A 16nm 256-bit Wide 89.6GByte/s Total Bandwidth In-Package Interconnect with 0.3V Swing and 0.062pJ/bit Power in InFO Package


Taiwan Semiconductor Manufacturing Company, Ltd.
DTP
Taiwan
Outline

● Motivation
  ■ In-Package Interconnect Applications
  ■ Advance Package Solutions

● Introduction
  ■ InFO Process
  ■ Low Power Design Concept (2013-VLSI)

● System Architecture

● Circuit Descriptions

● Experimental Results

● Conclusion
In-Package Interconnect Applications

- **LIPINCON™**: Low-voltage-In-Package-INter-CONnect
  - High performance computing
    - Limit die size for yield
  - In package memory
    - Higher level cache
  - Heterogeneous integration
    - High speed SERDES

- Smaller form factor
- Shorter interconnect trace
High Performance Computing (HPC)

- Yield goes down exponentially as die size gets large
  - Building large dice is quite difficult and very costly

In Package Memory (IPM)

- Additional memory hierarchy between on-chip SRAM and off-chip DDR
  - Smaller capacity; higher bandwidth; faster latency

- Advantages
  - DRAM is more cost-effective than eDRAM
  - Non-TSV allows fine pitch RDL and shorten trace between AP and IPM
  - Short trace enable termination-less IO design

✓ Save main DRAM bandwidth by 60%

Memory Interface Development Trend

- Higher Bandwidth; Lower Power; Lower Cost (TSV-less)?

Advanced Package Solutions

Why we choose FO-WLP?
- Ultra-fine pitch RDL (W/S < 5um/5um)
- Ultra-thin package (~0.6mm including BGA)
- Smaller die size (no-TSV)
- Suit for smartphones, tablets, and wearables

TSMC Package Solution

- **InFO-WLP (INtegrated Fan-Out Wafer-Level-Package)** vs FC-BGA

  - More pin counts with smaller die size
  - Lower parasitic resistance (thicker copper traces)
  - Lower substrate loss (molding compound)
  - Better thermal behavior (smaller form factor)

Project Target

- Demonstrate an in-package interconnect for in-package memory in InFO package
  - Less capable of memory process is considered
- Low-Power ➔ Low-Latency ➔ Small Area
  - Try to be transparent as possible
- 2Gbit/s/pin; 64Gbyte/s total bandwidth
- Power efficient IO
- Prompt and automatic timing-calibration scheme
Introduction

- TSMC InFO side-by-side
  - 560um thickness of packaged chip including 3*RDL & BGA

![Diagram showing TSMC InFO side-by-side packaging with BGA balls, D2P Bump, D2D Bump, InFO Trace, and RDL layers.](image)

Height: ~560um
Introduction

- Low power design concept

Reduced-power TX

- Low swing with lower VDDQ
- Termination-less
- Dynamic power only

[7] JESD8-28
Introduction

- Low power design concept 2013-VLSI by TSMC

Reduced-power RX

- Clock-based sense amplifier
- No balance buffer
- Termination-less
- Dynamic power only

Introduction

- Low power design concept (2013-VLSI by TSMC)

Simplified design in DRAM

- Get rid of DLL/PLL in DRAM

System Architecture

- 64GByte/s
- 256-DQ
- Bi-directional DQ/DQS
- Modular design
- Easily scalable
- Physical balance

![Diagram of System Architecture]

- SOC
  - WRDATA[63:0]
  - RDATA[63:0]
  - DFI_CA[43:0]
  - DFI_CLK
  - Data Align
  - SUB-CA Channel
  - SUB-CAL Channel
  - PLL
  - BIST
  - IIC

- Interconnect
  - DQ[15:0]
  - DMI[1:0]
  - DQS_v/DQS_c
  - CA[10:0]
  - RST_n/CS_n/CKE
  - CK_t/CK_c
  - CAL[9:0]

- MEM
  - P2L_CA[43:0]
  - P2L_CLK
  - P2L_DQ[63:0]
  - L2P_DQ[63:0]
  - P2L_CA[15:14]
  - Slice
  - IIC
  - BIST

- SRAM
  - (Memory Size 524288 = 512 address * 64 bit * 4 cell * 4 ch)

