRAM Die
  - WFBI
  - Hot/Cold Test
  - Repair
- Stack Process (KGSD)

KGSD
- B/I Stress
  - Dynamic Stress (BISS)
- Hot/Cold Test
- Speed Test
- KGSD Speed Test

Base Die
- Logic Test
  - Logic Test (IEEE1500)
- KGSD (DRAMs + 1 Logic)
- Bump & Stack Process
Quality and Reliability Features

HBM Features enable high quality and reliability at post 2.5D assembly

1. *Cell Repair*
2. *Error Correcting Code Storage*
3. *PMBIST*
4. *BISS (Built In Self Stress)*
5. *Microbump Repair*
6. *Proxy Package*
## Collaterals Available from HBM vendors

<table>
<thead>
<tr>
<th>Item</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functionality</strong></td>
<td>Datasheet (Jedec/Vendor)</td>
</tr>
<tr>
<td></td>
<td>Verilog (mission mode and DFT)</td>
</tr>
<tr>
<td></td>
<td>IBIS</td>
</tr>
<tr>
<td></td>
<td>Hspice</td>
</tr>
<tr>
<td><strong>Mechanical/Interposer design</strong></td>
<td>GDS</td>
</tr>
<tr>
<td></td>
<td>Bump pad netlist</td>
</tr>
<tr>
<td></td>
<td>Bump Ballout</td>
</tr>
<tr>
<td><strong>Thermal Simulation</strong></td>
<td>Flotherm</td>
</tr>
<tr>
<td></td>
<td>Icepak</td>
</tr>
</tbody>
</table>
Future of HBM Solution

HBM would penetrate various market segments in the short future.
Thank you