Hybrid & Fusion Bonding: from Heterogeneous Integration to 3D Front-End of Line Technologies

By Thomas Pleschke and Viorel Dragoi, EV Group

For decades, progress in the semiconductor industry was defined by More Moore [1]: the relentless scaling of transistor dimensions to deliver higher performance, lower power consumption and reduced cost per function. With device geometries shrinking and technology nodes and their flavors becoming increasingly application-optimized, economic tradeoffs became increasingly significant since the early 2000's. This trend gave rise to the More than Moore (MtM) paradigm, an era putting focus on heterogeneous integration, novel materials and functional diversification as a second complementary vector to shrinking.

With this development, wafer bonding has continuously grown as a foundational MtM enabler, and several breakthroughs in precision alignment, low-distortion wafer bonding, and thin wafer handling are now the foundation for the next transition. Back-end-of-line technologies are starting to grow into front-end-of-line processes, enabling "More Moore" transistor scaling following the concept of "from shrinking to stacking".

After an overview of the history and applications driving hybrid and fusion bonding, this article intends to provide insight into the driving factors and key advancements towards this next paradigm shift.

History and Applications driving Hybrid and Fusion Bonding

Due to its suitability to build 3D architectures, wafer bonding gained significant attention over the past two decades. The process technique joins two substrates by bringing their surfaces into contact. Various principles govern the wafer bonding processes: molecular bonds for fusion bonding [2, 3, 4], atomic bonds and grain growth in the case of metal thermocompression bonding (TCB) [5], alloying in the case of eutectic [5] or intermetallic bonding (known also as transient liquid phase – TLP, or Solid-Liquid-Interdiffusion-SLID bonding) [5], or adhesion mediated bonding by a low melting temperature glass [6] or a radiation- or thermally-cured polymer layer [7].

Due to specific process requirements (e.g., thermal budget, contamination restrictions), the processes selected for 3D interconnect applications were metal thermo-compression and low-temperature hybrid and fusion bonding.

Metal thermo-compression associated with through-silicon vias (TSVs) was the first choice for such applications (via-first approach). The resulting bond is conductive, and the thermal budget can be tuned to stay within the CMOS thermal budget, so that metal pads from the two wafers can be bonded to form an electrical interconnection. However, the bonding process is typically followed by thinning one wafer (e.g., by grinding and polishing), and due to the very low area of the metal bonded pads (typically less than 10-15% of wafer area) the bonded pairs are mechanically very fragile and thus major yield loss and scrapping occurs during the thinning step.

To overcome this issue, attempts were made to increase the bonding area to 100% by filling the space between the metal bonding pads with a flowable polymer [8] and allowing for a metal/polymer hybrid interface formation. This solution worked in principle as the polymer filled the space between the metal bond pads with the thermal budget staying below 400°C, but in real life fabs were reluctant to adopt it due to the additional polymer process steps, as well as the high risk of production line contamination from using a polymer layer.

The next development in this direction took place in the early 2000s, and was based on the adoption of a rigid dielectric at the interface to maximize the total bonding area [9, 10]. This was possible once low-temperature bonding processes were available in the market [11] and hybrid bonding entered the stage.

Hybrid bonding has rapidly evolved into a critical technology for heterogeneous integration. By providing ultrafine pitch (as small as hundreds of nm), high performance interconnects far below the ~20 to 30 µm micro-bump limit, hybrid bonding enables orders of magnitude more interconnects between dies, compared to earlier packaging technologies (see Figure 1).

