Hybrid Bonding Process Flow - Advanced Packaging Part 5 – February 9, 2024

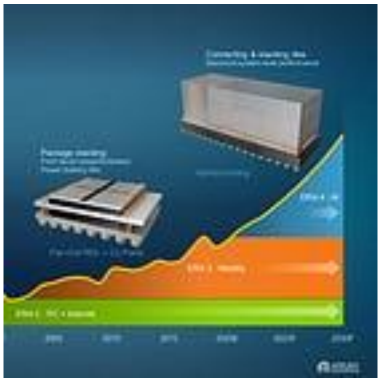
BESI, EV Group, AMAT, TEL, ASMPT, SET, Shibaura, SUSS Microtec

Hybrid bonding is going to be the most transformative innovation to semiconductor manufacturing since EUV. In fact, it will have an even bigger impact on the design process than EUV itself, branching from package architecture down to cell design and layouts. The IP ecosystem will be dramatically reshaped, but so will manufacturing flows. The era of shrinking transistors in 2D will continue, but at a muted pace, but hybrid bonding will bring in a new age, where chip designers think 3D.

With that hype-filled ballad finished, we should note there are many major engineering and technical challenges for bringing hybrid bonding to market in high volume, as today it is only reserved to a few AMD chips, CMOS image sensors, and some vendors' 3D NAND. This transition will reshape supply changes and design flows.

We will go from **the basics** all the way through to advanced aspects of hybrid bonding from process flow, tooling, design use cases, challenges, costs of chip on wafer versus wafer on wafer. We will also walk through our proprietary adoption modeling that captures use, tooling requirements, and volume by market (mobile, client PC, datacenter CPU, AI Accelerator, HBM, etc) as well as firm level adoption through to the end of the decade.

Hybrid bonding is a continuation of our advanced packaging series. In part 1 of this series, we discussed the need for advanced packaging and an overview of the economics on process node shrinks and why advanced packaging is needed so badly.



Advanced Packaging Part 1 – Pad Limited Designs, Breakdown Of Economic Semiconductor Scaling, Heterogeneous Compute, and Chiplets

DYLAN PATEL
·
DECEMBER 15, 2021

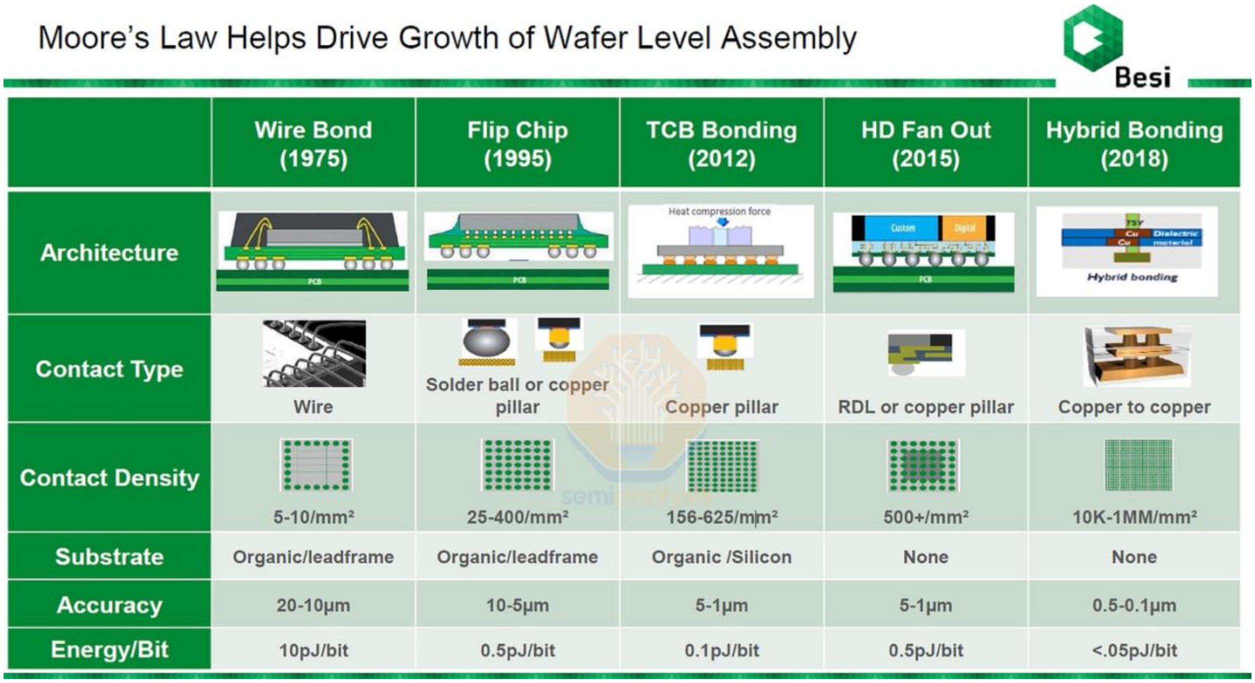
Read full story

In part 2, we went through all the various packaging technologies and their use cases. In part 3, we discussed Intel's bet on thermocompression bonding (TCB) and the tool landscape. In part 4 we talked about fanouts, organic interposers, and silicon bridges as a path past expensive passive interposers. As the AI supply chain heated up, we also covered CoWoS supply chains, allocations, and HBM. We also covered the upstream equipment supply chain and capacity adds. SemiAnalysis offers a detailed model from production to chip SKUs to volumes, ASPs, and revenue by fabless design, semicustom partner, and hyperscaler while also tracking accelerator install base at over 30 firms. Our model preceded beats and/or guidance increases at Nvidia, Broadcom, AMD, and Marvell, and the market is still misunderstanding the capex ramps.

In part 5 and 6 of this series on advanced packaging, we will first discuss the process flow of hybrid bonding, difficulties related to manufacturing, major players in the equipment tool chain as well as the major players and designs using and involved in hybrid bonding, such as TSMC, Intel, Samsung, SK Hynix, Micron, CXMT, Sony, Omnivison, YMTC, Kioxia, Western Digital, Besi, Shibaura Tokyo Electron, Applied Materials, ASM Pacific, EV Group, SUSS Microtec, SET, Bosch, Adeia (formerly Xperi), and many more. We have a very non-consensus view on BESI.

Then we will cover the battle waging between chip-on-wafer (D2W) and wafer-on-wafer (W2W) hybrid bonding. In our series, we will directly break down adoption rates for various product lines and end markets for AMD, Apple, Nvidia, Marvell, Broadcom, etc. and marry that with data around tool throughput, utilization, bond step counts, evolving use cases, and more. We will also dig into the design flow, custom customer-specific modifications for TSMC's N2 process technology, future use cases as hybrid bonding pitches continue to shrink, and how the current cost barrier gets conquered.

This is part 5, primarily focused on the process flow, so let's get started.



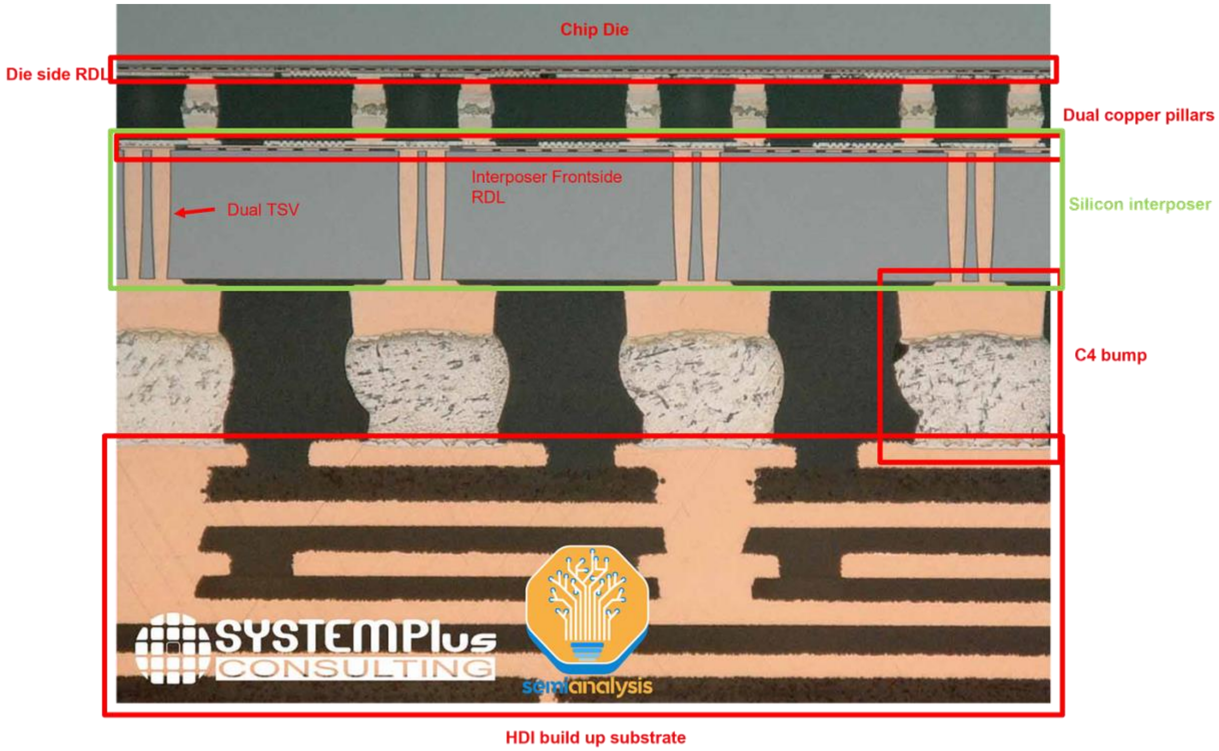
Moore's Law Helps Drive Growth of Wafer Level Assembly

Besi

	Wire Bond (1975)	Flip Chip (1995)	TCB Bonding (2012)	HD Fan Out (2015)	Hybrid Bonding (2018)
Architecture					
Contact Type	Wire	Solder ball or copper pillar	Copper pillar	RDL or copper pillar	Copper to copper
Contact Density	5-10/mm²	25-400/mm²	156-625/mm²	500+/mm²	10K-1MM/mm²
Substrate	Organic/leadframe	Organic/leadframe	Organic /Silicon	None	None
Accuracy	20-10µm	10-5µm	5-1µm	5-1µm	0.5-0.1µm
Energy/Bit	10pJ/bit	0.5pJ/bit	0.1pJ/bit	0.5pJ/bit	<.05pJ/bit

In the history of packaging, the last major paradigm shift was from wire bonding to flip chip. Since then, more advanced forms of packaging such as wafer level fan outs and TCB have been incremental improvements of the same core principles. These packaging methods all use some kind of bump with solder as the interconnect between silicon and package or board. These technologies can scale all the way down to ~20-micron pitch.

