Circuit and Package Co-design for 3D Integration (Invited)

Tadahiro Kuroda

Univ. of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
Phone: +81-3-5841-6561

Abstract

This paper presents a ThruChip Interface using inductive coupling for data communications, and Highly Doped Silicon Via using deep well for power delivery. Circuit and package co-design, applications, and future directions are discussed.

I. Introduction

IC was invented against the backdrop of a connection problem - the challenge of wiring - in large-scale systems. Since then, the IC underwent exponential improvement following Moore's Law, with computer performance increasing dramatically in lockstep. However, due to large data movement between memory and processor, inter-chip communication has become the main reason for the erosion of energy efficiency. This is known as the von Neumann bottleneck.

With the combined effect of data explosion, the industry has fallen into a situation where there is "no computing performance improvement without improvement in energy efficiency."

In order to increase the energy efficiency of computing, we need to shorten the distance between memory and processor, as well as increasing the number of connections to avoid pushing up data rates too aggressively. In other words, we should stack chips in 3D to minimize inter-chip distances and to allow the entire chip surface to be used for interconnections to enable a moderate data rate. This is the reason why we are transitioning from 2D to 3D chip integration.

Moving away from relying solely on on-chip integration, we have been evolving from 2D to 3D chip integration, which calls for a breakthrough solution to the connection problem.

Through-Silicon Via (TSV) has been investigated. Although many die stacking approaches rely on TSV as the fundamental means for 3D integration, TSV is not generally available. It still has reliability/yield issues and its cost adder is limiting acceptance of 3D stacking. Replacing the mechanical approach, a low-cost electrical solution is developed, namely ThruChip Interface (TCI) for data communications, which uses inductive coupling, and a Highly Doped Silicon Via (HDSV) for power delivery, which uses deeper and highly doped well.

II. Circuit Design

1) ThruChip Interface (TCI)

TCI [1] is depicted in Fig.1. A coil is formed by using multi-layer standard wires and vias, so that digital wires can go across it. As the coil should not resonate, narrow wires and small vias can be used for low *Q* factor. Data communication uses baseband, not carrier, as the conventional digital on-chip communication. A transceiver is implemented by digital CMOS circuits of as small as 36 2NAND gates and it scales down by device miniaturization. TCI is a digital CMOS circuit solution, and hence eventually zero additional cost.

Comparison with TSV is summarized in Fig.2. TCI is cheaper than TSV while bearing comparison in performance.

TCI performs high speed (30 Gb/s/ch), low energy (0.02 pJ/b)[2], high integration (128-die stacking)[3], and low stacking height (bump less). Data rate can be raised by increasing number of channels while keeping energy efficiency. The energy is lowered significantly, because an ESD protection device is eliminated.

To further improve area and energy efficiency, circuit techniques have been developed. The coils formed by using multi-layer wires and vias allow overlapping. Crosstalk can be suppressed by phase division multiple access [4]. Using dual Tx coils eliminates PMOS, resulting in 0.55 V operation and 0.01 pJ/b.

2) Highly Doped Silicon Via (HDSV)

Power can be delivered by conventional means (wire bond, TSV) or a new way with HDSV (Fig. 3) [1]. A deeper than normal and more highly doped well is used to make a low resistance HDSV pathway directly through a thinned wafer using the silicon itself. Highly doped regions for power vias are first created by implants, followed by nominal process for transistors and wires, and metal caps are added on the HDSV. Wafer is then thinned to ~4 μm. The HDSV on one die and the electrodes on the next die are connected by pressure using a room-temperature wafer bonding machine to create larger stacks. It is reported that DRAM chip substrate was thinned to 4 μm and yet no degradation of retention characteristics was found.

TCAD simulations indicate the front-to-back resistance can be made lower than 3 mΩ when substrate thickness is below 5 μm and the HDSV net area is 0.7 mm^2, under conditions of $1x10^{16}$ cm^{-2} dose, 200 keV ion implantation, and 50 hour annealing at 1050°C. The HDSV can be divided into small regions and distributed. As large area is required to reduce resistance, the HDSV is not usable for high speed data, but TCI should be used instead. The HDSV should be low cost as it is made by implants.

III. Package Design

An Ultra-Thin Fan-Out Wafer Level package [4] enables package on package (Fig.4), which allows using the conventional supply chain. Power can be delivered by Thru Mold Via (TMV), and Re-Distribution Layer (RDL) which is fan out from the edge of a chip and connected to the center located pads for power and ground. Known-good-dies are placed to reform a wafer. The silicon can be thinned to 40μm to make the TCI area efficient. This wafer level package technology is available for mass production.

