Pivotal Memory Technologies Enabling New Generation of AI Workloads

Tien Shiah
Memory Product Marketing
Samsung Semiconductor Inc.
This presentation is intended to provide information concerning the 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.

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 of developments in future periods.
Applications drive Changes in Architectures

1st Wave
MS-DOS

2nd Wave
PC Era

3rd Wave
Internet

4th Wave
Mobile

NOW
AI

CPU-centric

Data-centric

x86 Processors

Apps Processors

Non-x86 processors & platforms

MS-DOS...

Apps
Processors

Processors

GPU/TPU

x86

FPGA's

Non-

FPGA's

GPU/TPU

Data-centric
Artificial Intelligence → MAINSTREAM

Speech, Natural Language
- Amazon Echo & Alexa
- Google Smart Home Devices
- Siri & Cortana Smart Assistants

Deep Learning
- Screening
- Genomics
- Prediction
- Game Theory

Image / Facial Recognition

Autonomous Driving
AI – What has Changed?

Deep Learning algorithms require **high memory bandwidth**
Faster Computation → Multi-core

High performance compute requires high memory bandwidth
Memory Bandwidth Comparison

* Based high performance configurations of HBM, GDDR, and DDR
HBM: High Bandwidth Memory

- Stacked MPGA (micro-pillar grid array) memory solution for high performance applications
- Samsung launched HBM2 in Q1 2016
- Uses DDR4 die with TSV (Through Silicon Vias)
- Available in 4H or 8H stacks
- Key Features:
  - 1024 I/O’s (8 Channel, 128bits per channel)
  - Per stack: 307GB/s (current generation)
    - 77X the speed of a PCIe 3.0 x4 slot, or
    - 77 HD movies transferred per second

** Announced HBM2E: +33% throughput (410GB/s), 2X density (16GB stack) **
HBM Basics: 2.5D System In Package

- A typical HBM SiP consists of a processor (or ASIC) and 1 or more HBM stacks mounted on a Silicon Interposer.
- The HBM consists of 4 or 8 DRAM die mounted on a buffer die.
- The entire system (Processor + HBM stack + Si Interposer) is encapsulated into one larger package by the customer.

Samsung manufactures and sells the HBM stack.
MPGA: Micro-Pillar Grid Array

Four High Stack (4H)

Eight High Stack (8H)

~ 720um
Not just about speed: Space Efficiency

**GDDR5**
- Density: $1 \text{ GB} \times 12 = 12\text{GB}$
- Speed/pin: $1 \text{ GB/s}$
- Pin count: 384
- B/W: $384 \text{ GB/s}$

**HBM2E**
- Density: $16 \text{ GB} \times 4 = 64\text{GB}$
- Speed/pin: $0.4 \text{ GB/s}$
- Pin count: 4096
- B/W: $1,640 \text{ GB/s}$

Real estate savings
AI: Compute vs. Memory Constrained

Roofline Model for TPU ASIC

Memory constrained

Roofline Model
- Point below slope = memory bandwidth constrained
- Point below horizontal = compute constrained

Neural Network | Characteristic             | Use Case
--------------|---------------------------|--------
MLP           | Structured input features | Ranking
CNN           | Spatial processing        | Image recognition
RNN           | Sequence processing       | Language translation

* LSTM (Long Short-Term Memory) is subset of RNN

Many Deep Learning applications are MEMORY bandwidth constrained → Need High Bandwidth Memory

Source: Google ISCA 2017
Memory Drives AI Performance

- **Faster Training, More Bandwidth**
- **Better Accuracy, More Capacity**

### Required Memory BW (GB/s)

<table>
<thead>
<tr>
<th>TFLOPS, # Core, Product</th>
<th>5.2</th>
<th>7</th>
<th>10</th>
<th>15</th>
<th>HBM2</th>
<th>HBM2E</th>
</tr>
</thead>
<tbody>
<tr>
<td>K110</td>
<td>2880</td>
<td>3072</td>
<td>3584</td>
<td>5120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M200</td>
<td>3072</td>
<td>3584</td>
<td>5120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P100</td>
<td>3584</td>
<td>5120</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V100</td>
<td>5120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Memory allocation size (GB)

- 8H HBM 32GB
- 4H HBM 16GB

- 10 layers
- 110 layers
- 210 layers
- 310 layers
- 410 layers

**Deeper Network**
HBM Presence – Some Examples

**NVIDIA**

**Datacenter (Acceleration, AI/ML)**
- Tesla P100, V100
- DGX Station, DGX1, DGX2
- GPU Cloud
- Titan V

**Professional Visualization**
- Quaddro GP100, GV100

**AMD**

**Datacenter (Acceleration, AI/ML)**
- Radeon Instinct MI25
- Project 47

**Professional Visualization**
- Radeon Pro WX, SSG, Vega

**Consumer Graphics**
- Radeon Rx Vega64, Vega56

**Intel**

**Datacenter (Acceleration, AI/ML)**
- Nervana Neural Net Processor
- Stratix10 MX (FPGA)

**Consumer Graphics**
- KabyLake-G

**Google**

**Datacenter (Acceleration, AI/ML)**
- TPU2

**TPU2**: 4 ASICs, 64GB HBM2

**TPU POD**: 4TB HBM2

**Samsung**

**Datacenter (Acceleration, AI/ML)**

**Sources**: Tom’s Hardware, Anandtech, PC World, Trusted Reviews

Al Cities
Healthcare
Retail
Robotics
Autonomous cars

Architecture
Engineering/Construction
Education
Manufacturing
Media & Entertainment

Traffic sign recognition
Image synthesizer
Object classifier
Model conversion

VR content creation
Graphics rendering
Gaming, AR/VR

Cloud TPU for Training & Inference

Thin/light
Extended battery life
HBM2: Market Outlook

- Bandwidth needs of High-Performance Computing/AI, High-end Graphics, and new applications continue to expand

Bandwidth and market for HBM growing rapidly

HBM adoption started with HPC, expanding into other markets

Source: Samsung
AI Inference: GDDR6

- Inference less computationally & memory intensive than AI Training

- GDDR6 is a good option – double the bandwidth of GDDR5
  - Up to 16Gbps per pin → 64GB/s per device

- Samsung is first to market with 16Gb GDDR6

- Nvidia T4 cards
  - 16GB GDDR6
  - AWS G4 Inference
Foundry Services

- Latest process nodes, testing, packaging, design services
- WW partners to complement solutions with IP and EDA tools
Summary

- AI workloads rely on Deep Learning algorithms that are memory bandwidth constrained

- HBM has become the memory of choice for AI training applications in the data center

- GDDR6 provides an “off-the-shelf” alternative for AI inference workloads

Make the smart choice: AI hardware powered by these technologies
Thank You…

Samsung - The Heart of Your Data

March 19-21, 2019
SAN JOSE CONVENTION CENTER - BOOTH 726
Contact: t.shiah@Samsung.com