HBM: Memory Solution for Bandwidth-Hungry Processors

Joonyoung Kim and Younsu Kim

SK hynix Inc.
August 2014
Memory requirement

HBM: Memory Solution for Density & Bandwidth-Hungry Processors

< Exa-scale Roadmap >

<table>
<thead>
<tr>
<th>Systems</th>
<th>2009</th>
<th>2011</th>
<th>2015</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Peak Flops/s</td>
<td>2 Peta</td>
<td>20 Peta</td>
<td>100-200 Peta</td>
<td>1 Exa</td>
</tr>
<tr>
<td>System Memory</td>
<td>0.3 PB</td>
<td>1 PB</td>
<td>5 PB</td>
<td>10 PB</td>
</tr>
<tr>
<td>Node Performance</td>
<td>125 GF</td>
<td>200 GF</td>
<td>400 GF</td>
<td>1-10 TF</td>
</tr>
<tr>
<td>Node Memory BW</td>
<td>25 GB/s</td>
<td>40 GB/s</td>
<td>100 GB/s</td>
<td>200-400 GB/s</td>
</tr>
<tr>
<td>Node Concurrency</td>
<td>12</td>
<td>32</td>
<td>0(100)</td>
<td>0(1000)</td>
</tr>
<tr>
<td>Interconnect BW</td>
<td>1.5 GB/s</td>
<td>10 GB/s</td>
<td>25 GB/s</td>
<td>50 GB/s</td>
</tr>
<tr>
<td>System Size (Nodes)</td>
<td>18,700</td>
<td>100,000</td>
<td>500,000</td>
<td>0(Million)</td>
</tr>
<tr>
<td>Total Concurrency</td>
<td>225,000</td>
<td>3 Million</td>
<td>50 Million</td>
<td>0(Billion)</td>
</tr>
<tr>
<td>Storage</td>
<td>15 PB</td>
<td>30 PB</td>
<td>150 PB</td>
<td>300 PB</td>
</tr>
<tr>
<td>I/O</td>
<td>0.2 TB/s</td>
<td>2 TB/s</td>
<td>10 TB/s</td>
<td>20 TB/s</td>
</tr>
<tr>
<td>MTTF</td>
<td>Days</td>
<td>Days</td>
<td>Days</td>
<td>0(1 Day)</td>
</tr>
<tr>
<td>Power</td>
<td>6 MW</td>
<td>~10 MW</td>
<td>~10 MW</td>
<td>~20 MW</td>
</tr>
</tbody>
</table>

Memory bottleneck & solution - Speed, Density, Power & SFF

TSV is a revolutionary technology for overcoming the bottleneck

- **High Bandwidth**
  - solution beyond DDR4/GDDR5/LPDDR4

- **High Capacity**
  - solution to overcome DRAM Scaling Limit

- **Power Efficiency**
  - solution to meet system power budget

- **Small Form Factor**
  - solution to meet Size & Cost limitation
TSV (Through Silicon via)

TSV is the technology of 3D Stack
(High Density / Small size PKG / High speed)

< Wire bonding PKG >  

< TSV PKG >
## Bottleneck 1) Bandwidth

<table>
<thead>
<tr>
<th>Config.</th>
<th><strong>DDR3</strong></th>
<th><strong>TSV(HBM)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IO</strong></td>
<td>64 DQ</td>
<td>1024 DQ</td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td>1.6G bps</td>
<td>1~2Gbps</td>
</tr>
<tr>
<td><strong>Bandwidth</strong></td>
<td>64 Gbps ➔ 12.8GBps</td>
<td>1024 Gbps ➔ Max 256GBps</td>
</tr>
<tr>
<td><strong>Compare</strong></td>
<td>Long Length ➔ RLC increase</td>
<td>Short Length ➔ RLC decrease</td>
</tr>
</tbody>
</table>
Bottleneck 2) Technology Limit

