Multi-Chiplet Heterogeneous Integration Packaging for Semiconductor System Scaling

Surya Bhattacharya, Vempati Srinivasa Rao
Institute of Microelectronics, Agency for Science, Technology and Research (A*STAR), Singapore. bhattass@ime.a-star.edu.sg

***Abstract*:** Since the invention of the transistor, we have enjoyed tremendous impact of semiconductors on electronic systems. Transistor scaling has played a critical role in achieving increased functionality of semiconductor systems in main-frames, personal computers, and mobile phones by enabling lower power, cost and area per function through monolithic System-on-Chip (SoC). However, over the past decade, the diverse system requirements from wide ranging markets have driven the industry to use heterogeneous integration of multiple chiplets enabled by advanced packaging as a key new toolbox for System-in-Package scaling. This paper provides an overview of multi-chiplet heterogeneous integration (MCHI) packaging platforms to address system scaling needs in coming decades.

I. Introduction

In recent years, a multitude of applications such as 5G, Hyper-scale Data Centers, AI, and High Performance Compute have driven diverse Power-Performance-Formfactor-Cost (PPFC) demands from semiconductor systems in a manner that can no longer be met solely by SoC approach. 5G and 6G connectivity drives the need for tightly integrating phased-array antennas with mmWave ICs. Hyperscale data centers need optical chiplets to be integrated with drivers, switch/GPU within the same package. AI/ML systems need large packages with dense interconnects to connect compute chiplets with large amounts of nearby memory chiplets. To overcome reticle-size limitations, large SoCs and terabyte memories are disintegrated and reintegrated as chiplets. This is driving the transition from SoCs integrated on PCB to heterogeneous chiplets integrated in advanced PPFC optimized System-in-Package (SiP). MCHI packaging of chiplets from diverse technologies (CMOS, Silicon Photonics, GaN, SiC, SiGe etc) and diverse nodes is proving to be a cost-effective approach to achieve innovative market-driven system scaling (Fig. 1).

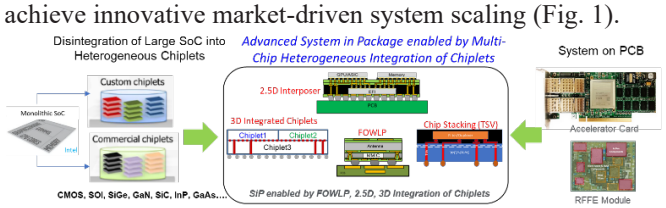


Figure 1: Driving force for MCHI packaging

We categorize market-driven MCHI packaging platforms as (a) High-Density Fan-Out Wafer Level Packaging (HD-FOWLP) (b) 2.5D Interposer enabled by micro-bump and TSV (c) 3D-Integrated Chiplets (3DIC) enabled by chip-to-wafer (C2W) and wafer-to-wafer (W2W) to hybrid bonding (Table 1). One can observe that going from FOWLP to 3DIC, the interconnect density increases by over 4 orders of magnitude.

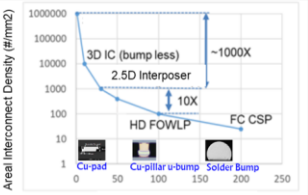
Heterogeneous Integration Platforms	Target Applications.	Interconnect Density (#/mm2)
High Density Fan-Out Wafer Level Packaging	mmWave Antenna in Package, Radar AP+Memory, Optical Engines	
2.5D Interposer	Data Centre, HPC, AI, Chiplets	
3D Integrated Chiplets	SOC Disintegration + Chiplet reintegration, 3D Memory	• Pad pitch drives chiplet bond-pad or microbump areal interconnect density (#/mm2) • Wire pitch and #layers drive chiplet-to-chiplet linear interconnect (RDL or Cu-damascene) density (#/mm)

Table 1: Heterogeneous Integration Packaging Platforms

II. High Density (HD) FOWLP Platform

Millimeter-wave applications (30-300GHz) such as 5G, 6G, and radar sensors demand cost-effective, small form-factor and short interconnects to avoid high signal loss between mm-wave ICs (MMIC) and the beam-forming antenna array. These requirements are achieved by FOWLP. In FOWLP technology, chiplets are embedded into an epoxy mold compound (EMC) to form a reconstituted wafer for further processing (for example: re-distribution layers (RDL) etc.) to connect the chiplets. FOWLP can be achieved by Mold-1st or RDL-1st approaches (Fig. 2), and the FOWLP package can be mounted directly on the PCB without a substrate.