So far, the major packaging types and process flows we have discussed in our multi-part advanced packaging series have been at that 220-micron to 100-micron scale and mostly used solder as the medium between the various chiplets' copper interconnects. To scale further there needs to be another paradigm shift: bumpless interconnects with hybrid bonding. Hybrid bonding scales beyond 10-micron interconnect pitch with a roadmap to the 100's nanometer regime, and it does not use any intermediary such as solder that has higher resistance.



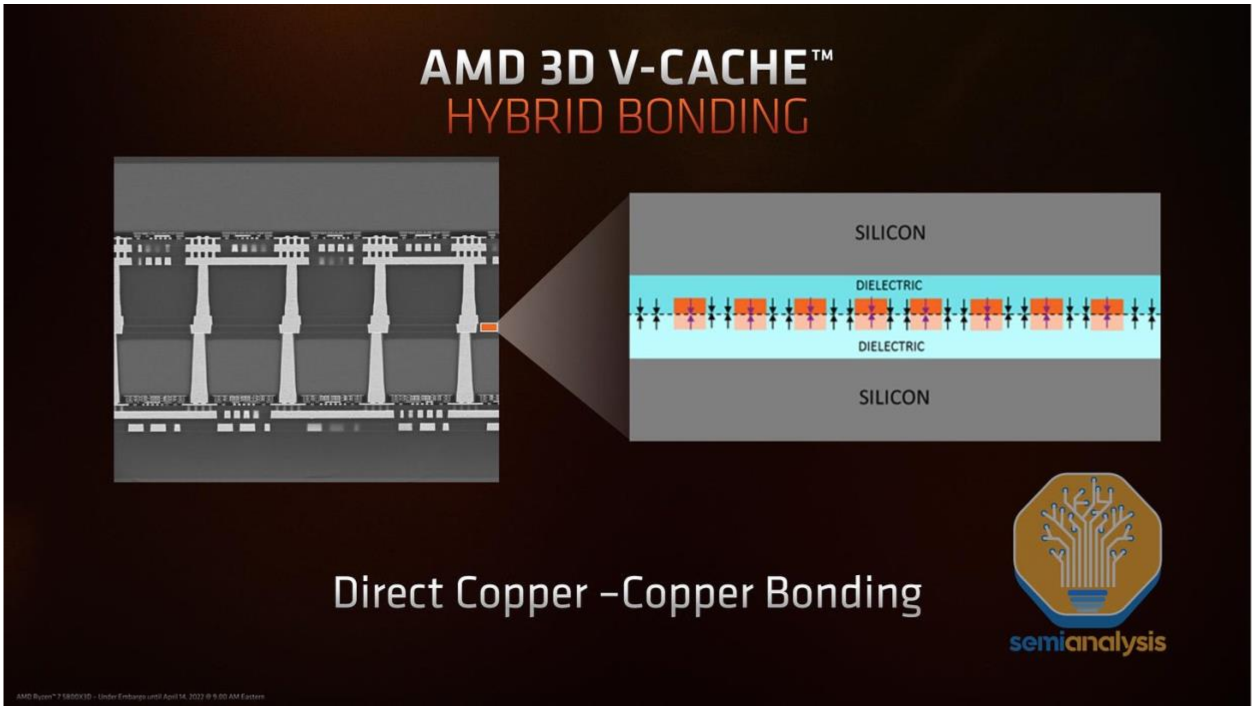
Nvidia A100 cross section, Silver blobs on the C4 bumps (~130-micron pitch) and on the copper pillars (~50-micron pitch) are solder.

Instead, the interconnects of different chips or wafers are directly connected with copper vias. Direct copper connections translate to much lower resistance and therefore lower power when sending data across to various chips. When combined with the orders magnitude increase in number of connections, there is a complete rethinking of design.

To recap from part 1, what is the whole point of Advanced Packaging? We can see that the progression of packaging technology aims to achieve greater interconnection density (more interconnects per area), reduce trace length to lower latency and energy per bit of transfer. We can see how hybrid bonding addresses both: trace length is reduced significantly so latency is as low as it can be without being on-die, and in some cases shorter than global rounding on chip, and interconnection pitch can scale well below 10-micron to increase density.

What exactly is Hybrid Bonding?

Hybrid bonding is used for the vertical (or 3D) stacking of chips. The distinguishing feature of hybrid bonding is that it is bumpless. It moves away from solder-based bump technologies to direct copper-to-copper connections. This means the top die and bottom die are flush against each other. Instead of bumps, both dies just have copper pads that can be scaled to ultra-fine pitches. There is no solder, so the problems related to solder are avoided.



AMD 3D V-CACHE™

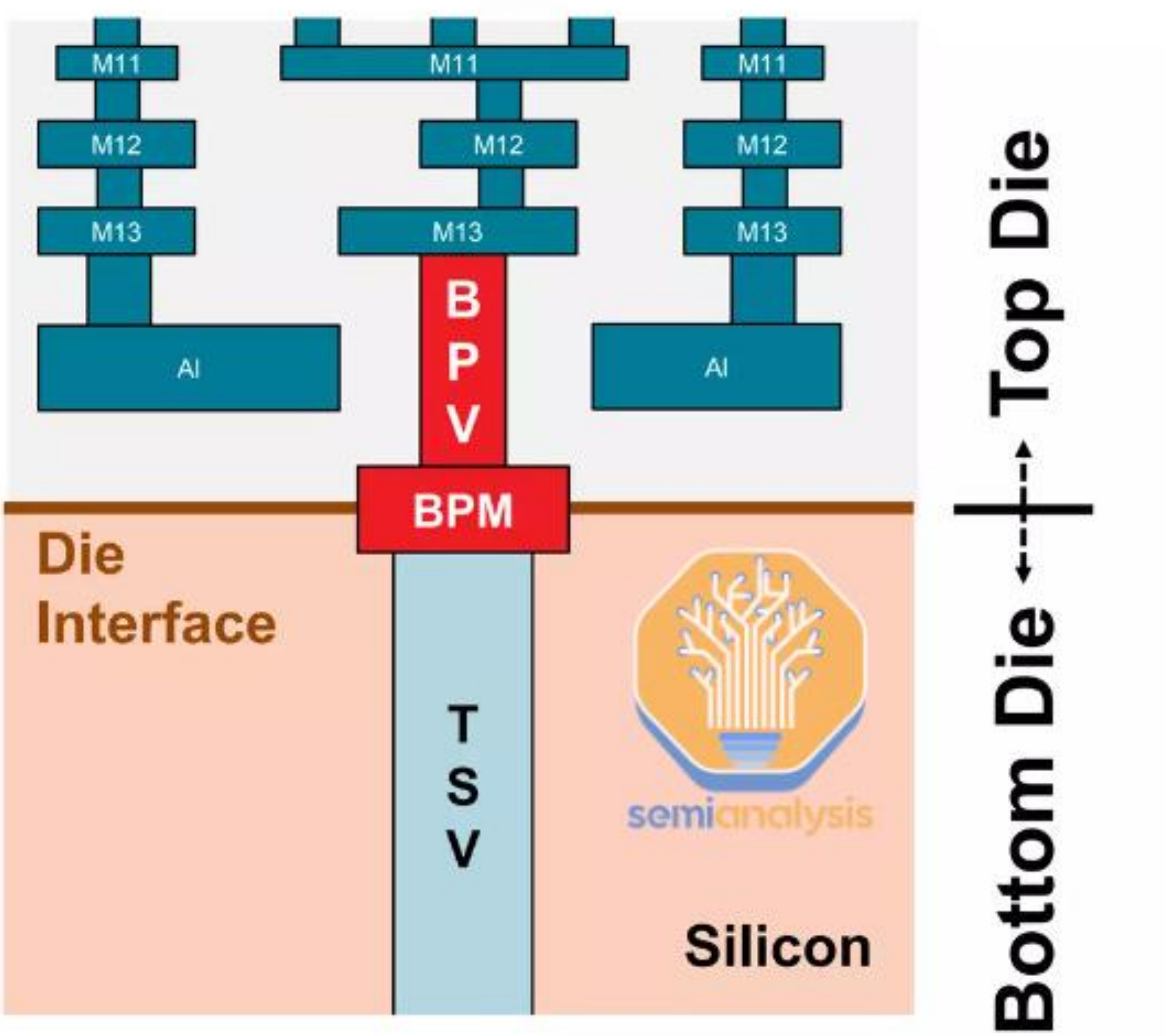
HYBRID BONDING

Direct Copper –Copper Bonding

AMD Ryzen™ 7 5800X3D - Under Embargo until April 14, 2022 @ 9:00 AM Eastern

From the above image, we can see a cross-section of AMD’s 3D V-Cache, which uses TSMC’s SoIC-X die-to-wafer hybrid bonding. The bond interface between the top and bottom silicon is a hybrid bonding layer that is present on top of the metal layers of the silicon dies. The hybrid bond layer is a dielectric (now most commonly SiO or SiCN) that is patterned with copper pads and vias that are usually sub-10-micron pitch.

The purpose of the dielectric is to insulate each pad so that there is no signal interference between the pads. The copper pads are connected to the chip metal layers with Through Silicon Vias (TSVs). TSVs are needed to deliver power and signal to the other die in the stack. As the bottom die is placed “face down” these vias are needed to connect the metal layers on the top die, passing through the transistor layers to the metal layers on the bottom die.



It is these copper pads that the signals pass through for die-to-die communication. The reason why the bond is “hybrid” is because it is a combination of a dielectric-dielectric bond, and a direct copper-to-copper bond. There is no additional adhesive or materials used in between the bond interfaces.

