The future of graphic and mobile memory for new applications

August 21st, 2016 | JIN KIM | Samsung Electronics
Disclaimer

This presentation is intended to provide information concerning memory industry. 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.
Contents

• Memory technology trend
• High speed graphic technology (>10Gbps)
• Low power mobile technology (>20%)
• Conclusion
Memory technology trend
Memory is at the core of new applications

Higher Performance
- GDDR5: 30GB/s
- HBM2: 256GB/s

x10 Bandwidth

Lower Power
- LP3: 1
- LP4: 0.7
- LP4X: 0.5

x0.5 Power Efficiency

Source: Samsung
Memory-centric system evolution

- Extreme B/W, performance/power, data processing, cost effective solutions
Memory technology trend

- GDDR6 with over 14Gbps, beyond 10Gbps GDDR5
- LP5, 20% more power-efficient than LP4X

Source: ISCA2016, Samsung
High Bandwidth Memory: HBM

1TB/s High Bandwidth

8H stacked 20nm 8GB HBM

Benefits

- **Performance**: HBM 1TB/s vs. GDDR5 0.8TB/s (X 2.7)
- **Power Efficiency**: HBM 0.8 vs. GDDR5 1 (X 0.8)

Source: Samsung
Processing In Memory: PIM

- Fill the performance gap and deliver energy-efficient solutions

Processing In-Memory

Better parallelism and lower bus traffic

Memory off-loading for lower frequency and power
High speed graphic technology ( >10Gbps)

- Graphic application requirement
- Asymmetric System, Crosstalk, EQ tuning
- GDDR6, Low cost HBM, PIM
High speed memory requirement

- For 4K real infographic virtual reality, 13.2GB, 1TB/s memory needed
- For 4K 3D mixed reality, +3.5GB, 151GB/s memory needed

### Gaming Virtual Reality memory

<table>
<thead>
<tr>
<th></th>
<th>QHD</th>
<th>4K UHD</th>
<th>8K UHD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gfx Capacity, GB</strong></td>
<td>2</td>
<td>8</td>
<td>13.2</td>
</tr>
<tr>
<td><strong>B/W, GB/s</strong></td>
<td>90</td>
<td>462</td>
<td>1064</td>
</tr>
</tbody>
</table>

### Mixed Reality memory

<table>
<thead>
<tr>
<th></th>
<th>QHD</th>
<th>4K UHD</th>
<th>8K UHD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Added Capacity, GB</strong></td>
<td>1.0</td>
<td>1.6</td>
<td>3.5</td>
</tr>
<tr>
<td><strong>B/W, GB/s</strong></td>
<td>28</td>
<td>101</td>
<td>527</td>
</tr>
</tbody>
</table>

Source: Samsung

**Variable Assumption**
Poly count, fps, # of texture per fragment, cache hit rate, tri-linear filtered,
# of virtual light source, Reflection/refraction ratio, ray bounce depth
Asymmetric system for higher data rate

- Focus on the respectively dedicated features to maximize data rate
  - Smart GPU: Training (Per-bit Timing/EQ) for minimizing static offset/noise
  - Noise immune DRAM: minimizing dynamic noise (Jitter, ISI/x-talk, clock duty/skew)

Source: Samsung
X-talk reduction for Board/PKG design

- Small X-talk Package : reduction of X-talk with better return path
- Crosstalk Reduction with coding : 3B4B, 8B9B

Source: Samsung
DFE for return-loss reduction on system

- Single ended signaling requires noise immune equalizer
  - DFE* is more suitable than CTLE**

CTLE & DFE

Quarter rate DFE with summer in sampler

Adopt merged summer/sampler for fast feedback

CTLE and DFE Periodically Calibrated by GPU

Source: Samsung

- Decision Feedback Equalization
- Continuous Time Linear Equalization
**GDDR6 ideas**

- **High Speed Signaling, 14Gbps ~ 16Gbps, 1.35V**
  - Low jitter clocking with WCK/byte, Per-bit RX/TX equalizer training, X-talk reduction
  - 2 channel with BL16, same Clock/ADD freq., twice of WCK/DQ freq.

---

**WCK Clocking**

- 7GHz ~ 8GHz
- Word → Byte
- 14Gbps ~ 16Gbps

**Target Timing**

- GDDR5
  - CK : 1.75Gbps
  - CMD : 1.75Gbps
  - ADDR : 3.5Gbps
  - WCK : 3.5Gbps
  - DQ : 7Gbps

- GDDR6
  - CK : 1.75Gbps
  - CA : 3.5Gbps
  - WCK : 7Gbps
  - DQ : 14Gbps

Source: Samsung
Low cost HBM for consumer segment

- ~200GBps with smaller # of TSV compared to HBM2
  - Cost competitiveness; remove buffer die, reduce # of TSV, organic interposer, etc..
  - Need inputs from Client segment for specific features

Challenges
1. IO reduction, Smaller # of TSV
2. Remove buffer die
3. Master/Slave structure
4. Remove ECC
5. Si or organic Interposer

Comparison

<table>
<thead>
<tr>
<th></th>
<th>HBM2</th>
<th>Low cost HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O</td>
<td>1024</td>
<td>~512</td>
</tr>
<tr>
<td>Pin speed</td>
<td>2Gbps</td>
<td>3Gbps ~</td>
</tr>
<tr>
<td>BW (GB/s)</td>
<td>256</td>
<td>~200</td>
</tr>
<tr>
<td>Cost/GB</td>
<td>1</td>
<td>0.X</td>
</tr>
</tbody>
</table>

Source: Samsung
PIM, Deep Learning in DRAM

- Parallel processing in buffer to reduce extreme-bandwidth
  - convolution, subsampling, matrix calculation
- Collaborate with accelerator for performance/cost

**Extreme B/W Requirement**

**Processing in Buffer**

* * xHBM: Extreme HBM
Low power mobile technology ( >20%)

• Motivation for low power mobile
• LP4X / LP5
• PIM
Motivation for low power mobile

- PC-level graphic performance and mobile power budget
- Power is continuously increasing with limited thermal budget

Source: Samsung
**Lower power solution, LP4X**

- **LP4X**: 4266Mbps, VDDQ/VDD = 0.6V/1.1V
  - IO power reduction with 0.6V VDDQ, Good example of small change but big gain

\[ V_{OH} = V_{DDQ}/3 \]
\[ V_{REF} = V_{OH}/2 \]

**LP4X Idea**

**LP4X Power Reduction**

- **18% Total Power Saving!!!!**
- **-45%**

**Conditions**: IDD4R(VDDQ+VDD2) Spec Value / 50% Data change each burst transfer / Included process node contribution

Source: Samsung
LP5 target & ideas

- LP5: 6400Mbps, VDDQ/VDD < 0.6V/1.1V
  - Extremely high band-width (~6.4Gbps) and smart power reduction (~20%)

Power Efficiency Trend

LP5 ideas

- CMD Based Data CLK (WCK)
  - IDD2N reduction
  - IDD4W/R reduction

- WCK Center-tap term
  - IDD6 reduction

- Deep Sleep Mode
  - IDD6 reduction

Source: Samsung
PIM, Lower power processing

- **Memory off-loading for reduced power consumption**
  - Reduce the unnecessary data transfer and frame rate control
- **Collaborate with SoC/AP for performance/power**
  - PoC with special memory for post/pre-processing
Conclusion
Conclusion

• Memory requirements have become more strict in time with respect to performance, power, and cost

• Keeps innovating technology to correspond to those requirements
  – Make efforts to extend the value of traditional memory
  – Figure out innovative memory solution

• Close collaboration with partners is essential for delivering the right memory solution.
Thank You!