1. Capacity Limit

2. Scale Down Limit
## Bottleneck 3) Small Form Factor

<table>
<thead>
<tr>
<th></th>
<th>Wire bonding - DDR3</th>
<th>TSV - HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image</strong></td>
<td><img src="http://www.shmj.or.jp/english/packaging/pac90s.html" alt="Image 1" /></td>
<td><img src="http://www.shmj.or.jp/english/packaging/pac90s.html" alt="Image 2" /></td>
</tr>
<tr>
<td><strong>PKG Size@die</strong></td>
<td>100% (117mm$^2$)</td>
<td>36% (42mm$^2$)</td>
</tr>
<tr>
<td><strong>mm$^2$@128GB/s</strong></td>
<td>100% (3744mm$^2$)</td>
<td>11% (42mm$^2$)</td>
</tr>
<tr>
<td><em><em>Power Consumption</em> @128GB/s</em>*</td>
<td>100% (6.4W)</td>
<td>51% (3.3W)</td>
</tr>
</tbody>
</table>

* Power Cal = IMPT
Bottleneck 4) Low Power

- Lower Speed/pin and x1024 Wide IO → low power consumption per Pin.
- Lower Cio(0.6pF), No Termination → small IO current consumption
- Power consumption is decreased by 42% compared with GDDR5
HBM Overall Architecture

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

HBM, What are the differences?

**KGSD**

**Memory supporting for System in package**

- **FBGA**
- **KGSD**

* KGSD (Known Good Stacked Die)
HBM Market Segments

HBM market will scale-out to various segments (Over 21 Design-Ins In Progress)

- **Low Power**
  - VDD=1.XV
  - Good I/O power eff.
  - Digital Consumer, Automotive, etc.

- **HPC**
  - Density flexibility
  - >256GBps BW

- **Network**
  - ECC
  - Post-package Repair
  - MBIST
  - 128GBps BW
  - JEDEC std

- **Client**
HBM Overall specification

1\textsuperscript{st} Gen HBM
- 2Gb per DRAM die
- 1Gbps speed /pin
- 128GB/s Bandwidth
- 4 Hi Stack (1GB)
  - x1024 IO
  - 1.2V VDD
  - KGSD w/ µBump

2\textsuperscript{nd} Gen HBM
- 8Gb per DRAM die
- 2Gbps speed/pin
- 256GBps Bandwidth/Stack
- 4/8 Hi Stack (4GB/8GB)
HBM Gen1 5mKGSD Structure

5mKGSD (molded Known Good Stacked Die)
- 1 Base + 4 Core (DRAM) with Side Mold -

<table>
<thead>
<tr>
<th>Item</th>
<th>Value</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) uBump Diameter</td>
<td>25um (± 3 um)</td>
<td></td>
</tr>
<tr>
<td>(b) uBump Height</td>
<td>35um (± 3.5 um)</td>
<td>Cu/Ni/SnAG (17/3/15um)</td>
</tr>
<tr>
<td>(c) uBump Pitch</td>
<td>55um</td>
<td></td>
</tr>
<tr>
<td>(d) uBump Array (MPGA)</td>
<td>JEDEC</td>
<td>JC11-2.883, JC11-4.884</td>
</tr>
</tbody>
</table>
### Comparison of HBM and other DRAMs