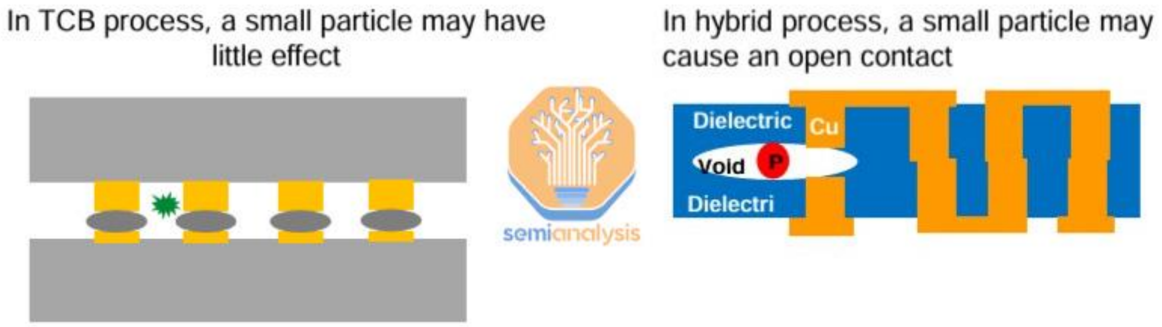
Key Process Conditions

Compared to previous bump-based interconnects, a whole new set of technology and process challenges are introduced. There are very stringent requirements on: surface smoothness, cleanliness, and bond alignment accuracy to achieve a high-quality bond. We will describe some of these challenges first as the process flow is designed around mitigating these challenges. Having these in mind will help you better understand why the flow is the way it is, as well as the pros and cons of different methods.

Particles and Cleanliness

In any discussion about hybrid bonding, particles will come up. This is because particles are the enemy of yields in hybrid bonding. As hybrid bonding involves 2 very smooth and planar surfaces being bonded flush together, the bond interface is **very** sensitive to the presence of any particles.

A particle of just 1 micron height results in a bond void of 10 millimeters in diameter which would result in a defective bond. With a bump-based interconnect, there is always a gap between the device and substrate which can tolerate some particles, as underfill or non-conductive films are used.


In TCB process, a small particle may have little effect
semianalysis
In hybrid process, a small particle may cause an open contact
Dielectric
Cu
Void
P
Dielectri

Maintaining cleanliness is critical, and this is very challenging. Particles come from many steps such as wafer dicing, grinding, and polishing. Any kind of friction also generates particles, which is an issue, especially because hybrid bonding involves mechanically picking up dies and placing them on top of the other chips. There is a lot of movement in the tool coming from the die bond-head as well as the die flipper. Particles are unavoidable, but there are several techniques to mitigate the yield impact.

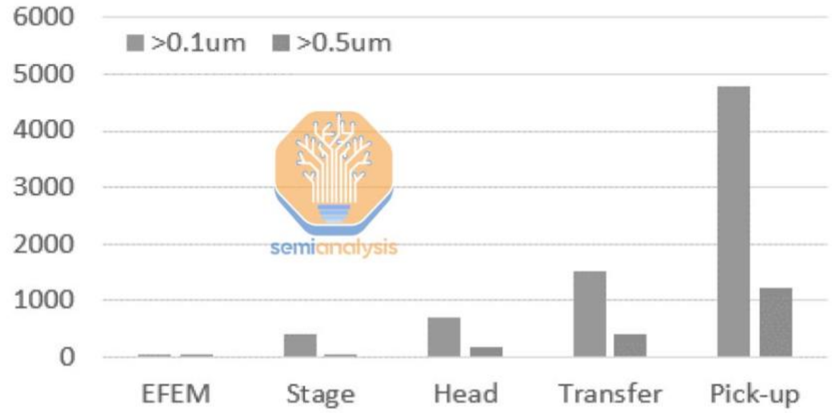
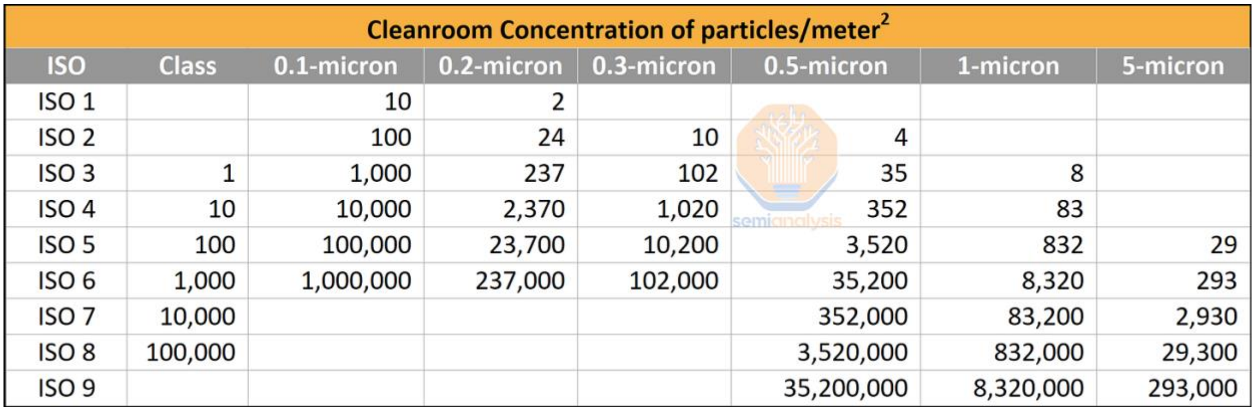


Fig.10 the number of particles generated by area in C2W Bonder

Of course, wafer cleaning is performed regularly to get rid of contaminants. However, cleans are imperfect and do not get rid of 100% of the contaminants in a single pass, so it is better to avoid contaminants in the first place. Cleanrooms required for hybrid bonding are far more advanced than those required for other forms of advanced packaging.



Cleanroom Concentration of particles/meter^2

ISO	Class	0.1-micron	0.2-micron	0.3-micron	0.5-micron	1-micron	5-micron
ISO 1		10	2				
ISO 2		100	24	10	4		
ISO 3	1	1,000	237	102	35	8	
ISO 4	10	10,000	2,370	1,020	352	83	
ISO 5	100	100,000	23,700	10,200	3,520	832	29
ISO 6	1,000	1,000,000	237,000	102,000	35,200	8,320	293
ISO 7	10,000				352,000	83,200	2,930
ISO 8	100,000				3,520,000	832,000	29,300
ISO 9					35,200,000	8,320,000	293,000

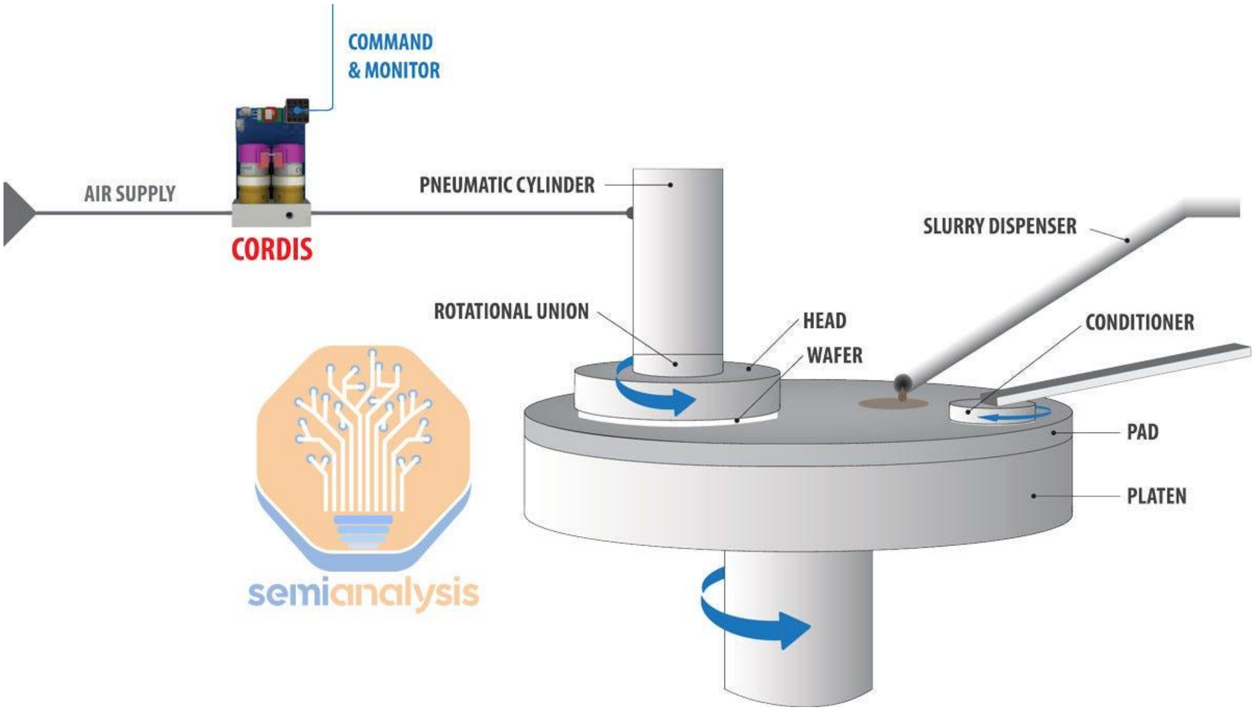
Therefore, hybrid bonding generally requires class 1 / ISO 3 clean rooms and equipment or better. TSMC and Intel, for example, are going all the way to ISO 2 or ISO 1 classes. This is one major reason why Hybrid Bonding is considered a “front-end” process, i.e it happens in environments similar to fabs rather than those of traditional packaging players (OSATs). It is very difficult for OSATs to pursue hybrid bonding given this upgrade in cleanliness requirement. Most OSATs would need to build newer more advanced cleanrooms if they wanted to be involved in hybrid bonding, whereas firms like TSMC and Intel can use older fabs or build to similar standards as their existing fabs.

The process flow of hybrid bonding also involves many tools that have traditionally only been used exclusively by fabs. Outsourced assembly and test firms (OSAT) such as ASE and Amkor have relatively little experience with chemical vapor deposition (CVD), etch, physical vapor deposition (PVD), electrochemical deposition (ECD), chemical-mechanical planarization (CMP), and surface preparation/activation.

The combination of cleanliness required and increase in tools contributes to a huge cost increase. The process of hybrid bonding is not a cheap one relative to other forms of packaging. We will walk through the process flow below.

