Expanding the Boundaries of the AI Revolution:
An In-depth Study of High Bandwidth Memory

Nayoung Lee & Sung Lee | March 2018
Table of Contents

1 THE MEMORY CHALLENGES of DEEP LEARNING
2 WHY HBM?
3 HIGH BANDWIDTH MEMORY DEEP-DIVE
THE MEMORY CHALLENGES of DEEP LEARNING
Deep Neural Network Fundamental Concepts

Deep Neural Network

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

Simple View

Weights x Input => Multiply & Accumulate sum

Layer

Output

Source: Standford
The Need for High Bandwidth Memory

1) In-Datacenter Performance Analysis of a Tensor Processing Unit, N. P. Jouppi et al. (Google)
HBM, What’s the difference?

**GDDR/DDR/LPDDR**

- FBGA

![GDDR/DDR/LPDDR Diagram]

**HBM**

- KGSD

![HBM Diagram]

- HBM in 2.5D SiP

![HBM in 2.5D SiP Diagram]

Soldered on PCB directly
Or
Use as DIMM Type

Substrate
High Bandwidth Memory Delivers Small Form Factor

HBM provides highest bandwidth compared to other DRAM memories per unit area

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
High Bandwidth Memory Delivers Small Form Factor

<table>
<thead>
<tr>
<th></th>
<th>GDDR5(X)</th>
<th>HBM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>8Gb x 12 = 12GB</td>
<td>8GB x 4 = 32GB</td>
</tr>
<tr>
<td>IO speed</td>
<td>8Gbps - 11Gbps</td>
<td>2Gbps</td>
</tr>
<tr>
<td># of IO</td>
<td>384 bits</td>
<td>1024*4 = 4096</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>384 – 528GB</td>
<td>1TB</td>
</tr>
</tbody>
</table>
High Bandwidth Memory Delivers Unprecedented Bandwidth

HBM overcomes all DRAM bandwidth challenges
High Bandwidth Memory Delivers Power Efficiency

HBM low speed per pin & Cio reduces power consumption and increases power efficiency

Power Efficiency

<table>
<thead>
<tr>
<th>Gbps</th>
<th>DDR 1</th>
<th>DDR 2</th>
<th>DDR 3</th>
<th>DDR 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR</td>
<td>0.4</td>
<td>0.8</td>
<td>1.6</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Power Consumption (mW/Gbps/pin)

- DDR x 16: 1.00
- DDR4 x 16: 0.96
- GDDR5 x 32: 0.58
- HBM1: 0.35
- HBM2: 0.33
Next Generation System Architectures Leveraging HBM

HBM and 2.5D integration unlock new system architectures

**HPC & Server**
(B/W & Capacity)

**Network & Graphics**
(B/W)

**Client-DT & NB**
(B/W & Cost)
HIGH BANDWIDTH MEMORY DEEP-DIVE

1) Innovative Design
2) Revolutionary Technological Features
3) Next Generation Line-up Considerations

High bandwidth, high power efficiency and compact form factors have propelled HBM collaboration engagements covering all IT sectors.
e.g. Graphics, AI/Deep Learning, HPC, SVR, NTW Router/Switches etc.

Total HBM (+HMC) market expected to increase from $922.7M in 2018 to $3,842.5M by 2023, resulting in CAGR 33%.  
(Source: RESEARCH AND MARKETS)
Innovative Design

HBM KGSD Architecture

- 11.87x7.75x0.72mm PKG dimension
- 9Gb per cell array (Optional 1Gb ECC cell)
- 4/8GB density per mKGSD stack
- Max 2.4Gbps data transmission speed enabling 307GB/s B/W performance
Innovative Design

HBM Gen2 Core Die

- 10.63mm x 6.65mm
- Supports Pseudo CH mode
- 2 individual sub-CH of 64bits I/O, 16 banks
- Two seamless array access w/ Burst Length 4
- 256b Prefetch per PCH
- **11.87mm x 8.87mm**
- **Programmable Memory Built-In Self Test**
- **Direct Access**
- **IEEE1500**
- **PHY**
PKG Stacking & Interconnection

- TSV Formation
- Wafer Molding
- Underfill
- Temporary Bond/Debonding
- Vertical Chip Stacking

Revolutionary Technical Features
PKG Stacking & Interconnection

Revolutionary Technical Features

Wire Bonding

Through Silicon Via
## Revolutionary Technical Features

### Wafer & KGSD PKG Level Reliability

<table>
<thead>
<tr>
<th>Wafer-level Process Qualification</th>
<th>PKG-level Product Qualification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Dependent Dielectric Breakdown</td>
<td>EFR, HTOL, LTOL (Lifetime)</td>
</tr>
<tr>
<td>Hot Carrier Injection</td>
<td>TC, THB, HAST, uHAST, HTS w/ Preconditioning (Environmental)</td>
</tr>
<tr>
<td>Negative Bias Temp Instability</td>
<td>Electrostatic Discharge</td>
</tr>
<tr>
<td>Electro Migration</td>
<td>Latch-up</td>
</tr>
<tr>
<td>Stress Migration</td>
<td>Package Construction Analysis</td>
</tr>
<tr>
<td><strong>TSV, uBump Electromigration</strong></td>
<td>Electrical Characterization</td>
</tr>
</tbody>
</table>
## Wafer & KGSD PKG Level Reliability

