A high-density logic-on-logic 3DIC design using face-to-face hybrid wafer-bonding on 12nm FinFET process

S. Sinha, S. Hung, D. Fisher†, X. Xu, C. Chao, P. Chandupatla, F. Frederick, H. Perry, D. Smith†, A. Cestero‡, J. Safran‡, V. Ayyavu, M. Bhargava, R. Mathur, D. Prasad, R. Katz‡, A. Kinsbruner‡, J. Garant‡, J. Lubguban‡, S. Knickerbocker‡, V. Soler‡, B. Cline, R. Christy, T. McLaurin, N. Robson‡, D. Berger‡
Arm Inc., 5707 Southwest Parkway, Austin, TX, 78735
†GLOBALFOUNDRIES, Malta, NY 12020 USA. ‡GLOBALFOUNDRIES, Hopewell Junction, NY 12533, USA.
Email: *saurabh.sinha@arm.com / daniel.fisher@globalfoundries.com*

***Abstract*—A high-density-3D test-vehicle showcasing a synchronous cache coherent mesh interconnect design (Arm Neoverse® CMN-600) operational at frequencies up to 2.4 GHz and partitioned in 3D using 5.76µm pitch face-to-face wafer-bond 3D connections on a 12nm FinFET process is presented. The test-vehicle is designed using an industry tool compatible innovative physical implementation flow and serves as the first known industry demonstration of the IEEE 1838 3DIC Design-for-Test (DFT) standard. We demonstrate a 3D aggregate bandwidth of 307 GB/s, a record bandwidth density of 3.4 TB/s/mm^2, and an energy efficiency of 0.02 pJ/bit for the 3D-stacked dies. We present measurement and analysis data from 945 dies where a total of 13.5 million signal 3D wafer-bond nets and 20 million power-delivery 3D wafer-bond nets on multiple wafer-bonded pairs are tested showing robust functionality, paving the path for 3D-stacked high performance logic-on-logic applications.**

I. INTRODUCTION

3D stacking technologies have the potential to augment Moore's Law scaling through improved bandwidth, lower interconnect latency and power consumption, reduced cost through assembly of "Known-Good-Die" and better cost efficiency due to heterogeneous integration. High-density 3D stacking technologies with sub-10µm wafer-bond pitch expand the System-on-Chip (SoC) design space and enable partitioning across three dimensions and cross-3D-tier timing optimization. This work uses the 3D design space to improve system-level interconnects and illustrates the strong collaboration between design, foundry and Electronic Design Automation (EDA) that is required to demonstrate performance and power improvement with 3D stacking (Fig. 1).

II. 3D HYBRID WAFER BONDING AND TEST

Fig. 2 shows the 3D face-to-face hybrid wafer bonding and test flow. Prior to bonding, wafers are tested using top-metal layer pads to sort and pair wafers with similar device performance. Hybrid bonding terminals are fabricated at 5.76µm pitch and bonded in a face-to-face configuration [1]. One of the wafers is thinned to reveal the Through-Silicon Vias (TSVs) required for I/O or test signals and power delivery to the 3D chip and pads are fabricated on the revealed TSVs for C4 bumps. At this stage, post-bond wafer-level tests of the entire 3D design are conducted before dicing and package assembly.

III. 3D DESIGN

The Neoverse® CMN-600 product family is Arm's second-generation, highly configurable, mesh-based coherent interconnect based on the AMBA® CHI (Coherent Hub Interface) cache coherent protocol specification [2] (Fig. 3(a)). The coherent interconnect is a vital component to enable many-core systems to scale without compromising latency and available memory bandwidth. In this 3D test-vehicle a 2x2 CMN-600 interconnect mesh is implemented in 3D, wherein 2 mesh routers (the "XP" blocks in Fig 3(b)) are located in each 3D tier, as shown in Fig 3(b). A 3D mesh can be superior to a 2D mesh since it increases the total number of links in the system and reduces the average number of hops between mesh router points, as shown in Fig. 3(c), which can improve the overall throughput of the system [3].

Fig. 4 shows the logic block diagram of the functional components in the test-chip. The Register-Transfer Level (RTL) functional units are partitioned according to the 3D layers. The test-vehicle consists of 4 CMN-600 cache coherent mesh routers as well as multiple 2D and 3D test-structures to monitor the 3D process and yield. A key feature of this test-vehicle is that it is a *completely synchronous 3D design with a single clock domain*. Additionally, no synchronizing circuits are used at the 3D interface between the two tiers. Fig. 5 describes an industry standard, EDA-tool-compliant physical design flow which includes a 'multi-tier co-placement' step that allows us to co-optimize the location of gates and 3D connections across both 3D layers simultaneously, enabling cross-tier timing closure in a single design database [4].