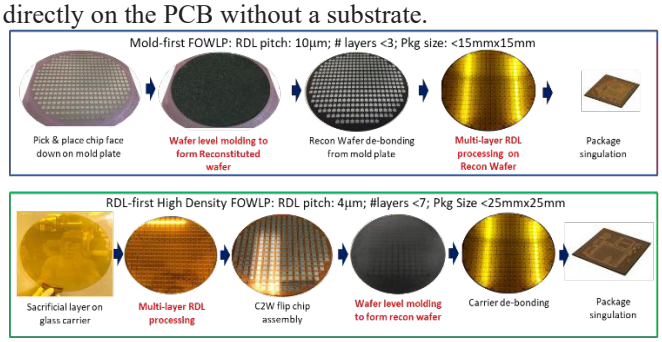
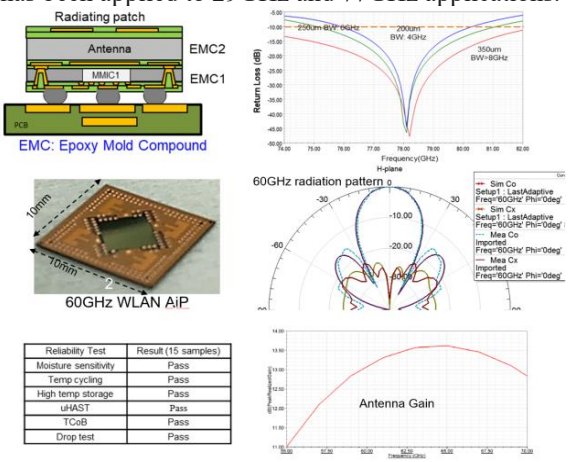


Figure 2: Mold-1st FOWLP and RDL-1st HD-FOWLP.

In Mold 1st, the chiplets are attached to a carrier in a face-down manner and embedded in compression molding. RDL at wire pitch >10µm are formed on the reconstituted wafer to connect chiplets and form the package. In RDL 1st approach, high density RDL (pitch: >4µm) are initially fabricated on the carrier followed by attaching of micro-bumped chiplets in flip-chip manner to the under bump metallization (UBM) pads, and then compression molded to form wafer level packages. RDL 1st is ideal for HDFOWLP for chiplets with high I/O counts (500 pads/mm^2) that have high inter-chiplet bandwidth needs, and for package sizes up to 25x25 mm^2 [1].

The mold 1st FOWLP based Antenna-in-Package (AiP) configuration (Fig. 3) that houses MMIC(s) in the lower level

of the package and the phased array antenna in the upper layer allows the antenna to be designed with minimal interference from the MMIC. Fig. 4 depicts the process-flow and the AiP cross-section. After the 1st mold (EMC1), the reconstituted wafer is bonded to a temporary carrier, back-grinded and laser-drilling is used to make the vias (TMV). After a metal layer is plated, the 2nd mold (EMC2) is applied, and a top Copper layer is plated to form Antenna patches. EMC1/2 are optimized to minimize package warpage. The package has been realized and tested to verify radiation performance and reliability. The AIP has been applied to 29GHz and 77GHz applications.