### Revolutionary Technical Features

<table>
<thead>
<tr>
<th>Type</th>
<th>Direction</th>
<th>10.1% Lifetime</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Die</td>
<td>VDD</td>
<td></td>
<td>• $\Delta R/R_0 x 100 &gt; 20%$</td>
</tr>
<tr>
<td></td>
<td>VSS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base Die</td>
<td>VDD</td>
<td>&gt;&gt; 10 years</td>
<td>• $F(10\text{yrs}) &lt; 0.1%$  @ use condition</td>
</tr>
<tr>
<td></td>
<td>VSS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSV</td>
<td>VDD</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VSS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The diagram illustrates the reliability testing and statistical distribution (CDF) of the different components under various conditions. The CDF graph shows the distribution of failure rates over time, crucial for understanding the long-term reliability of the packaging solution.
**Revolutionary Technical Features**

**Wafer & KGSD PKG Level Reliability**

### Direct Access Bump

- **Method**: Human Body Model
- **Target**: \( \geq 2,000\text{V} \)

- **Method**: Charged Device Model
- **Target**: \( \geq 500\text{V} \)

### PHY Bump

- **Method**: VF-TLP (CDM-like)
- **Target**: \( I_{t2} \geq \sim 1.\text{mA} \)

*Very Fast Transmission Line Pulse*
### Revolutionary Technical Features

#### Wafer & KGSD PKG Level Reliability

**KGSD HBM Test Flow**

<table>
<thead>
<tr>
<th>Core Die</th>
<th>Base Die</th>
</tr>
</thead>
<tbody>
<tr>
<td>WFBI</td>
<td>Logic Test</td>
</tr>
<tr>
<td>Hot &amp; Cold Test</td>
<td></td>
</tr>
<tr>
<td>Repair</td>
<td></td>
</tr>
</tbody>
</table>

**KGSD**

- TSV Scan
- Built-In Stress
- Hot & Cold Test
- Speed Test
<table>
<thead>
<tr>
<th>Area</th>
<th>Type</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHY</td>
<td>Function Test</td>
<td>RD/WT, CL, BL</td>
</tr>
<tr>
<td></td>
<td>Margin Test</td>
<td>Speed, VDD, Setup/Hold Timing</td>
</tr>
<tr>
<td>TSV</td>
<td>Function Test</td>
<td>RD/WT, CL, BL, TSV interface</td>
</tr>
<tr>
<td></td>
<td>OS Check</td>
<td>TSV Open/Short Check</td>
</tr>
<tr>
<td>Logic</td>
<td>Function Test</td>
<td>IEEE1500, Function, BIST, Repair</td>
</tr>
<tr>
<td></td>
<td>Margin Test</td>
<td>VDD, Speed, Setup/Hold</td>
</tr>
<tr>
<td>Core</td>
<td>Function Test</td>
<td>RD/WT, Self Ref, Power Down</td>
</tr>
<tr>
<td></td>
<td>Margin Test</td>
<td>Speed, VDD, Async, Refresh</td>
</tr>
<tr>
<td></td>
<td>Repair</td>
<td>Cell Repair</td>
</tr>
</tbody>
</table>
Next Generation Line-up

Key Performance Considerations

- Transistor performance between DRAM process and Logic Process (2.8Gbps~3Gbps may be the realistic max speed on DRAM)
- TSV lines to be doubled to secure valid window
- Speed increasing makes worse power consumption
- All possible solution should be considered for power reduction
- Additional HBM cubes
- DRAM density and process are limited by SiP size
- Higher DRAM stack has to be considered to increase density
Next Generation Line-up

Key Performance Considerations

Cost Effective Solutions

- **TSVless Si-Interposer**
  - Removing Si to expose BEoL layer (as RDL)

- **2.1D SiP**
  - Fine pitch organic substrate allows direct interconnection w/o interposer

- **Fan Out SiP on Sub.**
  - Removing Si-interposer thanks to fine pitch RDL trace of Fan Out Package

High Speed Signal Transmission

- **Si Photonics in 2.5D SiP**
  - Chip to chip optical signal transmission through embedded wave guide in Si-interposer

Source: CEA-Leti

Low Power and Small Form Factor

- **Hetero-generous 3D Stack**
  - More chips in a package with TSV stack
Thank you

Come visit us at Booth #711 and learn more about SK hynix memory solutions