In future, Wafer-on-Wafer bonding will work for homogeneous stacking for memory devices, while Chip-on-Wafer or Chip-on-Chip bonding will be required for heterogeneous stacking for system integration.

IV. Applications

TCI can be used for applications where TSV is expected for use. It can also be applied to contactless memory, contactless wafer testing, and bus proving through a package for debugging.

1) NAND stacking for SSD

One-package SSD is made possible with TCI [3]. I/O power dissipation is reduced to 1/5. TCI consumes constant I/O energy and delay regardless of number of IOs that are connected, enabling low energy broadcasting.

2) DRAM/SoC interface

A 352Gbps DRAM/SoC interface was developed [4]. The overlapping coils with Quadrature Phase Division Multiplexing (QPDM) are employed. It outperforms WIO2 (TSV) in cost and LPDDR4 in power dissipation and latency.

3) 3D Processor

A 3D processor can adjust its performance by changing number of dies in the stacking. Linux OS was installed, and several applications were performed and demonstrated.

4) AI Accelerator

A deep neural network (DNN) inference engine was stacked with multi-vault SRAMs using TCI (Fig.5) [5]. Energy consumption for memory access was reduce to half compared to the case with DRAM.

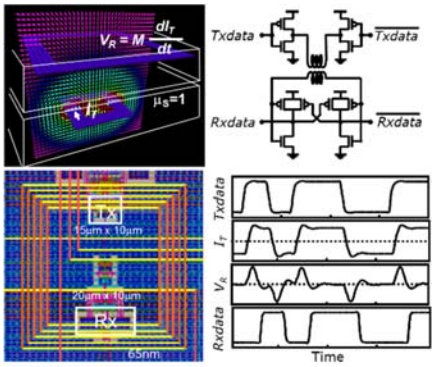


Fig. 1 ThruChip Interface (TCI).

	TSV	TCI
Connection by	Wire	Wireless
Area/channel	TX/RX: 50x 2NAND TSV: 50μm x 50μm (area below/above TSV cannot be utilized for circuits)	TX/RX: 36x 2NAND Coil: 1μm x 250μm x 12 (area below/above TCI can be utilized for circuits)
# of channel	>250 ch	<50 ch
Data rate	<1Gb/s/ch	>5Gb/s/ch
Delay	40x 2NAND FO4	7x 2NAND FO4
Energy	40x 2NAND + 0.5pJ/b	80x 2NAND
Manufacturing	Additional Steps Req.	Standard CMOS Process
Additional Cost	>40%	<0.1%
Supply Chain	OSAT involved	Conv. business model

Fig. 2 TCI vs. TSV.

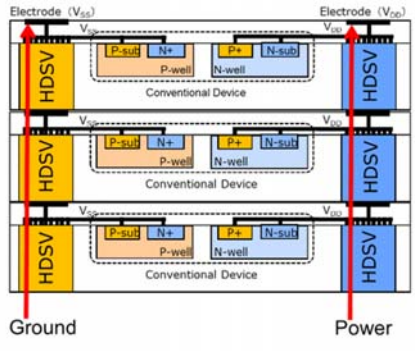


Fig.3 Highly Doped Silicon Via (HDSV).

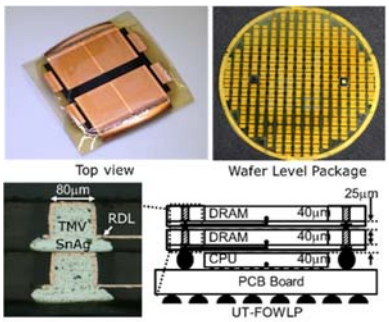


Fig. 4 Ultra-Thin Fan-Out Wafer Level Package.

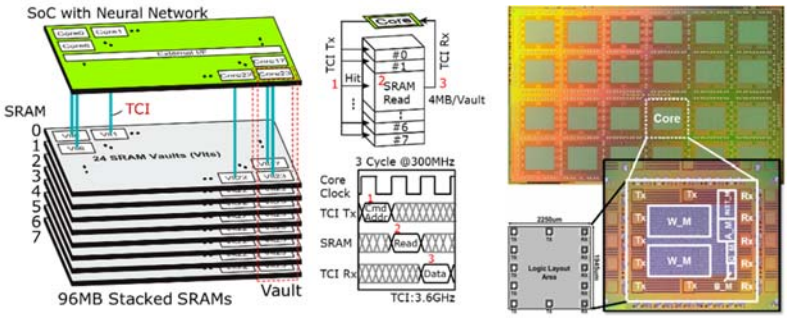
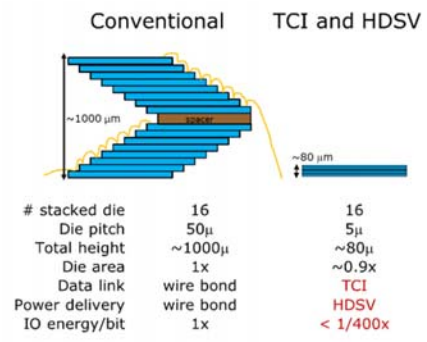
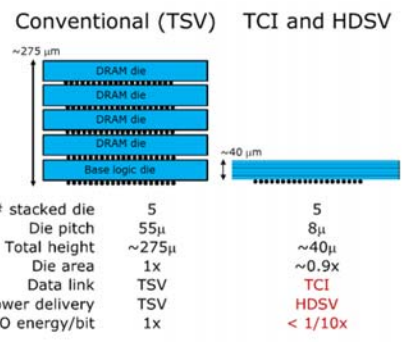