Reliability Test	Result (15 samples)
Moisture sensitivity	Pass
Temp cycling	Pass
High temp storage	Pass
uHAST	Pass
TCoB	Pass
Drop test	Pass

Figure 3: mmWave Antenna-in-Package Structure, Performance

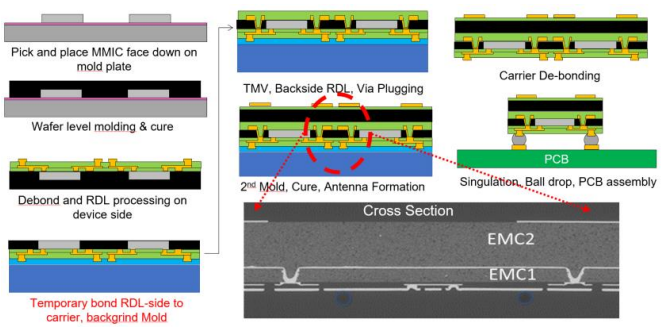


Figure 4: mmWave Antenna-in-Package Process Flow, XSEM

As hyper-scale data center top-of-rack switches reach > 12.5Tb/s bandwidth, industry is targeting >800Gbps optical engines (OE) to convert electrical output into optical signals. For compact form-factor and low-power, the switch and OE are integrated on the same package to realize shortest electrical links between the switch and the optical-driver (EIC) and from EIC to Photonic ICs (PIC). (Fig. 5 option c).

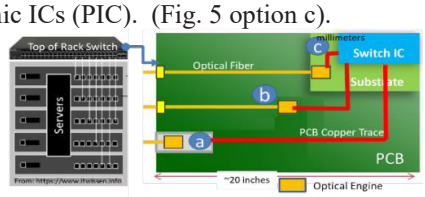


Figure 5: Optical Engines in Top of Rack application. (a) Board-edge (b) Mid-board (c) Co-packaged Optical Engines

Fig. 6 shows the co-packaged OE structure that integrates the PIC and EICs in a Mold 1st FOWLP which enables optical edge coupling to the PIC. This approach has been used to design an 800Gbps OE in a 15mm x 9mm package with less than 1dB insertion loss up to 60GHz. Fig. 7 depicts the mold 1st FOWLP process flow with protection for sensitive PIC waveguides and couplers, where the key step is when the dicing lane runs through the mold compound and the packaged PIC to expose the optical couplers [2]. Fiber-to-waveguide optical coupling losses <1.7dB has been achieved.

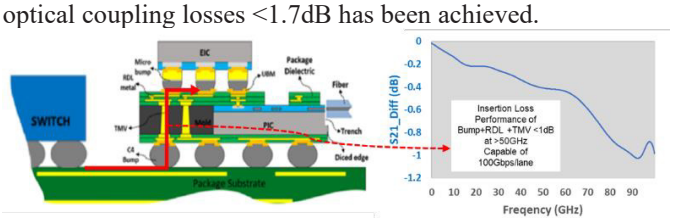


Figure 6: Heterogeneous Integrated OE Package for CPO

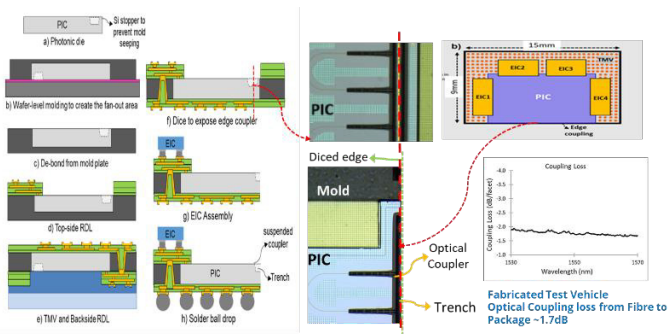


Figure 7: Process flow, design and optical coupling for 800G OE

The rapidly growing size and complexity of Neural network (NN) models makes it a challenge to design and manufacture monolithic SoCs to keep up with requirements. Thus modular chiplets integrated in an advanced SiP is an ideal scalable solution based on model size. We have used RDL 1st HD-FOWLP (5 RDL layers; <4µm wire pitch) to integrate 4 CNN-based perception acceleration chiplets that use Advanced Interface Bus (AIB) interface (Fig. 8). The multi-chiplet package is formed with five RDLs which enable a total bandwidth density of 256GB/s/mm. The AIB interface consists of 3000+ lines that are routed using the upper three RDLs. The lower RDLs are used to distribute the ground and power. The measured results are displayed in Fig. 8 indicating 2.14 TOPS/W at 0.6V [3]. This approach can be extended to > 16 chiplets in a 25x25mm^2 RDL 1st package.

