SAINT-S: 3D SRAM Stacking Solution based on 7nm TSV technology

Kyoungsun Cho, Jinhong Park, Billy Koo, Sunkyong Seo, Yoonjae Hwang, Sungcheol Park and Mijung Noh
Samsung Electronics
1 July 2020
Data movement of emerging devices hit the memory wall. AI devices require higher memory bandwidth and higher memory capacity. AR/VR devices also require the lowest latency possible. To fulfill the memory requirements, Samsung Foundry proposed a three-dimension (3D) static random access memory (SRAM) stacking solution (SAINT-S, Samsung Advanced INterconnection Technology with SRAM). The solution is implemented by Samsung 7LPP process and leverages stacking technology using TSV (through silicon via) to achieve high bandwidth and low latency interface between logic die and SRAM die, and small form factor. Implementation results present read/write memory latency of 7.2/2.6 ns and 24.3 GB/s memory bandwidth with 0.156W average power consumption per channel at 760MHz frequency. The memory bandwidth per power of SAINT-S is 6.2x higher than GDDR6’s and 2.2x higher than HBM2e.
AI devices require higher memory bandwidth and low latency with small form factor. To fulfill the memory requirements, Samsung Foundry proposed a three-dimension (3D) static random access memory (SRAM) stacking solution (SAINT-S, Samsung Advanced INterconnection Technology with SRAM).
Logic die and SRAM die are combined using TSV

Logic Die includes custom 3D SRAM controllers as well as CPU and DMA to verify performance and power

3D SRAM Controller
- Access to SRAM (64Mbits) on SRAM die through TSV
  - 256bit @ 760MHz per channel
- Source synchronous interface and asynchronous FIFO
- Double data rate (DDR) conversion to reduce the number of signals
- Controllable delay lines to compensate the clock to data and data to data skew
Source Synchronous DDR Interface

TSV IO

Low Power Features for TSV IO

<table>
<thead>
<tr>
<th></th>
<th>Current work</th>
<th>Future work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Driver</strong></td>
<td>0.07pj/bit</td>
<td>0.05pj/bit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Eliminating the on-die termination (ODT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Minimizing the # of blocks using VDDQ power</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Using source termination in main driver to improve signal integrity</td>
</tr>
<tr>
<td><strong>Receiver</strong></td>
<td>0.03pj/bit</td>
<td>0.02pj/bit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Un-terminated low voltage swing signaling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Using the only core logic power (VDD) and it is designed to minimize the # of stages</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No level-shifter is needed in receiver path</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• CMOS base receiver scheme instead of single-ended type receiver</td>
</tr>
</tbody>
</table>
Any die on the stack can be a master by MASTER_SEL

MASTER = Die 0  OUTPUT
SRAM control & SRAM DIN signals

MASTER = Die 0  INPUT
SRAM DOUT signals

Die 3

SRAM only active

Die 2

SRAM only active

Die 1

SRAM only active

Die 0

SRAM & Controller are active

SRAM only active

SRAM only active

SRAM only active

SRAM & Controller are active
Chip on wafer (CoW) Bonding
Based on Mass Manufacturing Infra with >1.33 CpK
Passed JEDEC Package Reliability Standard
Package Flexibility; Multi-memory Stacking (≥ 2H)
Memory Side by Side Bonding
Face to Face Bonding
### Budget-based Flow

- **[Constraint Budgeting] Tier1:** set delay constraints on register-to-IO input paths
- **[Constraint Budgeting] TSV:** IO-to-IO Path
- **[Constraint Budgeting] Tier2:** set delay constraints IO-to-register paths
- Add the constraints and run STA
- Run SPICE simulation using TSV paths and check the constraints

### By running STA of each die concurrently, TAT and resource can be optimized as in a conventional design with maintaining SPICE-level accuracy of jitter/DCD in TSVs

*Considering clock skew of both Tier1 and Tier2 at a time*

*Designers have better to assign other tier’s context (clock latency, data delay and TSV delay) at one tier’ in SDC*
PSI Analysis on TSV Interconnects

- **IO decap for Power Noise design guideline at pre-layout stage**
  - Criteria: below 5% of Duty Cycle
  - Insertion guideline: 30nF for enabling 760MHz interface

- **PSI Analysis for 1-stack SRAM at 800MHz**
  - Focusing on only 5-coupled lanes (Total IO: 256ea) means the same effect to consider all SSO noise from other 251 IO lanes

![Diagram showing SI only and 5-coupled with SSO effect](image-url)

**Table:**

<table>
<thead>
<tr>
<th>Decap[nF]</th>
<th>Duty↑, Jitter ↑</th>
<th>Vpp[mV]</th>
<th>DCD[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>5%</td>
</tr>
</tbody>
</table>
Concurrent multi-die analysis flow

- Showing how hot spots have an effect on the other die concurrently
- According to this result, the power meshes of both logic and SRAM die are reinforced

Difference from single chip IREM

- Chip to chip connection
- TSV modeling
3D-IC P&R Challenges

- New TSV-related rules restrict floor-planning and placement work
- The number of TSVs is more than 1000ea
- Efficient TSV signal/power placement architecture is necessary to minimize design overhead

P&R Solutions

- Custom scripts for automated placement & routing
# Chip Summary

<table>
<thead>
<tr>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
</tr>
<tr>
<td>Die Size</td>
</tr>
<tr>
<td>Package Size</td>
</tr>
<tr>
<td>SRAM Capacity</td>
</tr>
<tr>
<td>Supply Voltage</td>
</tr>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td># of channel</td>
</tr>
<tr>
<td>Power consumption</td>
</tr>
<tr>
<td>Bandwidth</td>
</tr>
<tr>
<td>Latency</td>
</tr>
</tbody>
</table>

# Performance Comparison

<table>
<thead>
<tr>
<th>Memory Bandwidth per Power (GB/second/Watt)</th>
<th>Memory Latency (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDDR6</td>
<td>25</td>
</tr>
<tr>
<td>HBM2e</td>
<td>70</td>
</tr>
<tr>
<td>SAINT-S</td>
<td>156.1</td>
</tr>
</tbody>
</table>

- 45 (70% column hit)
- Read 7.2/Write 2.6

---

[Results & Future work]

SAINT-S PKG SEM picture

SAINT-S Test board
Thank You
samsung foundry