Fig. 5 DNN Inference Engine with Stacked SRAM.



	Conventional	TCI and HDSV
# stacked die	16	16
Die pitch	50μ	5μ
Total height	~1000μ	~80μ
Die area	1x	~0.9x
Data link	wire bond	TCI
Power delivery	wire bond	HDSV
IO energy/bit	1x	< 1/400x

Fig. 6 NAND stacking with TCI & HDSV.



	Conventional (TSV)	TCI and HDSV
# stacked die	5	5
Die pitch	55μ	8μ
Total height	~275μ	~40μ
Die area	1x	~0.9x
Data link	TSV	TCI
Power delivery	TSV	HDSV
IO energy/bit	1x	< 1/10x

Fig. 7 DRAM stacking with TCI & HDSV.

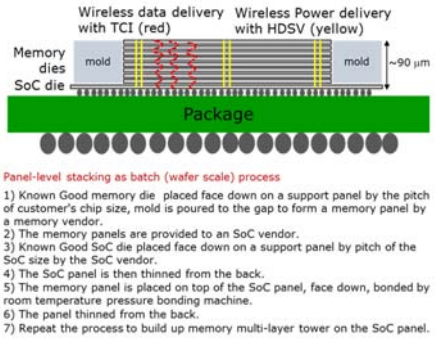


Fig. 8 Ultra-thin lowest cost 3D packaging.

V. Future Directions

1) Constant Magnetic Field Scaling Scenario

Just like the constant electric field scaling scenario in scaling a Field Effect Transistor, a constant magnetic field scaling scenario will improve cost performance of the TCI significantly [6]. Thinning wafers/chips is the future directions of competition.

2) System in a Package

With the TCI and the HDSV, IO energy per bit will be reduced to 1/400 in NAND stacking (Fig. 6), and 1/10 in DRAM stacking (Fig. 7) [7]. Panel-level stacking as batch (wafer scale) process is a future scenario (Fig. 8).

3) Academic and Industrial Collaboration

A Systems Design Lab (d.lab) was launched at the Univ. Tokyo for academic collaboration [8], and a Research Association for Advanced Systems (RaaS) was established as industrial consortium [9].

VI. Conclusion

The connection problem for 3D integration can be solved by electrical connection using inductive coupling. One of the future directions of circuit and package interactions is found in thinning chips, stacking, and connecting them by electrical connection.

Acknowledgement

The author is grateful to K. Johguchi for support.

References

[1] "ThruChip Interface (TCI) for 3D Integration of Low-Power System (Invited)," *IEDM*, p.17.1.1, 2010.

[2] "A 0.7V 20fJ/bit Inductive-Coupling Data Link with Dual-Coil Transmission Scheme," *Symp. VLSI Circuits*, pp.201-202, 2010.

[3] "A 2Gb/s 1.8pJ/b/chip Inductive-Coupling Through-Chip Bus for 128-Die NAND-Flash Memory ...," *ISSCC*, pp.440-441, 2010.

[4] "A 352 Gb/s Inductive-Coupling DRAM/SoC Interfaces Using Overlapping Coils with Phase Division Multiplexing and Ultra-Thin Fan-Out Wafer Level Package," *Symp. VLSI Circuits*, pp.C44-45, 2014.

[5] "QUEST: A 7.49TOPS Multi-Purpose Log-Quantized DNN Inference Engine Stacked on 96MB 3D SRAM Using Inductive-Coupling Technology in 40nm CMOS," *ISSCC,* pp. 216-217, 2018.

[6] "Constant Magnetic Field Scaling in Inductive-Coupling Data Link," *SSDM*, pp. 606–607, 2006.

[7] "Low-Cost 3D Chip Stacking with ThruChip Wireless Connections," *Hot Chips*, 2014.

[8] d.lab homepage http://www.dlab.t.u-tokyo.ac.jp/

[9] RaaS homepage https://raas-cip.org