New Directions in Memory Architecture

Bob Brennan  
Senior Vice President  
Memory System Architecture Lab  
Samsung Semiconductor, Inc.  
bob.brennan@samsung.com
Legal Disclaimer

This presentation is intended to provide information concerning memory industry trends. We do our best to make sure that information presented is accurate and fully up-to-date. However, the presentation may be subject to technical inaccuracies, information that is not up-to-date or typographical errors. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of information provided on this presentation. Samsung reserves the right to make improvements, corrections and/or changes to this presentation at any time.

The information in this presentation or accompanying oral statements may include forward-looking statements. These forward-looking statements include all matters that are not historical facts, statements regarding the Samsung Electronics' intentions, beliefs or current expectations concerning, among other things, market prospects, growth, strategies, and the industry in which Samsung operates. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements contained in this presentation or in the accompanying oral statements. In addition, even if the information contained herein or the oral statements are shown to be accurate, those developments may not be indicative developments in future periods.
Agenda

» Environment – BW & Capacity growth

» DRAM – BW & Capacity -> Tiering

» Flash – Scales, Becomes Intelligent, Tiers

» STT-MRAM: New “Persistent Performance”
Environment: Escalating Demand for DRAM and Storage

In-Memory Analytics for Big Data

Escalating Memory-Intensive Workloads

- HPC
- Graphics
- Financial
- Gaming

Data Center Processor Growth

2x Volume Growth

- Network
- Ent.Storage
- Workstation
- HPC
- Public Cloud
- Enterprise
- Small Scale

Source: Intel

Growing x86 Server Virtualization Density

- VMs per Host
- VM Density per Host
- % of Installed Workloads Running in a VM

Source: Gartner and IDC

Unstructured Data vs. Structured Data

Source: EMC and IDC

Source: Intel

© Samsung
Environment – Bandwidth Demand

**Mobile:**
Display/GFX/Camera

Exponential Bandwidth Demand

**Server:**
Core Scaling

Linear to Exponential Bandwidth Demand

---

**FHD (1920x1080)**
13MP
1080p
F-HD

**UD (3840x2160)**
20+MP
4K
UHD

---

**Memory Bandwidth Requirements**

<table>
<thead>
<tr>
<th>Year</th>
<th>Peta-flops</th>
<th>20Peta-flops</th>
<th>Exa-flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Now</td>
<td>400~600 Mbps</td>
<td>10~20GB/s</td>
<td>7.5x (~100GB/s)</td>
</tr>
<tr>
<td>2018</td>
<td>12.5x (~5.3Gbps)</td>
<td>100x (~1.4TB/s)</td>
<td>12.5x (~1.4TB/s)</td>
</tr>
</tbody>
</table>

**Source:** “Memory systems for PetaFlop to ExaFlop class machines” by IBM, 2007 & 2010
Environment – Capacity Demand

Memory Capacity Requirements

Now
- Memory Capacity/System
- Memory Capacity/Node

2018
- >70x (≈10PB)
- >32x (≈128GB)

Memory Capacity Requirements

Now
- >5x (≈750TB)
- >4x (≈16GB)

2018
- >70x (≈10PB)
- >32x (≈128GB)

[Source: “Memory systems for PetaFlop to ExaFlop class machines” by IBM, 2007 & 2010]

Mobile:
Display/GFX/Camera
~Linear Capacity Demand

Server:
Core Scaling
Linear - Exponential Capacity Demand
Agenda

» Environment – BW & Capacity growth

» DRAM – BW & Capacity -> Tiering

» Flash – Scales, Gets Intelligent, Tiers

» STT-MRAM: New “Persistent Performance”
The “Trade-off Triangles”

- **DRAM**
  - Bandwidth
  - Power
  - Latency
  - Capacity

- **Non-Volatile**
  - IOPs
  - Power
  - Endurance
  - Capacity

© Samsung
## DRAM: Bandwidth Scaling

<table>
<thead>
<tr>
<th>Bandwidth [Mbps]</th>
<th>Latency</th>
<th>Power</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1333</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1866</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2133</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2400/2667</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2667</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3600</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4200</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Multi-Drop Bus Challenge:** Higher BW, Lower VDD

**DDR Wall?**

**New Solution Needed!**

Subject to cost/energy efficiency, scaling, ...

- DDR5 (?) & New I/F (?)
- Optical (?)