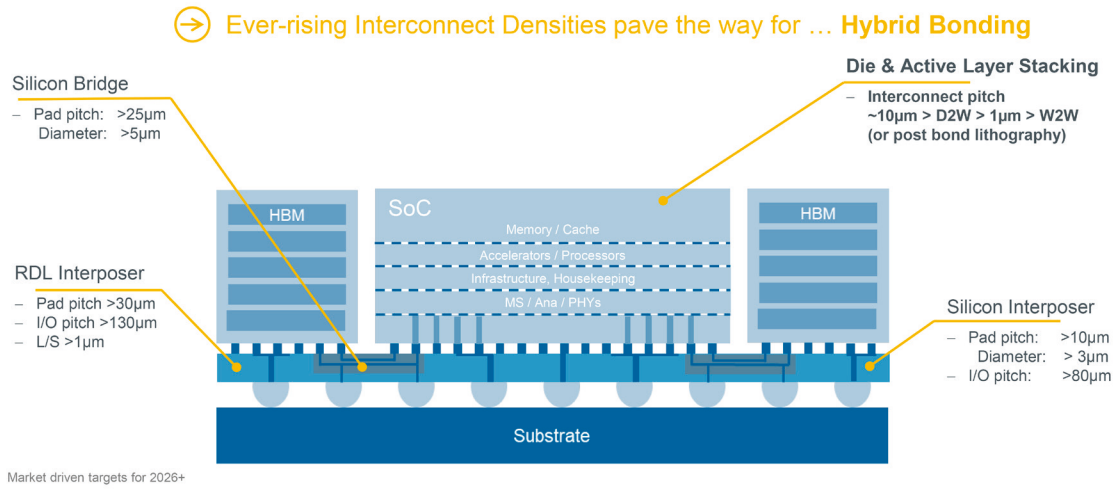

Figure 1
Heterogeneous Integration Hierarchy: Interconnect Densities pave the way for Hybrid Bonding

Wafer-to-wafer (W2W) hybrid bonding involves stacking and electrically connecting wafers from different production lines and has a proven track record of success for image sensors and various memory and logic technology applications

Backside Illuminated CMOS Image Sensors (BSI CMOS) became one of the earliest high volume applications for hybrid bonding and still are today. The architecture requires extremely tight alignment between photodiode and logic wafers. Hybrid bonding allows more compact stacking, better low light performance, and reduced noise in BSI sensors, becoming a standard technique in smartphone imaging technology.

As the demand for storage density increased, 3D NAND as a first memory application pushed beyond the limits of planar scaling. Hybrid bonding provides an efficient path to stack memory layers vertically with improved electrical and thermal characteristics. By eliminating micro-bumps and allowing direct Cu Cu contact, hybrid bonding improves interconnect density and reduces parasitics, enabling higher layer counts and improved performance. The success in NAND memory has prepared the grounds for even more demanding high-bandwidth memory (HBM) and DRAM stacking deployments that define today's high performance computing landscape by tearing down the "memory wall" and letting processors unfold their full performance.

Heterogeneous integration for the disaggregation of SoCs (System on Chip) has become the next frontier in driving hybrid bonding application KPIs. Chiplet based architectures — where a system is composed of multiple dies or die-stacks — enabled by hybrid bonding, almost behave like monolithic systems (see Figure 1). The chiplets, however, are not necessarily the same size and/or height and consequently die-to-wafer (D2W) process flows and their variants have become established alternatives [12, 13]. Figure 2 provides an overview and summary of D2W technology.

A major advantage of D2W processes is the support for "Known Good Die" (KGD) concepts, where die maps are generated by running probe card testing on complete wafers prior to die singulation. Subsequently, only good dies are further processed, improving the overall yield. While Direct Placement (DP) D2W is the most versatile method, its die-by-die handling obviously is the most time intensive, with operating speed competing against precision.