<table>
<thead>
<tr>
<th>Item</th>
<th>DDR3 (x8)</th>
<th>GDDR5 (x32)</th>
<th>4-Hi HBM (x1024)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O</td>
<td>8</td>
<td>32</td>
<td>1024</td>
</tr>
<tr>
<td>Prefetch (Per IO)</td>
<td>8</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Access Granularity (=I/O x Prefetch)</td>
<td>8Byte</td>
<td>32Byte</td>
<td>256Byte</td>
</tr>
<tr>
<td>Max. Bandwidth</td>
<td>2GB/s</td>
<td>28GB/s</td>
<td>128~256GB/s</td>
</tr>
<tr>
<td>tRC</td>
<td>40~48ns</td>
<td>40ns (=1.6v, 1.5v)</td>
<td>40~48ns</td>
</tr>
<tr>
<td>tCCD</td>
<td>4ns (=4tCK)</td>
<td>2ns (=4tCK)</td>
<td>2ns (=1tCK)</td>
</tr>
<tr>
<td>VPP</td>
<td>Internal VPP</td>
<td>Internal VPP, (Opt. Ext. VPP)</td>
<td>Ext. VPP</td>
</tr>
<tr>
<td>VDD</td>
<td>1.5, 1.35</td>
<td>1.6, 1.5, 1.35</td>
<td>1.2</td>
</tr>
<tr>
<td>CMD Input</td>
<td>Single CMD</td>
<td>Single CMD</td>
<td>Dual CMD</td>
</tr>
<tr>
<td>Refresh Single Bank</td>
<td>X</td>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>DBI mode</td>
<td>X</td>
<td>O (DBI_DC)</td>
<td>O (DBI_AC)</td>
</tr>
</tbody>
</table>
Dual CMD interface

CMD efficiency increased by Semi-independent row/column input

Row/column input through different pins

- **RAS [0:5]**
- **CK_t**
- **CK_c**
- **CAS [0:7]**

Conventional DRAMs share RAS/CAS CMD:

- **ADD**
- **Bank**
- **CMD**
- **CK_t**
- **CK_c**

RA, CA
RAS/CAS/WE/CS

RA, CA
REF Single Bank

Single bank refresh and programmable tRAS
Concurrent read/write operation with single bank refresh allows data bus to remain active.
REF Single Bank

Command bus efficiency can be maximized
- Dual Command & REF single bank -

<table>
<thead>
<tr>
<th>CLK</th>
<th>T</th>
<th>T+1</th>
<th>T+2</th>
<th>T+3</th>
<th>T+4</th>
<th>T+5</th>
<th>T+6</th>
<th>T+7</th>
<th>T+8</th>
</tr>
</thead>
<tbody>
<tr>
<td>BANK0</td>
<td>ACT</td>
<td></td>
<td>tRCD</td>
<td></td>
<td>WT</td>
<td></td>
<td>PCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BANK1</td>
<td>tRRD</td>
<td></td>
<td>ACT</td>
<td></td>
<td></td>
<td></td>
<td>WT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BANK2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BANK3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>WT</td>
<td></td>
<td></td>
<td></td>
<td>RD</td>
</tr>
<tr>
<td>BANK4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>tRRD</td>
<td>WT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BANK5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BANK6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REFSB</td>
</tr>
<tr>
<td>BANK7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HBM Core Architecture

HBM single die has 2 channels
1 channel consists of 128 TSV I/O with 2n prefetch

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

HBM Base Die Architecture

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

Base Die Customization – Future HBM Concept

Logic Layer ➔ Host I/F + Memory I/F + Base Logic/IP Block

Customization to meet various requirements

Overcome Memory Scaling
- Timing
- Refresh

- Parallel-to-Serial(P2S)/S2P
- JTAG, PMBIST
- Configuration Registers
- Error Handling
- ...
HBM Thermal Management

Thermal dummy bumps as well as well-designed device architecture are helpful for thermal dissipation → No mechanical reliability issues by thermal dummy bumps.
HBM Long-term Roadmap\(^{(2)}\) (Preliminary)

HBM product longevity is critical in several applications
SK hynix plans to address longevity requirement

Bandwidth: starts at 128GB/s & plans for higher

• Note 1 – anticipated future HBM density
• Note 2 – roadmap is subject to changes without prior notifications
HBM Summary

Perfect memory solution for various application requirement

- **High Bandwidth**
  - ~256GB/s

- **High Density PKG**
  - Up to 8GB

- **Smaller Form Factor**
  - -65%

- **Good Power Efficiency**
  - 68%

1) KGSD

1 ~ 4 Cubes per GPU

4 ~ 8 Stacks

205.132.242.85 / 2014. 07. 18  16 : 51 / B34047 / 2057897
Thank You!