Smoothness

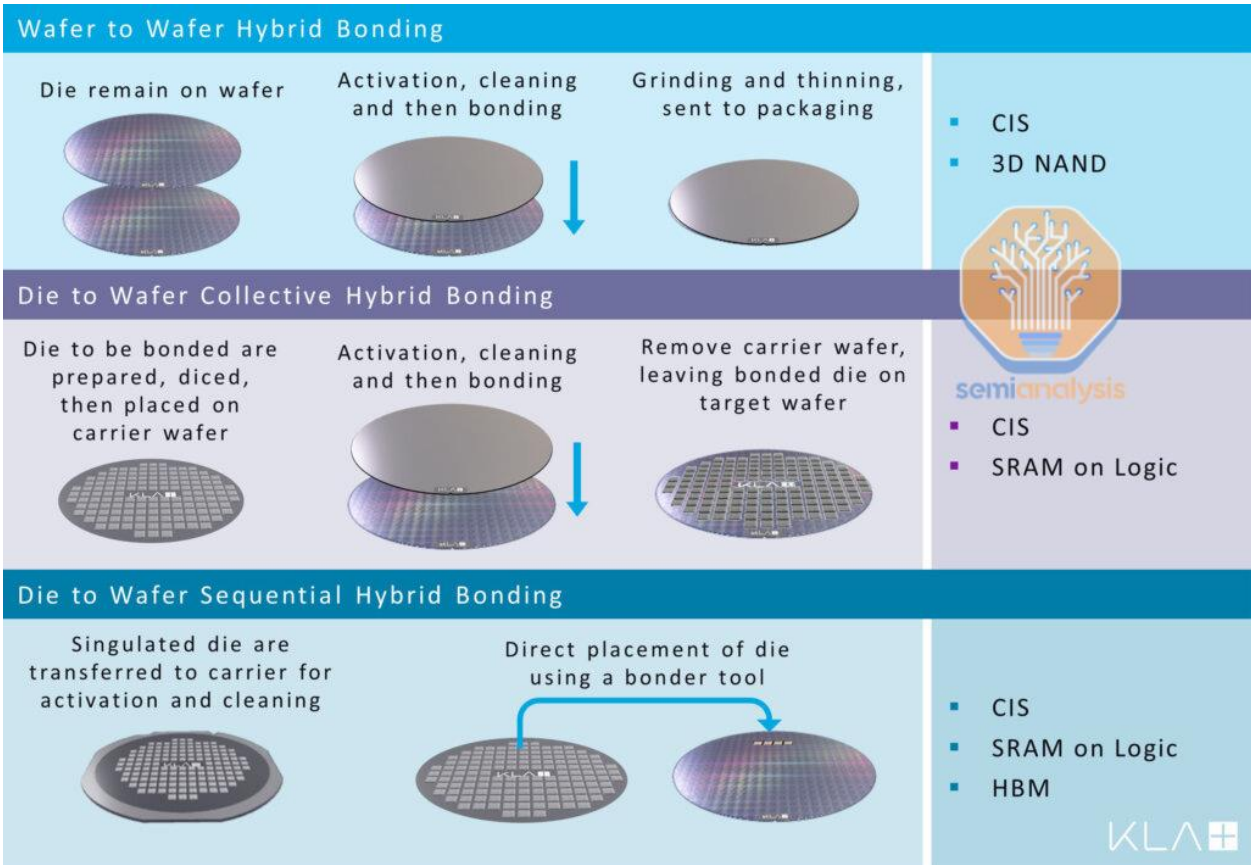
The surface smoothness of the hybrid bonding layers is also extremely critical. The HB interface is similarly sensitive to any kind of topography, which would create voiding and an ineffective bond. The surface roughness threshold is generally said to be 0.5nm for the dielectric and 1nm for the copper pads. To achieve this level of smoothness, chemical mechanical planarization (CMP) is performed and is highly a critical process for hybrid bonding.



After polishing, this smoothness needs to be maintained all the way through the flow. Any steps that could damage this surface, such as harsh cleaning, are avoided. Even probing for wafer sort needs to be adjusted so the surface does not become damaged.

Wafer-to-Wafer (W2W) or Die-to-Wafer (D2W)

First, a discussion on W2W or D2W. Hybrid bonding can be completed via wafer-to-wafer (W2W) or die-to-wafer (W2W) processes. W2W means two fabricated wafers are bonded directly together. W2W offers higher alignment accuracy, throughput, and bond yields. The vast majority of hybrid bonding is currently completed via W2W, given its relative ease.


Wafer to Wafer Hybrid Bonding
Die remain on wafer
Activation, cleaning and then bonding
Grinding and thinning, sent to packaging
CIS
3D NAND
Die to Wafer Collective Hybrid Bonding
Die to be bonded are prepared, diced, then placed on carrier wafer
Activation, cleaning and then bonding
Remove carrier wafer, leaving bonded die on target wafer
CIS
SRAM on Logic
Die to Wafer Sequential Hybrid Bonding
Singulated die are transferred to carrier for activation and cleaning
Direct placement of die using a bonder tool
CIS
SRAM on Logic
HBM
KLA
semianalysis

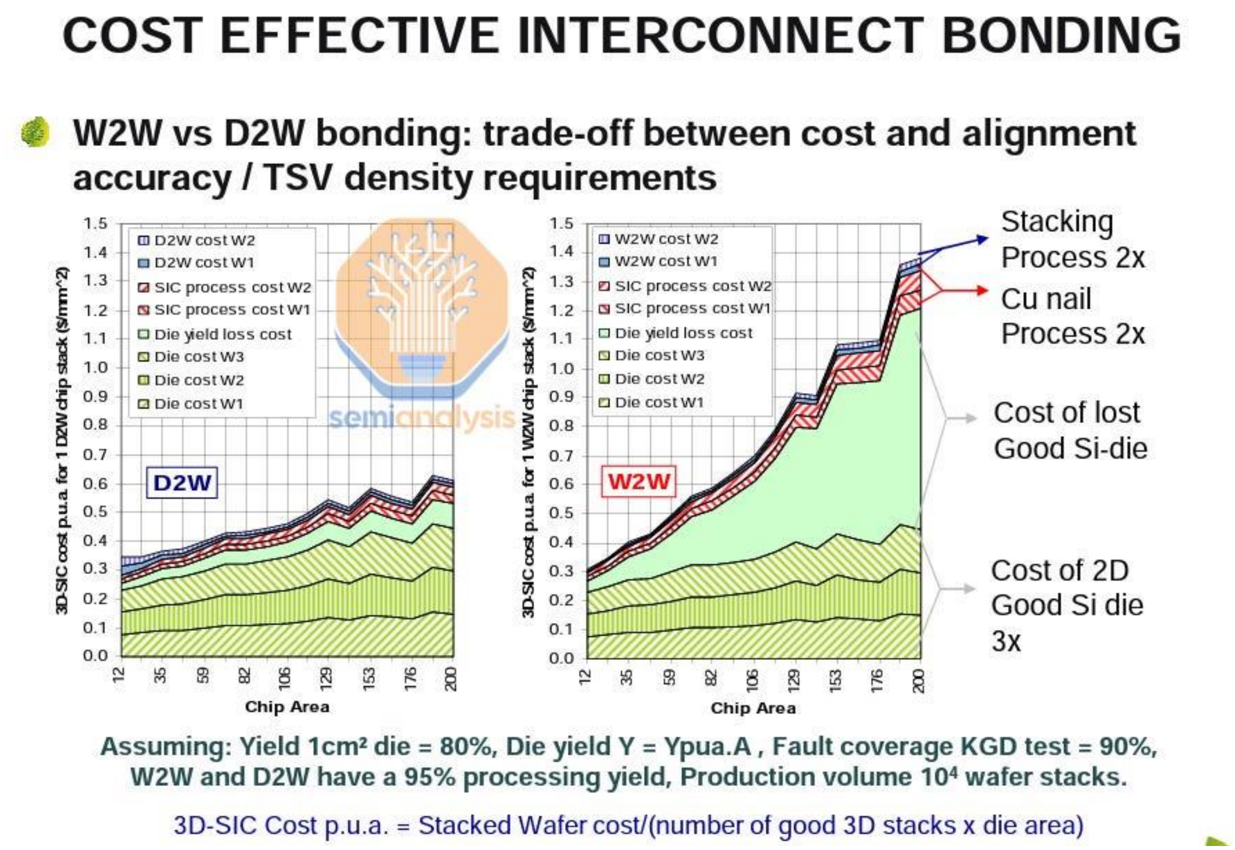
What makes W2W bond yield higher is that the alignment and bonding steps are separated. In W2W tools there is a separate chamber that performs the alignment. Once the top and bottom wafers are aligned, they are moved into the bonding chamber (which is in a vacuum), where they are pressed together with a bit of force and after a period of around 20 minutes, the initial pre-bond is formed.

The key to W2W is that it is a much cleaner process with fewer steps. Before the alignment and bonding, the wafers can be cleaned to remove most of the particles. Die singulation, a source of particle contamination, only happens after the bond. More time is also afforded for the alignment step as it is a wafer-level process so longer alignment time does not hurt throughput as much as it would for a chip-level process.

There is not much movement that happens in the chambers either so there are fewer contaminants in the chamber itself. Currently, W2W bonders can achieve alignment accuracy of sub-50nm. W2W bonding is already a mature process and not particularly expensive. The proof is that we see it widely adopted in mass-market products like 3-layer image sensors and NAND.

W2W bonding is great, but one major limitation is the inability to perform a wafer sort to select known good dies (KGD). This results in the unwanted outcome of bonding defective chips to good chips which can results in a waste of the good silicon.

Given this, W2W is used for wafers where yields are high which generally means on smaller designs. In the below chart we can see how the relationship between chip area and cost for W2W and D2W. W2W is cheaper at smaller die sizes as wafer yields will be higher. However, the W2W cost curve is much steeper as we go to higher die sizes and that is primarily driven by the cost of lost good dies. As chip size goes up the portion of good die per wafer decreases, resulting in more bonding of defective dies and good dies.



COST EFFECTIVE INTERCONNECT BONDING

- W2W vs D2W bonding: trade-off between cost and alignment accuracy / TSV density requirements

Stacking Process 2x

Cu nail Process 2x

Cost of lost Good Si-die

Cost of 2D Good Si die 3x

Assuming: Yield 1cm² die = 80%, Die yield Y = Ypua.A , Fault coverage KGD test = 90%, W2W and D2W have a 95% processing yield, Production volume 10^4 wafer stacks.