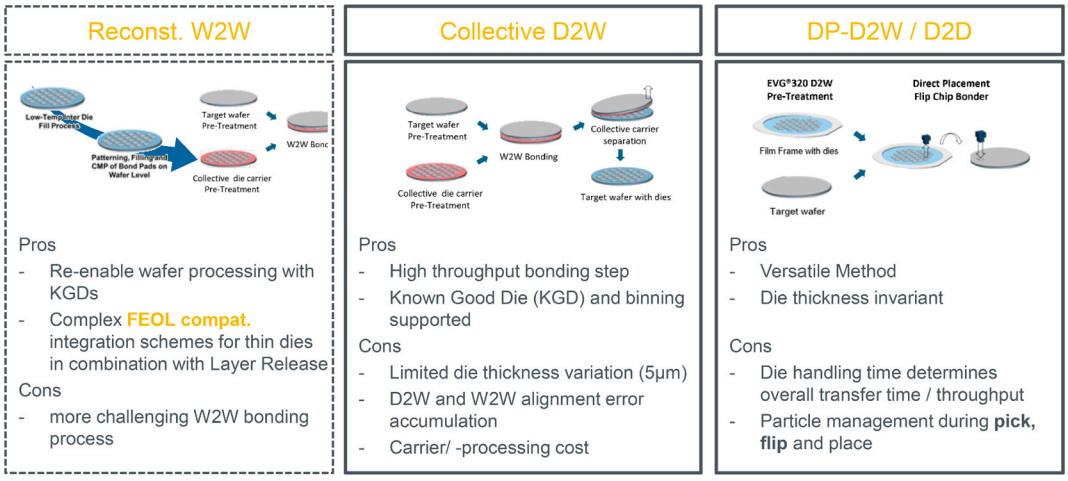

Figure 2
Die to Wafer process flow options

Collective (Co-) D2W in comparison, reduces bonding to a single step at the wafer level, by first temporarily bonding the chiplets to a carrier, thereby decoupling placement and bonding steps. The reduced time window for the bond step decreases potential process variations and relaxes die logistics management with the downside of accumulated alignment inaccuracies of die to temporary carrier and wafer level bond-steps.

Wafer Reconstitution as a further variation of Co-D2W allows for additional process steps at the wafer level. Inter-die filling with thick oxide or other materials provides additional mechanical stability and sometimes is required for thinning the reconstituted wafers. This opens the potential to massively improve thermal performance, while at the same time reducing necessary dimensions and hence cost of TSV processes or enabling even more advanced backside routing for power supply and input/output signals. A combination of this process flow with purely inorganic materials based on temporary bonding / de-bonding technologies like LayerRelease® is especially advantageous as it allows for thinning down to as far as single-digit micrometers and provides front end of line compatibility [14].

Figure 3 shows an overview of today's hybrid and fusion bonding applications.

	BS Illum. Image Sensor	µLED Displays	Memory				Logic		Comms
			3D NAND Flash	HBM Stacks	DDR6+	Next Gen. Memory	SoC Partitioning	Scaling	CPO
Device Stack	Photo Diode + DRAM + Logic	EPI layer + Logic	NAND Block + Periphery	12+ layer stacking	Peri under DRAM	Peri on MRAM, FeRAM, PCM	Application specific: Logic, MS, Ana, Mem.	Backside PDN (5nm node)	III-V on Si, LN, X-OI SiN, BTO
Bonding Process	W2W	W2W and/or D2W	W2W	W2W and/or D2W	W2W	W2W	W2W and/or D2W	W2W	D2W
	Hybrid	Hybrid or fusion	Hybrid	Hybrid	Fusion	Fusion & hybrid	Hybrid	Fusion	Fusion
Pitch	2µm → <1µm	2µm → 1µm	2µm → 1µm	5µm → 3µm	2µm → <1µm	2µm → <1µm	9µm → <1µm	Scanner dependent	Optical coupling <1µm
Maturity	HVM	Ramp Up	HVM	R&D	R&D	R&D	Ramp Up	Ramp Up	R&D

Figure 3
Hybrid and Fusion Bonding Applications

Front-end-of-line layer stacking enabled by backside power and signal routing are the next steps on the hybrid bonding application roadmap with the ultimate goal of bonding CMOS transistors on top of each other (CFETs).

Hybrid and Fusion Bonding as a solution

As stated above, the hybrid bonding concept was developed based on low-temperature fusion bonding and fulfilled the need to provide not only excellent electrical interconnections, but at the same time offer a strong, reliable bond. It allows for both high wafer-to-wafer alignment accuracy and high mechanical robustness to survive grinding processes.

The process flow and its main steps are shown in Figure 4.