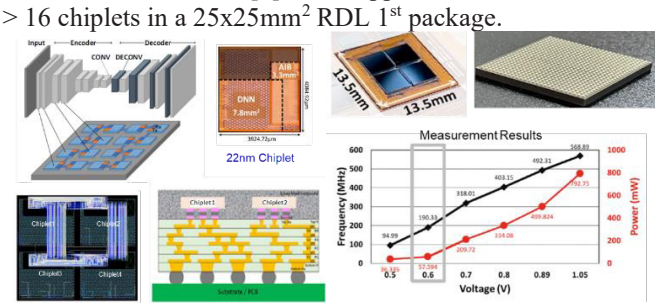


Figure 8: CNN based perception accelerator using Chiplets: Package realization and performance of the CNN Chiplet SiP

III. 2.5D Interposer Platform

To integrate FPGA/GPU SoC with multiple high density memories or SERDES chiplets (with >50µm µbump pad-pitch) in datacenter-compute, AI/ML applications, which need packages that are >40mm x 40mm in size with multi-layer wire pitch (<4µm) that HD-FOWLP cannot support, the 2.5D interposer is the go to choice for the industry. The options to manufacture interposers include (a) Passive Through Silicon

Interposer with TSV (b) Active Silicon Interposer (c) Embedded Fine-pitch Interconnect (EFI) Bridge interposer.

Passive TSV-interposers (Fig. 9) can interconnect micro-bumped ICs/Chiplets (~50µm bump-pitch, <2µm wire-pitch) on large interposer up to 3X reticle size (~2500mm^2).

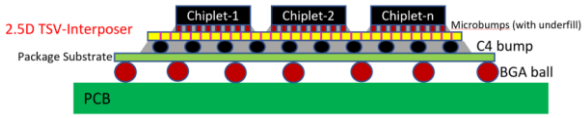


Figure 9: Passive Through Si Interposer schematic

Active interposers are suitable for SoC-partitioning where Analog, I/O circuits, decoupling capacitors are off loaded from the SOC to the interposer fabricated in a mature node CMOS (130nm-28nm). We have developed a via-last from back-side approach (Fig. 10) to convert a foundry wafer into an active interposer. Fig. 11 shows the X-SEM, the finished test vehicle package [4], and the feasibility of transferring I/O protection circuits from chiplets to the active interposer.

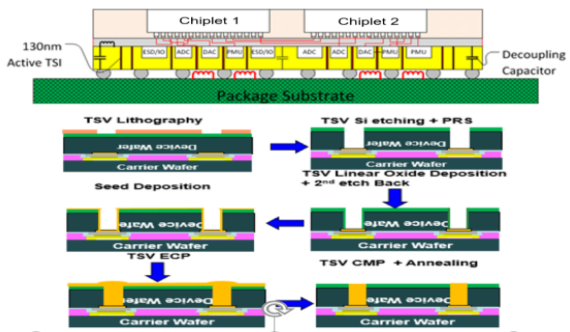
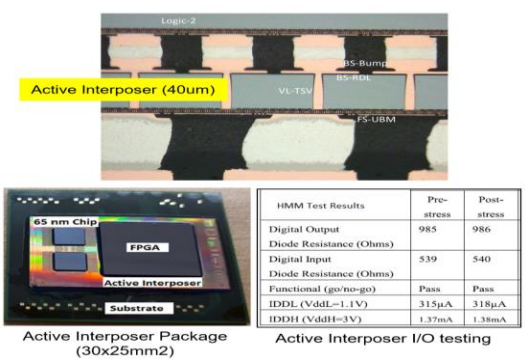


