HBM (High Bandwidth Memory) for 2.5D

Dr. Hongshin Jun
SK hynix Inc.
ICT Industry Outlook – Changes & Preparations

ICT Trend & Value-chain

IoT, IoE
Big Data
Networks
Data in Motion
Analytics
Vertical Integration toward service-centric Industry

Dr. Center Traffic
3.1ZB @’13 → 8.6ZB @’18 (CAGR 23%)

System Architecture

Computing Intensive
Data Intensive

Database
Servers
in-Memory Computing

HPC Memory Bandwidth/node
100–500GB/s @’15 → 2~4TB/s @’19

Data in Motion
Analytics

Evolutionary

Follow the scalability requirement of system with existing infra-structure
→ DDR4E, P-DDR4, P-LPDDR4

Revolutionary

Architecture Breakthrough to meet future system performance requirement
→ Total Optimization
→ Memory hierarchy tiering
→ Stacking technology

Role Sharing of Memory Solutions: B/W vs. Capacity

HBM* (B/W)
Capacity

simultaneous Preparations

simultaneous Changes

Technology Complexity ∝ CAPEX

DRAM Scaling Limits

# of Process Steps
CAGR > 5%

# of Process Steps
CAGR < 5%

2006 (~um)
2016(1x nm)

~9$
ASP CAGR -30%
(1Gb Eq.)

~0.x$

* HBM: High Bandwidth Memory

Computing Intensive

System Architecture
System & Memory Architecture Projection

HPC & Server (B/W & Capacity)

Network & Graphics (B/W)

Client-DT & NB (B/W & Cost)

Mobile & Wearable (LP, Small Form Factor, B/W & Cost)
General Memory Requirements

Each application has different memory requirement, but most common are high bandwidth and power based on real time random operation.
Technology Challenges: Bandwidth

HBM can overcome all DRAM’s challenges for high bandwidth

Bandwidth Challenges

High B/W with many I/O

Source: SK hynix
Low Speed/pin and Low Cio of HBM reduce power consumption and increase power efficiency.
Various TSV development experiences have resulted in real production of HBM1

SK hynix’s Plan on year 2015
- Volume Production of HBM1
- HBM2 Universal Daisy Chain
- 9mKGSD HBM2 Development
Mass Production of the World 1\textsuperscript{st} HBM

SK hynix completed the qualification for mass production (March’15)

- Worldwide first HBM provider
- Mass Production Start from Apr.’15
- HBM2 design wins in progress with major SoCs in multiple markets

“It is shipping mass production volumes of 1st generation High Bandwidth Memory (HBM1)”

# HBM2 Product Configuration

![HBM2 Diagram]

<table>
<thead>
<tr>
<th>Density/Cube (GB)</th>
<th>8Gb based</th>
<th>9mKGSD</th>
<th>5mKGSD</th>
<th>3mKGSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>8GB</td>
<td>1024</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>4GB</td>
<td>1024</td>
<td>1.0</td>
<td>1.6</td>
<td>2.0</td>
</tr>
<tr>
<td>2GB</td>
<td>1024</td>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speed/pin (Gbps)</th>
<th>128</th>
<th>204</th>
<th>256</th>
<th>128</th>
<th>204</th>
<th>256</th>
<th>128</th>
<th>204</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>IO</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bandwidth (GB/s)</th>
<th>8 Cores</th>
<th>4 Cores</th>
<th>2 Cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>HPC, Server</td>
<td>HPC, Server, Graphics, Network</td>
<td>Graphics, Cache</td>
</tr>
<tr>
<td>Config. / system</td>
<td>8 / 6 / 4 Cube</td>
<td>4 / 2 / 1 Cube</td>
<td>2 / 1 Cube</td>
</tr>
</tbody>
</table>
Industry is moving to HBM solution as TSV memory

The number of design win means number of company (More number of projects)
HBM Architecture Overview

- 4 Core DRAM + 1 Base logic die (Chip on Wafer)

Source: D.U Lee, SK hynix, ISSCC 2014
HBM1 Base Die Architecture

- Base die consists of 3 Areas – PHY, TSV, Test Port Area

HBM ballout area
6,050x3,264 μm

Source: D.U Lee, SK hynix, ISSCC 2014
HBM1 Core Design Architecture

- Each core die has 2 channels
- 1 channel consists of 128 TSV I/O with 2n pre-fetch

