Chip Scale Review®

ChipScaleReview.com

The Future of Semiconductor Packaging

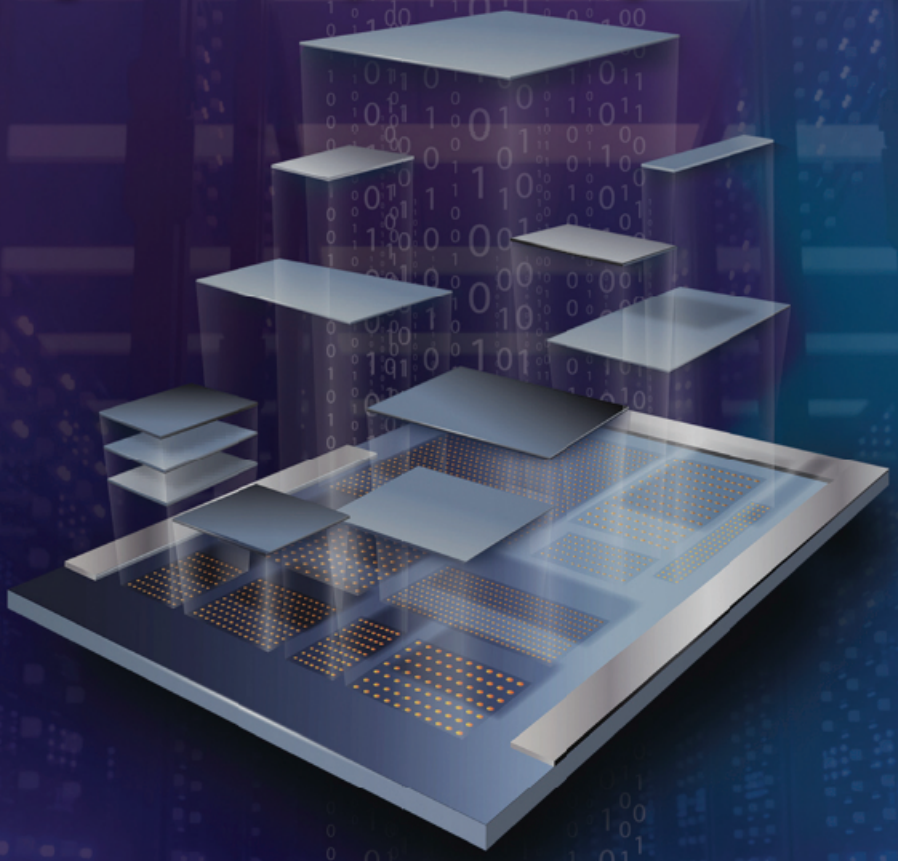
Volume 27, Number 2

March • April 2023

CELEBRATING 25 YEARS

3D chiplet integration with hybrid bonding

page 7



- Heterogeneous integration and chiplets and dielets: why the hype?
- Hybrid bonding bridge for chiplet design and heterogeneous integration
- Electrical design challenges of multi-layered fan-out RDL MCM packaging
- Accelerating 3D and heterogeneous integration with high-volume D2W hybrid bonding
- Wafer-scale integration for graphene-based optoelectronics, sensors, and imaging devices

3D chiplet integration with hybrid bonding

By Laura Mirkarimi [Adeia, Inc.]

Market demand for high-performance computing (HPC) in server, gaming, artificial intelligence, and machine learning applications is growing. In 2021, HPC was a $35B business and is predicted to be $65B in 2030 with a compound annual growth rate (CAGR) of 7.2% [1]. At the same time, the semiconductor industry has experienced more than a decade of slowing of Moore's Law as the cost and technical challenges to produce the next transistor node have risen sharply. In response, the industry embraced advanced packaging with vertical stacking in 2.5 and 3D platforms to achieve higher compute performance, overcome the advanced node slowdown, and maintain product release timelines.

Silicon interposer-based 2.5 and 3D packaging – leveraging through-silicon via (TSV) technology – has been in high-volume manufacturing for more than 10 years. The ecosystem developed with the boost from companies such as Samsung, SK Hynix, Xilinx, and AMD that brought stacks of memory on logic into products with assistance from foundries like TSMC and outsourced semiconductor assembly and test providers (OSATs) like ASE. The chiplet approach in 2.5D was shown to cost one-half of the comparable monolithic structure [2]. However, the adoption rate of these products has been relatively slow and limited to a few companies, in part due to technical challenges with the interconnect density and overall cost [3].

While the semiconductor industry is anticipated to enjoy a healthy CAGR of 5.7% from $605B (2022) to $735B in (2026), cost management is a central theme for advanced packaging adoption and proliferation within the industry [4]. The monolithic nature of today's system-on-chips (SoCs) requires escalating design and development costs that are not suitable for small-volume manufacturers and entities like the U.S. Department of Defense. Apparently, a tipping point was reached when the Defense Advanced Research Projects Agency's (DARPA) Common Heterogeneous Integration and Intellectual Property (IP) Reuse Strategies (CHIPS) program was born. The goal of this program is to create a paradigm shift, "to enhance overall system flexibility, reduce design time for next-generation products, with significant IP reuse [5]." At the first Chiplet Summit Conference in January 2023, Yole shared that the chiplet-based processors market will grow from $62B in 2022 to $180B in 2027—a CAGR of approximately 24% (Figure 1) [6]. The promise of further standardization within the supply chain for IP and/or interconnect guidelines has brought much optimism to electronics companies.

Chiplet concepts

The technical success and learnings from the 2.5D and 3D packaging with TSVs have built a foundation of excitement and vision for the possibilities of a new chiplet era [2]. A chiplet is a portion of an integrated circuit (IC) with a specific functionality

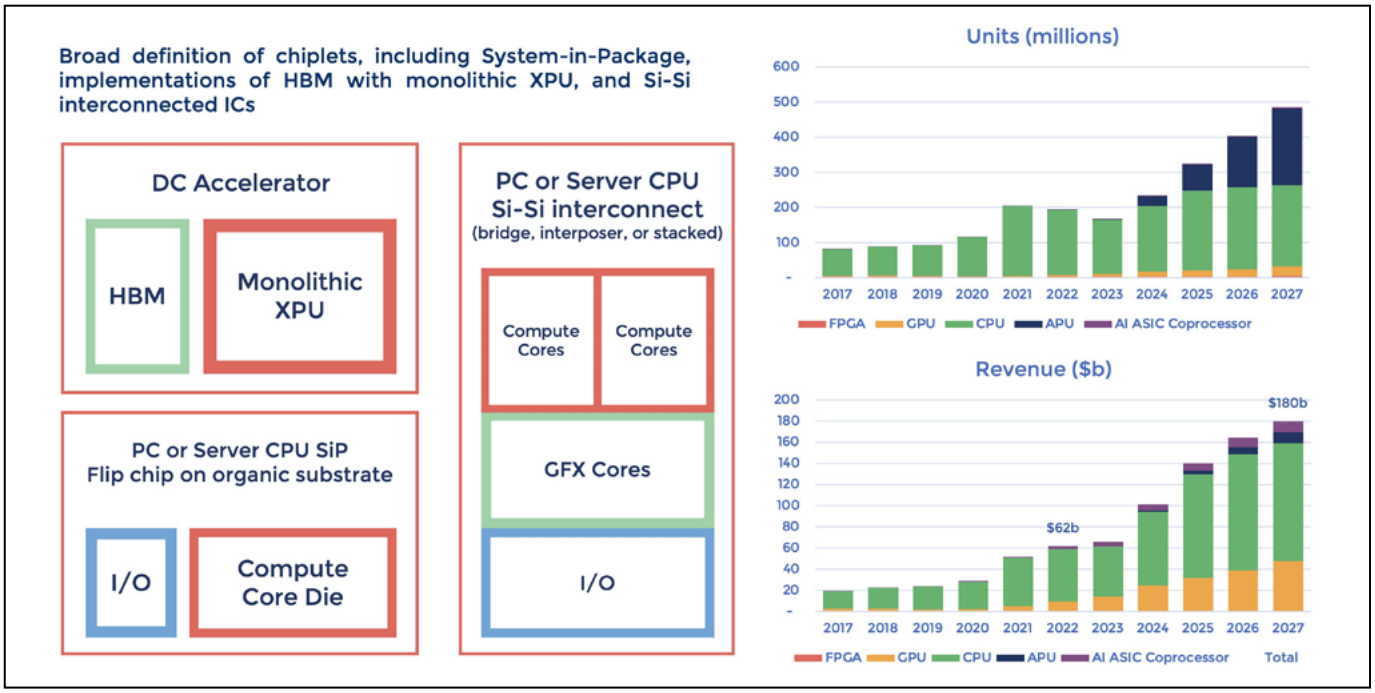


Figure 1: Market growth in chiplets. SOURCE: Yole Group, "Chiplet Market Update" presentation, Yole Intelligence - Chiplet Summit 2023

that is designed to be combined with other chiplets to complete a fully-functional module within a package or system. Chiplets require an I/O controller die to bring the multi-chiplet module together to function as an IC. The advantages touted over traditional SoC packaging are heterogeneous integration, known good die and reusable IP for a multitude of applications.

The value proposition of the new chiplet era is to fundamentally reduce cost, while delivering enhanced features in electronic products. The main themes for cost management are die size shrink, disaggregation, reduced time to market, standardized I/O protocol, and increased IP reuse. At the Intel-Architecture Day 2020, the vision of creating multiple chiplet cores that connect to memory in the substrate showed a major shift toward distributed computing [7]. Recently, Ventana announced its partnership with Intel Foundries and shared a customer case study where the development costs could be reduced by a factor of 8 by using the distributed chiplet compute architecture [8]. Smaller chips have higher yield because of lower defect density physics. Disaggregation is important for the ability to use legacy chips as appropriate, and optimally process the various silicon circuit parts within the supply chain. For example, depending upon the specific application, A/D converters are better with legacy nodes, while some processors are better suited at the latest node. Design and fabrication of ICs in the most favorable node or process offer more options for cost savings. Reuse of chiplets reduces the development time of product families while delivering performance differentiators to the market in a timely fashion, thereby reducing development costs.

It sounds like a utopia for IC architects. The packaging reaggregation brings the reality of integration challenges including lack of scaling and performance with the conventional Cu microbump and non-standard IP among various packaging approaches within the industry. Products today have limited I/Os at 25µm pitch; however, many chiplets and 2.5D modules would benefit from finer pitch interconnection between memory and logic or logic/logic interfaces delivering high bandwidth and low latency, all critical for computation.

Hybrid bonding

The industry has been manufacturing fine-pitch direct bond interconnect (DBI®) hybrid bonding in wafer-to-wafer applications such as image sensors (~2.5-8µm) since 2016, and more recently, NAND memory (~1µm) because the manufacturing ecosystem was ready [9,10]. Hybrid bonding requires a level of cleanliness (i.e., ISO-5 to ISO-4) like back-end-of-line (BEOL) wafer fabrication; therefore, the wafer bonding process line had an immediate home. In contrast, die-to-wafer and die-to-die hybrid bonding manufacturing readiness have been in development for many years. The advanced packaging OSAT

companies typically operate in an ISO-7 environment and require an upgrade to their infrastructure for the hybrid bond advanced interconnect technology. As the pitch of the interconnect continues to scale in die-to-wafer applications, the micro-environment cleanliness specifications will tighten. Cluster tool platforms are now being considered for efficiencies of scale and throughput for the packaging houses and other manufacturing facilities to usher in this new technology.

Another gap in the infrastructure addressed in the past 5 years was the die-to-wafer bonder equipment alignment accuracy and local environmental cleanliness. HVM tools were specified at about +/-3µm to 5µm for a throughput of ~2000 units/hour. Pick and place equipment manufacturers began aligning their roadmaps with the cleanliness and alignment accuracies required for the hybrid-bonded chiplet at pitches below 20µm. Several pick and place companies report submicron placement accuracy tools on their roadmap to support further pitch scaling in generations to come.

During this time, Adeia, Inc. (formerly Xperi) worked with its customers to ensure that the die-to-wafer process being developed would scale to high-volume manufacturing. The customer requirements that stood out include hybrid interconnect with a flexible layout, high assembly yield and reliability with all die handling on the tape frame. In early 2017, we began developing the DBI® Ultra assembly process for die-to-wafer, which was launched at the 2019 ECTC. This process is shown in **Figure 2**. The hybrid bond interconnect is formed with a standard BEOL Cu damascene process that includes dielectric deposition, etch, barrier layer, Cu seed, Cu plate and chemical mechanical polish (CMP).

After obtaining nanoscale topographic control across 300mm wafers, the wafers must be diced. Die handling is completed on a tape frame and the die surface must emerge from dicing with the cleanliness specification after CMP. We have demonstrated equivalent performance among all three singulation techniques: mechanical saw, stealth and plasma dicing. Activation, bond and anneal are the final steps in the process. Dozens of test vehicle modules were assembled with the process shown in **Figure 3**. Single die stacks in memory-logic interface configurations with the interconnect pitches ranging from 40-4µm pitch with 30k to 1.6M interconnects (**Figures 3a-b**) were assembled and tested to JEDEC environmental stress test standards. In parallel, the efficacy of 4- and 8-die stacks designed with 6k I/Os in an HBM-like format with a 35µm interconnect pitch TSV was demonstrated (**Figure 3c**). Theil, et al., reported that the yield per layer is consistent between 1 and 8 die, which is critically important to confidently develop the technology for 3D stacking [11]. Additionally, Gao, et al., showed the reliability performance of hybrid bond interconnect test vehicles with and without TSVs was enhanced compared to the microbump [12]. Given an all-Cu interconnect, there is no driving force for intermetallic formation or Kirkendall voids that lead to electrical failure and mechanical weakness in Cu microbumps. Instead, the resistance in the hybrid-bonded daisy chain test structure reduces ever so slightly because of an enhancement of Cu-Cu diffusion across the bond interface. The interconnect in a direct bond is surrounded by a strongly-bonded dielectric that holds the multiple die together. The mechanical stress delivered to the hybrid interconnect during functional operation is much less than in a Cu microbump.

A hybrid bond interconnect is well aligned with the new chiplet era roadmap for several reasons. The hybrid bond interconnect formed with a standard BEOL Cu damascene process is scalable with the semiconductor supply chain and the fundamentals were demonstrated at 1µm pitch in wafer-to-wafer configurations. At a pitch of 1µm and below, the maximum interconnect density is greater than $1x10^6$ interconnects/mm^2. The small form factor of a hybrid bond pad interconnect maintains a low inductance as well as a capacitance ideal for signal integrity performance.

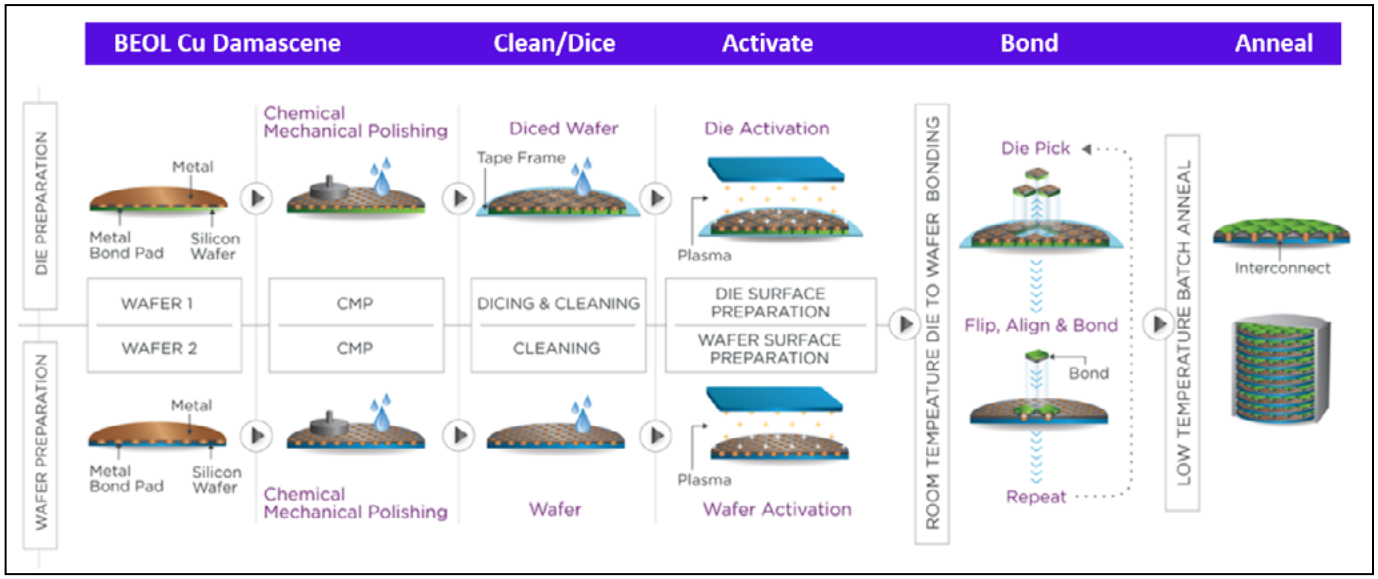


Figure 2: DBI® Ultra die-to-wafer hybrid bonding process flow.