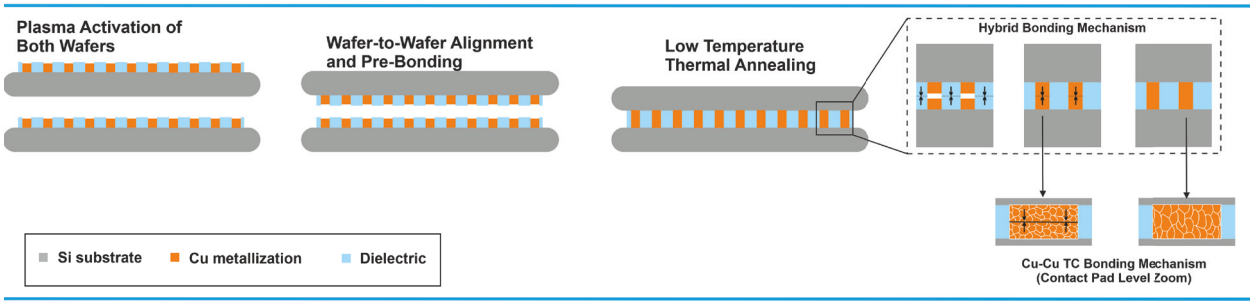

Figure 4
The principle of Hybrid Bonding

In a first step, the fully patterned device wafer must be planarized. Planarization processes used as a final step of the CMOS manufacturing flow are not the same as the planarization required for fusion/ hybrid bonding: in CMOS manufacturing, planarization aims for topography reduction, while for fusion/hybrid bonding, a planarized surface is completely flat with low microroughness.

The planarization process consists of a dielectric deposition on the wafer surface (usually by Chemical Vapor Deposition – CVD) followed by a densification annealing to remove the gas precursors trapped into the deposited layer. The densification is usually performed at ambient pressure and a temperature equal or slightly higher than the final thermal annealing of the bonded wafers. If densification is not performed, the dielectric will outgas during the thermal annealing and gas bubbles will form at the interface between the two wafers.

After densification, the surface must be prepared for bonding. The requirements for surface quality refer to the flatness and the microroughness (<0.5 nm) defined as the root mean square measured by an Atomic Force Microscope on a 2 x 2 µm2 sampling area [15], as well as to cleanliness (extremely low particle levels, no organics on the surfaces). The hybrid interface brings additional challenges, as a very special topography must be achieved by Chemical Mechanical Planarization (CMP): the metal pads surface must be recessed below the dielectric surface to allow for the hybrid bond formation. Based on an empirically determined Cu pad expansion of ~1 nm/1 µm Cu thickness per 50°C, the Cu recess must be in the range of 3 – 5 nm from the dielectric level [16] and also exhibit a microroughness in the range of a few nanometers [17]. The surface preparation is one of the most critical steps as the Cu recess must be achieved with high uniformity across the surface of 300-mm diameter silicon wafers.

Once the surface is ready, it is further processed to facilitate the dielectric-to-dielectric bond. In a low-temperature fusion bonding process flow, one or both wafers are first treated in a low-frequency plasma to condition the surface for bonding. The plasma treatment results in a surface activation and at the same time in subsurface implantation (in the nanometer-range) of various molecules useful for the interface chemistry to produce a good bond, such as water, oxygen, etc. [18]. The content of the created sub-surface reservoirs contributes actively to the change in interface chemistry during thermal annealing, lowering the thermal budget compared to non-activated surfaces (400°C vs. >1000°C). The bond is completed with a thermal annealing for 1-2 hours at a temperature typically ranging from 300°C to 400°C. After annealing, the maximum bond strength is reached and at the same time Cu-Cu bonds at pad level are completed.

An important aspect of fusion/hybrid bonding is the procedure of bringing the surfaces into contact: as the process is performed at ambient conditions, the gap between the two surfaces is filled with air. As air must not be trapped between the two surfaces, the bond (contact) must be initiated at a single location by gentle pressing with a pin (usually in the wafer center) [19]. As a result, a bond wave is generated between the two surfaces. The bond wave represents the spontaneous propagation of the bond towards the wafers' edges, and squeezes out the air from between the two surfaces.