All measurements presented in this paper were done using IEEE 1838 3DIC design-for-test (DFT) standard [5] compliant implementation (Fig. 6). Primary test access ports (PTAP) and associated controllers on each tier enable DFT scan tests of each layer independently. Additionally, the die-wrapper registers enable cross-tier scan tests of the interface. To the best of the authors' knowledge, this test-vehicle is the first known implementation of the IEEE 1838 3DIC DFT standard.

Fig. 7 shows the GDS view of the final 3D design. The I/O gates are connected to both the top metal layer pads for pre-bond 2D tests as well as the C4 bumps through TSVs, for post-bond 3D test. Fig 7(e) shows the 3D cross-section image from the test-vehicle with 3D wafer-bond pads visible from both 3D-stacked dies. Table I describes the key metrics for the 3D stack design. A total of 13.5 million 3D signal-nets were tested on multiple wafer-bond pairs.

IV. Results and Discussion

Pre-bond testing using the top metal layer pads allows us to sort wafers based on device performance and match wafers for bonding as shown in Fig. 8. Fig. 9 shows correlation between post-bond 3D stacked wafer probe measurements and post-diced packaged parts. In order to test the yield of 3D connections across the wafer bonding sites, various 3D ring oscillators (ROs) are implemented on the chip. Fig. 10 shows MUX-based ROs that enable us to locate failing wafer-bond pads on the die as well as 'chain ROs' that test multiple wafer-bond pads between gate stages. No failures are observed during the RO testing of 945 dies, measuring a cumulative 10 million 3D signals over multiple wafer-bond pairs.

3D gate-to-gate delay is measured using 3D ROs where every subsequent stage alternates between the two tiers (Fig. 11). It is observed that 3D fan-out-of-1 (FO1) delay can be as low as 2D fan-out-of-3 (FO3) delay (<20ps). At high supply voltages the delay reduces to FO2-equivalent delay. Simulations show that optimized gate sizing can enable lower delays (~10ps or less) at nominal supply voltages. Fig. 12 shows the distribution of measured delay of 2D and 3D ROs from multiple wafers and their correlation with simulation. 3D delay is well matched with simulation, giving confidence in the 3D-ready process-design-kit (PDK) used for timing simulations.

Fig. 13 plots the distribution of 2D FO1 gate-delay from the top and bottom die from multiple wafers. The delay skew (absolute difference in RO stage-delay between top and bottom die on the same die location) distribution has a μ =-1.34% and σ =2%, which demonstrates that the process skew between top and bottom tier is well controlled. In Fig. 14 we utilize the skew data to show that a small supply voltage difference (~10mV) can eliminate the measured skew between the two 3D layers. These results highlight the feasibility of a single clock domain design in 3D.

In a 2D mesh interconnect, cross-point routers (XPs) are located about 1 or 2mm apart to account for the placement of the CPUs and system-level caches (SLC) that attach to the XPs. For a 3D-mesh interconnect, the XPs are co-located in the same X-Y location but on different 3D tiers. Fig. 15 shows the energy and delay difference between 2 adjacent XP links in 2D versus 3D. 3D enables orders of magnitude improvement in both energy and delay when you compare short 3D vertical links to 1-2mm 2D links.

The 2x2 3D mesh interconnect is implemented with a single clock source as shown in Fig. 4. Fig. 16 shows the results of static-timing-analysis (STA) before sign-off using the physical implementation flow described in Fig. 5. Fig. 16 plots the number of paths versus timing slack for a target frequency of 1.5GHz (a clock period of 667 ps). As seen, most timing critical paths are 2D (blue) with a few 3D critical paths (red) since XP-XP connections in 3D are vertical with minimized wire-routing. The timing closure includes standard margins used for on-chip device and wire variation in a 2D design but *no special margins are added for the 3D tier-crossings*. Fig. 17 is a Shmoo plot that shows the design operational at 1.58 GHz at 0.8V supply voltage (V_{DD}) and 2.3 GHz at 1V V_{DD}, validating the 3D design and sign-off methodology. Fig. 18 plots a distribution of maximum measured frequency for the 3D mesh at 0.8V and 1V V_{DD} from multiple wafer-bonded pairs. All measured samples show operational frequency exceeding the 1.5GHz design point. At 1V V_{DD}, most measured samples operate at 2.4 GHz or higher. Transition delay ATPG (Automatic Test Pattern Generation) patterns and at-speed MBIST (Memory Built-In Self Test) patterns are used as stimulus when measuring the operational frequency of the die.