Figure 10: Active Interposer structure and process flow



HMM Test Results	Pre-stress	Post-stress
Digital Output Diode Resistance (Ohms)	985	986
Digital Input Diode Resistance (Ohms)	539	540
Functional (go/no-go)	Pass	Pass
IDDL (VddL=1.1V)	315µA	318µA
IDDH (VddH=3V)	1.37mA	1.38mA

Figure 11: Active Interposer XSEM, Package, and I/O performance

A key challenge with Through-Si-Interposer is the cost of manufacturing. To address this challenge, the industry is looking for alternate solutions. Using an embedded silicon bridge, we have demonstrated the Embedded Fine-pitch Interconnect (EFI) bridge interposer (Fig. 12). The EFI bridge can support short fine pitch interconnects (pitch<2µm; >6 layers achieved by Cu damascene) for chiplet-to-chiplet interfaces, while the RDL on interposer can support longer traces required for a large package. The EFI approach allows for up to 50% savings compared to the through-Si-interposer. Fig. 13 shows the process flow for fabricating an EFI-bridge interposer. First multi-level RDL, UBM, and Cu pillars are formed on a sacrificial layer on a carrier wafer. Next the EFI silicon bridge is flip-attached to the UBM. The carrier wafer is molded and back-grinded to reveal the Cu pillar through mold via, followed by the backside RDL and UBM formation. The carrier is de-bonded, and the GPU and HBM chiplets are assembled to the interposer followed by underfilling, and interposer assembly. Component level tests and signal integrity of 8mm long interconnects shows the reliability and performance of the EFI bridge interposer (Figs. 12, 14).

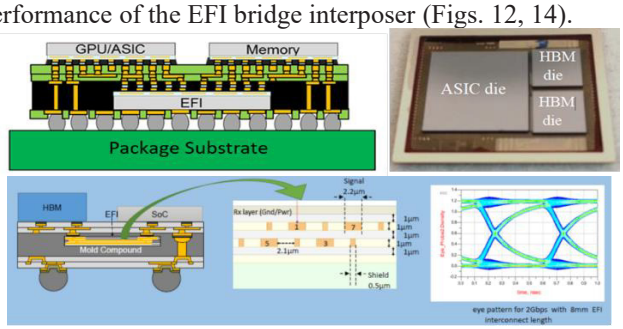


Figure 12: EFI Bridge interposer, package, signal integrity

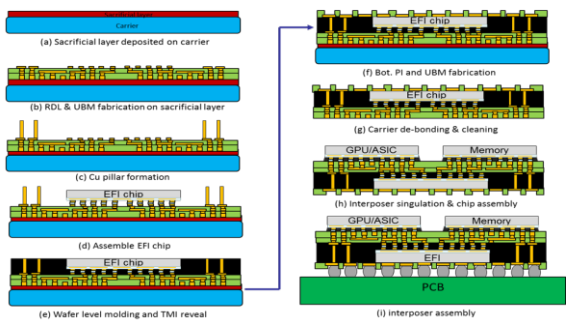


Figure 13: Process flow for EFI Bridge interposer

		TCOB testing (Board level)			
Group	Wafer info	250 cyc	500 cyc	750cyc	1000cyc
1	Material A	0 / 7	0 / 7	0/6	0/6
2	Material B	0 / 5	0 / 5	0/4	0/4

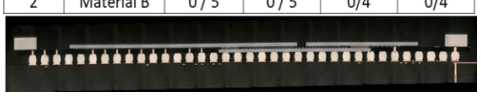


Figure 14: EFI Bridge Interposer X-section, Reliability data.

IV. 3D INTEGRATED CHIPLETS

Recently, large SoCs for complex GPU and AI/ML applications have reached the reticle-size limit, and the industry is looking to disintegrate SOC/GPU into chiplets and reintegrate into 3D integrated chiplet systems where the pad pitches can be <10µm to enable tight, seamless almost-monolithic re-integration. As pad pitches reach sub 10µm, bridging between closest micro-bumps prevents usage of traditional bumping. Thus, solder-free hybrid bonding (Cu-Cu + Oxide-Oxide) is emerging as key technology to form 3DIC.