While the main goal of single-point generation of the bond wave initially was related to effectively removing the air between the surfaces, in recent years the bond front was found to have an impact on various parameters influencing process success.

The mechanics of the bond initiation are important (Figure 5).

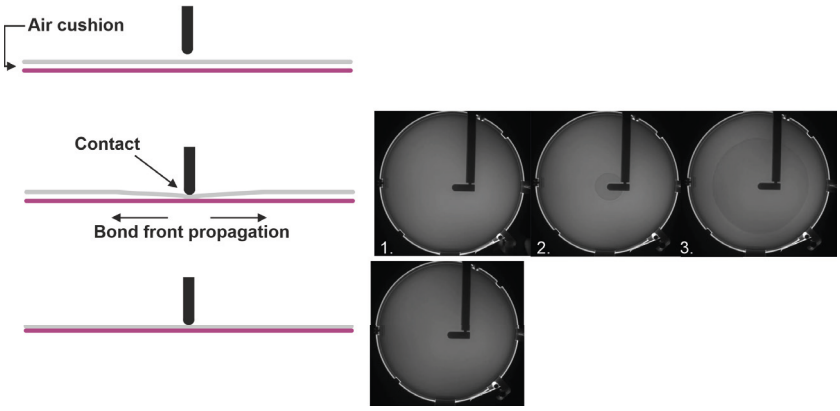

Figure 5
Bond Wave propagation: left – schematic drawing showing the bond initiation and wafer local deformation at contact point, and right – sequence of three IR images showing the bond propagation (top) an IR image showing bond completion (bottom). Bonded area shown in light gray contrast.

The bond wave initiation results in a local deformation at the pin contact point, while its propagation results in a light deformation of the wafers from center to edge, opposed by the atmospheric air pressure. In cases where the bond is performed under low pressure (vacuum) or the wafers are surface activated prior to bonding, the bond wave speed increases either due to the missing opposing air pressure (vacuum bonding) or due to the change in surface chemistry by activation. The bond speed has been determined to correlate with negative effects such as misalignment or pattern distortion, making its monitoring and controlling an important process optimization topic.

After bonding and thermal annealing, alignment accuracy is a very important first quality assessment. Through alignment key measurements, one may determine the three misalignment components: translation, rotation, and scaling (runout). This measurement is important as translation/rotation may be compensated relatively easily by adding offsets to the process, while for scaling compensation various hardware and process solutions can be used. Wafer-to-wafer alignment is performed using high-performance optical systems integrated with precise motion stages for positioning the wafers during alignment. Usually, alignment keys located at two sites positioned along the wafer diameter close to its opposite edges are observed with microscopes.

Previously for misalignment correction, the verification of the same two sites was employed. Based on the misalignment measurements, corrections were then applied globally to the wafer. However, with the decrease of alignment specifications to less than 100 nm, a global correction is no longer sufficiently accurate as it neglects local variations in distortion. For this reason, even though alignment still uses two sites, the alignment accuracy verification is extended to multiple measurement points across the bonded wafer area, and the data is fed back to an overlay model. In theory, the number of verification points defines the accuracy of the method. In practice, however, there is no generally applicable value. The number of measurements is decided based on die size, wafer shape and, of course, single measurement duration, as the alignment verification step must meet both technical and commercial requirements.