3D-SIC Cost p.u.a. = Stacked Wafer cost/(number of good 3D stacks x die area)

We can see that W2W is used for smaller chips that have high yields: CMOS image sensors, 3D NAND, and so far only in logic for the Graphcore Bow IPU.

FER ON WAFER PROC

Advanced silicon wafer stac technology co-developed b Graphcore and TSMC

World's first commercial de using TSMC SoIC-WoW™ te in BOW IPU

Enabling technology for clo coupled power delivery die maximize application perfo

Graphcore Announces World's First 3D Wafer On Wafer Hybrid Bond Processor

DYLAN PATEL

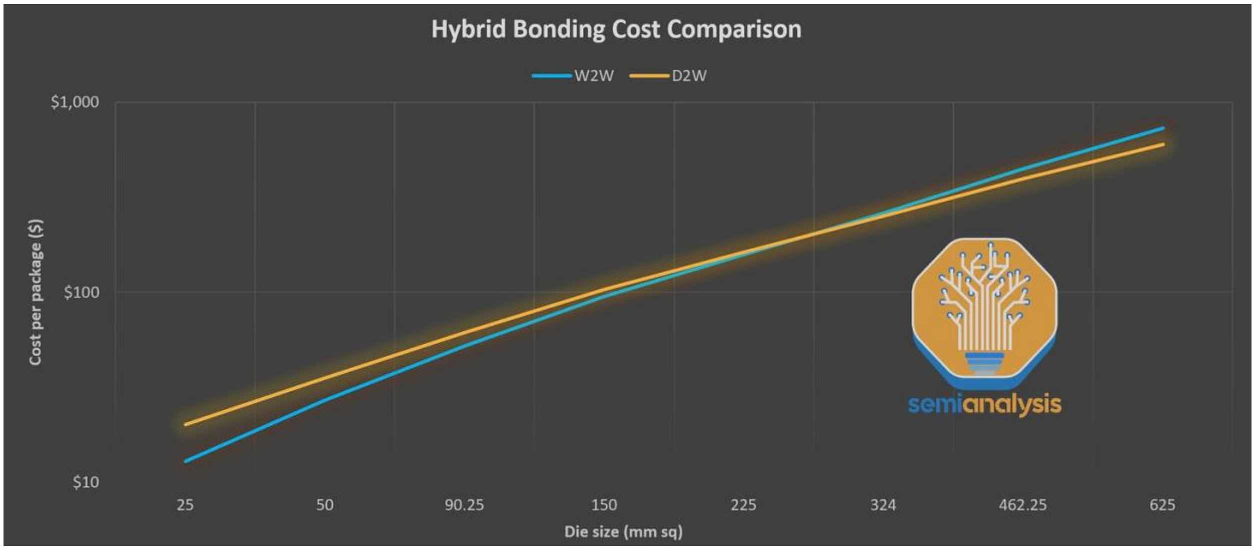
·

MARCH 3, 2022

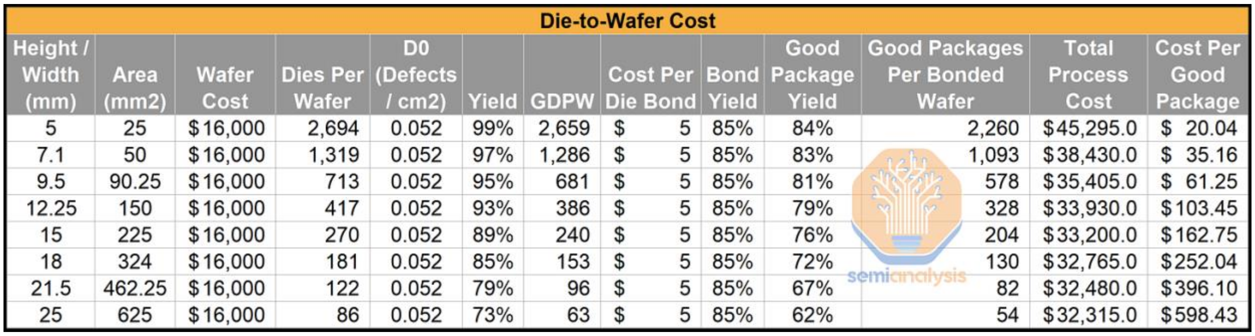
Read full story

While the Graphcore Bow IPU is a larger HPC chip, the top die isn't leading edge logic, it is a passive capacitor die used for power delivery, hence yields for this should be quite high and is much cheaper silicon. The other drawback with W2W is that both the top die and bottom die must be uniform in size, so this limits flexibility of heterogeneous integration options.

There are multiple levers with regard to cost. The main ones are wafer cost, D0 (defect density), and bond yields. Each of these levers can result in higher or lower costs. Note these are example figures to stress the point. Do not use the chart below as it does not show the actual cost for bonding. For actual costs for today's products, contact us for our AMD MI300X costing report or the Zen 3, Zen 4, and Zen 5 hybrid bonding costing report.

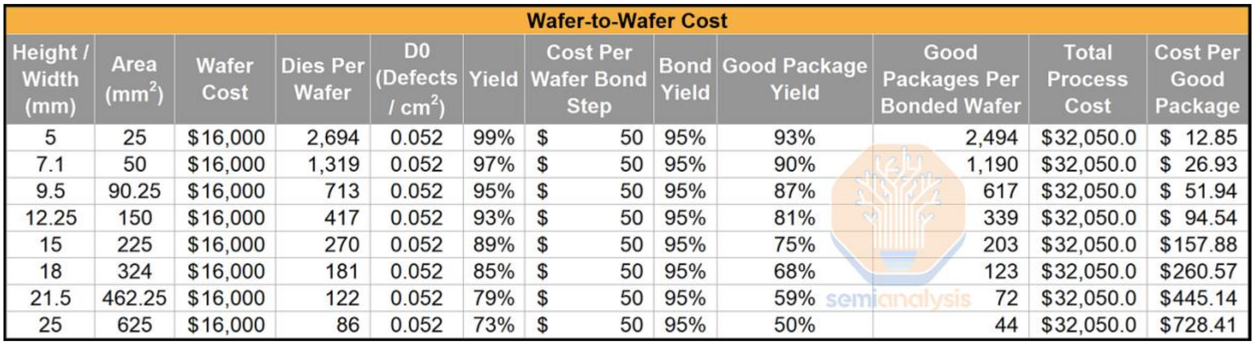


As can be seen, D2W is more expensive on small dies, but with large dies, this flips. W2W is more expensive. The ability to test and bond only known good die (KGD), instead of risking defects stacking and wasting good silicon, is critical and why die-on-wafer (D2W) is the first to be productized. It can deal with worse yields but still have a commercially viable product.



Die-to-Wafer Cost

Height / Width (mm)	Area (mm2)	Wafer Cost	Dies Per Wafer	D0 (Defects / cm2)	Yield	GDPW	Cost Per Die Bond	Bond Yield	Good Package Yield	Good Packages Per Bonded Wafer	Total Process Cost	Cost Per Good Package
5	25	$16,000	2,694	0.052	99%	2,659	$ 5	85%	84%	2,260	$45,295.0	$ 20.04
7.1	50	$16,000	1,319	0.052	97%	1,286	$ 5	85%	83%	1,093	$38,430.0	$ 35.16
9.5	90.25	$16,000	713	0.052	95%	681	$ 5	85%	81%	578	$35,405.0	$ 61.25
12.25	150	$16,000	417	0.052	93%	386	$ 5	85%	79%	328	$33,930.0	$103.45
15	225	$16,000	270	0.052	89%	240	$ 5	85%	76%	204	$33,200.0	$162.75
18	324	$16,000	181	0.052	85%	153	$ 5	85%	72%	130	$32,765.0	$252.04
21.5	462.25	$16,000	122	0.052	79%	96	$ 5	85%	67%	82	$32,480.0	$396.10
25	625	$16,000	86	0.052	73%	63	$ 5	85%	62%	54	$32,315.0	$598.43