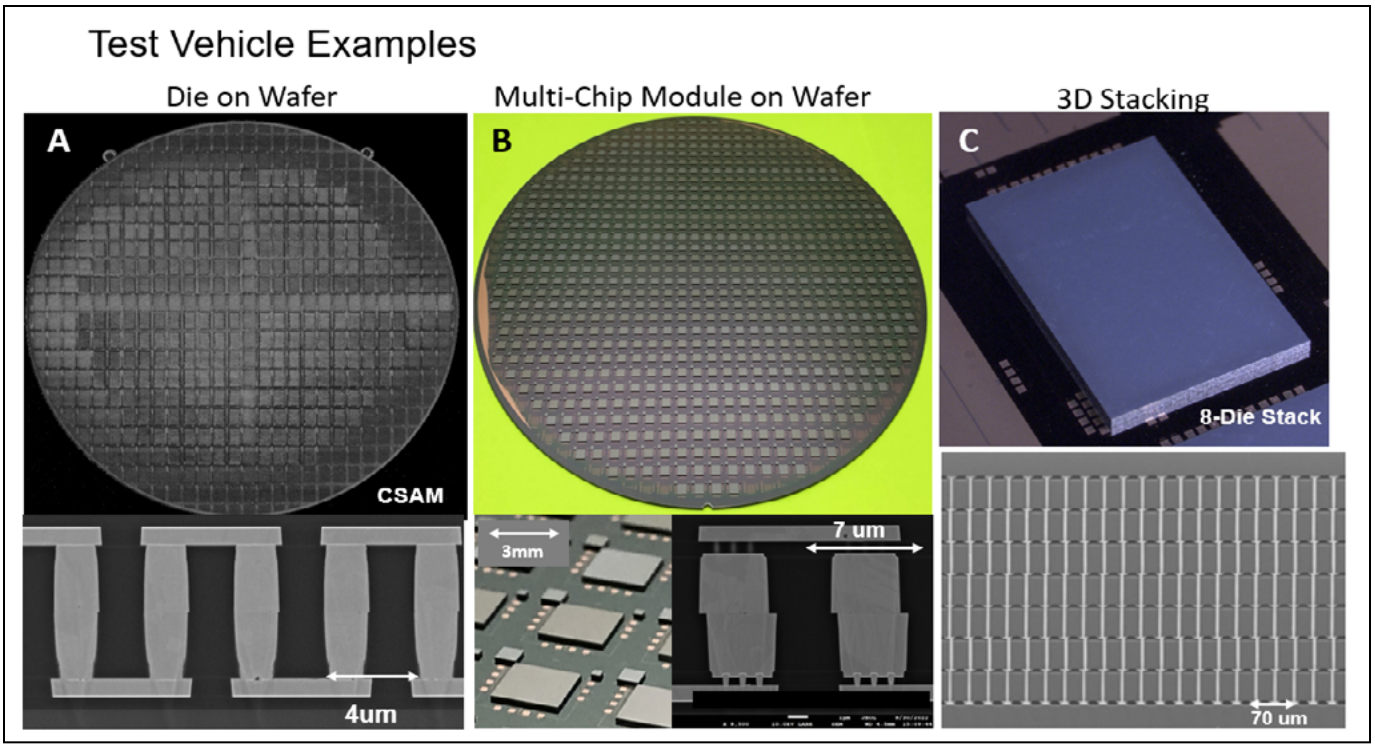
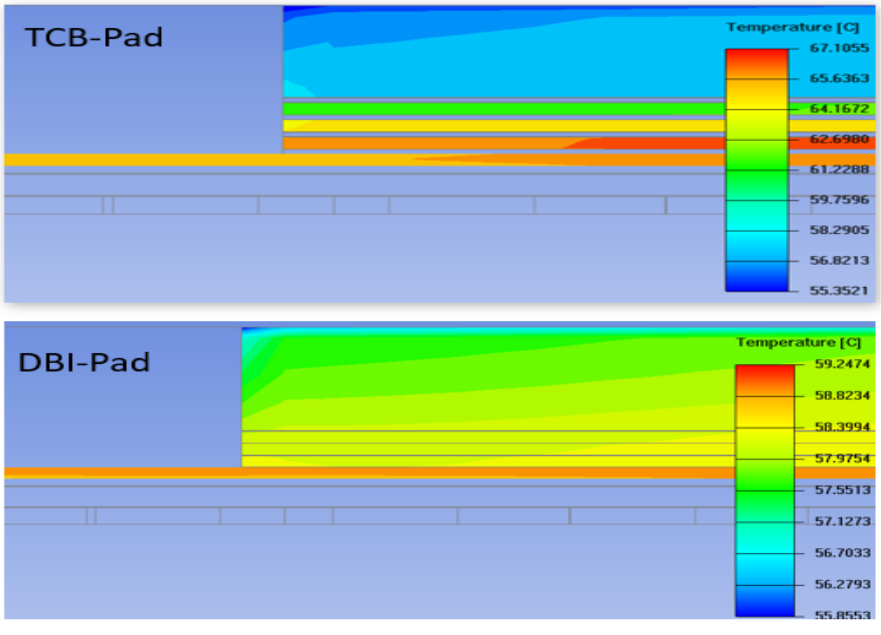


Figure 3: Examples of die-on-wafer configurations with hybrid bond interconnects: a) die on wafer; b) multi-chip module on wafer; and c) 3D stacking.

In Agrawal, et al., a 5µm hybrid bond pad is shown to have 1/50th the size, 96% less capacitance, 92% less inductance and 64% less resistance than a typical 10µm thermocompression bond (TCB) pad making it ideal for reduced latency [13]. Another feature of a DBI® is the inorganic dielectric surrounding the metal pads. The inorganic dielectric brings enhanced thermal performance to the module compared to the conventional microbump with underfill material. More uniform thermal conductivity between the die can reduce exacerbation of hot spots and allow for cooling solutions to positively impact the entire die stack more effectively. In the simulation of 4- and 8-high dynamic random access memory (DRAM)-like configured stacks, the differential temperature between die 1-4 and die 1-8 was compared for the TCB and DBI® interconnects. The temperature differential (ΔT) between die 1 and die 8, (4°C), in the hybrid-bonded stack is much lower than the TCB structure (28°C) (Figure 4). The lower ΔT between die within the stack is a significant advantage for high-speed devices that have temperature sensitive performance, such as DRAM [14].



Results Summary Table

Feature	Unit	4 Die HBM-Like Stack		8 Die HBM-Like Stack	
Forced Convection 3m/sec at 25C		TCB	DBI	TCB	DBI
Total Power	Watts	10	10	18	18
Temp. Junction	oC	67	59	95	70
DT btwn Die #1-N	oC	9	1	28	4

Figure 4: Simulation schematics for TCB and DBI® interconnects with 4-die and 8-die stacks.

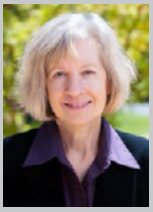
The distributed computing concept for the new chiplet era is driven by reduction of yield loss due to defect density. The same defect density cost drivers that moved the industry away from monolithic die to chiplets is also important for hybrid bond interconnect technology that requires a clean environment. This alignment of shrinking components for a distributed architecture is advantageous for hybrid bond yield enhancements too. The combined enhancements of die yield, electrical performance and thermal performance is a compelling argument to integrate the 3D chiplet with a hybrid bond interconnect. The conservative nature of the semiconductor industry demands that we invest in technologies that will serve multi-generations of product enhancements that echoes the value of the scalable hybrid bond interconnect. L. Cao of ASE explained the significant value advanced packaging brings to the semiconductor industry by offering numerous options to achieve higher performance modules [15]. More importantly, the OSATs appear to be evaluating hybrid bonding technology and the appropriate timing to provide that service, which signifies the expectation of high-volume customer interest [15].

Summary

AMD, an industry leader in 2.5D chiplet architecture, released its 3D chiplet technology and first hybrid-bonded module for the Ryzen series 5900 [16]. The L3 cache is bonded to a 5000 series processor. The interconnect pitch of ~9µm is only achievable via a hybrid-bond interconnect and represents a 200x times the density of 2D chiplets. Similarly, other thermal enhancing die were bonded in this module to obtain the 15% average performance improvement, which is equivalent to an advanced node. After this announcement, Intel discussed the use of hybrid bonding technology for its product roadmaps for high-end performance enhancements without the need to wait for the next advanced node transistor release. The ability to use advanced packaging technology to achieve the equivalent performance of an advanced node – in a shorter development time – has the chiplet industry exhilarated about a ubiquitous heterogeneous integration supply chain. While the future will unveil the proliferation rate of this high-performance interconnect through the supply chain and market, it appears this is only the beginning of a new generation of packaging innovation with hybrid bonding.

References

1. Global High-Performance Computing (HPC) market size by component (solutions, services) Deployment type (on-Premise, Cloud); Report ID 6826. By server Price Band; Sept. 2022.
2. S. Naffziger, "Chiplet architecture for high-performance server and desktop products," International Solid State Circuits Conf. (2020).
3. L. Mirkarimi, A. Nuruzzaman, "A new era of computing performance with hybrid bonding," *Chip Scale Review*, July-Aug 2021.
4. "The Semiconductor and Packaging Report," Prismark Partners, 2022.
5. CHIPS: https://www.darpa.mil/program/common-heterogeneous-integration-and-ip-reuse-strategies
6. T. Hackenberg, "The chiplet market," Chiplet Summit Conf., Jan. 2023 (San Jose, CA).
7. P. Kennedy, "Intel enters a new era of chiplets that will change everything," Serve the Home Forum, Aug. 2022.
8. B. Baktha, "RISC-V and chiplets powering the next leap forward in compute systems architecture," Chiplet Summit, Jan. 2023 (San Jose, CA).
9. Y. Kagawa, et al., "Novel stacked CMOS image sensor with advanced Cu2Cu hybrid bonding," IEEE IEDM, Dec. 3-7, 2016.
10. P. Clark, "China's YMTC takes lead in 3D-NAND memory," EETimes, Technology News, April 13, 2020.
11. J. A. Theil, et al., "Analysis of die edge bond pads in hybrid bonded multi-die stacks," 72nd IEEE ECTC (2022).
12. G. Gao, et al., "Low-temperature hybrid bonding for die to wafer stacking applications," 71st IEEE ECTC (2021).
13. A. Agrawal, et. al., "Thermal and electrical performance of direct bond interconnect technology for 2.5 and 3D integrated circuits," IEEE 67th ECTC, May 2017.
14. K. Heyman, "DRAM thermal issues reach crisis point," June 9, 2022, *Semiconductor Engineering.*
15. L. Cao, "Advanced packaging technology for chiplets and heterogeneous integration," Chiplet Summit Conf., 2023.
16. L. Su, Computex 2021, May 2021; https://www.youtube.com/watch?v=gqAYMx34euU



Biography

Laura Mirkarimi is SVP of Engineering at Adeia, Inc., San Jose, California. She earned a PhD in Materials Science at Northwestern U. She leads the 3D Technology Team at Adeia and focuses on hybrid bonding, advanced packaging and thermal management technologies for future generations of electronic products. Prior to Adeia, she developed electronic devices including ferroelectric memory, transparent conductors and photonic crystal sensors at Hewlett Packard Laboratories for 12 years. Email laura.mirkarimi@adeia.com

Accelerating 3D and heterogeneous integration with high-volume D2W hybrid bonding

By Thomas Uhrmann [EV Group] and Nelson Fan [ASMPT]

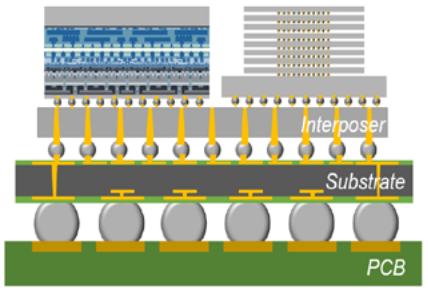
The semiconductor industry is undergoing a revolutionary transformation with the adoption of heterogeneous integration and chiplet-based design, marking a fundamental turning point. Monolithic 2D scaling options often come with complex and costly issues and limited scaling benefits for a system. Chiplets, therefore, are an inevitable solution to meet the demands of the scaling roadmap and performance, power, area-cost and time-to-market (PPACt) requirements. High-performance applications, including artificial intelligence (AI), augmented/virtual reality, and autonomous driving, require specialized processors for each task, making chiplet integration necessary. This design approach is already being used in various forms, from hybrid bonding to 2.5D interposers, and is equally critical for consumer and mobile devices to keep up with performance and flexibility requirements.

The shift to chiplet integration, however, requires a complete overhaul of the semiconductor manufacturing process. While 2D transistor scaling remains relevant, the rising costs and complexity of scaling have prompted the industry to embrace 3D and heterogeneous integration. This approach involves assembling and packaging different components or dies with varying sizes and materials into a single device or package, thereby enhancing performance on new device generations that support these new applications and leading to more precise and customized mapping of customer and application requirements.

There are two different chiplet approaches: partition and add-on. Which approach is used depends on the application and purpose. The partition scheme involves breaking down the original monolithic die into two or more smaller chiplets and stacking them on top of each other in a 3D-integrated circuit (IC) configuration. In contrast, the add-on scheme uses a base die as one of the chiplets with little to no partitioning and adds another chiplet (or multiple chiplets) with additional features, such as extra memory. The add-on chiplets are stacked above or below the original monolithic chiplet in the 3D-IC configuration [1]. The partition scheme is focused more on cost and footprint savings, while the add-on method is geared towards performance and power improvement, such as for high-performance computing applications. Cost savings are greater as the monolithic 2D die area increases and wafer costs become more expensive. In both cases, savings are optimized when the partitioned two chiplets have the same size because it improves the yield for each die individually. In addition to cost savings, partitioning is also expected to lead to effective capacity improvement due to higher yields in smaller chiplets. Technology considerations of 3D-ICs and the various component flavors of 3D-ICs are depicted in Figure 1.

Heterogeneous integration relies heavily on wafer-to-wafer (W2W) hybrid bonding, which involves stacking and electrically connecting wafers from different production lines. This process has proven successful for complementary metal-oxide semiconductor (CMOS) image sensors and various memory and logic technologies. W2W hybrid bonding has been mature for over a decade, with equipment and process now well established. It enables contact pitch of less than 1µm in production, but die size and grid matching are required. Each bonding layer consists of only one node, and cumulative yield decreases the overall stack yield for high layer count. However, W2W bonding offers high-throughput capabilities.

Die-to-wafer (D2W) hybrid bonding is a relatively new technology, and its process and equipment maturity are still evolving, resulting in many challenges. The contact pitch for this bonding method is currently at 9µm in production, but this is expected to decrease rapidly to 2µm. One advantage of D2W bonding is that there are no limitations on die size or system segmentation. Additionally, chiplets of different nodes can be combined, providing a high level of flexibility. However, binning may be necessary because of the varying yields of individual dies. The throughput of D2W bonding is dependent on the size of the chiplets and the number of chiplets integrated into a system.



L0	3D stacking	D2W | W2W
	Interconnect	Hybrid Bonding | µBump
	Technology	Same Node | Different Node
	Die Size	Bottom > Top | Same Size | Top > Bottom
L1	Die Direction	Face Up (F2B) | Face Down (F2F)
	Integration	Die First | Die Last
	Interconnection	TSV middle | TSV last | Via height
L2	Package Type	Fan Out WLP | Interposer | Laminate

Figure 1: Heterogeneous integration and connection options along different packaging levels from chip- to board-level.

There are several D2W bonding methods available for heterogeneous integration, each with its own advantages and disadvantages, as shown in Table 1. Selecting the best approach for a given application depends on factors such as die size, thickness, total stack height, and interface considerations like contact design and density.

	Hybrid W2W Bonding	Hybrid D2W Bonding
Maturity	Wafer bonding equipment and process have been mature since 2010	Process and equipment maturity is starting to yield but still many difficulties
Contact Pitch	<1µm pitch is enabled in production	Currently 9µm pitch in production
Die Size	Die size and grid matching required	No limitations in die size and system segmentation
Segmentation	Each bonding layer consist of one node	Each chiplet can consist of a different node
Yield	Cumulative yield of each bonded layer	Cumulative yield can be avoided by binning
Throughput	>25 bonds per hour	Related to chiplet size and amount of chiplets per system

Table 1: Comparison between W2W and D2W hybrid bonding according to main decision criteria.

Collective D2W bonding process

Collective D2W bonding involves the bonding of multiple dies onto a wafer substrate in a highly accurate and reliable manner. The collective D2W bonding process typically consists of several key steps, shown in Figure 2, including carrier preparation with adhesive, die protection while handling, die population using a high accuracy D2W bonder, W2W die transfer of the carrier wafer to the product wafer, and finally, debonding of the die carrier and cleaning [2,3].

The first step in the process involves preparing the carrier wafer with a suitable adhesive material. The adhesive layer should be uniform and have a sufficient thickness to provide adequate bonding strength. The carrier wafer should also be compatible with the adhesive and should have a surface that can be easily cleaned and prepared for bonding.

In the next step, the dies are protected while being handled to prevent damage or contamination. This may involve the use of specialized handling equipment or techniques, such as vacuum or tweezers. The dies should be handled with care to avoid any potential damage, which can result in yield loss and reduced device performance.

Once the dies are protected, they are populated onto the carrier wafer using a high-accuracy D2W bonder. The bonder should be capable of achieving sub-micron alignment accuracy, and should also be able to handle a high volume of dies for efficient production. The next step involves transferring the dies from the carrier wafer to the product wafer. This is typically achieved through W2W bonding using high-precision alignment and bonding equipment. The bonding process should be performed under controlled conditions to ensure uniformity and reliability. After the dies have been transferred to the product wafer, the die carrier is debonded and removed. This involves separating the adhesive layer from the carrier wafer and cleaning any residual adhesive from the product wafer. The cleaning process should be carefully controlled to avoid damage to the dies or product wafer.

Direct placement D2W hybrid bonding

Direct die placement D2W hybrid bonding involves picking and placing a die onto a target wafer, followed by annealing to covalently bond the dies and electrically connect them [4]. The first step in this process is selecting a suitable die carrier, as shown in Figure 3. Depending on the requirements of the application, the carrier can be a film frame or a specially-designed and fabricated die carrier. The selection of the carrier should be based on factors such as die size, die thickness, and the number of dies to be bonded.

The next step is the carrier picking, which can be done from a completed singulated wafer or a reconstituted carrier consisting of different dies. Once the die has been picked, it is necessary to activate and clean the surface before bonding. Plasma activation is used to remove any contaminants and rehydrate the surface of the die, ensuring good bonding. The plasma activation process is critical to ensure strong bonding between the die and the wafer.

A high-accuracy pick-and-place process is essential for ensuring a high alignment accuracy with less than 200nm on opposite corners of the die. This high-accuracy pick-and-place process is achieved using advanced equipment and technologies. The controlled bond wave is initiated by contacting the die center, ensuring a stable and uniform bond between the die and the wafer.