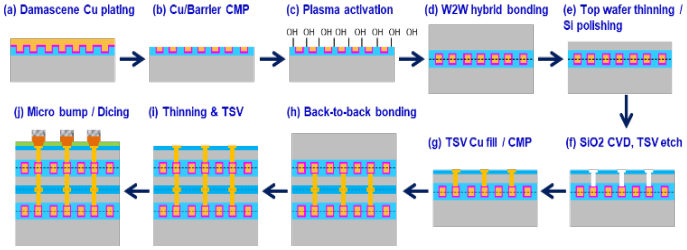


Figure 15: Process flow for Wafer-to-wafer hybrid bonding.

Fig. 15 shows the process flow for W2W hybrid bonding that is preferred in applications where the stacked die-sizes are

identical (e.g memory). Wafer surface flatness, cleaning and pre-bond preparation are key steps in hybrid bonding. Fig. 15 shows the hybrid bond interface with <5µm pad pitch corresponding electrical data.

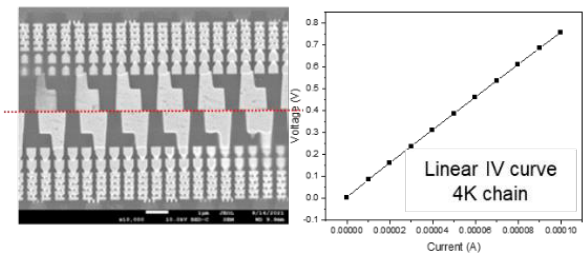


Figure 16: Cross-section of W2W Hybrid bond and Linear I-V

C2W Hybrid bonding is an ideal approach to integrate dissimilar sized chiplets to form 3DIC (Table 1). We developed Copper CMP process on damascene plated interconnects and achieved SiO_2 surface roughness <0.5nm and Copper dishing in the range of 5-6nm. Fig. 17 shows the process flow for C2W hybrid bonding. To achieve void free bonding, particle free singulation, and optimization of surface activation, die placement in clean chamber and annealing are critical. To control the particle count on singulated wafer, dicing process can be optimized using dicing before back grinding. Furthermore, plasma chemistries, pick-&-place and annealing processes for SiO_2 based and polymer-based hybrid bonding are optimized for good quality hybrid bonding (Fig. 18).

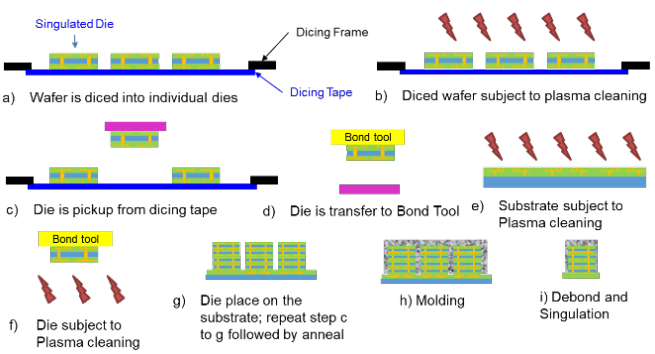


Fig. 17: Process flow for Chip to wafer hybrid bonding.

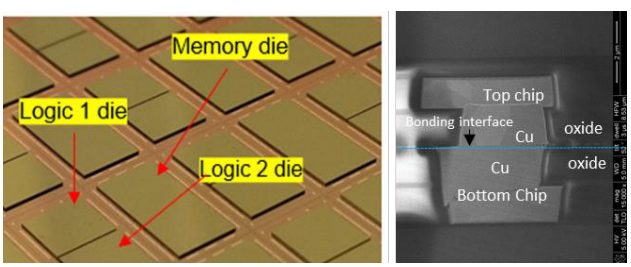


Fig. 18: Chip-to-wafer hybrid bonded wafer, XSEM of hybrid bond.

