A 0.296pJ/bit 17.9Tb/s/mm² Die-to-Die Link in 5nm/6nm FinFET on a 9µm-pitch 3D Package Achieving 10.24Tb/s Bandwidth at 16Gb/s PAM-4

Mu-Shan Lin, Chien-Chun Tsai, Shenggao Li, Tze-Chiang Huang, Wen-Hung Huang, Kate Huang, Yu-Chi Chen, Alex Liu, Yu-Jie Huang, Jimmy Wang, Shu-Chun Yang, Nai-Chen Cheng, Chao-Chieh Li, Hsin-Hung Kuo, Wei-Chih Chen, C.H. Wen, Kevin Lin, Po-Yi Huang, Kenny Cheng-Hsiang Hsieh, Frank Lee

Taiwan Semiconductor Manufacturing Company

Abstract

This paper presents a die-to-die link with a compute die in 5nm FinFET and a SRAM die in 6nm FinFET, with a face-to-back 3D stacking at a 9µm bond pitch. Modular design that supports full scalability is demonstrated, achieving a 10.24Tb/s aggregate bandwidth for 320 Tx lanes and 320 Rx lanes, at a PAM-4 16Gb/s per lane data rate. Each data cluster is designed with 80 Tx/Rx lanes in a 378µm*378µm footprint, achieving a bandwidth density of 17.9Tb/s/mm² and an energy efficiency of 0.296pJ/bit per link.

Introduction

2.5D/3D advanced packaging technology has undergone rapid growth in recent years. 3D stacking achieves the highest interconnect density with unprecedented bandwidth in smaller form-factor and better signal integrity, and it leads to future monolithic 3D integration. A higher number of channels with lower speed per pin can achieve the same bandwidth as a few channels with higher speed per pin, but with better energy efficiency, lower area and latency. Due to 3D package technology, the bump pitch is reduced from a 40~55µm in 2.5D to 9µm or below in 3D [1] [5] [6]. It is important to take advantage of the dense and smaller bump footprint and push up the data rate per pin, so that the bandwidth for die-to-die communication can be advanced without introducing system overhead.

Design and Implementation

As shown in Fig. 1, considering scalability, a 3D die-to-die (D2D) interconnect is designed as a small granularity (Data-cluster) as it can be deployed alone with the orthogonal track, serving for CPU-SRAM clusters cross-die computing as needed. The intersection is designed with a Common-cluster for local clock source/calibration circuits. The scalability across different stacking variants is also considered in Fig. 2, including face-to-face 2-tier stacking between one logic layer and one memory layer, and face-to-back 3-tier stacking between one logic layer and two memory layers or two logic layers and one memory layer. The signal bonds are designed symmetrically from chiplet perspective to ensure the transmit pin connect to the receive pin. The D2D high-speed signals must be driven through a front-side bond to another front-side bond, or a front-side bond cascading with multiple through-silicon vias (TSV) and bonds across several chiplets.

To push the speed envelope while keeping the design compact enough to fit into a 9µm pitch, PAM-4 signaling was adopted in this design. Fig. 3 illustrates the insertion-loss and crosstalk for signals through front-side bond or TSV. To achieve good transmit linearity on RLM (Ratio of Level Mismatch) for PAM-4 signaling, series resistors are used in each unit cell. The calibration on the driver MOSFET is done at the gate side to ensure a compact driver size without cascode device. The PAM-4 Rx is composed of 6 SAFFs (StrongARM sense-amplifier flip-flop) to detect 4-levels double-data-rate signals. The SAFF is designed with a wide input common-mode range by in paralleling with P-type/N-type amplifying stages to accommodate high/middle/low reference voltage individually. The worst case driver loading targets as 2*TSV + 2*bonds and 2*Tx + 1*Rx in a 3-tier stacking (i.e. two chiplets access one chiplet).