<table>
<thead>
<tr>
<th>CH-Left</th>
<th>CH-Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0</td>
<td>B0</td>
</tr>
<tr>
<td>XCTRL</td>
<td>XCTRL</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>YCTRL</td>
<td>YCTRL</td>
</tr>
<tr>
<td>B1</td>
<td>B1</td>
</tr>
<tr>
<td>B2</td>
<td>B2</td>
</tr>
<tr>
<td>B3</td>
<td>B3</td>
</tr>
<tr>
<td>B4</td>
<td>B4</td>
</tr>
<tr>
<td>B5</td>
<td>B5</td>
</tr>
<tr>
<td>B6</td>
<td>B6</td>
</tr>
<tr>
<td>B7</td>
<td>B7</td>
</tr>
<tr>
<td>DWORD 0</td>
<td>DWORD 0</td>
</tr>
<tr>
<td>32 I/O</td>
<td>32 I/O</td>
</tr>
<tr>
<td>DWORD 1</td>
<td>DWORD 1</td>
</tr>
<tr>
<td>32 I/O</td>
<td>32 I/O</td>
</tr>
<tr>
<td>DWORD 2</td>
<td>DWORD 2</td>
</tr>
<tr>
<td>32 I/O</td>
<td>32 I/O</td>
</tr>
<tr>
<td>DWORD 3</td>
<td>DWORD 3</td>
</tr>
<tr>
<td>32 I/O</td>
<td>32 I/O</td>
</tr>
</tbody>
</table>

1 bank : 2 sub-banks(64Mb) → non-shared I/O between sub-banks

Source: D.U Lee, SK hynix, ISSCC 2014
HBM2 Core Architecture: Pseudo Channel

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

- Pseudo channels share AWORD(CMD), but have separated banks & independent 64 I/Os
Restriction of tFAW in Legacy mode

- In Legacy mode, each channel has 2KB page size
- Restriction of Gapless Bank Activation by tFAW (4 Activate Window)
  - Suppose \( tCK=2\text{ns}, \ tFAW=30\text{ns}, \ tRRD=4\text{ns} \)
  - \( tFAW=30\text{ns} > 4\text{Bank}*tRRD=16\text{ns} \)
  ➞ Lower efficiency of Band Width
Benefit of Pseudo Channel

- Pseudo channel has reduced page size 1KB (2KB in Legacy mode)
- Lower Active Power (IDD0) by 1K Page size
- Define tEAW (1KB x 8 ACT) instead of tFAW (2KB x 4 ACT)
- Bandwidth improvement by more Activations during tFAW
Mechanical Outline: molded KGSD

- mKGSD (1 Base + 2/4/8 DRAM (Core); molded Known Good Stacked Die)

<table>
<thead>
<tr>
<th>Item</th>
<th>Value</th>
<th>Bump</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Gen1 - Package Dimension (X, Y)</td>
<td>5.48 mm x 7.29mm</td>
<td>CD 25um (As Reflow)</td>
<td></td>
</tr>
<tr>
<td>(a) Gen2 - Package Dimension (X, Y)</td>
<td>7.75 mm x 11.87mm</td>
<td>Pitch 55um</td>
<td></td>
</tr>
<tr>
<td>(b) Gen1 - Package Body Height (Z)</td>
<td>0.49 mm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) Gen2 - Package Body Height (Z)</td>
<td>0.72 mm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Micro Bump Array (MPGA)</td>
<td>JEDEC</td>
<td>-</td>
<td>JC11-2.883, JC11-4.884</td>
</tr>
</tbody>
</table>
HBM 2.5D SiP Structure

- System-in-Package implementation with KGSD

KGSD

HBM in 2.5D SiP

3D Memory (HBM)

Silicon die

Base die

Interposer

PKG Substrate

SoC
HBM KGSD Test

**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)

- **Base Die**
  - Logic Test
    - JEDEC IEEE1500

- **KGSD**
  - B/I Stress (BISS DFT)
    - Dynamic Stress (BISS)
  - KGSD Core Test (Productivity)
  - KGSD Speed Test

**KGSD (4 DRAMs + 1 Logic)**

- Bump & Stack Process
HBM KGSD Test

- DA pads for productivity
- KGSD Test covers TSV, DRAM cell, PHY, IEEE1500, and repairs TSV, DRAM cells
- TSVOS test and Repair
- uBump Test
  - Screen leakage failure from uBump (DC)
  - At-speed loopback test (AC)
- At-speed test on PHY, DRAM cells