This work demonstrates that a key value of hybrid bonding 3D technology is the ability to communicate across two dies without requiring special circuits such as PHYs that are needed for advanced packaging technologies that use bumps at the interface. We further demonstrate that it is possible to design a synchronous 3D design with robust process control and careful pre-bond matching of wafers. This enables 3D stacked designs to implement the on-chip interconnect bus across dies with simple logic gates as drivers, resulting in an order-of-magnitude lower die-to-die latency as shown in Fig. 19. Table II presents a comparison of state-of-the-art 2.5D and 3D packaging technologies with figures of merit (FOM) metrics such as bandwidth, bandwidth density and energy-per-bit. Hybrid bonded 3D stacking achieves an order-of-magnitude higher bandwidth density and lower energy-per-bit owing to a purely digital synchronous die-to-die interface. Designs with cross-tier connections exceeding the 3D stacking pitch or with large die-to-die variations will likely require synchronization circuits at the interface.

V. Conclusions

In this work a 3D test-vehicle built using a high-density face-to-face wafer-bonding technology at 5.76μm 3D pitch with 12nm FinFET devices is demonstrated. The featured cache coherent interconnect mesh is tested to be operational up to 2.4 GHz and achieves 10X lower bandwidth density and energy-per-bit compared to state-of-the-art 2.5D/3D bump-based technologies. Key contributions include an industry EDA-tool-compatible 3D physical design flow, the first known demonstration of IEEE 1838 3D DFT standard, and comprehensive device and 3D cross-tier delay characterization with a cumulative test of 13.5 million 3D connections across multiple wafer-bond pairs. This work is an essential step towards proving the viability of 3D technologies for next generation high-performance and energy efficient designs.

References

[1] D. Fisher, et. al., ECTC 2020, pp. 595., 2020.
[2] A. Pellegrini et al., IEEE Micro, vol. 40, no. 2, pp. 53-62, 2020.
[3] B. S. Feero et. al., IEEE Trans. on Computers, vol. 58, no. 1, pp. 32-45.
[4] X. Xu, et. al., ISLPED 2019, pp. 1-6, 2019.
[5] IEEE Std 1838-2019 , vol., no., pp.1-73, 13 March 2020.
[6] https://www.semiconductor-digest.com/2019/09/16/intel-updates-advanced-packaging-technologies-at-semicon-west-part-2/
[7] M. Lin et al.,VLSI Circuits 2019, pp. C28-C29, 2019.
[8] W. Gomes et al., ISSCC 2020, pp. 144-146, 2020.

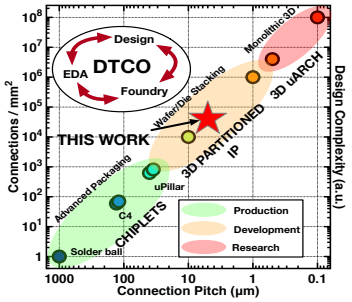

Fig. 1 The 3D integration roadmap. This work targets 5.76μm 3D face-to-face bonding pitch. Strong Design-Foundry-EDA collaboration is important for high-density 3D technologies.

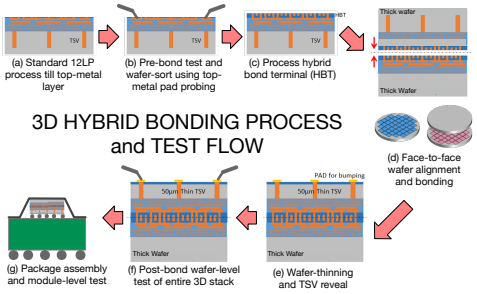

Fig. 2 3D process and test flow diagram showing (a-b) pre-bond tests using top metal test pads to enable wafer-sorting and matching, (c-d) hybrid wafer bonding at hybrid bonding terminal (HBT) layer and (e) TSV reveal with contact metal, (f-g) post-bond test through C4 bumps and TSV and packaging.

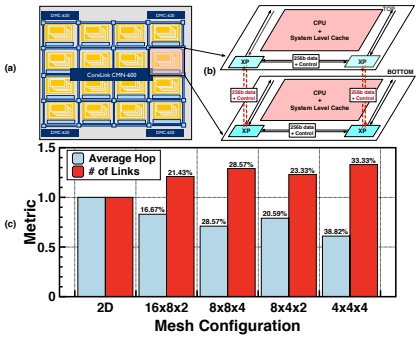

Fig. 3 (a) Arm CMN-600 interconnect, (b) a conceptual 3D mesh. The blue blocks (2x2 mesh cross-points) are implemented in the test-vehicle and (c) number of links and average hops in 2D versus 3D mesh.