The D2D interconnect is a parallel bus with forwarded clock, as shown in Fig. 4. To ensure modular design and system scalability, the common block (CB) is consolidated with clock generator, high/middle/low VREF generator, and the PAM-4 driver calibration circuit. All other blocks for data communication are included in the data channel block (DCH). Each DCH is composed of four sub-channels, with 20 Tx lanes/20 Rx lanes in each sub-channel. Each DCH (80 Tx lanes/80 Rx lanes) shares a forwarded clock (CKP/CKN) and a Valid (VLD) signal. Two DLLs (delay-lock-loops) are implemented. The DLL-R90 for the Rx can track any process-voltage-temperature (PVT) variation and provide the 90-degree phase to sample the data center at the Rx SAFF. The DLL-Deskew in Tx is used to synchronize the data transfer between the SoC to individual DCH, providing tracking capability of low-frequency phase drift as opposed to adopting a TXFIFO with hardware/power and latency penalty. A RXFIFO is designed to accommodate the long-term temperature/voltage/aging drift, due to clock domain crossing between chiplets. The use of PAM-4 signaling can benefit on simplified D2D interconnect, resulting in reduced latency of 1.875ns, which includes Tx, Rx, and RXFIFO, during 4GHz SoC operation.

Since 3D D2D stacking are connected through 9µm-pitch dense regularly bonds, defect detection and repairment become critical. To improve yield without much circuit overhead, a ‘Shift and Switching Repair’ concept is proposed in this design [2]. The "MUX-switch" to the one redundant lane can repair not only the 10 data lanes within the same cluster but also 10 data lanes from others. Accompanied by the "Shift" circuits, it provides multiple defect repair if multiple lanes fail in the same cluster.

The modular design concept is also implemented in physical perspective. As shown in Fig. 5, each DCH is designed in square-shape to enable dual-orientation expansion. Four sub-channels are in the middle, with a 9µm bond pitch and very dense D2D interconnects. The top and left sides have Global bus implementation, which includes a 4GHz clock tree from the PLL and high/middle/low reference voltage to Rx SAFFs from the VREF-generator. The right and bottom sides have an L-shaped synthesis logic to accommodate both E/W (east/west) and N/S (north/south) data traffic from the SoC. The common block (CB) is also in square-shape, and can be shared by multiple DCHs as bandwidth scales.

Measurement Results

Fig. 6 depicts the measurement results, where the shmoo plot is obtained with different workloads of 25%, 50%, 75%, and 100% active lanes. The plot shows that a higher workload requires a higher supply to achieve the same data rate, indicating a limitation on the power delivery network of the probe card. The 100% workload with all sub-channels active can achieve a data rate of 16.9Gb/s at 0.75V. A maximum speed of 24.6Gb/s is achieved under a 25% workload when the supply is raised to 0.9V. The amplitude and timing margin of D2D link are important since the link has a wide bus of 16Gb/s PAM-4 signals across chiplets through TSV/ bond. By properly selecting a programmable High/Middle/Low pattern through the built-in self-test circuit, the high/middle/low eyes can be scanned separately and combined to

emulate a real PAM-4 eye diagram. The results are depicted as an eye contour with BERs of 10-6, 10-9, and 10-12. Under BER 10-12, the eye width is 34%/33%/39% of 125ps and 37%/42%/23% of 250mV.

Fig. 7 shows the chip photo after 3D-stack packaging, along with layout views. Two chips, with a size of 9mm by 9mm, occupy the same footprint in a face-to-back stacking. The sub-channel layout breakdown is provided, and each IO-lane occupies only a 9µm by 27µm area (3 bond pitches), which includes the TSV keep-out, PAM-4 driver/receiver, clock tree, and buffer logic.

Summary

This test vehicle demonstrated a scalable solution for die-to-die link in 3DIC. Two sets of 2*DCH+1*CB are symmetrically implemented on each corner of the chip for yield learning, achieving a 10.24Tb/s aggregate bandwidth for 320 Tx lanes and 320 Rx lanes, at a PAM-4 16Gb/s per lane data rate. Each DCH is designed with 80 Tx lanes and 80 Rx lanes in a 378µm*378µm footprint, achieving a bandwidth density of 17.9Tb/s/mm^2 and an energy efficiency of 0.296pJ/bit per link. Fig. 8 provides a comparison table with recent publications.