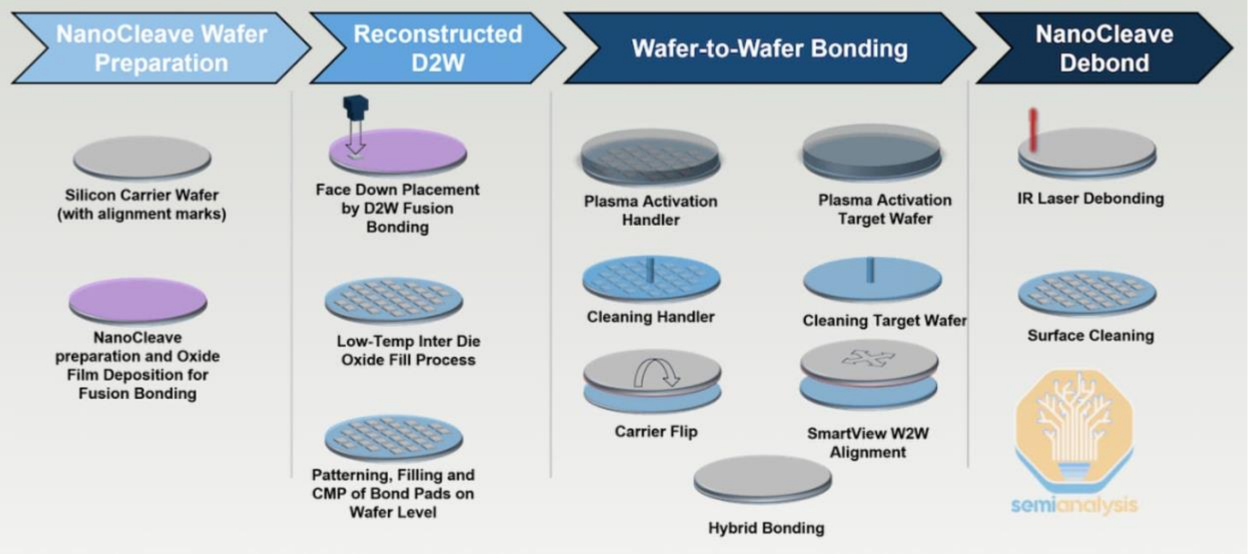


Wafer-to-Wafer Cost

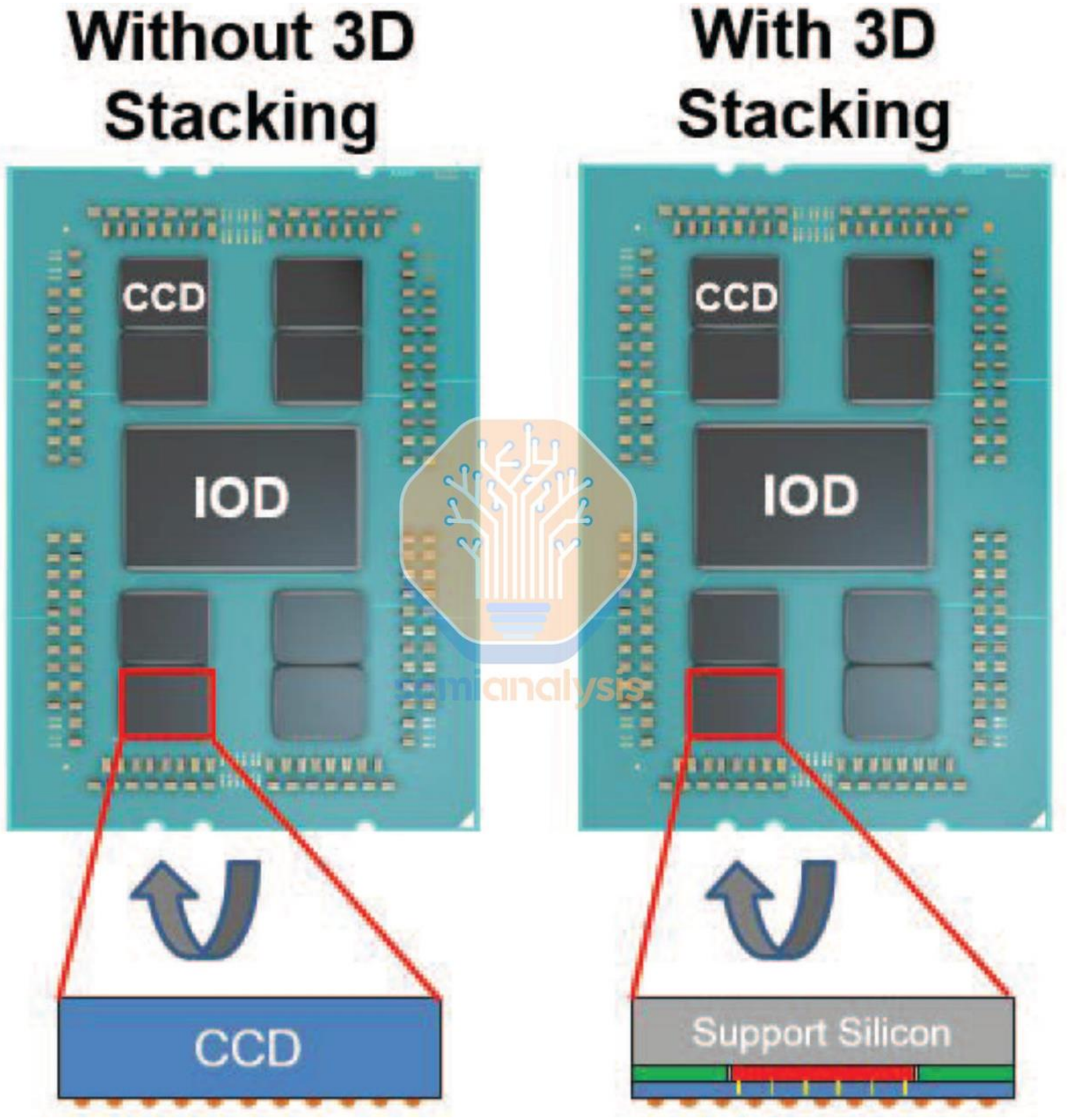
Height / Width (mm)	Area (mm^2)	Wafer Cost	Dies Per Wafer	D0 (Defects / cm^2)	Yield	Cost Per Wafer Bond Step	Bond Yield	Good Package Yield	Good Packages Per Bonded Wafer	Total Process Cost	Cost Per Good Package
5	25	$16,000	2,694	0.052	99%	$ 50	95%	93%	2,494	$32,050.0	$ 12.85
7.1	50	$16,000	1,319	0.052	97%	$ 50	95%	90%	1,190	$32,050.0	$ 26.93
9.5	90.25	$16,000	713	0.052	95%	$ 50	95%	87%	617	$32,050.0	$ 51.94
12.25	150	$16,000	417	0.052	93%	$ 50	95%	81%	339	$32,050.0	$ 94.54
15	225	$16,000	270	0.052	89%	$ 50	95%	75%	203	$32,050.0	$157.88
18	324	$16,000	181	0.052	85%	$ 50	95%	68%	123	$32,050.0	$260.57
21.5	462.25	$16,000	122	0.052	79%	$ 50	95%	59%	72	$32,050.0	$445.14
25	625	$16,000	86	0.052	73%	$ 50	95%	50%	44	$32,050.0	$728.41

To get around limitations we need to go to D2W. D2W bonding is much more challenging. After a wafer sort is performed, KGD are singulated from the top wafer, and attached individually onto the base wafer via a pick and place tool. This is more challenging in terms of bonding as there are several more bond steps per wafer. These additional steps introduce more particle contamination especially from die singulation and from the movement of the bond-head during pick-and-place.

D2W can be a “collective” process, where KGD are aligned and first temporarily bonded onto a reconstituted carrier wafer. Then the reconstituted carrier wafer is bonded onto the base wafer for the actual pre-bond. This is to separate the alignment and bond like with W2W and allows for a cleaning step before the final pre-bond to get rid of any contaminants that have been accumulated. The drawback is the additional steps involved and the additional W2W bond step creates more chance of alignment error.


NanoCleave Wafer Preparation
Reconstructed D2W
Wafer-to-Wafer Bonding
NanoCleave Debond
Silicon Carrier Wafer (with alignment marks)
NanoCleave preparation and Oxide Film Deposition for Fusion Bonding
Face Down Placement by D2W Fusion Bonding
Low-Temp Inter Die Oxide Fill Process
Patterning, Filling and CMP of Bond Pads on Wafer Level
Plasma Activation Handler
Plasma Activation Target Wafer
Cleaning Handler
Cleaning Target Wafer
Carrier Flip
SmartView W2W Alignment
Hybrid Bonding
IR Laser Debonding
Surface Cleaning
semianalysis

This is actually a simplistic flow as the bottom die can reconstituted on a carrier wafer too. So for both top and bottom chips are diced from the original silicon wafer and KGD are sorted. Both sets of chips are bonded onto precise locations on their respective carriers. Then, the 2 carrier wafers are bonded with a W2W process. This is done in TSMC SOIC. As such, there are 5 bond steps used for each AMD 3D V-cache chip (bottom CPU die to carrier, 3D V-cache chiplet to carrier, 2x dummy silicon to carrier), and wafer on wafer.



Reconstituted processes can also be used for more extreme options for heterogeneous integration. Intel demonstrated “quasi-monolithic chips (QMC)” at IEDM 2022. One example of a QMC application they showed off was a package with 2 dies each on the top and bottom heterogeneously integrated. For both the top and bottom, each die was attached to a carrier wafer. Then the wafer was molded with thick inorganic oxide such as SiO2. The W2W bond was performed. The molded chips were then singulated and attached to a package substrate to complete the flow.

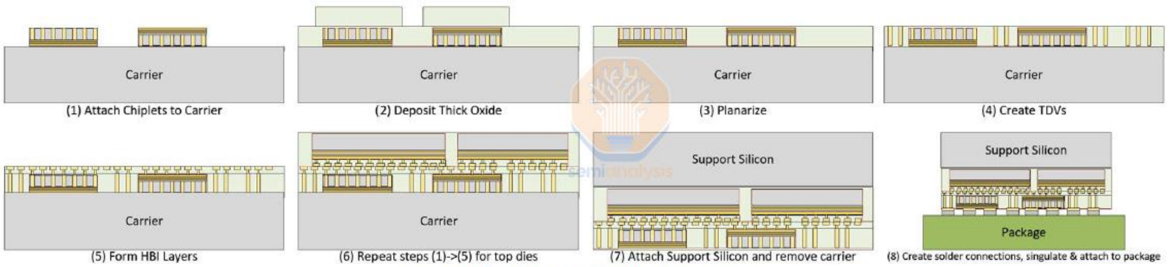

Carrier
(1) Attach Chiplets to Carrier
Carrier
(2) Deposit Thick Oxide
Carrier
(3) Planarize
Carrier
(4) Create TDVs
Carrier
(5) Form HBI Layers
Carrier
(6) Repeat steps (1)->(5) for top dies
Support Silicon
(7) Attach Support Silicon and remove carrier
Support Silicon
Package
(8) Create solder connections, singulate & attach to package

Fig. 11 QMC Simplified manufacturing process

Notice there can be TSV's in the area that is reconstituted.

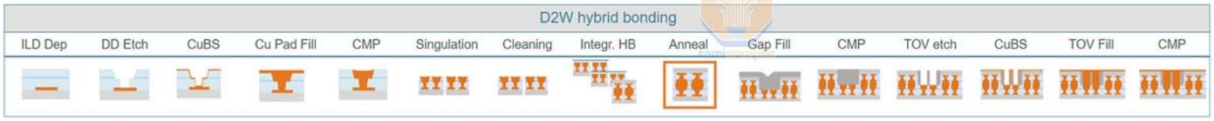
Direct D2W bonding is where the individual die are placed directly onto the target wafer for the pre-bond. Direct D2W is less mature but it seems going forward direct D2W will be used more due to process simplification. One benefit of collective D2W is that the clean can be performed then directly sent into the alignment chamber to reduce contamination. D2W cluster tools (will be discussed below) are now being introduced recently which can recreate this flow, reducing the benefit of this collective process. In addition, D2W is better for finer pad pitches as alignment becomes more challenging, so there is benefit in eliminating the W2W step which introduces additional risk of misalignment at the W2W step.

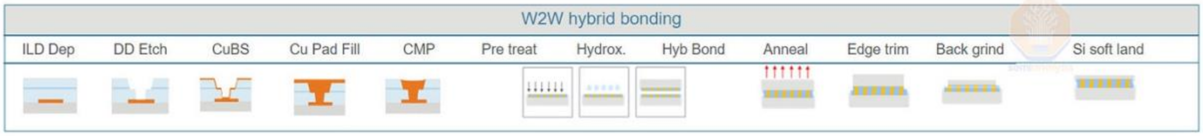
Given the process challenges and therefore cost of D2W hybrid bonding, current applications are limited. AMD was the first adopter in 2022 and remains the only adopter so far. We will discuss future applications, the rate of adoption by various firms, number of process steps, and more later.