Data Alignments:
- P2L_DQ[63:0]
- L2P_DQ[63:0]

Data Rates:
- 500Mbps/SDR
- 2Gbps/DDR
- 500Mbps/SDR
Circuit Description

- Sub-DQ-channel architecture

- 16-DQ share one differential DQS
- Per byte DMI combines Data_mask (DM) & Data_bus_inversion (DBI)
Circuit Description

- **Sub-DQ-channel architecture**

**WRITE-path**
- DLL-TX [6]

---

Circuit Description

- Sub-DQ-channel architecture

- READ-path
  - DLL-RX [6]
  - RX FIFO ➔ DLL-RPT

- Asynchronous clock between SCLK & RDQS

Circuit Description

- Sub-CAL(calibration)-channel architecture

- 3 DLLs (TX/RX/RPT)
- Prompt and automatic
  - Lock time < 1\text{us}
- 10-pin overhead shared by 64-DQ
  - But scalable

Note: “RPT”=“Read-Preamble-Training”
Latency-cost Degrades System Efficiency

- De-/Serialization ratio: 4
- DBI enable (data-bus-inversion)

(1T=500MHz)

- PLL
- BIST
- IIC

- Data Align
- DFI_CLK
- DFI_CA[43:0]

- Slice
- 1.875T

- SUB-CA Channel
- PLL
- BIST
- IIC

- PLL
- BIST
- IIC

- Slice
- 1.5T

- SUB-CA Channel
- PLL
- BIST
- IIC

- PLL
- BIST
- IIC

- Channel[0]

- P2L_CA[43:0]
- P2L_CLK

- P2L_DQ[63:0]
- L2P_DQ[63:0]

- Memory
- Size
- 524288=
- 512 address*
- 64 bit*
- 4 cell*
- 4 ch

4.75T

1.875T

1.5T

2T

(Memory Size
524288=
512 address*
64 bit*
4 cell*
4 ch)
Temperature Drift Monitoring Scheme

- Temperature drift ➔ Timing variation ➔ Performance degradation
- Leveraging DLL architecture
- Periodic check (1ms) **in background**
- Alert issue if drift exceed the tolerance

Temperature drift ➔ Timing variation ➔ Performance degradation

- Leveraging DLL architecture
- Periodic check (1ms) **in background**
- Alert issue if drift exceed the tolerance
Pad Plan

- Compact pad plan with perfect-match trace length
- RDL routing density: 1*RDL/10um
Boundary Scan Strategy

- **KGD (Known-Good-Die) – Contactless IO**
  - Compatible with IEEE1149.1
  - Each die builds in self TAP controller
  - Support contactless LIPINCON-IO open/short test individually
Boundary Scan Strategy

**KGS** (Known-Good-Stack) – **Interconnect**

- Cascade TAP structure and control in sequence
- Support inter-connectivity test through TAP/Bscan cells
- Capable of per pin diagnosis

![Diagram of Boundary Scan Strategy](image)
Packaged-Die Photo After InFO

- Fan-out to get the required direct-access BGA balls

[BGA ball side] [7mm*7mm]
KGD Shmoo Plot

- SOC/MEM-PHY loopback BIST (Logic VDD vs. Frequency)

- VDDQ=0.3V
- 256-DQ toggle
- DBI enable
- PRBS

Logic VDD Target=0.8V
Freq. Target=2Gbps
Freq. Max=2.8Gbps
@ VDD=0.8V

VDD -10%
Freq +10%

Freq. Max=2.8Gbps
@ VDD=0.8V
+40% Speed
Freq. Target=2Gbps

VDD -10%

Logic VDD Target=0.8V
KGS Shmoo Plot

- SOC-to-MEM Write/Read BIST (Logic VDD vs. Frequency)

- VDDQ=0.3V
- 256-DQ toggle
- DBI enable
- PRBS

Logic VDD Target = 0.8V
Freq. Target=2Gbps
Freq. Max=2.8Gbps
Freq. Max=3.6Gbps @ VDD=1.0V

Same speed as KGD
KGS Shmoo Plot

- SOC-to-MEM Write/Read BIST (IO-VDDQ vs. Frequency)

- VDD=0.8V
- 256-DQ toggle
- DBI enable
- PRBS
Capable of Eye-ploting

- 0.3V Swing is critical to SSO under wide bus application
  - How to ensure signal integrity on the un-probed interconnect IO?