	Co-D2W
Transfer Method	Collective die transfer by reconstituted carrier
Pros	• Proven technology • Die activation and cleaning equivalent to W2W hybrid bonding • Oxide management • Rework on carrier feasible
Cons	• Error propagation of D2W + W2W alignment • Cost of carrier prep, utilization and clean • Die thickness needs to be in narrow range
Maturity	Limited volume production proven for several years

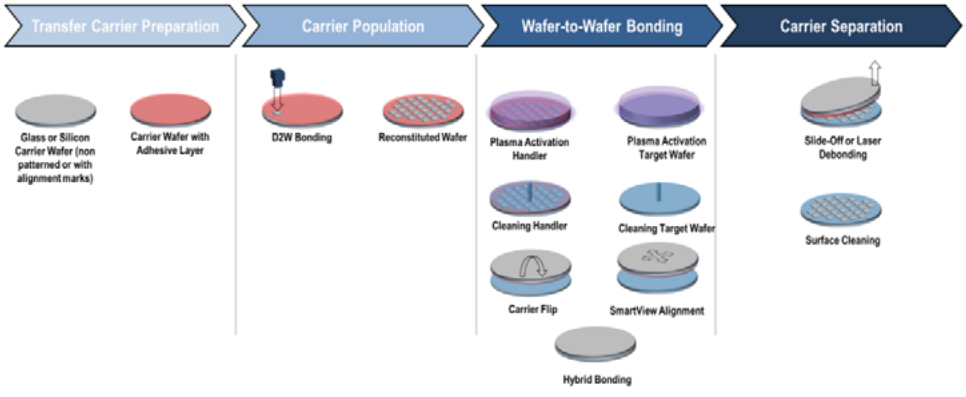


Figure 2: Collective D2W hybrid bonding process flow.

	DP-D2W
Transfer Method	Direct placement of activated dies using flip-chip bonder
Pros	• Versatile method • Die thickness invariant
Cons	• Bonding interface needs to be touched • Die handling especially for multi die stacks such as SRAM, DRAM • Particle management during die placement
Maturity	Scaling and in low-volume production

Figure 3: Direct placement D2W hybrid bonding process flow.

Reconstructed D2W hybrid bonding

A recent publication introduces a novel integration method called reconstructed D2W bonding (Figure 4) that combines direct and collective placement D2W bonding [5]. The direct placement approach is used to mechanically attach the dies to a carrier wafer, but it only provides a mechanical connection.

Once the chiplets are formed, they are permanently attached to a carrier wafer with the die either facing up or down, and the gap between them is filled with silicon oxide, which is a front-end compatible version of fan-out wafer-level packaging (FOWLP) that is inorganic. The challenge in this step is that the oxide thickness needs to be significantly higher to support the overall chiplet height, including the silicon substrate and metal interconnects. Afterward, the dies and oxide layer are planed thoroughly, and through-dielectric interconnects and hybrid bond pads are created at the wafer level. The actual hybrid bonding and electrical contact are done later in a W2W bonding process. Cleanliness must be strictly maintained throughout the process.

The major advantage of this process is its full front-end fab compatibility, and there are no materials that are incompatible with the fab present throughout the process flow. However, one of the main challenges is controlling and optimizing the oxide fill process. Analogies to FOWLP apply, where silicon content, die thickness, deposition temperatures, and oxide properties all impact the wafer shape and contact pitch scalability of the W2W hybrid bonding process. Despite this challenge, reconstructed D2W bonding has shown great promise in achieving full front-end fab compatibility while maintaining the highest level of cleanliness throughout the entire process flow. Further research is needed to optimize the oxide fill process and to address other potential challenges that may arise.

Self-aligned D2W integration

Currently, research institutes are focusing on self-aligned die bonding, which follows similar key steps to traditional die bonding such as cleaning and activation of the dielectric interface and copper bonding pad. Two guiding principles are being explored for self-aligned die bonding: 1) shaping the die to achieve ultra-precise dimensions; and 2) defining guiding pads on the die surface using hydrophilic and hydrophobic regions patterned with optical lithography at the wafer level. However, the singulation, cleaning, and activation processes still require the same level of precision as traditional die bonding. The placement of dies in self-aligned die bonding can be coarser, leading to potentially lower process costs, but the increased complexity in die preparation, guiding pad definition and combination with crucial chemical mechanical planarization (CMP) processes, must also be considered. Self-aligned die bonding has high potential, but further research and development is necessary to improve the integration process [6].

D2W bonding equipment status

The preparation and conditioning of surfaces for direct placement fusion and hybrid bonding of dies on wafers is a critical step. Challenges related to

	Reconstructed W2W
Transfer Method	W2W hybrid bonding using anorganic filled reconstructed wafer(s)
Pros	• Proven technology • Die activation and cleaning equivalent to W2W hybrid bonding • Oxide management • Rework on carrier feasible
Cons	• Error propagation of D2W + W2W alignment • Cost of carrier prep, utilization and clean • Die thickness needs to be in narrow range
Maturity	Scaling and in low-volume production

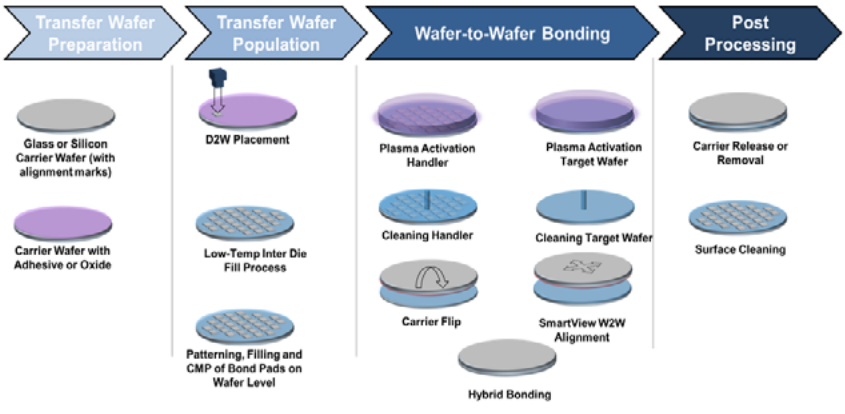


Figure 4: Reconstructed D2W hybrid bonding process flow.

cleanliness and activation are similar to those of other fusion bonding techniques. The dies require repopulation on a dedicated cleaning carrier wafer and may need optional cleaning during dicing before transport to the front-end clean hybrid bonding step. Surface coating on a readily CMP-treated wafer is crucial to preserve surface properties and cleanliness during dicing.

The EVG320 D2W is a flexible die preparation and activation system designed to seamlessly integrate with ASMPT's pick-and-place die bonding systems. It is equipped with a universal hardware/software interface and can be used as a stand-alone system. The system incorporates cleaning and plasma activation technology, and features an optional integrated metrology module that provides direct feedback to the die bonder on critical process parameters, such as die placement accuracy and die-height information, for improved process control.

After die preparation, the high-precision D2W die pre-bond step is carried out using ASMPT's LithoBolt—an automatic ultra-high-precision die bonding system. The system is designed to achieve high accuracy, throughput, and yield for volume production. Material handling for cleanliness control and alignment mechanism to achieve target alignment accuracy are crucial in this step. The cleaned and activated die and target wafer materials from the die preparation machine are transported to the load port of the equipment front-end module (EFEM) of the die bonding system through a cleaned front-opening unified pod (FOUP). The die is then bonded onto the target wafer under a compression force of 0.05 to 0.3MPa with a special curved collet. The force-controlled bond-arm of the bonder adjusts the compression force for bonding. Cleanliness must be well controlled throughout the entire material handling and bonding process inside the EFEM and bond chamber.

Die placement alignment accuracy is another crucial requirement in the die bonding step. Currently, the industry is calling for alignment accuracy below 200nm at 3σ for bond pad sizes below 10µm, and future bonding solutions should have accuracy down to 100nm or below. LithoBolt utilizes a new concept of alignment approach supported by a powerful optical system to assist in sub-micron-level pattern-recognition alignment. The bond head module incorporates intelligent design to enable true active alignment with real-time compensation. The alignment accuracy verification was performed using chip-on-glass (COG) for face-down mode bonding, and overall results achieved under 3σ were 106nm and 103nm (X-direction) and 131nm and 147nm (Y-direction) for the two corners accuracy as shown in **Figure 5**. Cleanliness assessment data showed that LithoBolt achieved ISO3 cleanroom standard both in idle and operation modes.

Accelerating process development

Manufacturers must undertake extensive development projects to determine the optimal bonding method for their devices. These projects must

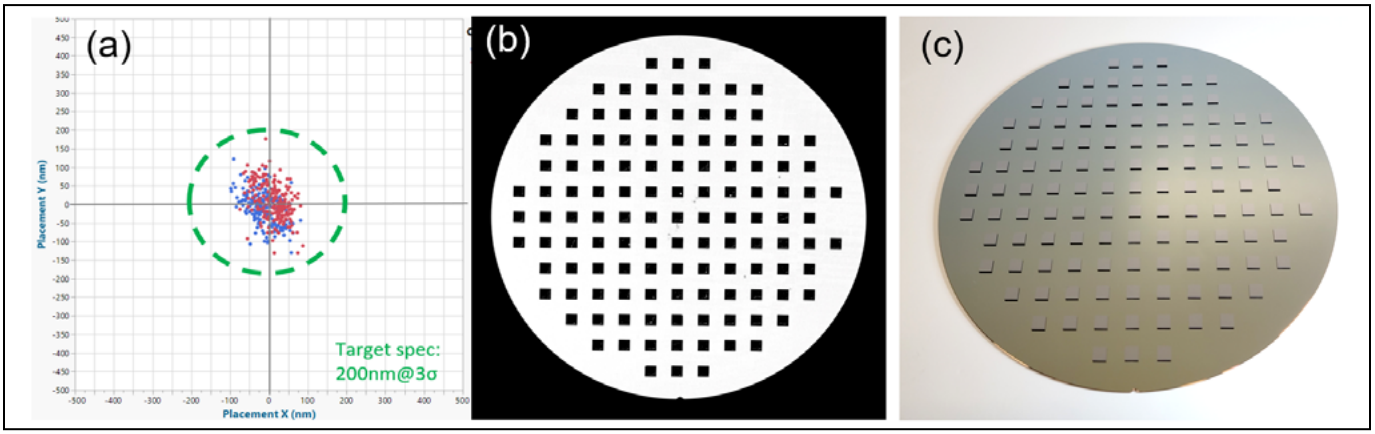


Figure 5: Recent DP-D2W bonding results using EVG320 D2W activation and cleaning followed by ASMPT LithoBolt die bonding: a) Alignment verification result using COG (9x9mm glass chip), face-down mode; achieving specification below 200nm at 3σ; b) post-bond scanning acoustic microscope; and c) photography of the bonded dies on the wafer.

consider not only the wafer bonding equipment, but also the materials involved in temporary and permanent bonding, as well as related processes such as die activation and cleaning, as well as subsequent die bonding. Process expertise and access to cutting-edge technologies are essential, but these systems are often already in use at customer sites and may be difficult to access for research and development purposes. To overcome these challenges, EVG established the Heterogeneous Integration Competence Center (HICC), which leverages EVG process solutions and expertise to enable new and improved products and applications driven by advancements in system integration and packaging, using ASMPT's latest-generation die bonding equipment. ASMPT has also established an advanced lab in Hong Kong, focused on overlay and including EVG's die cleaning and activation capabilities. These incubators were established to lower the barriers to development for customers.

Summary

The utilization of D2W hybrid bonding is crucial for the swift adoption of 3D/heterogeneous integration and the development of next-generation devices that offer superior performance, high bandwidth, and low power consumption. Even though the infrastructure for D2W hybrid bonding is still evolving, an increasing number of process solutions and collaborations across the supply chain are emerging and will be integral in establishing the best practices for D2W hybrid bonding. Close cooperation and seamless optimization between equipment design and process integration in appropriate testing labs are necessary for qualifying and refining D2W hybrid bonding.

Acknowledgements

The authors express their gratitude to Jürgen Burggraf and Mariana Pires from EV Group, as well as Hoi Ping Ng, Ming Li, and Siu Cheung So from ASM Pacific Technology, for their valuable contributions to this paper.

References

1. IEEE Heterogeneous Integration Roadmap, Ch. 2.
2. J. Burggraf, M. Pires, T. Uhrmann, "Collective die bonding - an enabling toolkit for heterogeneous integration," PRiME 2020 (ECS, ECSJ & KECS Joint Meeting), 2020.
3. T. Uhrmann, et al., "D2W hybrid bonding using high-accuracy carrier solutions for 3D system integration," 2023 IEEE ECTC.
4. C. H. Fan, et al., "Direct die-to-wafer Cu hybrid bonding for volume production," 2023 IEEE ECTC.
5. A. Elsherbini, et al., "Enabling next-generation 3D heterogeneous integration architectures on Intel process," 2022 International Electron Devices Meeting, pp. 27.3.1-27.3.4.
6. A. Bond, et al., "Collective die-to-wafer self-assembly for high alignment accuracy and high-throughput 3D integration," 2022 IEEE ECTC, pp. 168-176.

Biographies

Thomas Uhrmann is Director of Business Development at EV Group, Austria. He is responsible for overseeing all aspects of EVG's worldwide business development. Previously, he was Business Development Manager for 3D and advanced packaging, as well as compound semiconductors and Si-based power devices at EV Group. He holds an Engineering degree in Mechatronics from the U. of Applied Sciences in Regensburg, and a PhD in Semiconductor Physics from the Vienna U. of Technology (TU Wien). Email: T.Uhrmann@EVGroup.com

Nelson Fan is Vice President of Business Development, Advanced Packaging Technology, at ASM Pacific Technology (ASMPT), Hong Kong SAR, China, where he focuses on high-precision bonding and pick-and-place solutions for advanced packaging. Prior to joining ASMPT, he held multiple senior management roles in R&D and manufacturing in both OSAT and design houses. He has more than 40 US patents in semiconductor packaging technologies and holds a Bachelor's degree in Electrical Engineering from the U. of Colorado at Colorado Springs, USA.

Heterogeneous integration and chiplets and dielets: why the hype?

By Subramanian (Subu) Iyer [The UCLA Center for Heterogeneous Integration and Performance Scaling (UCLA CHIPS), Samueli School of Engineering, University of California, Los Angeles]

Over the last few years, there has been an over-abundance of attention paid to packaging driven by heterogeneous integration and chiplets. Actually, heterogeneous integration is not new – just look at any printed circuit board (PCB) and you will see a diversity of chips assembled and connected to one another on the board. Moreover, if you look closely at the chips that have been assembled, some of them are quite small—and if chiplet is the diminutive of chip, many of them would certainly qualify. In this article, I'd like to explore what makes the concept of heterogeneous integration and chiplets such a powerful concept in the context of recent developments in electronics packaging.

I like to think of the role of packaging to be akin to the role of the Los Angeles Police Department with the motto, "to protect and to serve." The package protects the chip from mechanical shock, environmental and corrosive ingress, and from thermal excursions, hotspots, and thermal runaway that the chip's operation may cause. The package serves the chip and system by supplying it efficiently with power—potentially in multiple voltage domains. The package also electrically connects the chip to other chips and the outside world with high efficiency, high bandwidth, and low latency connections. Equally importantly, the package provides a stable and well-controlled environment where the chip can be tested, and its functionality, performance, and reliability can be ensured.

Complementary metal-oxide semiconductor (CMOS) scaling, driven by the economics of miniaturization and manufacturing scale has, over the last several decades, allowed us to reduce the cost per transistor while simultaneously improving its performance – switching speed and power – at a smaller transistor footprint. This has fueled the well-known Moore's "Law" for miniaturization. While initially this scaling was straightforward, driven by Dennard's constant electric field scaling, and going to larger diameter wafers, since the early 2000s we have had to invoke new materials, strain engineering, and other innovations to keep up the "cost-performance per transistor" expectations of Moore's Law.

Figure 1 shows the relentless scaling of silicon CMOS (in blue on a log scale). In contrast, package scaling (shown in red on a linear scale) – measured, say by bump pitch (though any other metric could be used), has been the tortoise in this race. This has led to an apparent paradox illustrated in **Figure 2** where the die size of high-performance chips has been increasing even though the transistor size has been shrinking. In fact, today's high-performance dies may be as large as a reticle field.

The reason why we would like to grow die size has to do with the huge penalty of off-chip communication driven by the relatively modest scaling of packaging as shown in **Figure 1**. Making dies larger would minimize the number of times we would need to go off chip (although, this comes at the cost of huge wiring congestion, leading in turn, to as many as 19 wiring levels) and this led to Gene Amdahl's idea in the 80s to build a wafer-scale chip at the company he started, called Trilogy Systems. However, low yield (those were the days of 100mm-diameter wafers and ECL circuits) and the inability to manage long-distance communication on the wafer caused Trilogy to abandon that idea, only to be revisited recently by Cerebras Systems. Cerebras addressed the yield issue by making very small high-yielding cores that were pretested on the wafer and subsequently connected using a post-fab wafer-level interconnect

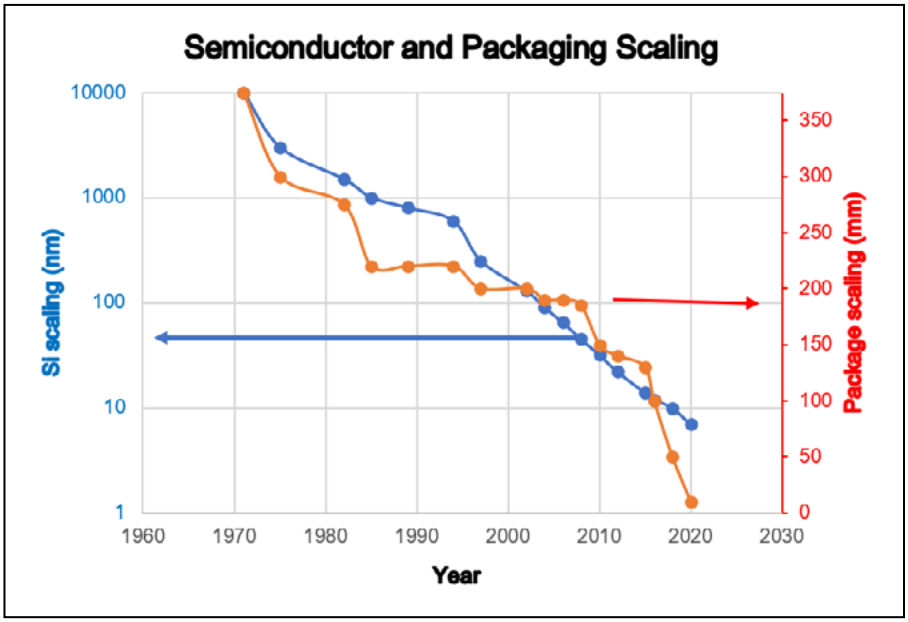


Figure 1: A comparison of CMOS scaling (blue, left axis, log scale) vs. time with package scaling (red, right axis linear scale). The typical feature size was used as the scaling metric.