One thing to note is that W2W is far ahead of D2W in alignment, so if your design isn't heterogeneous, and the wafer yield are high enough, it's actually going to be a more accurate higher yielding process. This finer pitch will also unlock many new use cases D2W cannot yet break into.

Hybrid Bonding Process Flow

Next lets go over the process flows for D2W and W2W in more detail.



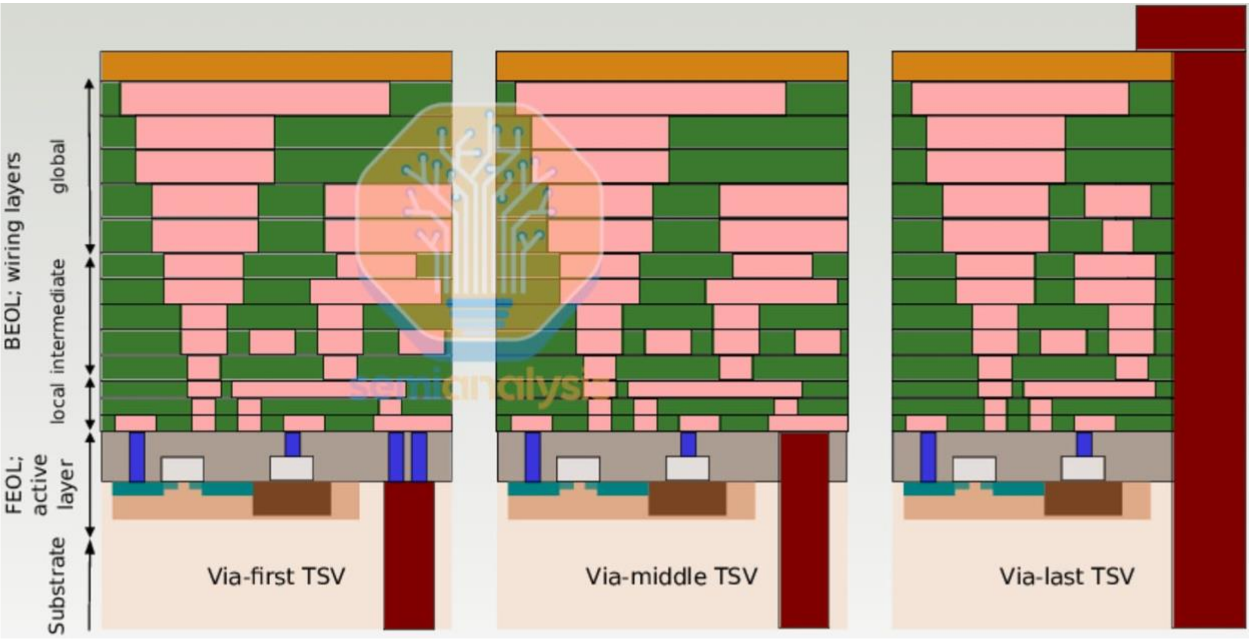


TSV formation

As we mentioned above TSVs are needed to deliver power and signal to all the chips in the package. Imagine a conventional flip chip package. The chip only needs interconnects on one side to receive power and communicate data with the package substrate. This interconnect layer has bumps that are connected to the passive wiring layers (also known as the “metal layers” or “back end of the line” / BEOL) that provide both power and signal to the transistor layers that switch and process data.

For a 3DIC, the die on the bottom needs to be able to communicate with the package substrate underneath it as well as the die on top of it so both sides of the die need interconnect. This is where TSVs come in. There are multiple variations of TSV depending on when in the flow they are fabricated. TSVs can be “via-first” where they are fabricated in the silicon first before the transistor layers, “via-middle” when they are fabricated after the transistor layers are complete and before the metal layers, or “via-last” which is after the BEOL.

Most common for 3DIC is the “via-middle” approach as the TSV runs between the metal layers, through to the transistor layers and revealing itself on the backside of the chip so that there is now a layer of interconnect on both sides of the chip which we will describe.



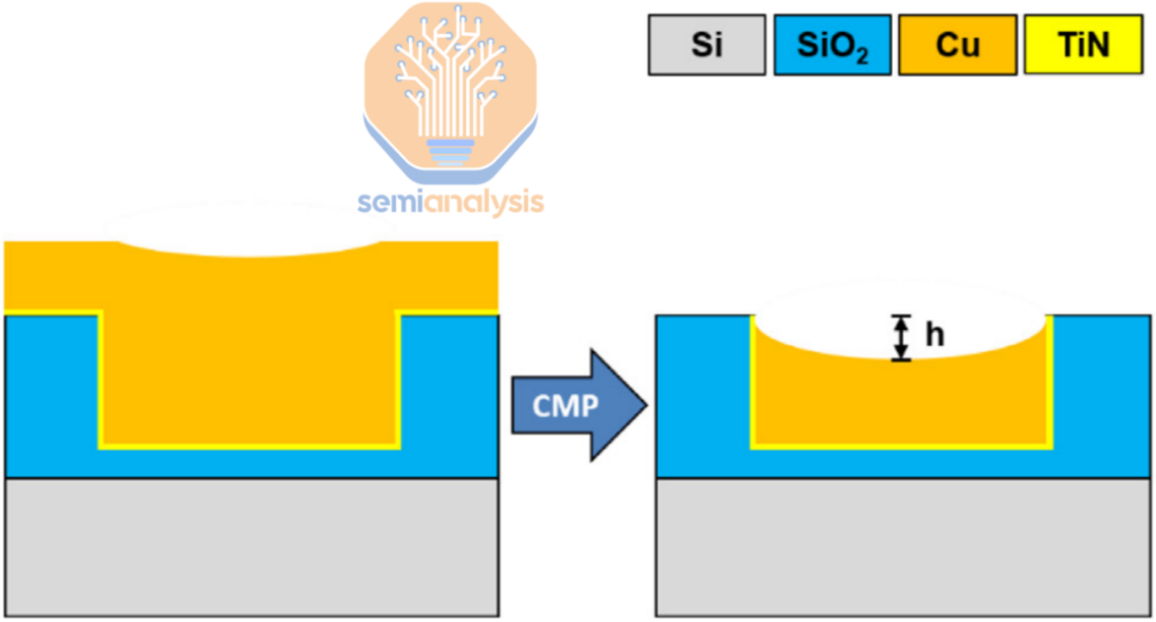
We went over the TSV flow here, but will recap it in this report.

The wafer is coated with photoresist and then patterned using photolithography. Then, TSV is then etched into the silicon by using a Deep Reactive Ion Etch (DRIE) to form a high aspect ratio trench deep into the wafer but this does not go through the entire wafer. An insulation (SiOX, SiNx) and barrier layer (Ti or Ta) is deposited using Chemical Vapor Deposition (CVD). These layers are to prevent the copper from diffusing into the silicon. Then, a copper seed layer is deposited using Physical Vapor Deposition (PVD). This seed layer is deposited in the trench, which is then filled using Electrochemical Deposition (ECD). This forms the TSV. However, the process is not complete as the via is not yet revealed on the backside. To reveal the TSV, the backside of the TSV is polished and in some cases etched to thin down the backside and subsequently reveal the TSV. Once this is complete, the wafer can then go on to formation of the BEOL.

TSV formation is not trivial and can be time consuming, especially due to the deep etch required. We understand that TSV formation is a step that is bottlenecking HBM and CoWoS production. One reason why some customers are switching to CoWoS-R from silicon interposers is to avoid the expensive TSV processes in the silicon interposer.

Hybrid Bond layer formation

After the wafers' bond interface, the hybrid bond layer is fabricated on top of the BEOL of a wafer. This is the same regardless of W2W or D2W. This is a layer of dielectric film patterned with fine-pitch copper vias. The dielectric, usually Silicon Carbon Nitride (SiCN) is deposited via PECVD. Then the pads are formed. Holes for the copper pads are patterned using lithography and etched away. A barrier and seed layer is deposited, then plated with copper using a typical copper damascene process.



Then, there is a CMP step to grind down and smoothen the dielectric surface and achieve the correct profile of the copper. A notable feature of the copper pads is that they are recessed until ~1-micron pitch. As mentioned before, a smooth surface is critical to forming a good bond. Roughness of the dielectric must be contained within 0.5nm, and within 1nm for the copper pads.

One feature for the HB interface is that the copper pads are initially recessed to about 5 nanometers below the dielectric layer. This is to ensure that the copper does not get in the way of the initial dielectric-dielectric bond during annealing. If the copper is too deeply recessed, then the Cu-Cu bond may not form properly.

With CMP of copper and other metals, dishing often occurs because of over-polishing and different levels of softness of the metal and dielectric. While not ideal, this phenomenon is something that is not critical and can be worked around. The exact profile of the dishing needs to be controlled to prevent copper over/undergrowth during the bond.

To achieve the correct dishing profile, multiple CMP steps combining low and high Cu removal slurries are required. CMP is a critical process for Hybrid Bonding to achieve a very smooth surface and the optimal profile.

At ECTC, Sony showed when they get down to 1-micron pitch, they are better of having the copper protrude rather than be recessed.

Wafer Sort / Singulation

For D2W only, a wafer sort is performed and the KGD are singulated and reconstituted on a carrier wafer or tape frame so that they can be further processed. As mentioned above, HB introduces new complications to the conventional wafer-sort process. Wafer sort involves probing the wafers bumps or pads with probe needles to perform electrical test.

Probing can create small amounts of damage on the copper pad surfaces, ruining the smoothness of the surface during the CMP process. While the damage to the pads is minor and usually acceptable in most cases, HB is far more very sensitive to small amounts of topographical changes as these would affect bond quality. One way to address this is to compensate for this in the initial CMP and then performing another round of CMP post-probing to polish out any damage caused by probing.