Not only the actual bonding, but all process steps that a wafer undergoes during its manufacturing can potentially add stress to it and are contributors to distortion, including:

- Front-end and mid-end-of-line processing / patterning
- Bonding to carriers
- Annealing for subsequent thinning
- Chemical mechanical polishing
- Etching
- Subsequent lithography steps after bonding and thinning

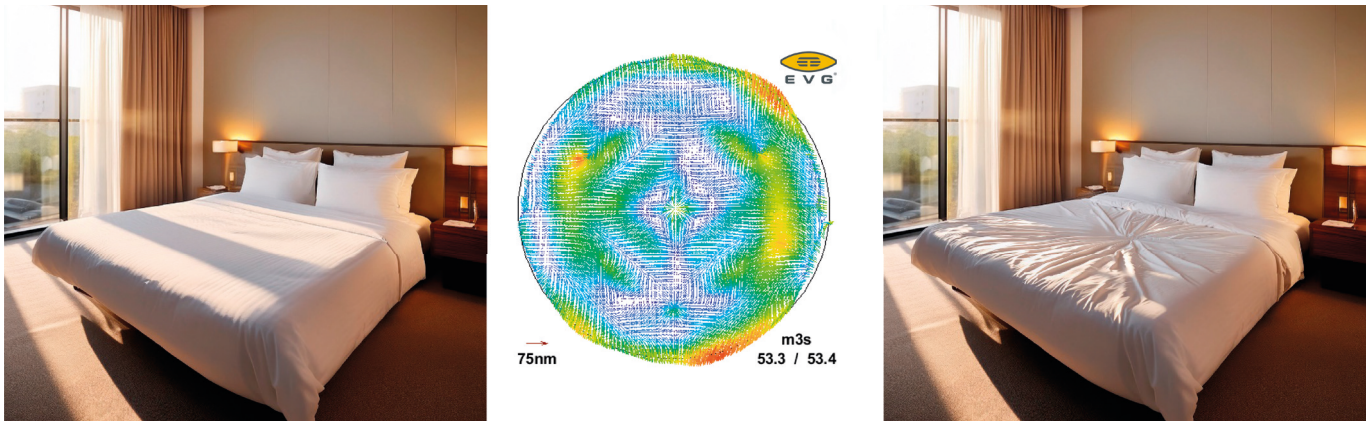

Figure 6
Wafer Distortion Fingerprint

All these stresses that are applied to the target wafer partly become visible as warpage, bow and other deformation, but also are "invisibly" represented by varying Young's moduli across the wafer's volume. Semiconductor wafers are stretchable and, if thin enough, behave like tissue (Figure 6). The detailed knowledge and capability of modelling physical wafer characteristics are key to advance overlay KPIs and enable new design concepts.

A prominent example of this is back side routing of power, also known as Backside Power Distribution Networks (BSPDNs). Further improved overlay accuracies even enable transistor gate contacting and thus also backside routing of signals. This relaxes single-sided routing congestions and thus enables overall denser designs [20].

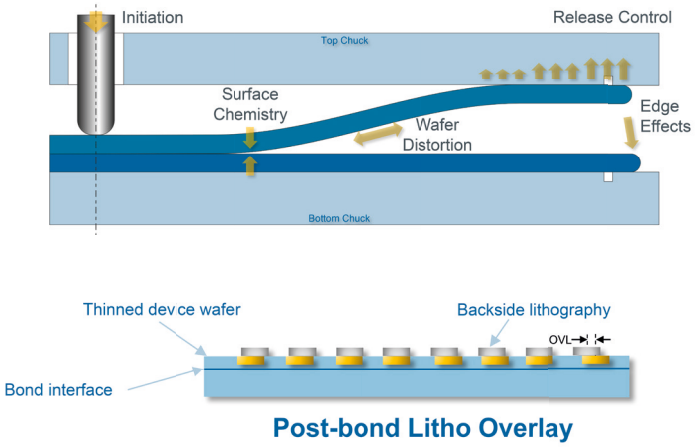

Figure 7
Wafer Distortion by Bonding and Post-bond Lithography Overlay

When pairing two patterned wafers for bonding, minimizing bond overlay variation is primarily driven by bond hardware design, including high accuracy alignment and sophisticated release control methodology (pre-compensation).

In comparison post-bond patterning provides additional optimization opportunities, as subsequent lithography steps can account for the distortion "fingerprint" and thus minimize post-bond litho overlay (Figure 7).