V. THERMAL COOLING FOR 3DIC

Thermal cooling is a major challenge for 3DICS where high-performance chiplets are stacked. We address this by forming micro pin-fins monolithically on the back side of the high-performance chiplet. Small formfactor is achieved by in-plane operation. Our testing data has shown 200 W/cm^2 with max chip temperature 70℃ at water flowrate 70 mL/min [5]. The thermal resistance is below 0.2 K/W, thus showing this can be a viable path for future 3DIC (Fig. 19)

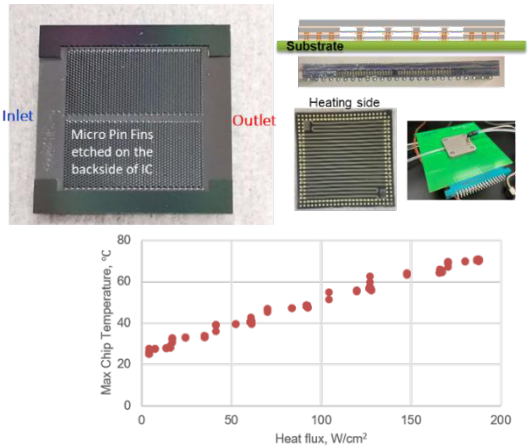


Figure 19:Active cooling for 3DIC

VI. WARPAGE IN WAFER LEVEL PACKAGES

Package mechanical modeling is a critical step in determining stress and warpage effects as warpage control is critical for 12-inch FOWLP processing. Usually, large wafer warpage hinders processability of wafers. Elastic mechanical stress models are not sufficiently accurate for high confidence prediction. We therefore extract viscoelastic properties of dielectric materials, molding compounds, and underfill materials and develop accurate models that allow us to make significantly better material choice/optimization (Fig. 20).

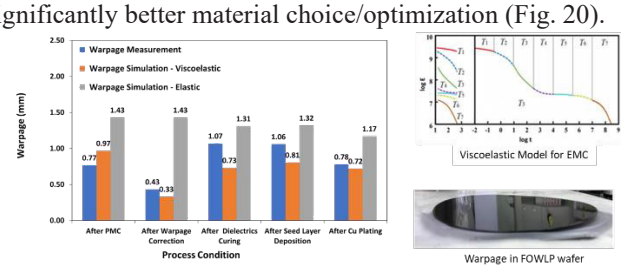


Figure 20: Advanced mechanical modeling of wafer warpage.

VII. SUMMARY

MCHI packaging has provided the industry with a powerful new toolbox to address PPFC-driven semiconductor system-scaling across multitude of end-markets. MCHI packaging will enable the industry to develop novel approaches to implement next generation systems for 5G, AI, HPC and other applications. This is driving new collaboration paradigms across the supply chain between research consortia, end-users, fabless, foundries, OSATs, equipment, materials and EDA companies to meet industry needs for the coming decades.

VIII. REFERENCES

[1] V. S. Rao, et al, "Development of high density fan out wafer level package (HD FOWLP) with multi-layer fine pitch RDL for mobile applications," in *Proc. of 66th ECTC*, USA, 2016, pp. 1522-1529.

[2] Sajay B. G, et al, " A Heterogeneously Integrated Wafer-level Processed Co-Packaged Optical Engine for Hyper-scale Data Centres," in Proc. of 73rd ECTC, Orlando USA, May 2023.

[3] T. Chou, et al, "NetFlex: A 22nm Multi-Chiplet Perception Accelerator in High-Density Fan-Out Wafer-Level Packaging" VLSI Technology and Circuits Symposia, June 12-17 2022, Honolulu, Hawaii.

[4] V. Chidambaram, et al, " Heterogeneous system level integration using active Si interposer," Journal of the Electronics Device Society, vol. 7, pp. 1209-1216, Sept. 2019.

[5] H. Feng, et al, "Embedded Micro-Pin Fin Heat Sink of Two-Phase Liquid Cooling for High Heat Flux 3D ICs," in Proc. of 73rd ECTC, Orlando USA, May 2023.