References

[1] M-S Lin et al., VLSI Symp. 2019
[2] I. Lee et al., Trans Reliab. 2019
[3] P. Vivet et al., ISSCC 2020
[4] U. Rathore et al., ISSCC 2022
[5] K. Chae et al., ISSCC 2023
[6] K. Seong et al., ISSCC 2023

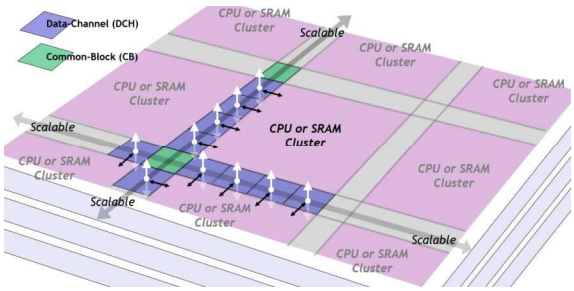

Fig. 1 Scalable D2D architecture in 3D-stacking.

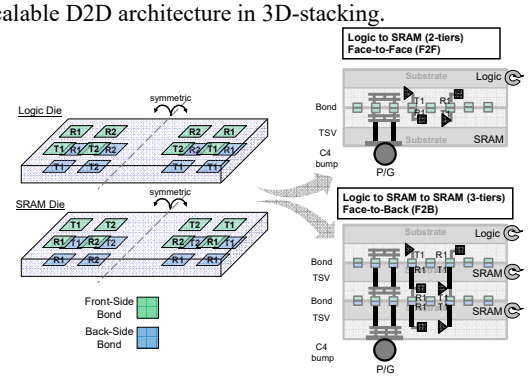

Fig. 2 Symmetric bond assignment for 2-tiers F2F/3-tiers F2B stacking.

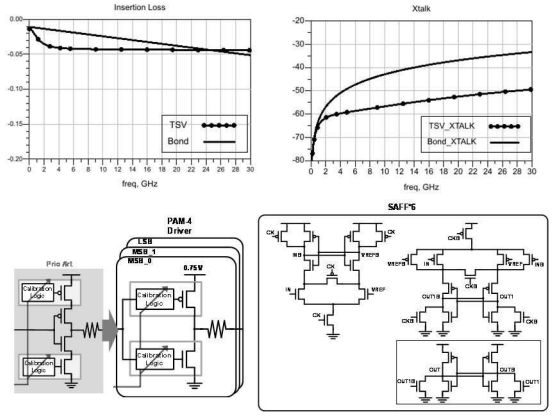

Fig. 3 (a) Insertion-loss/crosstalk of TSV/Bond (b) Die-to-die PAM-4 driver and Receiver SAFF.

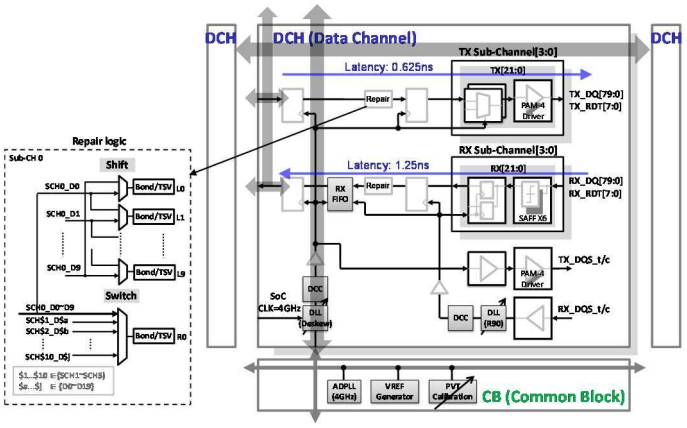

Fig. 4 Die-to-die interconnect architecture and the repaired logic.