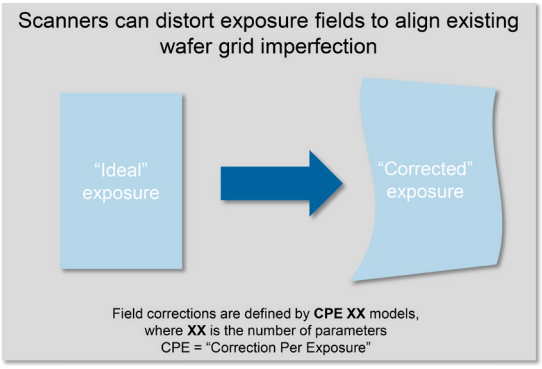

Figure 8
Scanner Corrections per Exposure

Distorting scanner exposure fields to align to the actual wafer grid imperfection (Figure 8) has achieved magnitudes down to 2.6 nm and even lower when neglecting edge regions [21], with edge improvements being "work in progress". It is noteworthy that research hints towards distortion fingerprint characteristics like vector field gradients and stability over time and batches of wafers as being key to push beyond today's limits. In other words, it is a matter of optimizing preprocessing and bonding steps in a way to enable lithography compensation to unfold its potential.

Summary

Over the past two decades, with the More than Moore paradigm taking shape, the enabling process technologies hybrid and fusion bonding have grown to also enable More Moore – transistor scaling - by going to the third dimension. Next to specific applications like image sensors and memory stacks, SoC (System on Chip) disaggregation enables next-generation, high-performance chip architectures, thereby opening up huge business potential, especially as these methodologies will eventually be migrating to lower-cost segments.

As the industry marches toward the ultimate goal of stacking complementary field-effect transistors (CFETs), co-optimization of bonding and lithography to further reduce overlay in ultra-thin die- and wafer stacks will be key to enabling this future device architecture.

Acknowledgement

This paper is based on an article which was first published in the May-August 2026 issue of Global Advanced Packaging magazine