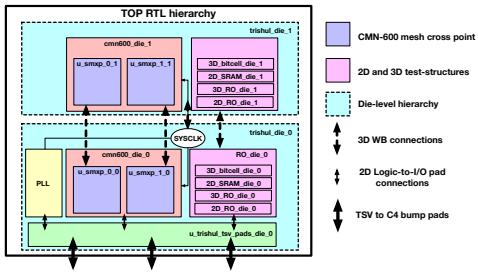

Fig. 4 3D test-vehicle logical block diagram. The logical blocks are pre-partitioned into top and bottom tiers.

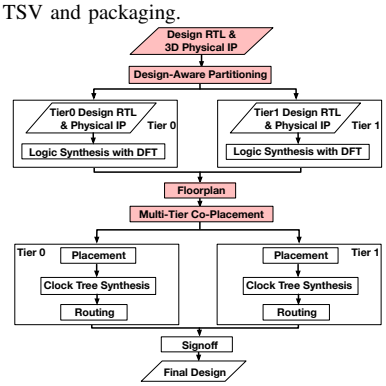

Fig. 5 A novel 3D physical implementation flow enabling multi-tier co-placement and timing optimization.

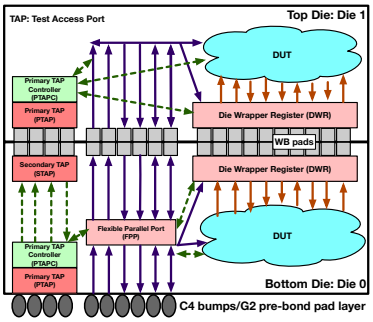

Fig. 6 IEEE 1838 compliant 3D DFT used in the test-vehicle.

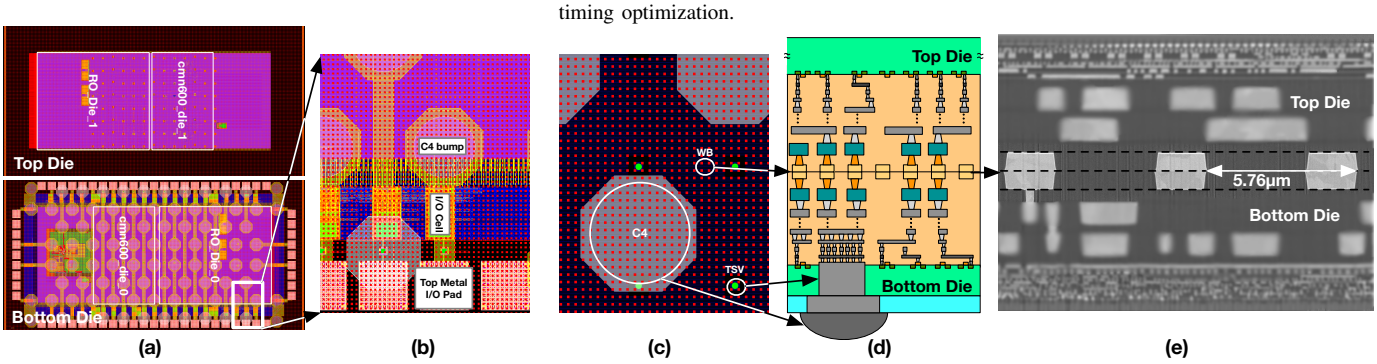

Fig. 7 (a) 3D design GDS view, (b) zoom-in view of the I/O cells, top-metal I/O pads for pre-bond testing and C4 bumps connected via TSVs, (c) zoom-in view showing C4 bumps, TSV and wafer-bond pads, (d) 3D-stack cross-section figure and (e) corresponding die cross-section image from the 3D test-vehicle.

Metric	Value
Process technology	12nm FinFET
Metal layers per die	11
3D stacking	Face-to-face hybrid wafer bond
3D pitch	5.76μm
TSV diameter	5μm
C4 bump pitch	150μm
Active die area	1.18mm²
3D signals for CMN-600	1600 per XP
3D signals/die	13800
3D pads for power delivery/die	22158
Cumulative 3D signal-nets tested from 945 wafer-bonded dies	**13.5 million**

Table I Key metrics of the 3D stacked test-vehicle. The vehicle demonstrates the feasibility of hybrid-wafer bonding for logic-over-logic high-density 3D design.