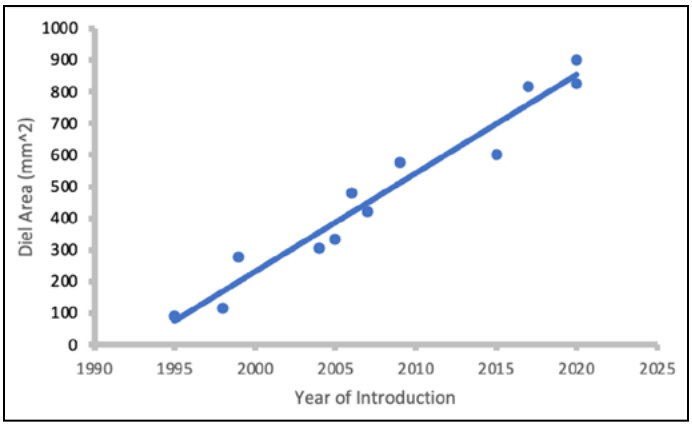


Figure 2: Typical die size for high-end microprocessors including graphics processing units (GPUs) as a function of year of introduction. For reference, the Intel Pentium introduced at the 0.8µm node was 300mm^2 and had about 3 million transistors. In 2021, the Nvidia A100 made at the TSMC 4N node was about 836mm^2 with about 54 billion transistors.

layer and requiring only nearest neighbor connections (sometimes called a systolic architecture). However, both these attempts employed a monolithic bipolar junction transistor (BJT) or CMOS technology and could not, for example, use high-density dynamic random access memory (DRAM) for its memory. Another thing to remember is something called Rent's rule. This rule suggests that as chip complexity increases, the number of I/Os should also increase as the more complex larger chip would need to send and receive larger amounts of data with lower latencies.

Tackling the package scaling challenge

Because package scaling has not kept up, chip designers have resorted to serializers and de-serializers to get around the problem. As shown in **Figure 3**, many signals are serialized and sent over the relatively few board channels at extremely high speed. We are reaching data rates of over a 100Gb/s per channel. However, sending signals at this high rate presents several challenges that result in these I/Os taking up more real estate (as high as 30-40%, or even more, of the chip area) and using up to 30-40% of the chip's power (chip area and power are correlated). This is unsustainable—and the poor control of wires on a PCB, surface roughness and

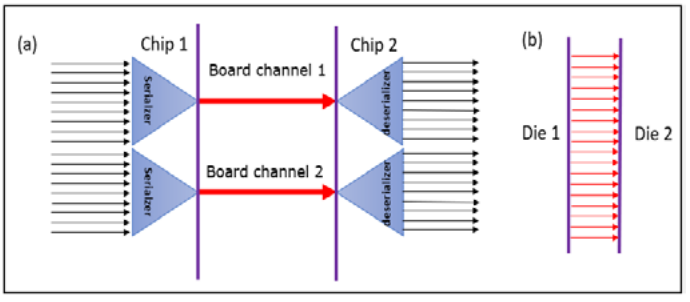


Figure 3: Inter-die signals can be handled in two different ways: a) Because of fewer inter-chip connections on PCBs, signals are serialized and sent at high speed over transmission lines on a PCB. They are deserialized at the receiving chip. These SerDes are large and complex and can consume as much as 40% of the chip power. b) If we had more wires between the dies, and the dies are closely spaced, we can send the native signals as-is over wires that resemble wires on a chip.

the skin effect only compound these difficulties. The dual realization that silicon scaling had slowed down and packaging had left much performance on the table has led to a feverish amount of work on how to address the packaging problem, and, as seen in **Figure 1**, there has been a significant scaling of key packaging metrics. For example, at UCLA CHIPS, we routinely connect dies to the substrate at sub-10µm pitches.

Why has it taken so long to address the problems noted above? Primarily, silicon CMOS scaling was more or less a sure thing, and one could get more performance and transistor density by scaling. Packaging, on the other hand, did not provide this kind of value and very often was an afterthought. The focus of classical packaging, therefore, was not on value, but more on cost. Now that calculus has changed and there is realization that, especially in high-performance computing, advanced packaging can provide extraordinary value. This value comes primarily from the increased bandwidth and reduced latency of inter-chip communication at lower energy per bit transmitted. This is accomplished in three ways: 1) Decreasing the pitch at which the dies are connected to the substrate (the so-called bump or pillar pitch); 2) Reducing the wiring pitch (also called trace pitch) on the substrate to sub-µm dimensions; and 3) Reducing the die-to-die separation on the substrate to sub-50µm dimensions. A major reason we have made such dramatic progress is that we have borrowed immensely from decades of silicon technology including the use of silicon as the substrate. It's no surprise that silicon fabs lead in this segment of packaging! It turns out that silicon is an incredibly versatile packaging material.

Si wafers are unbelievably flat and thermally-matched to the silicon dies. This allows for fine lithography (i.e., it's very difficult to write on a warped surface). The thermal conductivity of silicon is about a third that of copper, and unlike an organic PCB, offers a viable heat extraction path (typically only 10% of the power is extracted via the organic PCB). This is important as heat extraction is a serious issue for high-performance systems. Two arguments that have been made against using silicon are cost and the brittleness of silicon wafers.

Silicon with a few layers of fine-pitch wiring are indeed cost effective compared to FR4 boards where even sub-10µm pitch wiring is difficult to achieve (silicon becomes very expensive when you build tiny FinFETS or nanosheet transistors with 20 wiring levels). The cost can be reduced further by going to metallurgical-grade silicon used in the photovoltaics (PV) industry. Handling techniques in silicon fabs where breakage is almost non-existent can also be used in packaging. The elimination of solder in the assembly process has allowed us to scale the "bump" pitch. Solder is needed in classical packaging to provide compliance to the easily warped

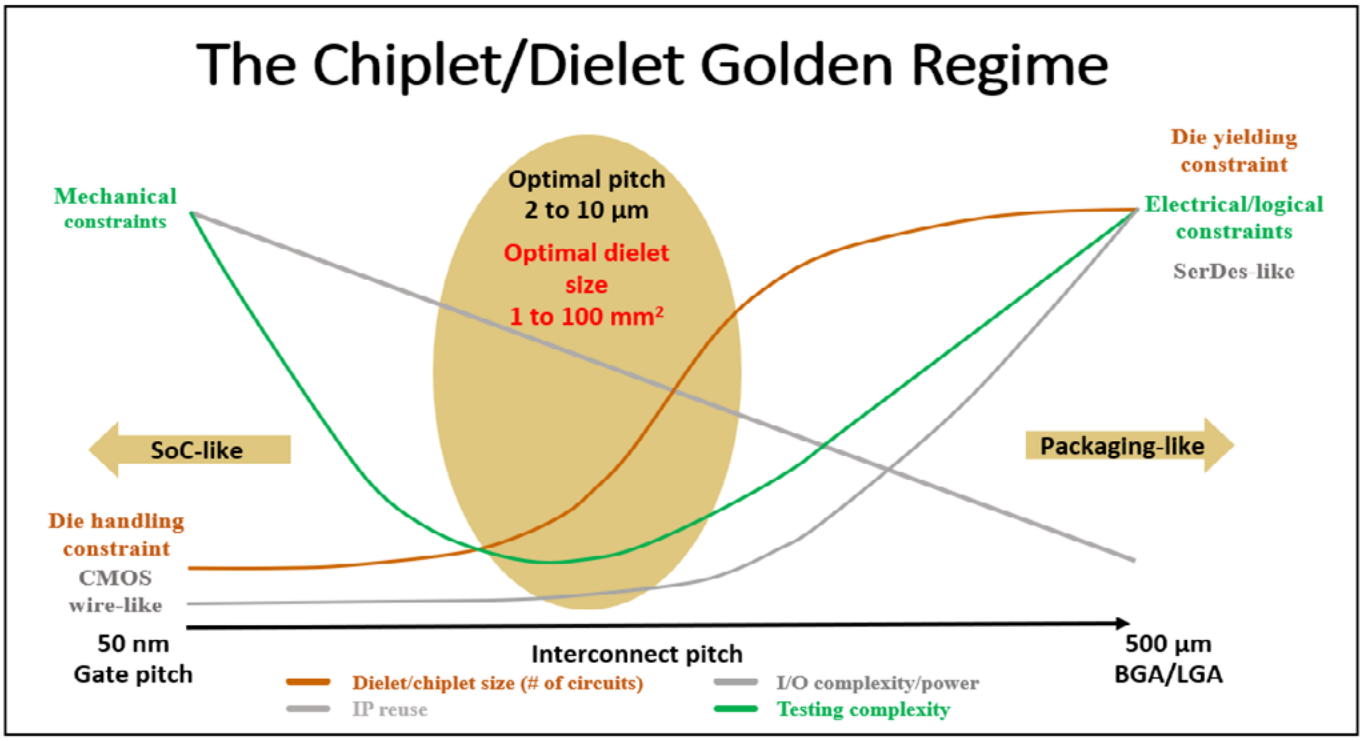


Figure 4: The chiplet golden regime has dielets of 1 to 100mm^2 that are connected to the substrate at a pitch of 2-10µm. There are many considerations that go into this analysis including dielet yield, dielet IP reuse, testing and I/O complexity.

organic substrates. With silicon substrates –where the metal lines and pillars are built using precision chemical mechanical polishing (CMP) – the pillars are co-planar and direct metal-metal bonding is possible with high yield. The elimination of solder also eliminates the effects of solder intermetallics and a host of reliability issues that they entail. We can also eliminate the use of underfill and can use thin (a few nm) atomic layer deposition (ALD) inorganic films such as Al_2O_3 to passivate and prevent moisture ingress.

Figure 4 shows a prototypical wafer-scale system where dielets are assembled at fine pitch with dies close together and passivated. A notable feature of this wafer-scale system and the homogeneous technology versions discussed earlier is the potential for heterogeneity. We can, therefore, assemble both memory dies (or even die stacks), logic dies, I/O dies and different kinds of accelerators, analog and mixed-signal dies, RF and such, on this substrate—and all these dies are connected at fine pitch. This capability allows us to optimize the technology nodes, material systems, and functionality in a manner that

is not possible in homogeneous monolithic chips, no matter how big. Furthermore, it is possible to scale these systems to wafer scale allowing us to pack high-compute power in very small form factors. We would not be able to do this were it not for the fine bump and trace pitch and the close die placement. The fine-pitch capability combined with the ability to place dies in close proximity, allows these heterogeneous systems to behave as if they were monolithic. This last point is crucial and differentiates today's heterogeneous systems from those of yore. One thing we must mention is that dielets, today, are assembled on a substrate by mechanical placement and this is a limitation on overlay tolerance (currently ~+/-1µm), but we expect both steady improvement in placement accuracy and perhaps true lithographically-defined ways of connecting the dies.

The newfound capability described above changes the way in which we can build system on chip (SoC) structures. Today's SoCs are typically synthesized with soft intellectual property (IP) design blocks using high-level synthesis tools. Over 80% of these IP blocks have been instantiated on some other SoC or on test sites. However, every time we tape out a new SoC, we re-synthesize these very IP blocks in new configurations, reduce them to schematics, and finally, lay them out physically and re-harden them. This is a time-consuming and expensive task and contributes enormously to the so-called non-recurring engineering (NRE) of building a new chip. Bigger chips take longer and cost more and limit the number of players who can play this game. In 2016, I published a paper [1] where I suggested that we could pre-build a bunch of commonly used dielets and stock them, so that we could build a complex SoC by assembling these prefabricated dies using these advanced packaging constructs. This vision is slowly, but surely, taking root.

Heterogeneous integration

Based on the above sections, let's explore this heterogeneous integration space. The goal is to determine the optimal dielet size and vertical interconnect pitch as shown in **Figure 4**. Let's start with the x-axis, which captures the pitch at which the dies are connected to the substrate. The right-hand side is bound by the ball grid array (BGA) pitches on today's PCBs (~100s of µm) and the left-hand side represents the contacted gate pitch at which transistors on the chip may be connected (~10s of nm). The y-axis depends on what is being plotted.

Let's consider the die size shown in brown. Large dies yield poorly, and as we decrease the die size, the yield increases; but below about 100mm^2, the yield increases—though slowly (this is why memory dies are typically this size). As the die size goes below 1mm^2, the yield approaches 100% in mature technologies (in fact, at these dimensions, we can blind-build these dies and expect them to work) and so the optimal die size is expected to be between 1 and 100mm^2. If dies go below 1mm^2, two effects begin to dominate: die handling becomes difficult (both for assembly and test), and kerf loss becomes appreciable.

Another consideration is IP reuse. Generally speaking, small dies have less content and are less likely to be specialized.

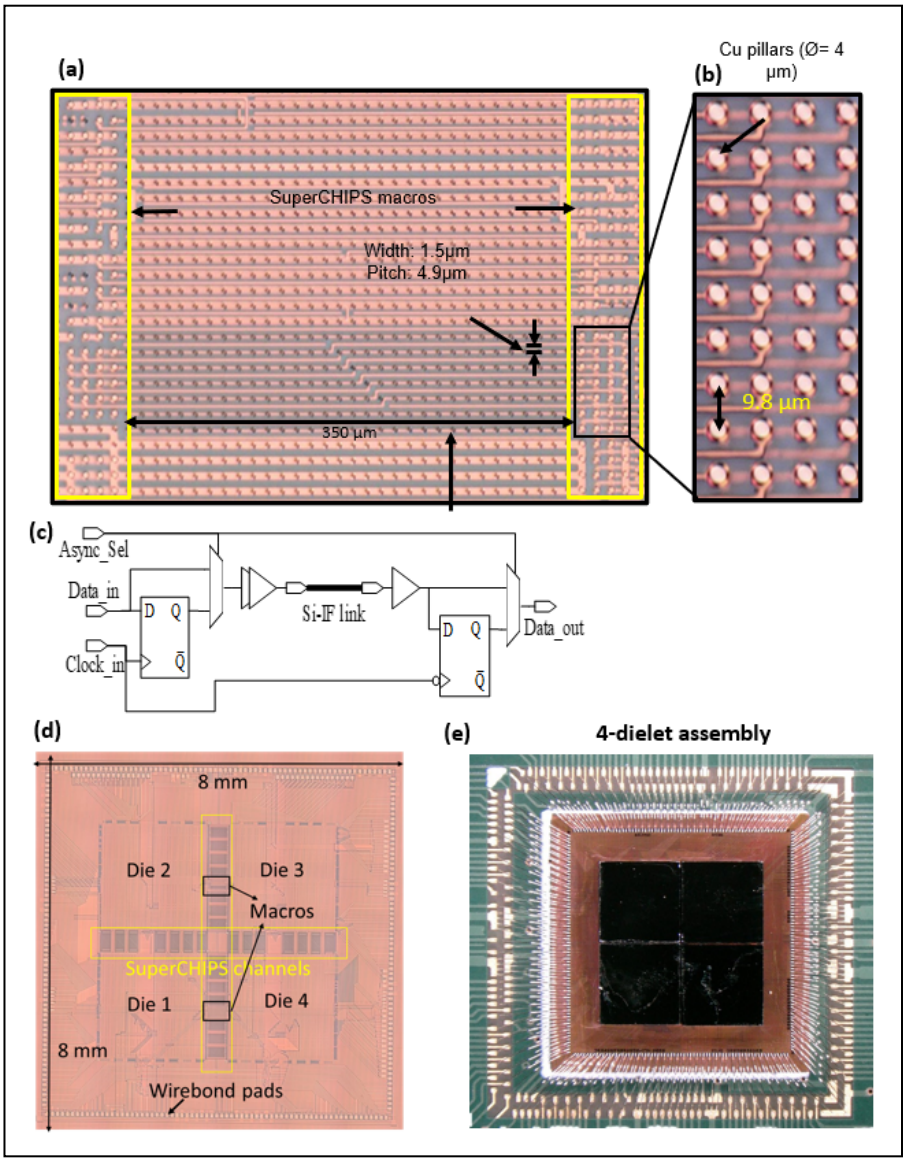


Figure 5: Various aspects of the Silicon Interconnect Fabric and SuperCHIPS platform. a) A chip micrograph showing the SuperCHIPS macros that drive fine-pitch connections between adjacent dies; b) A 4-deep row of pillars at 10µm pitch yields an effective "bump" pitch of 2.5µm. c) The I/O drivers are small and must be contained at the 10µm pitch, and consist of a cascaded set of inverters with the capability of both synchronous and asynchronous modes. d) The silicon interconnect fabric (Si IF) is a full-thickness silicon wafer with inorganic build-up layers similar to the backend wiring of a 90nm CMOS technology structure. e) An assembly of four 1mm^2 UDSP functional dielets on the Si IF at 9.8µm "bump" pitch, sub-2µm trace pitch, and about 30-50µm inter-die spacing.

This means the probability of reuse is high, while the opposite is true for large die. This trend is shown in the grey line. Testing complexity – shown in the green line – decreases as die content reduces up to a point, but very small dies present testing challenges. Finally, as dies become smaller, there is less space to accommodate pads (for testing or I/Os). This means the I/O sizes need to shrink, and by implication, the "bump" pitch as well. Large dies can also use small bump pitches, but smaller dies, or dielets, cannot afford to do so because there is not enough space available. Smaller dielets generate less data and can do with fewer I/Os. These arguments are explained in much more detail in [2].

If you put the considerations noted above together, you find there is a sweet spot – a golden dielet regime – where the dielet size is about 1-100mm^2 and the connection pitch to the substrate is between 2 and 10µm. These dimensions are not a coincidence. They can be derived from the size of typical I/P blocks on an SoC and the pitch of the wires connecting them—typically the fat wire vias.