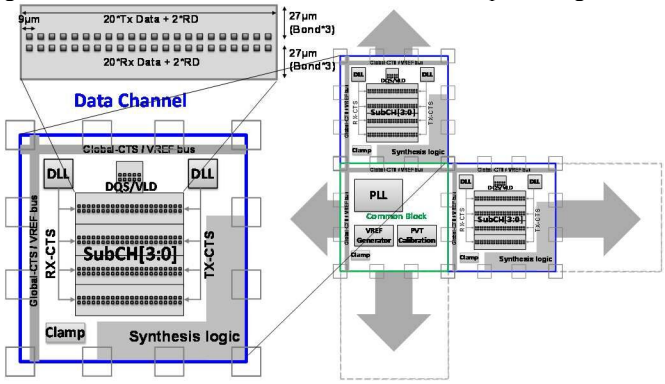

Fig. 5 Data channel floor-plan and dual-orientation abutment.

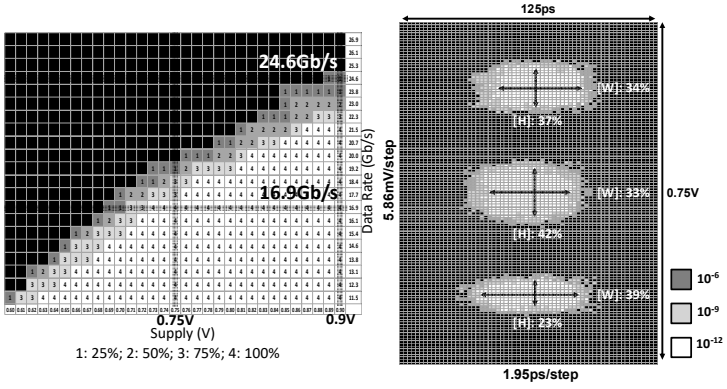

Fig. 6 Measured shmoo and PAM-4 eye contour plot.

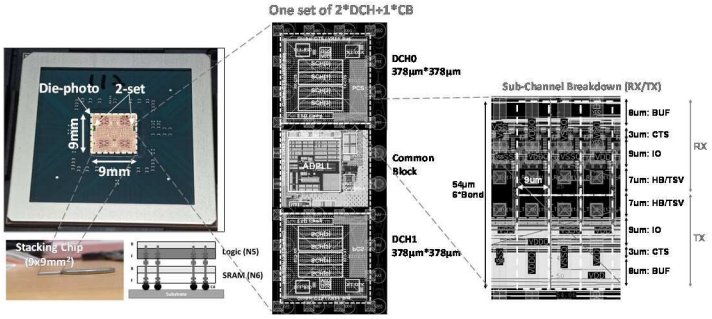

Fig. 7 Chip photo after 3D-stack package and the layout view.

Reference	VLSI'19 [1]	ISSCC'20 [3]	ISSCC'22 [4]	ISSCC'23 [5]	ISSCC'23 [6]	This Work
Technology	7nm FinFET	28nm FDSOI	16nm	4nm FinFET	4nm FinFET	6nm/5nm FinFET
Signaling	0.3V NRZ	1.2V	NRZ	NRZ	NRZ	0.75V PAM4
Channel	CoWoS 500µm	Active interposer 50µm (3D)	350µm	Interposer 6mm	Interposer 3mm	2*TSV+2*Bond
Die-to-Die Bump/Bond Pitch (µm)	40	20	10	55	50	9
Data Rate (Gbit/s/pin)	8	1.21	1.1	9	32	16
Energy Efficiency (pJ/bit)	0.56	0.59	0.38	0.3	0.44	0.296 (DCH) 0.136 (IO)
Bandwidth Density (Tbit/s/mm^2)	1.6	3.0	8.0	2.0	1.8	17.9 (DCH) 65.8 (IO-region)

Fig. 8 Chiplets interface performance comparison.