[1] G. E. Moore, Cramming More Components onto Integrated Circuits, Electronics 38(8), (1965).
[2] Q.-Y. Tong and U. Gösele, Semiconductor Wafer Bonding, John Wiley & Sons (1999).
[3] T. Plach, K. Hentinnen, T. Suni and V. Dragoi, Chapter 23: Silicon Direct Bonding, in "Handbook of Silicon Based MEMS - Materials and Technologies", Elsevier Inc., Burlington, 3rd ed., 567 (2020).
[4] C. Flötgen, N. Razek, V. Dragoi and M. Wimplinger, Conductive Semiconductor Interfaces Fabricated by Room Temperature Covalent Wafer Bonding, ECS Trans. 75 (9), 45 (2016).
[5] W. Reinert, A. Kulkarni, V. Vuorinen, F. Lofink and P. Merz, Chapter 26: Metal Alloy Seal Bonding, in "Handbook of Silicon Based MEMS – Materials and Technologies", Elsevier Inc., Burlington, 3rd ed., 609, (2020).
[6] R. Knechtel, S. Dempwolf and M. Schikowski, Chapter 25: Glass Frit Bonding in "Handbook of Silicon Based MEMS – Materials and Technologies", Elsevier Inc., Burlington, 3rd ed., 593, (2020).
[7] V. Dragoi, C. Flötgen, J. Burggraf, L. Oggioni and T. Suga, Chapter 27: Emerging Wafer Bonding Technologies, in "Handbook of Silicon Based MEMS – Materials and Technologies", Elsevier Inc., Burlington, 3rd ed., 631 (2020).
[8] J.-Q. Lu, 3-D Hyperintegration and Packaging Technologies for Micro-Nano Systems, IEEE Proc. vol. 97(1), 18 (2009), doi: 10.1109/JPROC.2008.2007458.
[9] Q.-Y. Tong, Room Temperature Metal Direct Bonding, Appl. Phys. Lett. 89, 182101 (2006).
[10] P. Enquist, High Density Interconnect (DBI) Technology for Three Dimensional Integrated Circuits Applications, MRS Proc. vol. 970, 0970-Y01-04 (2007).
[11] V. Dragoi, S. Farrens and P. Lindner, Low temperature MEMS manufacturing processes: plasma activated wafer bonding, MRS Proc. vol. 872, J7.1.1 (2005).
[12] J. Burggraf, T. Uhrmann and M. Pires, Collective Die Bonding: an Enabling Toolkit for Heterogenous Integration, ECS Trans. 98(4), 173 (2020).
[13] DV ref. D2W- R. Hung, G. See, Y. Wang, C. Bum Yong, K. Zheng, Y. Chang, A. Shantaram, R. Wang, A. Sundarrajan, J. Abdila, N. Hegde, S. Schmid, D. Bikalevic and M. Glantschnig, Enabling Die-to-Wafer Hybrid Bonding for the Next Generation Advanced 3D Packaging, IEEE Proc. 74th ECTC Conf, 778 (2024), doi: 10.1109/ECTC51529.2024.00127.
[14] P. Urban, B. Povazay, T. Uhrmann, M.J. Gruber, B. Thallner and V. Dragoi, Novel Inorganic IR Release Process for High Temperature W2W and D2W Integration, IEEE Proc. of 74th ECTC, 415 (2024), doi : 10.1109/ECTC51529.2024.00073.
[15] T. Plach, V. Dragoi, F. Murauer and K. Hingerl, Plasma Activation for Low Temperature Wafer Bonding,ECS Trans. 16(8) 549 (2008).
[16] S. Lhostis, A. Farcy, E. Deloffre, F. Lorut, S. Mermoz, Y. Henrion, L. Berthier, F. Bailly, D. Scevola, F. Guyader, F. F. Gigon, C. Besset, S. Pellisier, L. Gay, N. Hotellier, M. Arnoux, A.-L. Le Berrigo, S. Moreau, V. Balan, F. Fournel, A. Jouve, S. Cheramy, B. Rebhan, G.A. Maier and L. Chitu, Proc. IEEE 66th ECTC, 869 (2016), DOI 10.1109/ECTC.2016.202.
[17] B. Rebhan, S. Tollabimazraehno, G. Hesser and V. Dragoi, Analytical Methods Used for Low Temperature Cu-Cu Wafer Bonding Process Evaluation, J. Microsys. Techn. 21(5), 1003 (2015).
[18] T. Plach, K. Hingerl, S. Tollabimazraehno, G. Hesser and V. Dragoi, Mechanisms for Room Temperature Direct Wafer Bonding, J. Appl. Phys. 113, 094905 (2013).
[19] C. Mühlstätter, L. Koller, T. Plach, V. Dragoi, and M. Wimplinger, Multiphysics overlay modelling of monolithic 3D fusion and hybrid bonding processes, IEEE Proc. of 74th ECTC, 1548 (2024), doi: 10.1109/ECTC51529.2024.00252.
[20] A. Kedilaya, S. Oruganti, N. Gupta, X. Zhang, I. Karpov, M. Anders and J. Kulkarni (2025), Beyond Backside Power: Backside Signal Routing as Technology Booster for Standard Cell Scaling. IEEE J. on Exploratory Solid-State Comput. Dev. and Circuits. PP. 1-1. 10.1109/JXCDC.2025.3617784.
[21] R. van Haren, S. Li, M. Baars, B. Minghetti, L. van Dijk, I. Mendes, K. Abadie, M.-L. Pourteau, G. Mauguen, M. May, V. Balan, F. Fournel, L. Pain, T. Plach, G. Probst, M. Wimplinger, Characterization and mitigation of local wafer deformations introduced by direct wafer-to-wafer bonding, SPIE Proc. vol. 12956 – Novel Patterning Technologies, 1295606 (2024).

Contact

Companies and other parties interested in partnering with the Heterogeneous Integration Competence Center™ can contact EV Group by phone at +43 7712 5311 0 or e-mail.

Get in touch:

HeterogeneousIntegration@EVGroup.com