Reducing the "bump" pitch and the inter-die spacings has another beneficial consequence. The wire inductance between chips becomes negligible and the wire behaves not as a transmission line, as is the case in PCBs, but as a simple capacitive load similar to wires on the chip. This means we do not need to worry about terminations, reflections, inter-symbol interference, and such. The I/O, therefore, becomes very simple—just a bunch of cascaded inverter-drivers. It also turns out that we have a lot more wires available in this scenario compared to a PCB. So, we do not need to serialize the data and we can send it in its native parallel format, at much lower speeds (think DRAMS that activate an entire row, but we need to serialize this data to send it off-chip and at very high speeds). This reduces power significantly, as shown in **Figure 3b**. We have proposed a simple hard protocol called "SuperCHIPS" (the catchy acronym Simple Universal Parallel intERface for CHIPS) that allows one to do this and a figure of merit that allows us to compare different protocols beyond merely energy per bit, but also allows for the accounting of I/O area, drive strength, and bit error rate (**Figure 5**) [3].

Many of the concepts discussed above (i.e., <10µm bump pitch, 2µm traces, 50µm die-to-die spacing, 1mm X 1mm dielets all assembled on a silicon interconnect fabric and simple communication protocols) were demonstrated and verified at UCLA CHIPS in 2022 for the first time in a functional assembly that employed a scalable 1mm X 1mm DSP chiplet design (dielets were manufactured by TSMC and GlobalFoundries) as the building block [4]. It's only a matter of time before these concepts make it to the commercial market place.

To realize the full potential of the SuperCHIPS concept, however, we do need to increase the diversity of the chips used in these assemblies. Back in 2013, our group at IBM built (for Semtech) what was very likely the first commercial interposer [5] product that employed two 90nm SiGe transceiver chips and a 45nm application-specific integrated circuit (ASIC) with embedded dynamic random access memory (DRAM) and deep-

trench decoupling capacitors on a silicon interposer. (We did not call it 2.5D, which later became the abominable terminology that should be banned, in my opinion.) This work was presented at the now-defunct 3D ASIP conference in 2013 along with a product announcement. Since then, the industry has focused on large processor dies connected at ~40-50µm solder bump pitch to stacked memory dies via a silicon interposer. This is a good development that has driven the artificial intelligence/machine learning (AI/ML) applications.

Heterogeneous integration, however, is a lot more than a memory and processor chip tied together! Additionally, 500-836mm^2 dies are not chiplets or dielets either! We need to have a commercially-viable roadmap that reduces "bump" pitch, trace/wiring pitch, reduces dielet/chiplet size and inter-dielet spacing, so that dies can butt together on an extremely planar substrate. All these are difficult engineering problems, but well within the capability of the microelectronics industry. Heterogeneous integration with diversely-sourced dielets presents unique supply chain and trust challenges. For example, if you design a chip/chiplet, the foundry manufactures and delivers dies or dielets, and when singly-packaged, we call it a chip. When many chips are packaged together, however, we call it a module. All this calls for an independent packaging and test facility that can deal with multiple fabs, multiple nodes, multiple material systems, and the like, and an immensely trustworthy relationship with multiple fabs.

Today, there is no dielet marketplace (other than perhaps high-bandwidth memories [HBMs]—and try getting those!). We need to have a dielet discovery system based on aggregated IP that are selected and fabricated using a statistical usage model (e.g., which IPs occur together and how often they are used). This is going to require a lot of cooperation, transparency, and perhaps some I/P standardization. Universal communication protocols have eluded us so far. Perhaps, protocol translator dielets offer a pathway.

There are many issues that I have not talked about here, but which are very important. One issue is that many of these advanced packages do not allow for rework, therefore, redundancy and self-repair are going to be important. (We have proposed the use of utility dielets to perform these and other functions.) Another issue is thermal: as we bring the dies closer, there is no space to spread heat laterally so the heat needs to be extracted vertically. Additionally, power delivery is a challenge. There is active work going on in these areas and steady, incremental progress is being made for us to be optimistic.

Summary

To summarize, fine-pitch heterogeneous integration of a wide variety of small dielets offers a path to reduced size, lower power, and potentially lower cost at the functional system level. It's no longer about shrinking transistors, but more about miniaturizing systems with a diversity of components—not just transistors! The heterogeneous integration roadmap and its allied manufacturing roadmap, the Manufacturing Roadmap for Heterogeneous Integration and Packaging (MRHIEP) provide valuable guidance as we embark on the post-Moore's Law era. But the message is clear: feature scale down and system scale out. We will also need a more rigorous design methodology that borrows concepts, such as process design kits (PDKs) from silicon, and a predictable scaling roadmap for packaging.

Acknowledgments

I would like to thank the many colleagues in industry and academia who have been generous in teaching me about packaging (I am a relative latecomer to this field), my students, who unbeknownst to them, have also taught me well. I also wish to acknowledge DARPA, SRC, SEMI, NSF, NIST, the UCLA CHIPS Consortium members, and the UC system, who have generously sustained our work. This article is an abridged version of the short course I gave at IEDM 2022.

References

1. S. S. Iyer, "Heterogeneous integration for performance and scaling," IEEE Trans. on Components, Packaging and Manufacturing Tech., vol. 6, no. 7, pp. 973-982, July 2016.
2. S. S. Iyer, S. Jangam, B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for AI systems," in IBM Jour. of Research and Development, vol. 63, no. 6, pp. 5:1-5:16, 1 Nov.-Dec. 2019.
3. S. Jangam, S. S. Iyer, "Silicon-Interconnect fabric for fine-pitch (≤10µm) heterogeneous integration," IEEE Trans. on Components, Packaging and Manufacturing Tech., vol. 11, no. 5, pp. 727-738, May 2021, doi: 10.1109/TCPMT.2021.3075219.
4. S. S. Nagi, et al., "A 16nm 784-core digital signal processor array, assembled as a 2 × 2 dielet with 10µm pitch interdielet I/O for runtime multi-program reconfiguration," IEEE Jour. of Solid-State Circuits, vol. 58, no. 1, pp. 111-123, Jan. 2023, doi: 10.1109/JSSC.2022.3212685.
5. http://www.betasights.net/wordpress/?p=1216
6. C. Hornbuckle, S. S. Iyer, "Passive silicon interposers with deep trench decoupling for heterogeneous integration of disparate technologies," 3DASIP presentation, Dec. 2013. The charts that were presented are available from the author.

Biography

Subramanian S. Iyer (Subu) is Distinguished Professor and holds the Charles P. Reames Endowed Chair in the Electrical Engineering Department and a joint appointment in the Materials Science and Engineering Department at the University of California at Los Angeles. He is Director of the Center for Heterogeneous Integration and Performance Scaling (UCLA CHIPS). He is a fellow of IEEE, APS, iMAPS and NAI, as well as a Distinguished Lecturer of IEEE EDS and EPS. He is a Distinguished Alumnus of IIT Bombay. He received the IEEE Daniel Noble Medal for Emerging Technologies in 2012, and the 2020 iMAPS Daniel C. Hughes Jr. Memorial award, and the iMAPS Distinguished Educator Award in 2021. Email s.s.iyer@ucla.edu

Hybrid bonding bridge for chiplet design and heterogeneous integration packaging

By John H. Lau [Unimicron Technology Corporation]

The most important advantages of chiplet design and heterogeneous integration packaging, such as chip partition (driven by cost and technology optimization) and chip split (driven by cost and semiconductor manufacturing yield) are cost, cost, and cost [1-6]. Unfortunately, the most important disadvantages of chiplet design and heterogeneous integration packaging are the increase in the size of the package and the complexity of the package structure—both of which lead to higher packaging cost. The higher cost is because of the additional package area and the amount of packaging effort to design and manufacture the interfaces (so-called bridges) between those chiplets, which are the focus of this brief article.

An important consortium concerned with bridge technology is the Universal Chiplet Interconnect Express® (UCIe®). According to the consortium's website, the organization addresses customer requests for a more customizable, package-level integration—combining best-in-class die-to-die interconnect and protocol connections from an interoperable, multi-vendor ecosystem. This new open industry standard establishes a universal interconnect at the package level. The UCIe® board of directors and leadership (promoters) include founding members ASE, AMD, Arm, Google Cloud, Intel Corporation, Meta, Microsoft Corporation, Qualcomm Incorporated, Samsung Electronics, and TSMC, along with newly-elected members, Alibaba and NVIDIA.

In [7], Intel published the UCIe® 1.0 specification, which provides a complete standardized die-to-die interconnect with physical layer, protocol stack, software model, and compliance testing. **Figure 1** shows examples of standard packaging and advanced packaging with chiplet design and heterogeneous integration. It can be seen that there are three different kinds of bridges for advanced packaging: 1) bridge embedded in organic package substrate; 2) bridge embedded in Si-interposer; and 3) bridge embedded in fan-out epoxy molding compound (EMC) with redistribution layers (RDLs). The focus of this article is on bridges embedded in fan-out EMC with RDLs. There are three different kinds of fan-out processes [8]: 1) chip-first with chip face-up; 2) chip-first with chip face-down; and 3) chip-last.

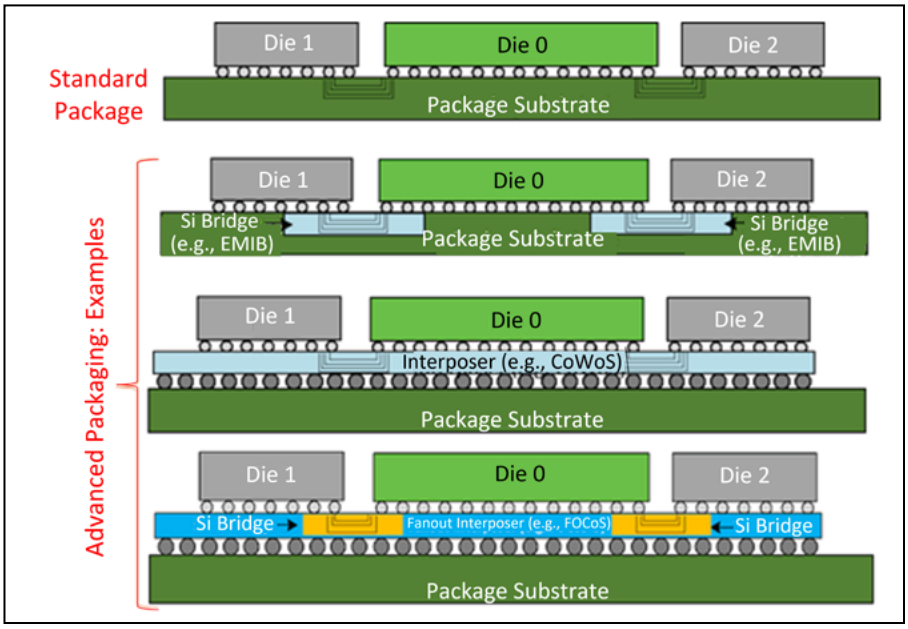


Figure 1: UCIe® bridges. SOURCE: IEEE

In addition to the three bridge designs noted above, a new kind of bridge with hybrid bonding has been proposed. There are two options in this proposal: 1) a hybrid bonding bridge with C4 bumps on the package substrate, and 2) a hybrid bonding bridge with C4 bumps on the chiplet wafer. The various examples noted in the above section are described in more detail below.

Applied Materials' bridge embedded in EMC with RDLs. Applied Materials filed its patent application on December 8, 2017 (**Figure 2**). It can be seen that the bridge is embedded using a fan-out chip (bridge) first with chip face-up [9] packaging method. The company obtained the patent (US 10,651,126) on May 12, 2020. This is the very first patent of bridge embedded in fan-out EMC with RDLs.

Unimicron's bridge embedded in EMC with RDLs. Unimicron filed its patent application on May 7, 2021 (**Figure 3**). This bridge is embedded using the fan-out chip (bridge) first with chip face-down [10] packaging method. The company obtained the patent (US 11.410,933) on August 9, 2022. This is the first bridge patent embedded in fan-out EMC with RDLs the chip-first with the chip face-down process.

IME's bridge embedded in EMC with RDLs. IME filed its patent application on March 17, 2017 (**Figure 4**). It can be seen that this bridge is embedded by using the fan-out chip (bridge) last [11] packaging method. The company obtained the patent (US 11,018,080) on May 25, 2021. This is the very first patent using a bridge embedded in fan-out EMC with RDLs with chip-last or RDL-first process.

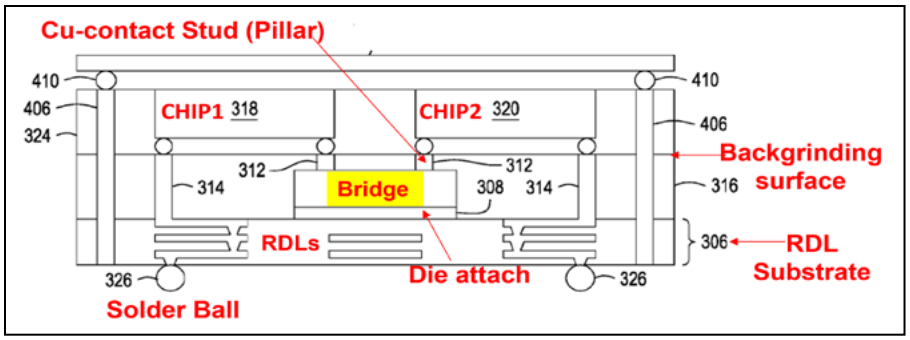


Figure 2: Applied Materials' bridge embedded in fan-out EMC with RDLs: chip (bridge) first with chip face-up process. SOURCE: Applied Materials, US patent 10,651,126 (May 12, 2020)

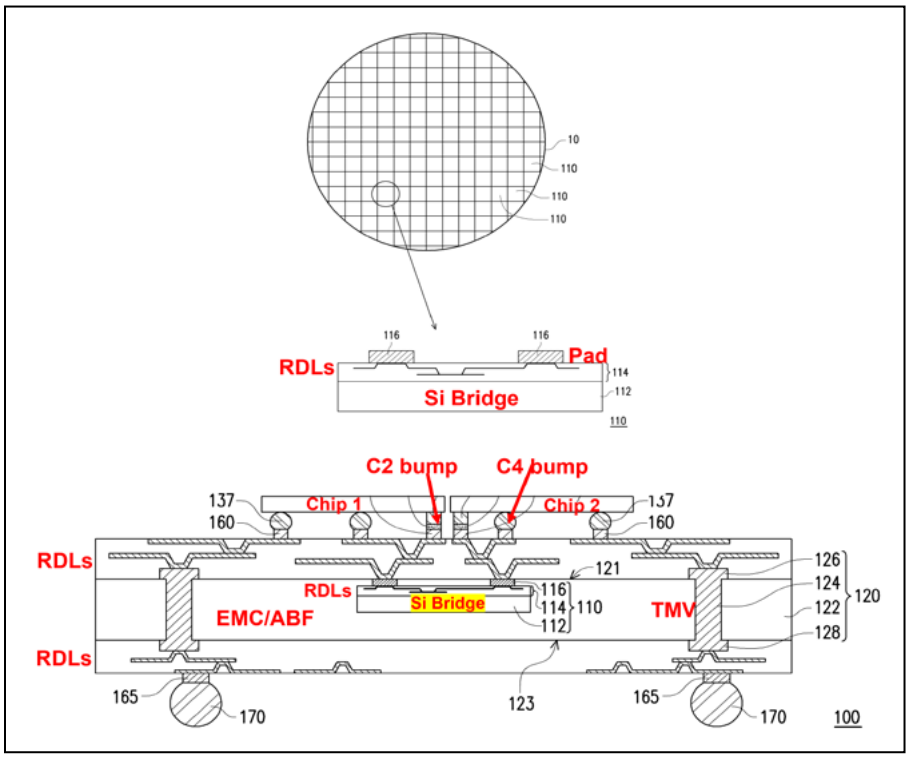


Figure 3: Unimicron's bridge embedded in fan-out EMC with RDLs: chip (bridge) first with chip face-down process. SOURCE: Unimicron, US patent 11,410,933 (August 9, 2022)

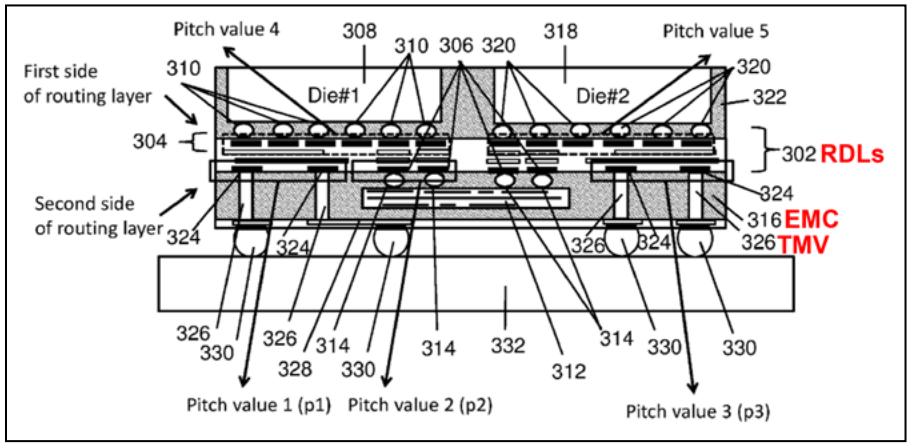


Figure 4: IME's bridge embedded in fan-out EMC with RDLs: chip (bridge) last or RDL-first process. SOURCE: IME, US patent 11,018,080 (May 25, 2021)

Papers published regarding bridge embedded in EMC with RDLs. Since the developments noted above, there have been many papers published by companies such as TSMC (local silicon interconnect) [12], ASE (Si bridge fan-out chip-on-substrate) [13], Amkor (S-Connect fan-out interposer) [14], SPIL (fan-out embedded bridge) [15], and IME (embedded fine interconnect) [16] in bridge embedded in EMC with RDLs.

Chiplet design and heterogeneous integration packaging

The following sections discuss various aspects of a hybrid bonding bridge for chiplet design and heterogeneous integration packaging.

Hybrid bonding bridge. Unimicron proposed the use of Cu-Cu hybrid bonding for the bridge between chiplets in chiplet design and heterogeneous integration packaging, (Figure 5). The advantages of this structure are: 1) higher density, 2) better performance, and 3) ordinary package substrate. There are at least two options: one is with C4 bumps on the package substrate, and the other is with C4 bumps on the chiplet wafer.