- X-axis: DCDL code (7ps)
- Y-axis: VREF code (9.5mV)
- VDD=0.8V
- 256-DQ toggle
- DBI enable
- PRBS

<table>
<thead>
<tr>
<th>VREF Code</th>
<th>DCDL Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.04V(Noise Ground)</td>
<td>0.15V</td>
</tr>
<tr>
<td>0.3V</td>
<td>0V</td>
</tr>
</tbody>
</table>

Org. Capture

Adjusted Capture

Eye Height (225mV)

Eye Width (420ps~0.84UI)
Power Breakdown

- Base on simulation results with 100% toggling conditions

- LIPINCON-IO power efficiency: **0.062pJ/bit** (consider one single-end IO)

- LIPINCON-PHY power efficiency: **0.424pJ/bit**

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Power Consumption</th>
<th>Power Efficiency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital (0.8V)</td>
<td>110.00</td>
<td>0.215</td>
<td>50.7%</td>
</tr>
<tr>
<td>LIPINCON-IO (0.8V/0.3V)</td>
<td>53.97</td>
<td>0.105</td>
<td>24.9%</td>
</tr>
<tr>
<td>DCDL (0.8V)</td>
<td>31.52</td>
<td>0.062</td>
<td>14.5%</td>
</tr>
<tr>
<td>NPL (0.8V)</td>
<td>20.00</td>
<td>0.039</td>
<td>9.2%</td>
</tr>
<tr>
<td>Replica cell (0.8V)</td>
<td>1.65</td>
<td>0.003</td>
<td>0.8%</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td><strong>217.141</strong></td>
<td><strong>0.424</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Power Consumption</th>
<th>Power Efficiency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Digital (0.8V)</strong></td>
<td><strong>110.00</strong></td>
<td><strong>0.215</strong></td>
<td><strong>50.7%</strong></td>
</tr>
<tr>
<td><strong>LIPINCON-IO (0.8V/0.3V)</strong></td>
<td><strong>53.97</strong></td>
<td><strong>0.105</strong></td>
<td><strong>24.9%</strong></td>
</tr>
<tr>
<td><strong>DCDL (0.8V)</strong></td>
<td><strong>31.52</strong></td>
<td><strong>0.062</strong></td>
<td><strong>14.5%</strong></td>
</tr>
<tr>
<td><strong>NPL (0.8V)</strong></td>
<td><strong>20.00</strong></td>
<td><strong>0.039</strong></td>
<td><strong>9.2%</strong></td>
</tr>
<tr>
<td><strong>Replica cell (0.8V)</strong></td>
<td><strong>1.65</strong></td>
<td><strong>0.003</strong></td>
<td><strong>0.8%</strong></td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td><strong>217.141</strong></td>
<td><strong>0.424</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>

- LIPINCON-IO power efficiency: **0.062pJ/bit** (consider one single-end IO)
- LIPINCON-PHY power efficiency: **0.424pJ/bit**
Conclusion

- An in-package interconnect for in-package memory application in InFO package has been demonstrated
  - Technology: TSMC 16FF + InFO

- 89.6GByte/s total bandwidth is achieved with 256-DQ operating in 2.8Gbit/s and 0.3V-swing
  - Low power: IO (0.062pJ/bit); PHY (0.424pJ/bit)
  - Low latency: Write (4.75T+1.5T=6.25T); Read (2+1.875=3.875T)

- 0.3V signal integrity on the un-probed IO has been clarified
  - 420ps (0.84UI) Eye width; 225mV (75%) Eye height

- Prompt and automatic timing-calibration scheme
Acknowledgement

The authors would like to thank TSMC “InFO-IP” team for InFO back-end support, TSMC “IPPM” team for measurement support, and “System-BD” team for business help. The authors would also like to specially thank “Mentor Graphics” for boundary scan design collaboration.
Reference

- [6] TSMC, 2013-VLSI, “An extra low-power 1Tbit/s bandwidth PLL/DLL-less eDRAM PHY using 0.3V low-swing IO for 2.5D CoWoS application”
- [7] JESD8-28 “300mV interface”