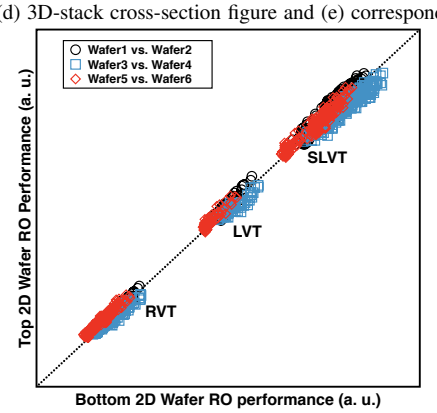

Fig. 8 Example pre-bond measurements from 6 wafers to match for bonding. Good correlation observed between selected wafers for bonding.

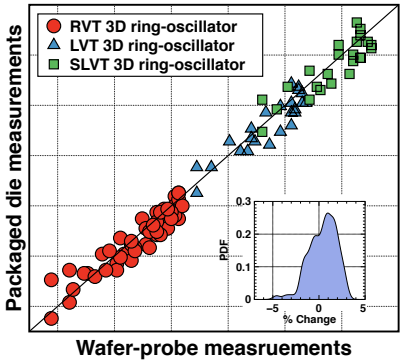

Fig. 9 Post-dicing and packaged device performance versus post-bond wafer-probe measurements.

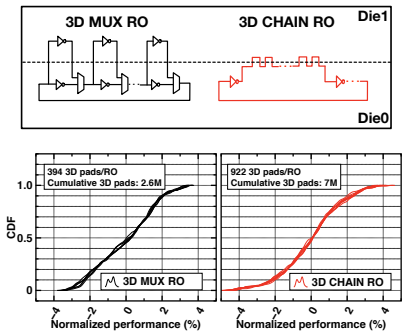

Fig. 10 CDF of 3D ROs measuring 3D connection yield. Each 3D MUX RO tests 394 3D connections and 3D CHAIN RO tests 922 3D connections on each die.

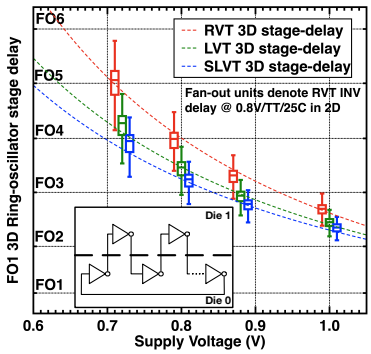

Fig. 11 Measured 3D RO stage-delay. 3D FO1 gate-delay can be as low as 2D FO3 gate-delay.

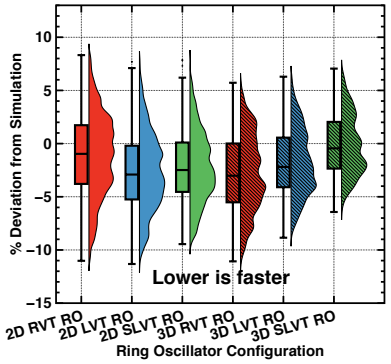

Fig. 12 Distribution of 2D and 3D RO delay from 945 dies. Most measurements are within +/- 5% of simulations.

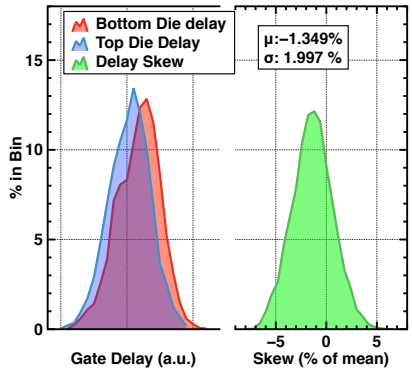

Fig. 13 Distribution of 2D FO1 gate-delay from top and bottom die. Measured delay skew between the layers is 2% of FO1 stage delay.

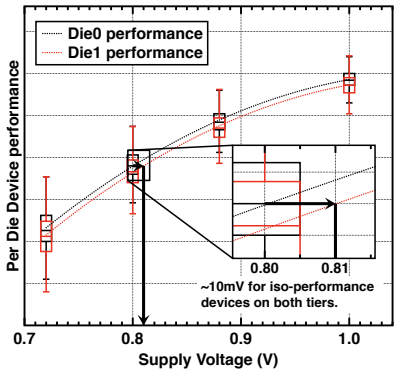

Fig. 14 Measured data showing that 10mV supply voltage bias can mitigate process-skew between 3D bonded wafers.