Hybrid bonding bridge with C4 bumps on the package substrate. Figure 6 shows the process flow of a hybrid bonding bridge with C4 bumps on the package substrate. For the bridge wafer, the processing starts off with chemical vapor deposition (CVD) to make a dielectric material such as SiO_2 and then it is planarized by an optimized chemical mechanical polishing (CMP) process to make the Cu dishing. Then, the bridge wafer is diced into individual chips (still on the blue tape of the wafer) after application of a protective coating layer on the wafer surfaces to prevent any particles and contaminants that may cause interface voids during the subsequent bonding process. These steps are followed by activating the bonding surface by using plasma and hydration processes for better hydrophilicity and a higher density of a hydroxyl group on the bonding surface.

To process the chiplet wafer, repeat the CVD process for the SiO_2, CMP for the Cu dishing, and plasma and hydration of the activation of the bonding surface. Then, pick and place the individual bridge chip on the chiplet wafer and perform the SiO_2-to-SiO_2 bonding at room temperature. These steps are followed by

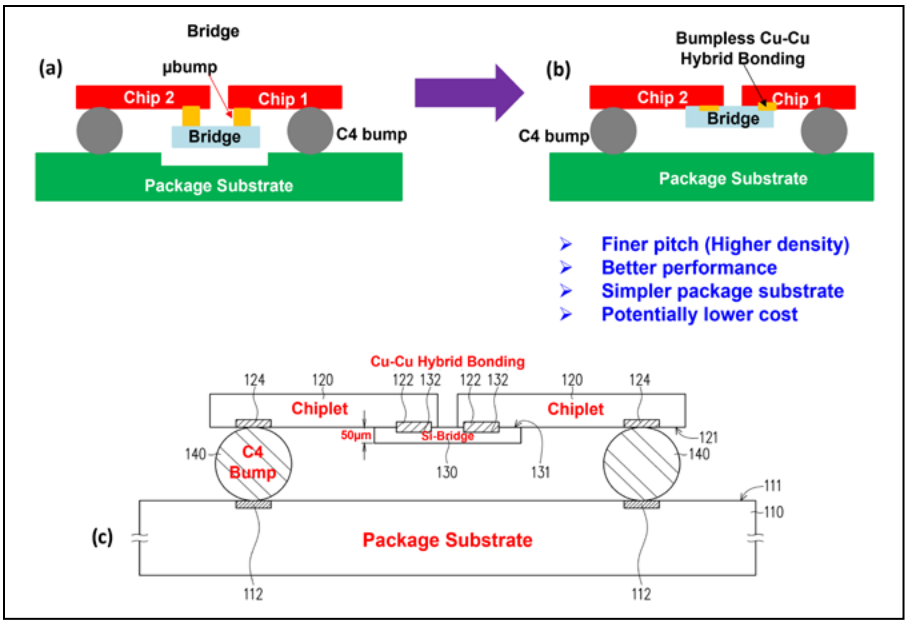


Figure 5: Hybrid bonding bridge. SOURCE: Unimicron

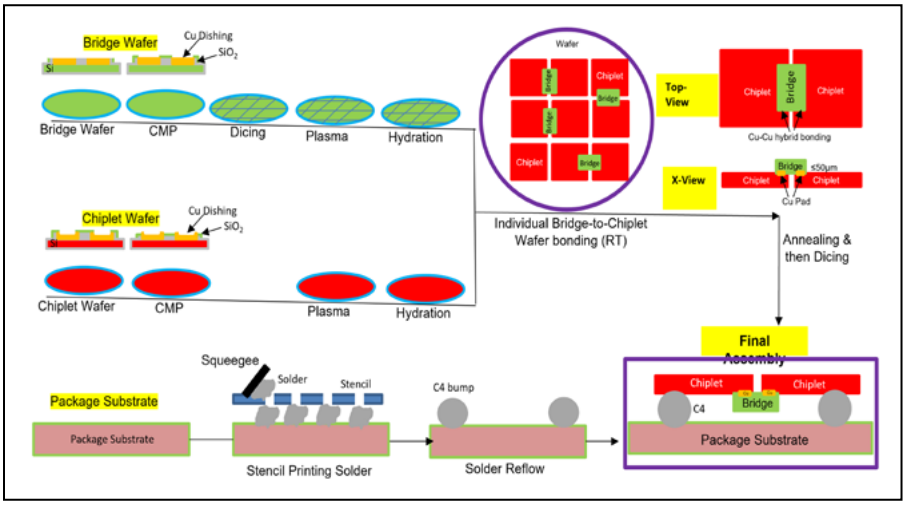


Figure 6: Hybrid bonding bridge with C4 bumps on the package substrate. SOURCE: Unimicron

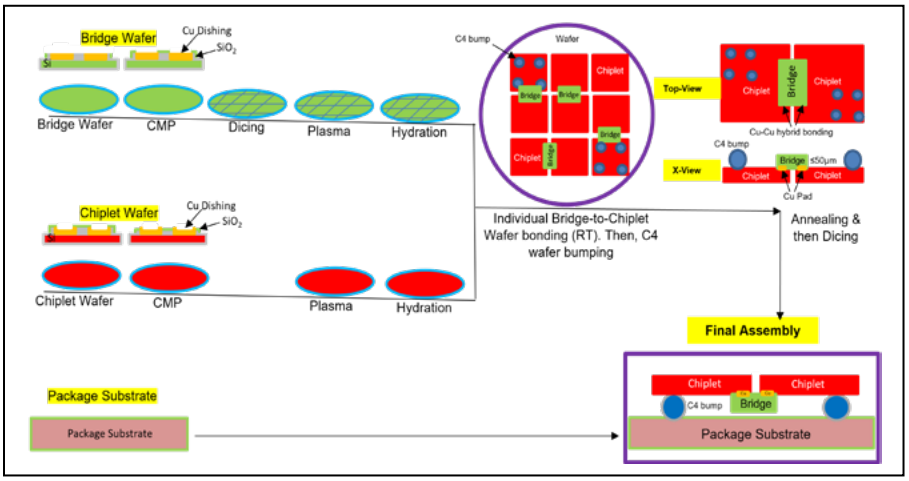


Figure 7: Hybrid bonding bridge with C4 bumps on the chiplet wafer. SOURCE: Unimicron

annealing to achieve covalent bonding between oxide layers and metallic bonding between Cu-Cu contacts and the diffusion of Cu atoms.

For the package substrate, the process is to stencil print the solder paste on the substrate and then reflow into C4 solder bumps. For the final assembly, the bridge + chiplets module is picked and placed on the package substrate, then the C4 bumps are reflowed.

Hybrid bonding bridge with C4 bumps on the chiplet wafer. Figure 7 shows the process flow of the hybrid bonding bridge with C4 bumps on the chiplet wafer. It can be seen that, compared with the C4 bumps for the package substrate case, the process steps for the bridge wafer and the chiplet wafer are the same up to the bridge-to-chiplet wafer bonding step. After that, the C4 bumps are fabricated by wafer bumping on the chiplet wafer. Then, the chiplet wafer is diced into individual modules (bridge + chiplets with C4 bumps). The final assembly is accomplished by picking and placing the individual module on the package substrate and reflowing the C4 solder bumps.

Summary

Some important results and recommendations are as follows:

- The key advantages of chiplet design and heterogeneous integration packaging are cost, cost, and cost.
- The key disadvantages of chiplet design and heterogeneous integration packaging are to increase package size and package complexity, which leads to higher cost.
- In general, the semiconductor cost is a few times the packaging cost, therefore, the savings that can be achieved with chiplet design and heterogeneous integration packaging are worth pursuing.
- The patents for bridges embedded in fan-out EMC with RDLs with chip (bridge) first and chip face-up, chip (bridge) first and chip face-down, and chip-last or RDL-first, have been provided.
- A new bridge with hybrid bonding has been proposed. Its advantages are: higher density, better performance, less process steps, and lower cost.

References

1. Y. Chiang, S. Tai, W. Wu, J. Yeh, C. Wang, D. Yu, "InFO-oS (integrated fan-out on substrate) technology for advanced chiplet integration," Proc. of IEEE/ECTC, May 2021, pp. 130-135.
2. J. H. Lau, *Semiconductor Advanced Packaging,* Springer, 2021.
3. J. H. Lau, *Chiplet Design and Heterogeneous Integration Packaging,* Springer, 2023.
4. J. H. Lau, "Recent advances and trends in advanced packaging," IEEE Trans. on CPMT, Vol. 12, No. 2, Feb. 2022, pp. 228-252.
5. J. H. Lau, "Recent advances and trends in multiple system and heterogeneous integration with TSV-less interposers," IEEE Trans. on CPMT, Vol. 12, No. 9, Sept. 2022, pp. 1271-1281.
6. J. H. Lau, "Recent advances and trends in multiple system and heterogeneous integration with TSV-interposers," IEEE Trans. on CPMT, Vol. 13, No. 1, Jan. 2023, pp. 3-25.
7. D. Sharma, G. Pasdast, Z. Qian, K. Aygun, "Universal chiplet interconnect express (UCIe®): an open industry standard for innovations with chiplets at package level," IEEE Trans. on CPMT, Vol. 122, No. 9, Sept. 2022, pp. 1423-1431.
8. J. H. Lau, "Recent advances and trends in fan-out wafer/panel-level packaging," ASME Trans., Jour. of Electronic Packaging, Vol. 141, Dec. 2019, pp. 1-27.
9. C. Hsiung, A. Sundarrajan, "Methods and Apparatus for Wafer-Level Die Bridge," US 10,651,126, date of patent: May 12, 2020.
10. J. H. Lau, C. Ko, P. Lin, T. Tseng, R. Tain, H. Yang, "Package Structure and Manufacturing Method Thereof," US 11,410,933, date of patent: Aug. 9, 2022.
11. R. Weerasekera, S. Bhattacharys, K. Chang, V. Rao, "Semiconductor Package and Method of Forming the Same," US 11,018,080, date of patent: May 25, 2021.
12. TSMC's Annual Technology Symposium, Aug. 25, 2020.
13. L. Cao, T. Lee, Y. Chang, S. Huang, J. On, E. Lin, O. Yang, "Advanced HDFO packaging solutions for chiplets integration in HPC application," IEEE/ECTC Proc., June 2021, pp. 8-13.
14. J. Lee, G. Yong, M. Jeong, J. Jeon, D. Han, M. Lee, et al., "S-connect fan-out interposer for next-gen heterogeneous integration," IEEE/ ECTC Proc., June 2021, pp. 96-100.
15. O. J. You, J. Li, D. Ho, J. Li, M. Zhuang, D. Lai, et al., "Electrical performances of fan-out embedded bridge," IEEE/ECTC Proc., June 2021, pp. 2030-2034.
16. C. Chong, T. Lim, D. Ho, H. Yong, C. Choong, S. Lim, et al., "Heterogeneous integration with embedded fine interconnect," IEEE/ ECTC Proc., June 2021, pp. 2216- 2221.

Biography

John H. Lau is a senior special project assistant at Unimicron Technology Corporation, Taoyuan City, Taiwan (ROC). He has more than 40 years of R&D and manufacturing experience in semiconductor packaging, 518 peer-reviewed papers, 43 issued and pending US patents, and 23 textbooks. He is an ASME Fellow, IEEE Fellow, and IMAPS Fellow. He earned a PhD degree from the U. of Illinois at Urbana-Champaign. Email John_Lau@unimicron.com

Electrical design challenges of multi-layered fan-out RDL MCM packaging

By Teny Shih, Sam Lin, Andrew Kang, Yu-Po Wang [Siliconware Precision Industries Co., Ltd.]

Historically, Moore's Law has meant that the semiconductor industry was able to maintain a cadence of next-node transistor development for processors roughly every two years. However, in going from the 5nm node down to the 2nm node, it has become more and more difficult to overcome the technical challenges to maintain that cadence. To maintain the momentum of improvement in processor performance, small chip-stacking technology has become one of the solutions. At a 2021 presentation by ASML, it was pointed out that further transistor development will encounter bottlenecks sooner or later in terms of the performance and area of the chip—to overcome these challenges will require a stacked architecture.

In the past, it was a challenge to produce processors with a stacked architecture. In addition to precisely controlling the process of each chip, it was also necessary to use interconnect technology to connect the chips. Now these problems can be solved through advanced manufacturing processes and packaging technology. Two individual chips with different functional blocks can be connected and integrated into a single package. This scheme not only reduces the bottleneck of data transmission, but also improves the operating efficiency, so that the performance of the processor is greatly improved.

With the advent of the era of big data, artificial intelligence (AI), and the Internet of Things (IoT), high-end process chips require features such as high-performance, low-power consumption, and multi-functionalities. With the increase of functions, the chip area is also getting larger and larger. To reduce the chip cost, advanced packaging technology is indispensable. The difficulty is that in the process of introducing advanced packaging technology, it is likely that the cost will be increased due to unstable yield rate. The larger the chip size, the lower the relative yield (**Figure 1**).

Background

To challenge the limit of Moore's Law, and to continue to evolve processor performance, chiplet stacking technology has been introduced as one way to address the problem. Engineers are using the stacking method to transform the traditional planar development processor structure into one that is three-dimensional. By integrating and stacking smaller chips with functions such as storage, graphics, and power management, these chips can then be connected by technology to improve performance and reach the goal either maintaining or enlarging the chip area.

Recently, the world's major chip manufacturers have been cooperating to define industry standards for chiplet technology. Companies participating in the plan include ASE, SPIL, AMD, Arm, Intel, Qualcomm, Samsung Electronics and TSMC. The new industrial standard has been named Universal Chiplet Interconnect Express (UCIe®), which will be a standard that could bring a transformation to the chiplet ecosystem. A chiplet is an increasingly common technology in modern semiconductor chips. The concept calls for many components originally included in a chip to be divided into small units one by one to form a system chip—with their functions strengthened using redesign and new manufacturing techniques, as well as through advanced packaging

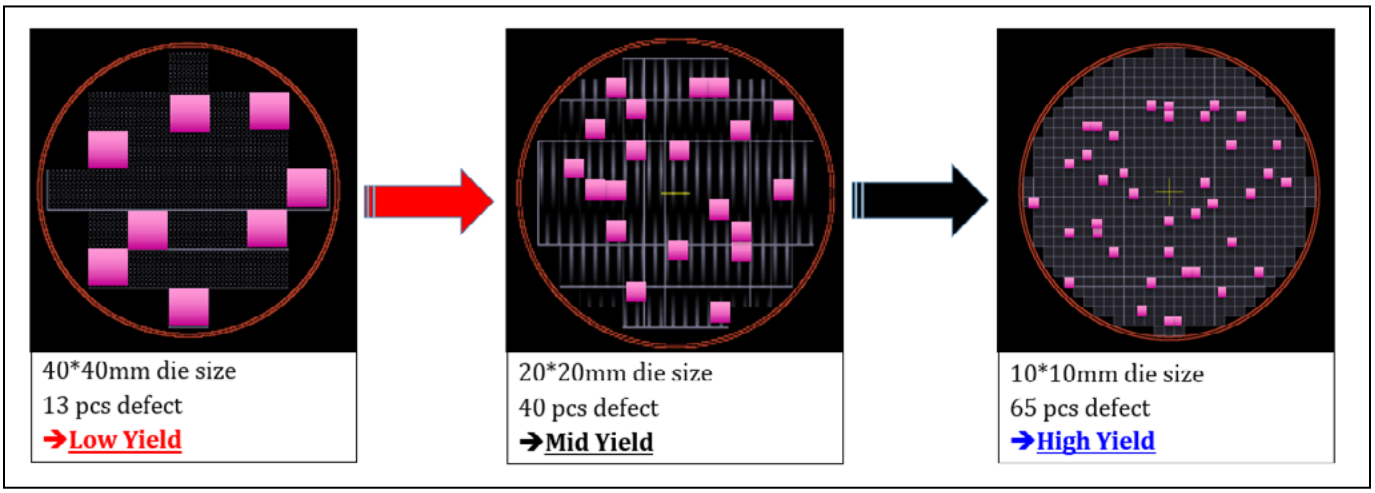


Figure 1: Smaller chip size can effectively improve yield.

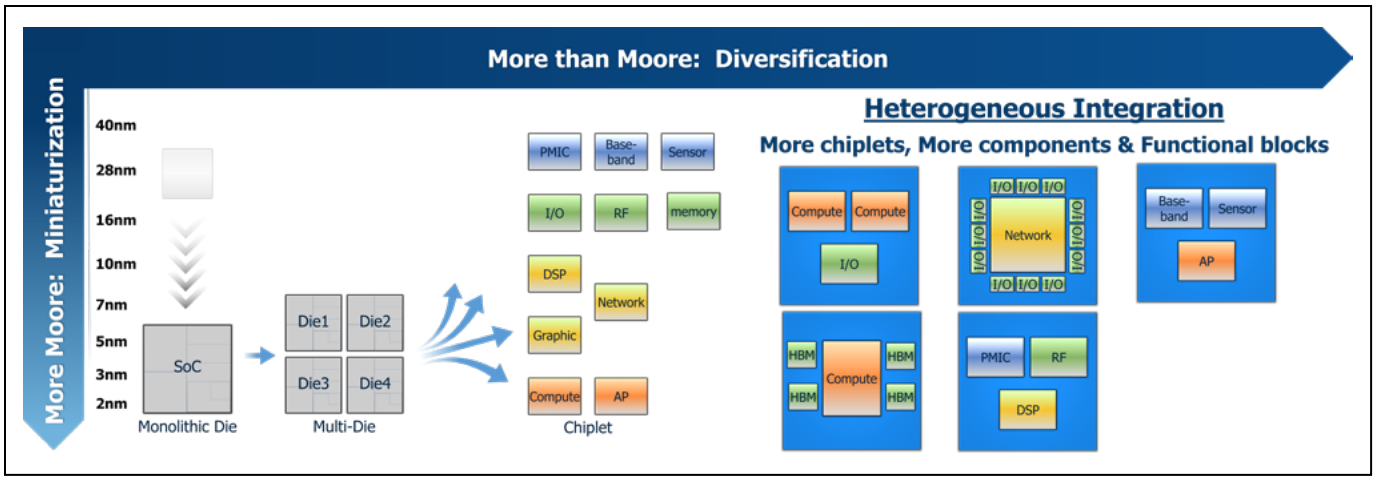


Figure 2: The use of chiplet technology is increasingly common in modern semiconductor chips. It divides many components originally included in a chip into small units one by one, and enhances their functionality. Furthermore, by using advanced packaging techniques, a system chip can be formed.

techniques. As shown in **Figure 2**, chiplets have been applied in many fields, including high-performance computer processing units (CPUs), graphics processing units (GPU), field-programmable gate arrays (FPGAs), and networking chips.