---

© Samsung
More & More Difficult: Disruptive Solution Needed
DRAM: Latency Challenge

More & More Difficult: Disruptive Solution Needed

~Constant

Low Latency Needed

Subject to cost/energy efficiency, scaling, ...
### DRAM: “Go Wide” for Bandwidth

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Mobile WIO2</th>
<th>HBM (High B/W Memory)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DRAM</td>
<td>Base die + DRAM</td>
</tr>
<tr>
<td>WIO2</td>
<td></td>
<td>Si Interposer</td>
</tr>
<tr>
<td>AP</td>
<td></td>
<td>HBM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Base</td>
</tr>
<tr>
<td>Bottom die</td>
<td>N/A</td>
<td>Buffering &amp; Signal re-routing</td>
</tr>
<tr>
<td>BW (GB/s)</td>
<td>25.6~51.2</td>
<td>128~256</td>
</tr>
<tr>
<td>Pin Speed</td>
<td>0.4~0.8 Gbps</td>
<td>1~2 Gbps</td>
</tr>
<tr>
<td># I/O</td>
<td>512</td>
<td>1,024</td>
</tr>
<tr>
<td># Bump Logic</td>
<td>1~2K</td>
<td>6K~8K</td>
</tr>
<tr>
<td>DRAM Logic</td>
<td>1~2K</td>
<td>~3K</td>
</tr>
<tr>
<td>Cube (GB)</td>
<td>1 / 2</td>
<td>1 / 2 / 4</td>
</tr>
<tr>
<td># TSV stack</td>
<td>1 / 2 / 4</td>
<td>1 / 2 / 4</td>
</tr>
<tr>
<td>DRAM density</td>
<td>8Gb</td>
<td>8Gb</td>
</tr>
<tr>
<td>Application</td>
<td>GFX card O</td>
<td>O</td>
</tr>
<tr>
<td>Application</td>
<td>ULT O</td>
<td>-</td>
</tr>
<tr>
<td>Application</td>
<td>HPC -</td>
<td>O</td>
</tr>
<tr>
<td>Application</td>
<td>Server -</td>
<td>O(Cache)</td>
</tr>
<tr>
<td>Application</td>
<td>Mobile O</td>
<td>-</td>
</tr>
</tbody>
</table>

**Good BW & Latency – Still Need Capacity**
DRAM: New Hierarchy Needed

Mobile

SOC

CPU

New Memory Controller

WIO

WIO

WIO

High BW DRAM

Server

CPU

New Memory Controller

DDR

DDR

New

New

High Bandwidth Tier

High Capacity Tier

“Wide” BW + Tiered Capacity
1st Step: Tiering DRAM

High Bandwidth Tier

High Capacity Tier
Agenda

» Environment – BW & Capacity growth

» DRAM – BW & Capacity -> Tiering

» Flash – Scales, Becomes Intelligent, Tiers

» STT-MRAM: New “Persistent Performance”
Flash: Capacity Scaling

Scaling Becomes Difficult – Need a New Solution
Flash: Endurance

NAND Flash Endurance

Years

Adaptive Algorithms on Controllers / Accelerators, Tuned for Process
Application & Usage Awareness

SSD Requirement
Shrink Rate Slows Down
Reliability Degrades
Performance Deteriorates

© Samsung
Flash: Performance

Latency & IOPS

- Rotational Latency
- AVG Seek
- IOPS

Interface & Performance

- Interface Unlocks Bandwidth: PCIeG2->G3->G4
- Solution needs to scale: Controllers, Algorithms, & Flash Organization

Increasing Intelligence & Sophistication
Flash: Inherent Intelligence

Intelligent IOPs

Intelligent Endurance

3D Scaling

© Samsung
2nd Step: Tiering Flash/HDDs

- **High Bandwidth Tier**
  - DRAM
  - DRAM
  - DRAM

- **High Capacity Tier**
  - DDR DDR DDR DDR DDR DDR DDR DDR DDR DDR DDR DDR DDR DDR DDR

- **Intelligent Flash Tier**
  - Flash Flash Flash Flash Flash Flash Flash Flash Flash Flash Flash Flash
  - HDD HDD HDD HDD HDD HDD HDD HDD HDD HDD HDD HDD HDD HDD HDD

- **HDDs**
  - HDD HDD HDD HDD HDD
Agenda

» Environment – BW & Capacity growth

» DRAM – BW & Capacity -> Tiering

» Flash – Scales, Becomes Intelligent, Tiers

» STT-MRAM: New “Persistent Performance”
Opportunity for New Technology

Bandwidth (GB/s)

Latency (ns)

HDD

Flash

Persistent Performance

DRAM

LLC
STT-MRAM

Promising Technology, Not Mature Yet
3rd Step: New possibilities

- High Bandwidth Tier
- DRAM
- DRAM
- DRAM
- Higher Bandwidth Tier
- DRAM
- DRAM
- DRAM
- High Capacity Tier
- HDD
- HDD
- HDD
- Persistent Performance Tier
- HDD
- HDD
- HDD

Intelligent Flash Tier
- HDD
- HDD
- HDD
- Persistent Performance Tier
- HDD
- HDD
- HDD

Higher Bandwidth Tier
- DRAM
- DRAM
- DRAM

Intelligent Flash Tier
- HDD
- HDD
- HDD

HDD

© Samsung
3rd Annual Samsung Memory and IT Solutions Summit 2013

“Architecting the Next Generation Data Center”

Join us: Tuesday, Oct. 22, 2013
Computer History Museum, Mt. View
Register at:
www.samsung.com/us/msf

THANK YOU!