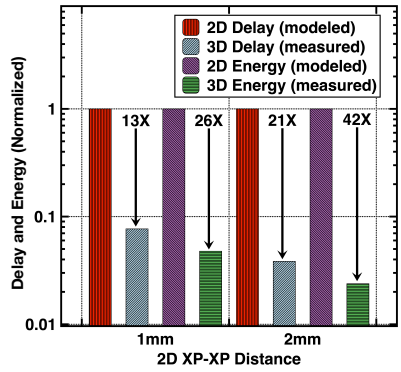

Fig. 15 Comparison of delay and energy-per-bit for signaling in a 2D versus 3D mesh.

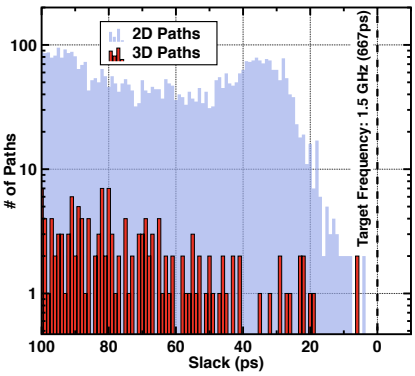

Fig. 16 Slack histogram from static-timing-analysis (STA) showing 2D and 3D critical paths. The timing of the design is closed at 1.5GHz.

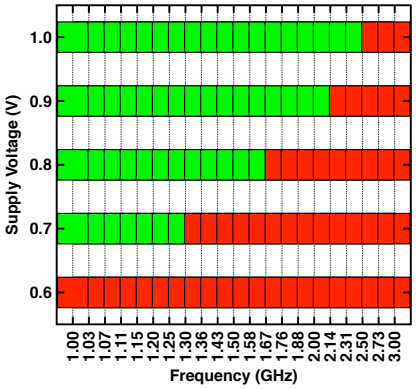

Fig. 17 Shmoo plot showing the 3D mesh operational at 1.58GHz at V_{DD}=0.8V and 2.3GHz at V_{DD}=1V supply as designed.

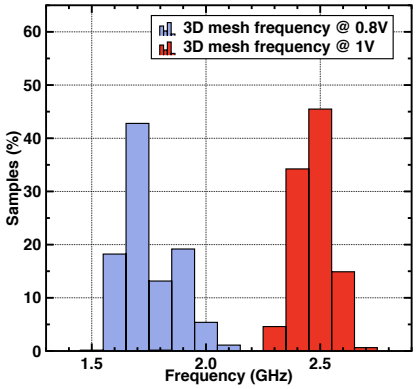

Fig. 18 Maximum frequency of 3D mesh for nominal (0.8V) and overdrive (1V) supply.

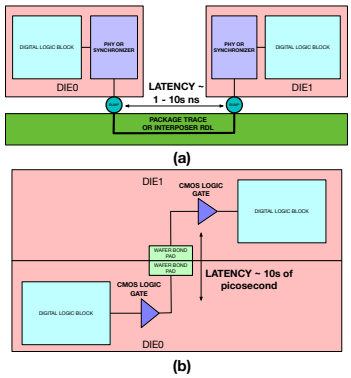

Fig. 19 (a) A typical 2.5D or 3D bump based interface and (b) a hybrid bonded 3D interface.

	[6]	[7]	[8]	This work	
Technology	Silicon bridge (2.5D)	Silicon Interposer (2.5D)	Foveros (3D)	Hybrid bonding (3D)	
2.5D/3D pitch	40μm	40μm	36μm	**5.76μm**	
Interconnect	MDIO	LIPINCON	-	AMBA® CHI	
Supply voltage	0.5	0.3	-	0.8	1
Pin speed (Gb/s)	5.4	8	0.5	1.6	2.4
Bits	-	320	200	512	
Aggregate Bandwidth (GB/s)	-	320	12.5	204.8	307.2
Bandwidth Density (GB/s/mm²)	198	198	not reported	**2276**	**3413**
Energy/bit (pJ)	0.5	0.56	0.2	**0.013***	**0.021***

*Energy/bit estimated based on 3D ring-oscillator current measurements.

Table II Comparison of state-of-the-art 2.5D and 3D die-to-die metrics. 3D hybrid bonding technology offers an order-of-magnitude better bandwidth, bandwidth density and energy/bit compared to existing 2.5D/3D bump-based techniques.