The use of chiplets breaks through the four design limits of system-on-chip (SoC) packaging: 1) it breaks through the scale limit of the mask area; 2) it breaks through the functional limit through the use of heterogeneous integration, so that it is no longer constrained by a single wafer node within a single chip; and 3) it improves the chip resilience through scalable computing power and performance; and 4) it greatly shortens the chip design and development cycle through agile development.

In addition to the above points, by using chip integration technology, chiplets can achieve better system integration, increased functional density, at a reduced cost. Coupled with the innovative technology of transmission circuits and devices, the value of electronic products can be further improved, and the cycle time of product development can be more efficient. At present, the application of chiplet technology has been adopted broadly, including in next-generation mobile communications, high-performance computing, autonomous vehicle technology, and the IoT.

Fan-out multi-chip module (FO-MCM) packaging technology

In general, fan-out multi-chip module (FO-MCM) packaging technology has two different process flows: chip first and chip last. Regarding the chip-first structure, the process flow applies silicon dies on a glass carrier with release tape and adds molding compound to build up redistribution layers (RDLs) directly on the silicon wafer. Next, C4 bumps are plated on the RDL module. In the next step, chemical mechanical polishing (CMP) is used to expose the back side and to reduce the molding thickness to achieve a specific target thickness, as necessary.

Another technology is the chip-last structure in which RDL layers are grown on a flat glass carrier. It should be noted that this process is independent from the micro-bumping process of the top dies. Dies with Cu pillar bumps are attached on a micro-pad of RDL and under-fill is added into the micro-bump space to protect the interconnect area; the RDL module is then covered by molding compound. Finally, C4 bumping is plated on the opposite side of the micro-bump joining interface. In comparison to the chip-first process, the chip-last approach has the advantage of controlling the organic interposer quality after the separated RDL process. There is no loss of known-good RDLs because the RDL yield and quality can be inspected before the die bonding process, so it can avoid the die loss risk. This is a particular benefit when using costly advanced wafer node die. With respect to yield performance, the chip-last process has a lower risk of having a failure caused by a non-coplanarity issue than the chip-first process due to the tolerance buildup when the surfaces of multiple dies are subjected to simultaneous grinding. In short, the assembly factory is capable of handling processes associated with RDL technology and can provide a turn-key solution for a design house. We conducted a study with respect to fan-out homogeneous silicon die integration and heterogeneous integration with high-bandwidth memory. **Figure 3** shows SPIL's FO-MCM solution. It should be noted that we have experience with 6-layer RDL, which can be used for die interconnect routing, where the micro bump pitch is 40µm. Multi-application-specific integrated circuits (ASICs) can support SoC+SoC, 1SoC+4HBM, 1ASIC+12 IO dies, 2SoC+8HBM, etc.

FO-MCM packaging technologies enable heterogeneous integration scaling with increased interconnect density along with increased bandwidth. FO-MCM also enables more effective die partitioning (e.g., heterogeneous integration) that, in turn, shortens the time to market. Several advanced packaging technologies have been developed to accelerate machine learning (ML), AI, and high-performance computing (HPC) applications.

All in all, the traditional monolithic die design that integrates multi-core processors into one SoC die architecture is facing a lot of challenges, such as increasing wafer costs, limited die size, and high power consumption. Consequently, the multi-chip module (MCM) structure is the alternative solution to reduce the packaging cost and provide the more flexible chiplet combination by die partition methodology. Therefore, the fan-out

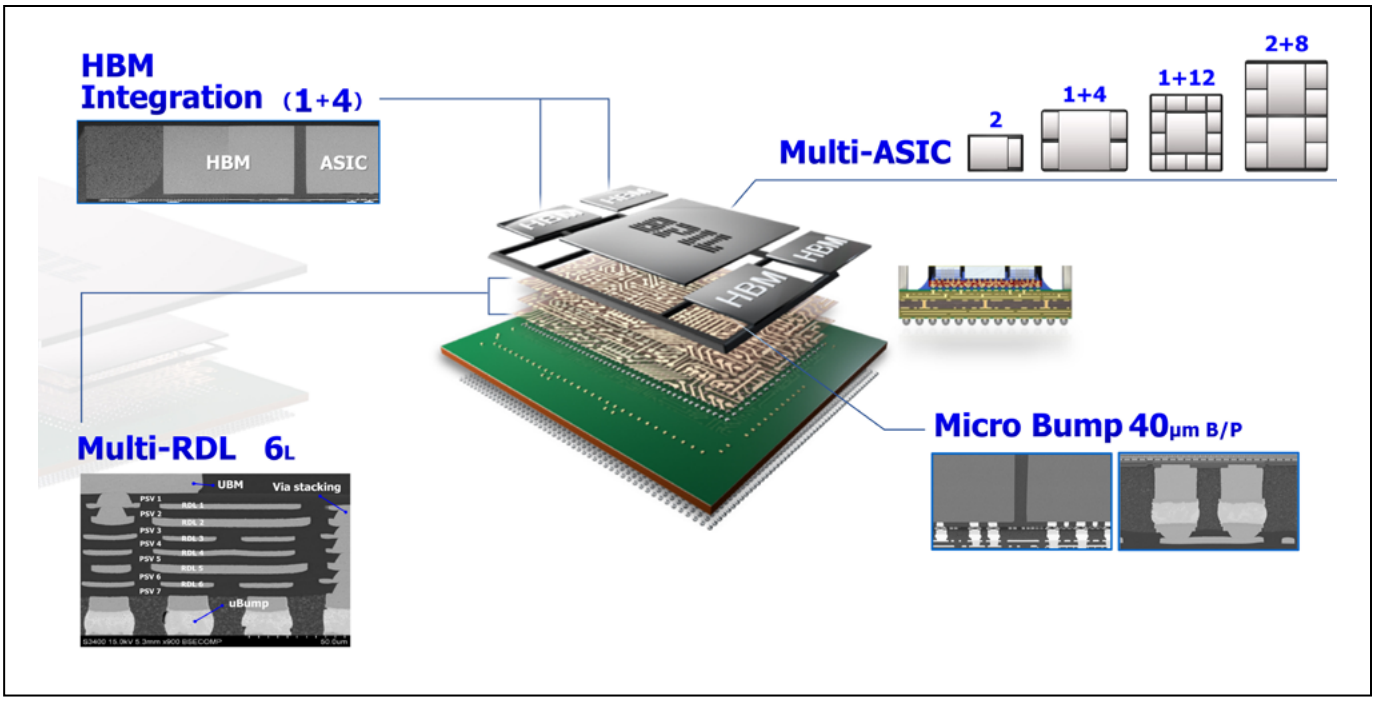


Figure 3: The FO-MCM solution.

redistribution layer (FORDL) became an essential routing technology between the die-to-die interconnection area. Additionally, because of high demand for processors to process and analyze big data at high speed and in large volume, the size of RDLs has evolved to finer lines/spaces. This evolution allows one to route more metal traces and layers to satisfy the design requirement.

The HBM interface is used to connect CPU/GPU/application-specific integrated circuits (ASICs) to large dynamic random access memories (DRAMs), especially for leading-edge high-performance computing, graphics, and networking applications that demand massive bandwidth and high power efficiency. The HBM standard has eight channels. Each channel is 128 interconnect pins, with a per pin data rate of 2Gbps. Higher data rates per pin (3.2Gbps) are supported in the recent updates to the HBM2E standard, hence, the eight channels could provide an aggregate bandwidth up to 409.6GBps. In the future, the speed of HBM3 will be increased to more than 6.4Gbps.

In this paper, the electrical design challenges using the polymer-based fan-out wafer-level package (FOWLP) RDL (see **Figure 3**) to implement these interfaces are discussed and presented. Increasing SoC-HBM interconnection length comes from the growing HBM package sizes. The package widths of HBM2 and HBM2E are 7.8mm and 10mm, respectively. Meanwhile, the package width of HBM3 has been increased to 11mm and the interconnection length between the HBM and SoC PHY (physical layer) is ~6mm (**Figure 4**). In order to understand which is the most suitable topology for routing HBM interfaces, a transient time domain simulation and the eye diagrams obtained from this simulation were used to qualify the physical structure. To qualify the suitability of the different configurations and the HBM specifications that were used, it

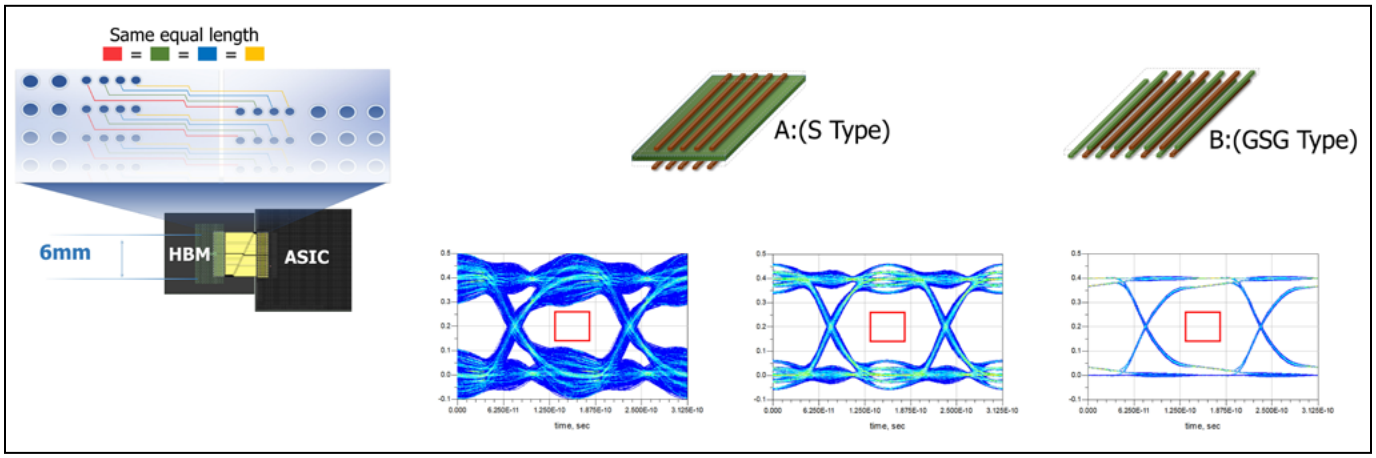


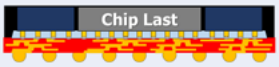
Figure 4: The current status of FO-MCM packaging development.

requires an eye mask with a width of 0.7UI (i.e., UI=unit interval, a unit of the eye diagram) and a height of 0.3VDDQ ~ 0.7VDDQ (VDDQ = the supply voltage to the output buffers of a memory chip). The length of the simulated bus in this paper study is around 4mm.

FO-MCM uses an organic interposer. The key components in FO-MCM include the RDL and through-silicon via (TSV)-less vertical interconnects. The organic interposer, including a good eye diagram and low insertion loss performance of multiple RDLs with a coplanar GSGSG isolation scheme, was demonstrated. RDL lines with a minimum linewidth/spacing of 2/2µm exhibit excellent robustness, ensuring the long functional lives of high-performance computing products.

Two possible stack-up configurations were considered (**Figure 4**). The topology (A) (S type) is a common form of wiring of the HBM interface. The signal trace width is 2µm and the gap between the adjacent traces is 2µm. Therefore, to satisfy the requirement for 1,700 interconnections one needs a 4-layer RDL, i.e., three of the layers are for signals and ground and one of the layers is for power. For the topology (B) (GSG type), the maximum winding density of one layer is only 750pcs. The full stack-up will require five layers: two layers for signals, two layers for ground, and one layer for power.

The eye diagrams of the topology (S Type) are shown in **Figure 4**. It can be seen that 1,700 signal lines of the HBM are distributed in 6mm of space. The S-type signal line is routed on layers 1 and 3, and the middle layer is the reference ground layer. When the trace width for the 2µm/spacing case is 2µm, in the HBM2/HBM3 analysis, it can be seen that the crosstalk between signals is large, and the isolation needs to be further optimized. When the trace width for the 2µm/spacing case is 4µm, it can meet the eye diagram requirements of HBM2/HBM3. When using the FO-MCM structure to implement the packaging process, it is recommended that a 5-layer structure, 2-layer signal line/2-layer reference ground/1-layer power supply design be used. Compared with the S-type trace, the GSG-type trace can improve the signal quality, so the linewidth of 2µm and the line spacing of 2µm can meet the electrical requirements of HBM2/HBM3. Relative routing, however, requires a larger area. Signal traces must use up to 3 layers. At the same time, by maintaining the signal dislocation of the upper and lower layers, the signal could have good reference ground layers. Compared with the S-type trace design, it is much more difficult.

Signal Integrity	I/L@5GHz (dB)	I/L@10GHz (dB)	I/L@15GHz (dB)	I/L@20GHz (dB)
	-0.848	-1.97	-2.67	-3.25
	-0.858	-1.985	-2.68	-3.28

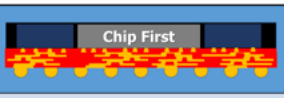
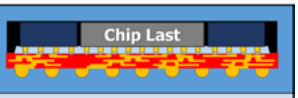
Power Integrity		
DCR	1X	1.02X
L(nH)	1X	1.02X
C(pF)	1X	0.97X

Figure 5: SI/PI ratio comparison between chip-first and chip-last structures.

Comparison of chip-first and chip-last structures

Fan-out packaging can be classified into two types: chip first and chip last. In the chip-first process, the chips are first embedded in a permanent material structure, followed by the RDL forming

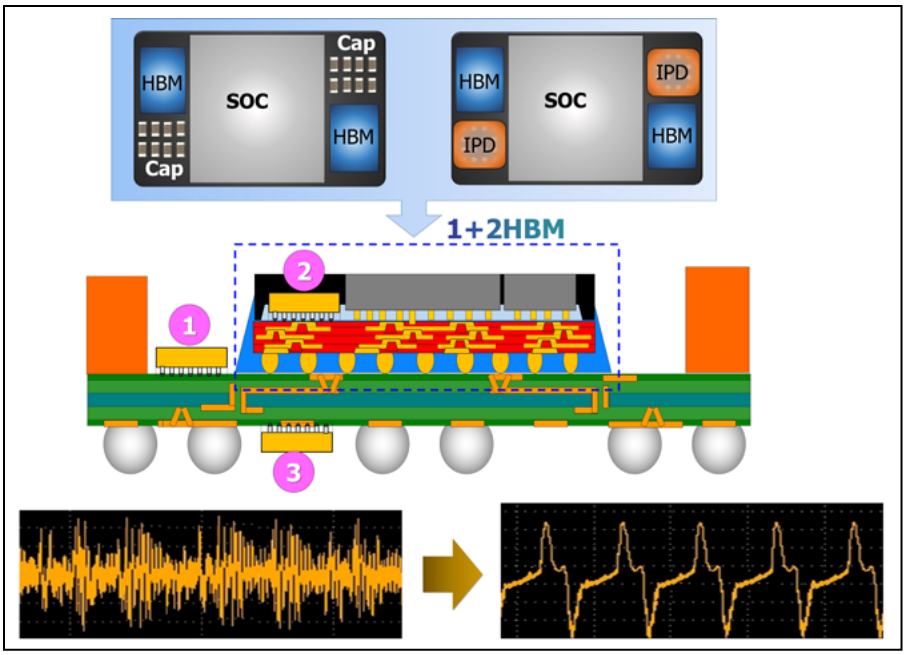


Figure 6: Adding capacitors at positions 1, 2 and 3 can reduce power noise by 30%.

processes. In the chip-last process – also known as RDL first – the chips are not integrated into the packaging processes until the RDL on the carrier wafer has been preformed. The chip-last process has less known-good die (KGD) yield concerns compared with the chip-first process. In cycle time comparison, the chip-last process also has the advantage of a shorter process time, although the electric capacity is almost the same (see **Figure 5**).

Power noise

For next-generation HBM, the power noise margin is a critical parameter to assess the system performance—and the power noise margin is decreased from previous generations because of the higher data rate and the lower operating voltage. The power supply voltage fluctuation degrades the signal output waveform performance. The output voltage at parallel simultaneous switching output (SSO) channels are affected by the impedance of the power distribution network (PDN) and the SSO pattern. In the FO-MCM package, traditionally adding capacitors at positions 1 and 2 can reduce power noise by about 15%. In the high-end integrated HBM2E package, a 2.5D package is usually used with an embedded deep trench capacitor (DTC). The DTC embedded in the interposer can be very close to the chip side and therefore, will have greatly improved electrical noise. Nevertheless, adding the capacitor configuration in position 3 can be comparable to the effect of a DTC, and can reduce the power supply noise by about 15%. Positions 1, 2, and 3 are connected to a traditional MLCC capacitor, or silicon capacitors, to reduce power noise by 30% (see **Figure 6**).

Summary

FO-MCM technology has been demonstrated to have the maturity for homogeneous silicon chip integration since 2020 and has been in mass production for switch applications at SPIL. Additional requirements have been imposed on assembly houses that need to come up with turn-key solutions for FO-MCM devices to be integrated with heterogeneous chips such as IPD or HBM. In addition, this paper summarized the reliability test results whereby all reliability conditions passed the MSL4, TCT700, u-HAST96, and HTSL1000 test conditions. The integrated de-cap capacitors suppress the power domain noise and enhance the HBM3 signal integrity at a high data rate. FO-MCM is one type of chiplet package platform that is designed for lower cost and a high data rate transmission package and meets the need for a high I/O count and a large number of Cu wires densely packed within a constrained package size.