<table>
<thead>
<tr>
<th>Test Area</th>
<th>Function</th>
<th>Detail item</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHY</td>
<td>Function Test</td>
<td>RD/WT, CL, BL</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Margin Test</td>
<td>Speed, VDD, Setup/Hold Timing</td>
<td>100%</td>
</tr>
<tr>
<td>TSV</td>
<td>Function Test</td>
<td>RD/WT, CL, BL, TSV interface</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>OS Check</td>
<td>TSV Open/Short Check</td>
<td>100%</td>
</tr>
<tr>
<td>Logic</td>
<td>Function Test</td>
<td>IEEE1500, Function, BIST, Repair</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Margin Test</td>
<td>VDD, Speed, Setup/Hold</td>
<td>100%</td>
</tr>
<tr>
<td>Core</td>
<td>Function Test</td>
<td>RD/WT, Self Ref, Power Down</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Margin Test</td>
<td>Speed, VDD, Async, Refresh</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Repair</td>
<td>Cell Repair</td>
<td>100%</td>
</tr>
</tbody>
</table>
HBM Test Features for 2.5D

- HBM has test/repair features for 2.5D accessible from host ASIC through IEEE1500
- Available at 2.5D SiP after assembly

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYPASS</td>
<td>ALL-Channel Bypass</td>
</tr>
<tr>
<td>EXTEST RX/TX</td>
<td>uBump boundary scan Rx/Tx test (Open/short)</td>
</tr>
<tr>
<td>MBIST</td>
<td>Memory Built-In Self Test</td>
</tr>
<tr>
<td>SOFT_REPAIR</td>
<td>Soft repair of failing DRAM bit cells</td>
</tr>
<tr>
<td>HARD_REPAIR</td>
<td>Hard repair of failing DRAM bit cells</td>
</tr>
<tr>
<td>DWORD MISR</td>
<td>Read back signature in DWORD MISR</td>
</tr>
<tr>
<td>AWORD MISR</td>
<td>Read back signature in AWORD MISR</td>
</tr>
<tr>
<td>SOFT/HARD_LANE_REPAIR</td>
<td>Perform Lane remapping</td>
</tr>
<tr>
<td>DEVICE ID</td>
<td>Read JTAG Device ID</td>
</tr>
<tr>
<td>TEMPERATURE</td>
<td>Read 8-bits binary temperature code</td>
</tr>
<tr>
<td>MODE REGISTER DUMP</td>
<td>Read/Write the DRAM's Mode Register</td>
</tr>
</tbody>
</table>
PPR: Soft/Hard Repair

- DRAM Cell Test and Repair through IEEE1500
  - After SiP Assembly
  - 4-Row Rep / 2-Bank (Repair DRAM) (64 row / channel @ 8Hi)
- Procedure
  1) Run MBIST: Pass/Fail, Report fail row addresses
  2) Run Soft Repair: Soft repair the fail row addresses by writing register
  3) Run Hard Repair: Hard repair the fail row addresses by cutting e-Fuse
Lane Repair

- HBM support interconnect lane remapping through IEEE1500 instructions
  - After SiP Assembly
  - Lane remapping is independent for each channel
- Procedure
  1) Test lanes between HOST and HBM using EXTEST and MISR instructions
  2) Run SOFT_LANE_REPAIR: Perform lane remapping by writing register.
  3) Run HARD_LANE_REPAIR: Perform lane remapping by cutting eFuse.

Mode 1: 1 fail per byte

Simple DBI lost

Mode 2: 1 fail per double byte

Complex DBI maintained

Note: circuit represents WRITE data path; same 2:1 MUX used in both modes!
Challenges in 2.5D Failure Analysis

- Traditional FA for DRAM Component
  - RMA Request
  - Failure Duplication @ System level
  - DRAM Component ATE Test
  - Customer System Based Application Test
  - Further Analysis & Root Cause

- New FA for HBM @ SiP Level
  - RMA Request
  - Failure Duplication @ System level
  - SiP Level ATE Test
  - SiP Level Application Test
  - Further Analysis & Root Cause

No good solution for evaluating the interface status between ASIC and HBM inside of SiP
Conclusion

- HBM provides breakthrough memory solutions for next generation high performance systems with higher bandwidth, lower power, and smaller form factor.
- SK hynix set up HBM 3D Stacking and Test flow to deliver KGSD.
- SK hynix is shipping mass production volumes of HBM.
- 2.5D Ecosystem requires close collaboration among all stakeholders.
Thank You

If you have further questions about SK hynix memory information, please ask to SK hynix DRAM Product Planning & Enabling Office (ppne@skhynix.com) or SK hynix Japan. Thank you.