References

1. B. Sabi, "Advanced packaging in the new world of data," Elec. Comp. and Tech. Conference (ECTC) 2017.
2. Y. L. Huang, "Challenges of large fan-out multi-chip module and fine Cu line space," ECTC 2020.
3. J. Li, "Large-size multi-layered fan-out RDL packaging for heterogeneous integration," IEEE 23rd Elec. Packaging Tech. Conf. (EPTC) 2021.
4. L. C. T. Lee, "Advanced HDFO packaging solutions for chiplets integration in HPC Application," IEEE 71st ECTC, 2021.
5. S-L. Liu, "Assembly technology for FO-MCM with HBM in HPC application," IMAPS 2022.
6. F. Kao, "Next generation fan out assembly technology in chiplet packaging to improve power loss and rout-ability," SPIL, IMPACT 2021.
7. M. Liao, "Heterogeneous integration fan-out package solution for HPC/AI application," SPIL, IMPACT 2022.

Biographies

Teny Shih is a Department Manager, Corporate R&D at Siliconware Precision Industries Co., Ltd., Taichung, Taiwan, R.O.C. He has over 25 years of industrial experience focusing on development of advanced packaging technology and analysis and measurement of electrical characteristics. He has published more than four conference papers and has 6 patents. Email: tenyshih@spil.com.tw

Sam Lin is a Manager, Corporate R&D at Siliconware Precision Industries Co., Ltd., Taichung, Taiwan, R.O.C. He has 12 years of industrial experience focusing on package electrical analysis, measurement and product application analysis. In recent years, he has focused on 2.5D, 3DIC, FO-MCM and FO-EB advanced packaging research.

Wafer-scale integration for graphene-based optoelectronics, sensors, and imaging devices

By Souvik Ghosh [imec vzw], Amaia Zurutuza [Graphenea], Alice Guerrero [Brewer Science]

The dawn of the 21st century kick-started an era of two-dimensional (2D) materials, with graphene in the forefront. Graphene is the most well-known 2D material and is often being referred to as the "wonder material." The exceptional properties of graphene, such as very high carrier mobilities and ballistic transport, promise a host of applications that range from sensors, optical modulators, and detectors to rapid-scan imagers, thermal management, and even room-temperature graphene-based spintronic devices. An interesting property of graphene is its linear energy-momentum dispersion that enables light absorption from the ultraviolet to the terahertz regime. This extreme broadband capability is a unique material property and makes it very interesting for on-chip optical communication because information can be multiplexed over a wider range of wavelengths and enable ultrafast communication.

The most high-end graphene-based applications (i.e., optical I/O devices), though in active development at laboratory scale, will need at least another decade to come to the market simply due to limitations of scalability, yield, and performance metrics [1]. In the near term, single-layer-graphene (SLG) is already seeing an increase in market opportunities mainly in the field of (low-cost) sensors. These applications will be enabled by the development of new graphene growth and transfer techniques. This article summarizes the near and long-term outlook using the practical benefits of graphene in view of its time-to-market. Further, the specific technological support structure needed to consider the material as a viable option to be compatible with existing semiconductor fabrication infrastructure will also be outlined.

Graphene synthesis

Graphene synthesis was first achieved via mechanical exfoliation from highly-oriented pyrolytic graphite (HOPG) [2] and can also be obtained via electrochemical spalling of graphitic membranes in solution. Both techniques facilitate easy and selective access to surface and edges of the nanosheets, thereby promoting nanodecoration and targeted functionalization of these nanosheets. Despite being a robust technique to manufacture graphene sheets, mechanical and electrochemical exfoliations remain a challenging approach to enable fab integration. The inherently small size of graphene flakes and the lack of control over the number of layers during graphene flake synthesis make it very difficult to position these flakes accurately on a target wafer.

The most promising graphene growth route is via chemical vapor deposition (CVD). In fact, soon after Geim and Novoselov's mechanical exfoliation method, CVD quickly caught up to enable large-scale controlled SLG synthesis. There are numerous catalyst substrates available that enable high-quality graphene growth. These substrates include transition metal foils such as Cu, Ni, Fe, Pt, and even alloys. Cu or alloys like CuNi are interesting due to their favorable catalytic properties, low C solubility and relatively low cost. The governing graphene growth mechanism relies on dissolution of the C species and subsequent saturation of the surface. Cu remains an interesting catalyst surface because of its low C solubility, facilitating monolayer, and even bilayer graphene growth control.

Additionally, research is ongoing to enable graphene growth on rigid template wafers. If Cu is sputtered directly on Si/SiO_2 wafers followed by graphene growth, the resultant SLG quality is relatively poor due to the polycrystalline nature and small grain structure of the Cu layer (see Figure 1). Metal epitaxy on template wafers such as sapphire can provide a single catalyst orientation and ideally a single-crystalline SLG is grown. In this respect, epitaxial graphene growth has been demonstrated on epitaxial Cu(111)/sapphire(0001) wafers, but also other wafer types like Ge(110) can be used for epitaxial graphene growth. The latter one is not straightforward because the process window to grow high-quality graphene is rather small but avoids transition metal contamination issues during graphene integration. Graphene growth on a thin epitaxial Ge layer is also very complicated because of diffusion of Si into the Ge layer, which increases the roughness of the Ge layer. Nevertheless, because SLG growth on epitaxial surfaces can follow a preferred orientation, this growth route is expected to give the highest graphene quality due to the absence of grain

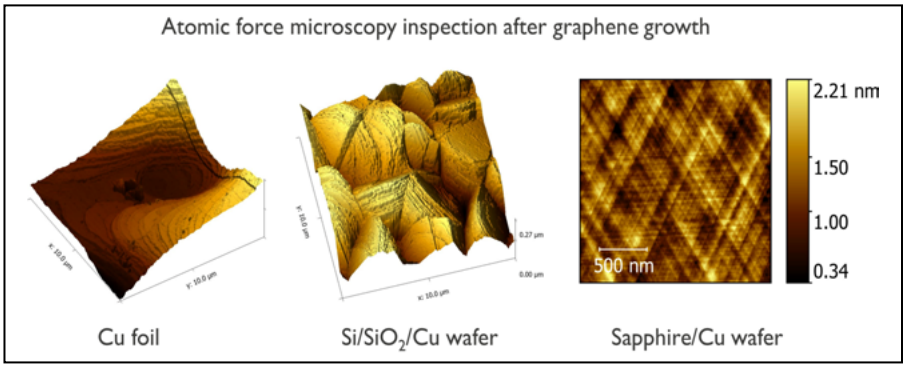


Figure 1: Atomic force microscopy inspection after graphene growth on a Cu foil, a Si/SiO_2/Cu wafer and a sapphire/Cu wafer. Clear topography variations are present after graphene growth on a Cu foil, but the observed grain size is much larger compared to a graphene growth on a Si/SiO_2/Cu wafer (the scan size of both AFM images is 10μm). A strongly reduced surface roughness can be obtained when graphene is grown on an epitaxial template wafer (e.g., Cu(111)/sapphire(0001)).

boundaries and its low surface roughness. A thermal decomposition route can also produce high-quality graphene on SiC, but is often considered to be less scalable compared to CVD, due to the need of large – and as a result – expensive SiC wafers.

Nowadays, CVD synthesis of SLG is mostly done on Cu foils at a manufacturing scale. However, SLG grown on Cu foils typically has high topography variations over a macroscopic scale because of the flexible nature of the foils and the high graphene growth temperature (typically well above 800°C). This foil roughness can likely be further improved by incorporating electropolishing and annealing before the actual graphene growth process. However, because of the polycrystalline nature of the foils themselves it will be difficult to completely avoid graphene grain boundaries. Nevertheless, high graphene mobilities have been reported for CVD graphene grown on Cu foil, which could make this technique ideally suited for several graphene-based sensor applications. To enable high-end graphene applications, a controlled and oriented growth on epitaxial catalyst template wafers is likely needed. However, this growth approach complicates transfer because a lateral wet-chemical etch-based transfer is more difficult to implement and recycling of the growth wafer is preferred from a cost perspective.

Layer transfer technologies applicable to 2D materials

Because the most mature graphene growth technique is CVD on a transition metal template, most graphene applications require the development of a layer transfer technology. Such a technology, especially for front-end-of-line (FEOL) applications, is the "new kid on the block." It is untested and historically unqualified to compete with mainstream fab integration processes such as selected-area growth, targeted etch, etc. While bonding and debonding are prevalent in back-end-of-line (BEOL) packaging techniques, proposing their usage for FEOL applications to transfer 2D layers invites skepticism. Industrial adoption and hesitation aside, existing layer transfer technology has already been demonstrated to enable new device possibilities. Furthermore, a layer transfer can potentially bring an epitaxial 2D layer on an amorphous dielectric surface that is otherwise impossible to achieve via direct growth.

The transfer of graphene from a growth substrate to a target device wafer typically requires an intermediate carrier to support fragile 2D materials. As often reported in literature, R&D-scale transfer approaches use poly-(methyl methacrylate) (PMMA) as a support layer. This is generally perceived as the golden standard aiding in transfer of 2D flakes and even CVD materials. Here, PMMA is dissolved in a solvent (e.g., anisole), spin-coated on the 2D material, and possibly adhered to a thermal release tape (TRT). Next, graphene is released from its growth substrate via Cu etching or even intercalation-based (i.e., electrochemical) release methods. The released graphene layer is then laminated or bonded to a target wafer. Finally, the adhesive is stripped to expose the 2D material. While a TRT-based transfer method is viable for proof-of-concept device demonstration, it remains challenging to implement this process in a production environment. Further, manual bonding and debonding steps introduce user-level variance that manifest in the form of wrinkles, surface-potential variation, and macroscopic cracks. A viable and scaled industrial approach to demonstrate transfer of graphene on 200mm, or even 300mm target wafers, likely requires the use of a rigid substrate as a temporary carrier instead of TRT. A rigid carrier will prevent nonuniform and excessive expansion typically observed for TRT-based transfers. Furthermore, a rigid temporary carrier is compatible with existing 200mm and 300mm (de)bonding and cleaning equipment, making it the preferred temporary carrier to transfer 2D materials.

In terms of temporary bonding materials (TBMs), PMMA has been the material of choice. In most cases, especially for small-scale demonstration of lab devices, this polymer has been applied in a successful and reproducible manner. It is imperative that alternate TBMs be identified to enable quality-controlled 2D transfer that meets industrial-scale requirements because uniform bonding of a temporary carrier system with a PMMA layer will be very difficult to achieve in a reproducible manner. Furthermore, such TBMs will need to be specifically selected to be compatible with a release process (e.g., laser debond approach). To enable this laser release mechanism, the TBM can be complemented with a second layer that specifically is compliant with the release mechanism (e.g., laser-absorbent materials). Finally, the last key element governing the selection of an appropriate TBM is its adhesion with the graphene layer and that with the carrier system. The adhesion between the TBM and the 2D of interest needs to be high enough such that during mechanical, chemical, or electrochemically assisted debonding, the 2D is spalled or discharged from the growth substrate while keeping the TBM-2D interface intact.

A manufacturable graphene transfer route from Cu foil

The highest quality graphene is grown on epitaxial template wafers, but debonding large-scale graphene from these wafers has not yet been demonstrated in a fab environment. Because graphene quality is sufficient for several applications when it is grown on a Cu foil, and the graphene release can be easily achieved via etching processes, it is likely that this foil approach combined with an etch-based release step is the preferred choice for introducing graphene in the BEOL. To achieve a reliable transfer, the use of a rigid temporary glass carrier is an option as it optimizes the bonding step to the target wafer and facilitates the etching and cleaning steps. **Figure 2** shows the different steps of a glass carrier-based graphene transfer when graphene is grown via CVD on a Cu foil. The different steps during a 200mm graphene transfer process are visualized in **Figure 3**. First, a laser release and

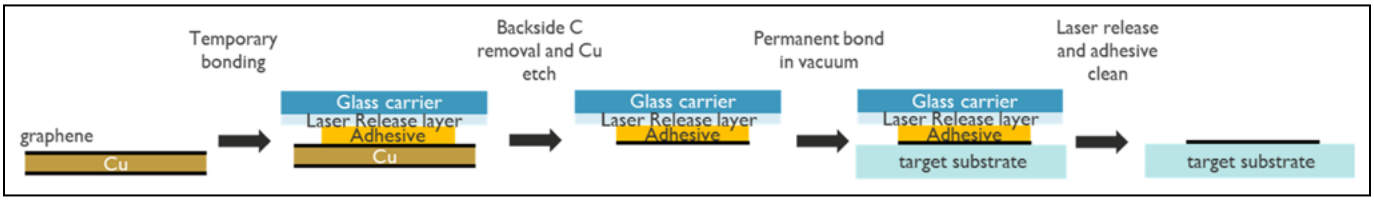


Figure 2: Schematic of a manufacturable graphene transfer process using a rigid glass carrier and a laser release approach.

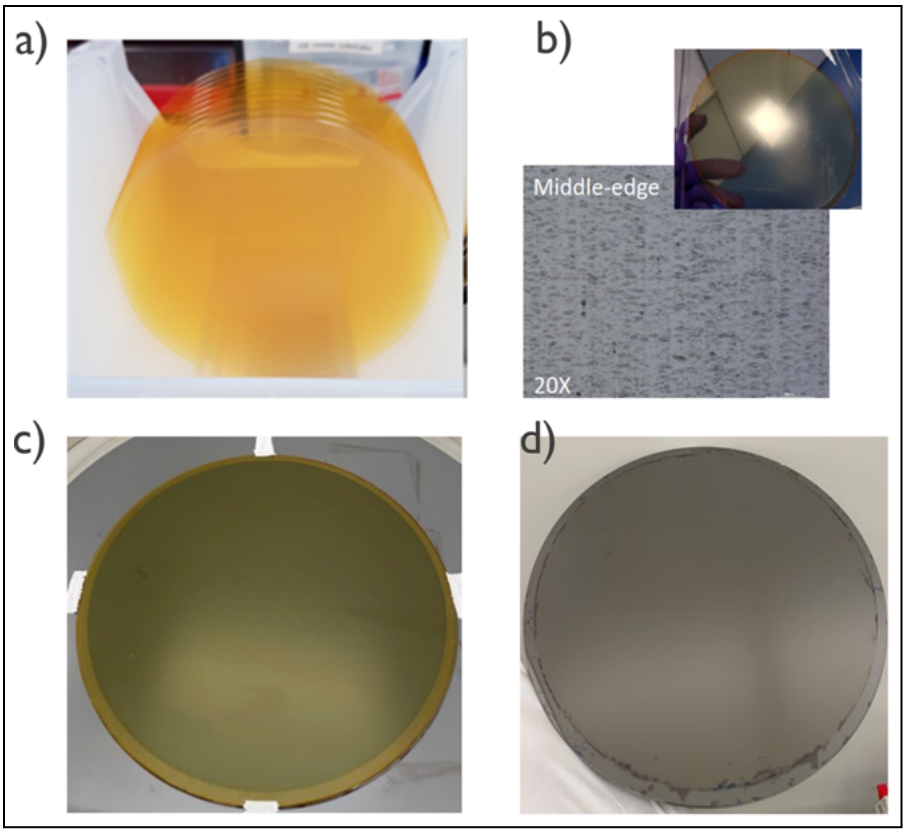


Figure 3: a) Glass carrier system with laser release and TBM; b) SLG on a glass carrier system obtained after Cu etch; c) Bonded glass carrier system with graphene on a 200mm target wafer; and d) Graphene on a target wafer after the full 200mm transfer process.

TBM layer are coated on the glass carrier (Figure 3a). Next, an edge bead removal step is implemented to avoid excessive TBM at the bevel of the glass carrier. After the temporary bonding step of the glass carrier on top of a Cu foil, C residues at the foil backside are removed and the Cu foil is selectively etched. After etching the Cu foil, graphene is exposed on the glass carrier. As can be seen in the microscopy picture in Figure 3b, the roughness of the Cu foil gets imprinted in the TBM material. As a result, the TBM layer still has considerable TTV variation, and improving the foil roughness will be important to improve the graphene transfer result. Next, the glass carrier that contains the graphene layer is bonded on top of a target wafer and the temporary carrier is subsequently removed using a laser release process (Figure 3c). The whole transfer approach ends with a solvent clean that removes the bulk TBM from the 2D material (Figure 3d). Typically, polymer residues are still observed after the solvent strip process, but these residues on top of graphene can be further removed with an additional annealing step or even the implementation of a remote hydrogen plasma clean. A protection layer on top of graphene could also be envisioned to avoid direct graphene/TBM contact.

Summary

To achieve a reliable transfer process and to maintain the intrinsic properties of graphene, it is known that device performance can improve when SLG is transferred on smooth target wafers. To avoid high surface roughness when integrating graphene in the BEOL (e.g., on a readout integrated circuit [ROIC] wafer), a planarization step of the top dielectric will have to be implemented before graphene transfer. After transferring graphene on such a wafer, it will have to be capped with a dielectric followed by graphene patterning and contact fabrication. Following this approach, a BEOL operational graphene transistor can be achieved that can serve a multitude of applications [3].

Acknowledgments

This work is supported by the imec IIAP optical I/O program and received funding from the European Union's Graphene Flagship grant agreement CORE 3 (No 881603) and 2D-EPL (No. 952792).

References

1. C. H. Wu, et al., "Graphene electro-absorption modulators integrated at wafer-scale in a CMOS fab," 2021 Symp. on VLSI Circuits, Kyoto, Japan, 2021, pp. 1-2, doi: 10.23919/VLSICircuits52068.2021.9492495.
2. K. S. Novoselov, A. K. Geim, et al., Science Vol. 306, 5696, pp. 666-669, Oct. 2004.
3. C. Huyghebaert, T. Scharm, et al., IEEE Inter. Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2018, pp. 22.1.1-22.1.4, doi: 10.1109/IEDM.2018.8614679.

Biographies

Souvik Ghosh is a Research Scientist at imec, Leuven, Belgium. He specializes in 2D materials transfer. Prior to imec he was a research engineer on chip packaging technologies for the Components Research division at Intel (Oregon, USA). His background is in flexible electronics, and he holds a PhD in Chemical Engineering from Case Western Reserve U. (Ohio, USA).

Amaia Zurutuza is the Scientific Director of Graphenea, San Sebastian, Spain, where she leads the R&D activities on graphene-based materials. Since joining Graphenea in 2010, she has filed for fifteen patents and published more than 85 publications in peer-reviewed journals, including *Nature* and *Science*. She received her PhD in Polymer Chemistry from the U. of Strathclyde (Glasgow, UK) in 2002.

Contact author: Alice Guerrero/Brewer Science; email aguerrero@brewerscience.com