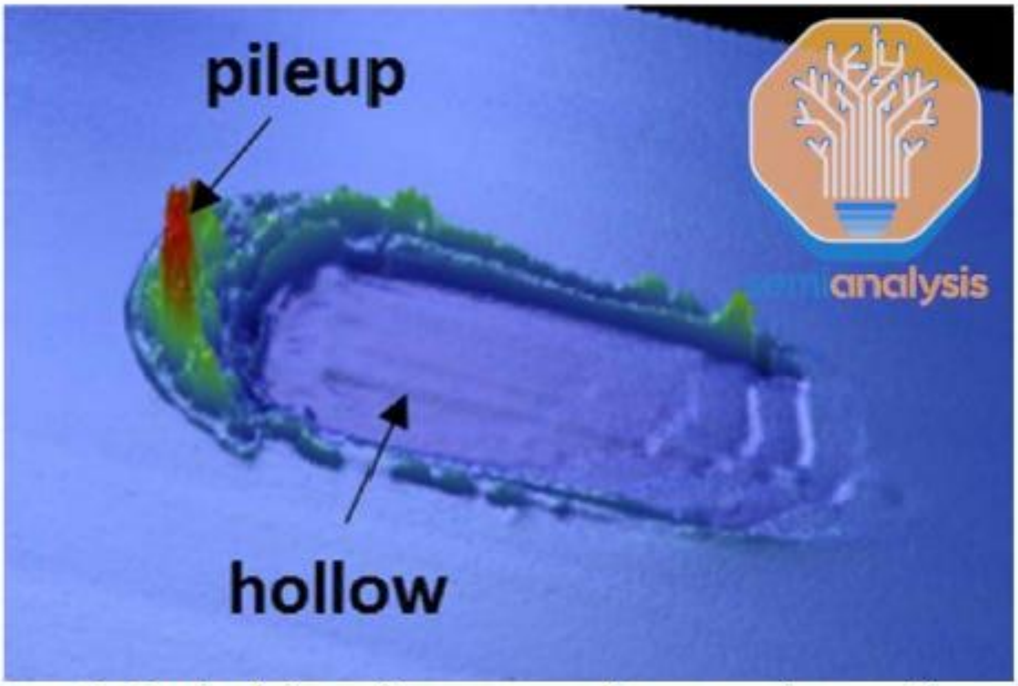
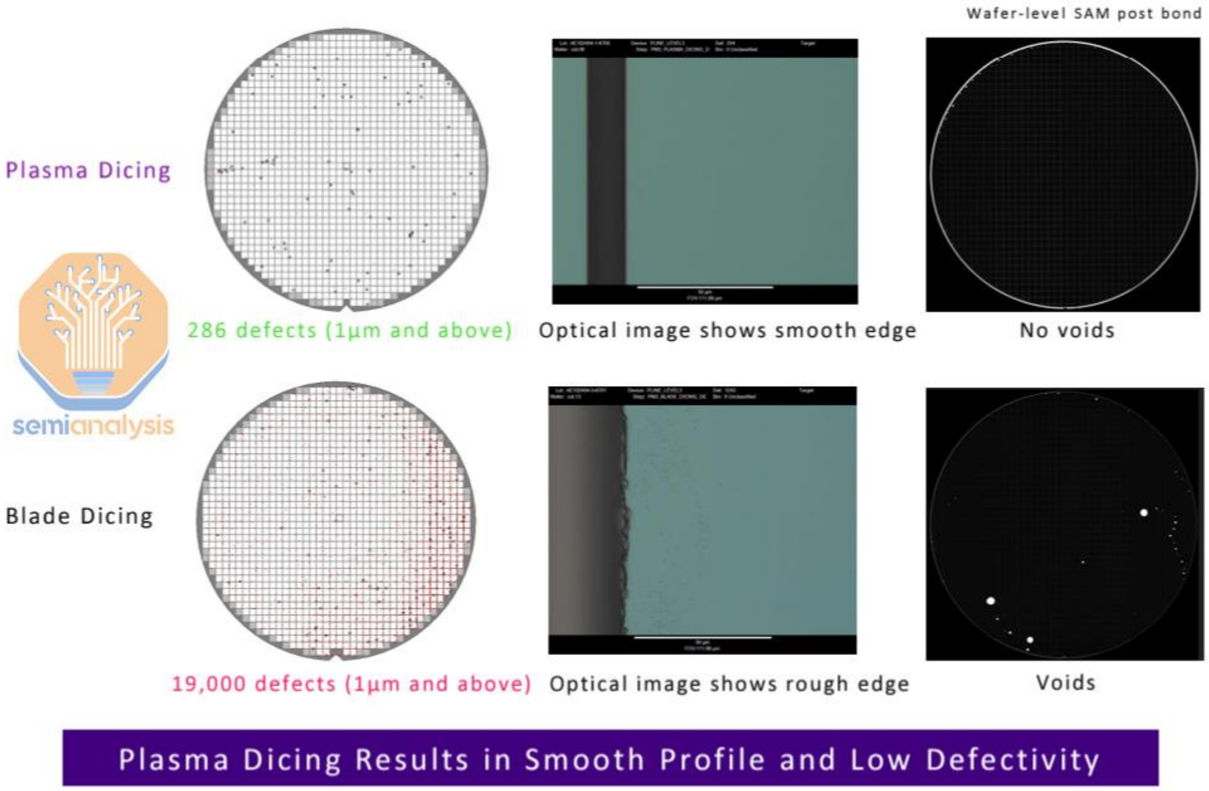


Figure 4. Optical interferometry of one probe mark on Cu.

For singulation / dicing, an issue is the particles that are generated from the process. Blade dicing is generally not used as it is the most dirty: resulting in a lot of particles and a lot of yield loss. Laser dicing and plasma dicing are preferred to blade dicing as they are much cleaner processes, but still particulate matter is generated. Plasma dicing is the most extreme method and is a similar mechanism to etching away the scribe separating the dies. However, this is much lower throughput given the time it takes to etch through the whole wafer.


Wafer-level SAM post bond
Plasma Dicing
286 defects (1μm and above)
Optical image shows smooth edge
No voids
semianalysis
Blade Dicing
19,000 defects (1μm and above)
Optical image shows rough edge
Voids

Plasma Dicing Results in Smooth Profile and Low Defectivity

Disco is the leader in this. their stock has more than tripled since we wrote about them.

DISCO Corporation, The World Leader In Semiconductor Capital Equipment For Cutting, Grinding, Polishing

DYLAN PATEL

·

JULY 19, 2022

Read full story

One mitigation technique is first applying a protective layer coating on the wafer. Particles land on the protective layer and can be removed alongside the protective layer when it’s stripped. While this helps solve the particle problem during singulation, there can be leftover residues of the protective layer and the stripping process may also cause some surface damage to the HB layer, increasing the surface roughness.

Plasma Activation and Cleaning:

The 2 wafers are now treated to prepare for the bond. They are treated with N2 plasma to activate the surface. The plasma treatment changes the surfaces’ characteristics to increase surface energy and make them more hydrophilic. Making both surfaces more hydrophilic allows the surfaces to facilitate hydrogen bonds. This helps enable the initial weak dielectric-dielectric pre-bond at room temperature which occurs at the next step.

After the treatment, a final clean is performed to get rid of any accumulated particles. It is important that prior to bonding, the incoming wafers are as clean as possible. Cleaning needs to be thorough but also non-damaging to maintain the integrity of the HB interface. The best way appears to be deionized water-based cleans assisted by megasonics. Cleans that use a scrubber or are plasma-based can be too damaging and/or introduce contaminants.

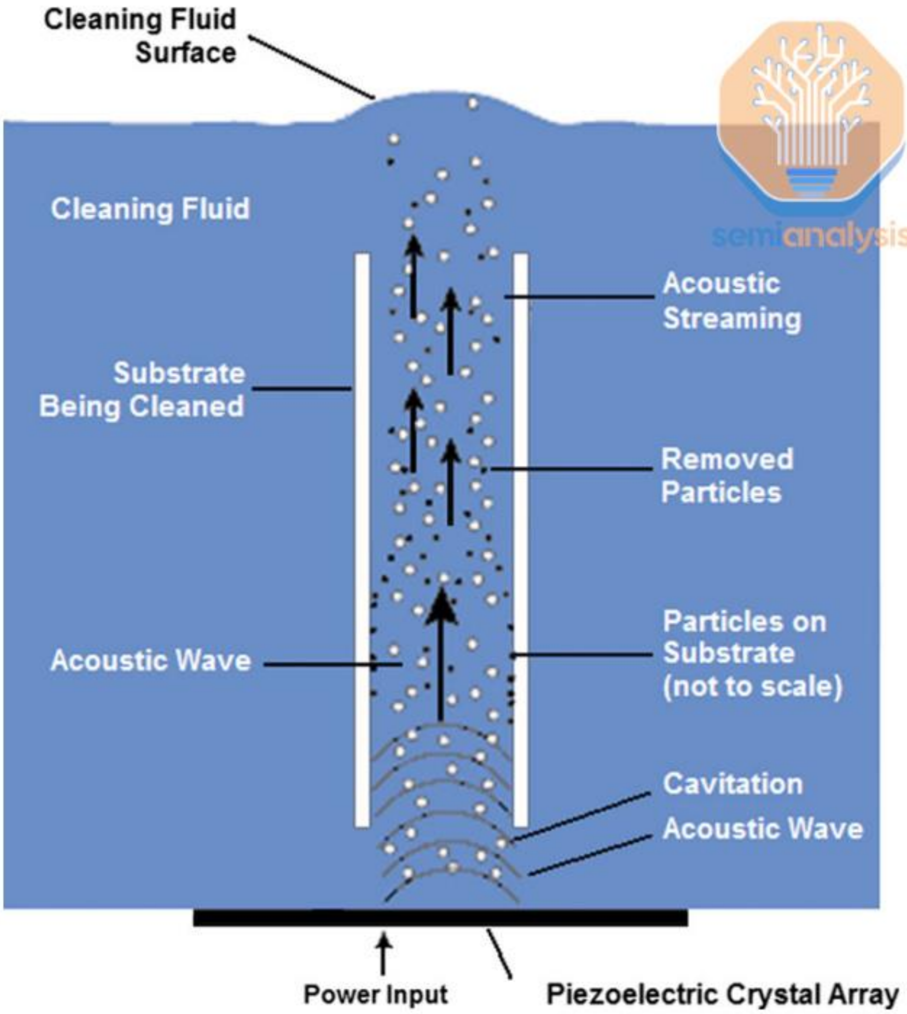


FIGURE 2.2 Micro-streaming in a megasonic field [15]

Bonding

Now the bond step. More accurately it is more of a “pre-bond” as this step only forms the initial dielectric-dielectric bond is only a weak van der Waals bond. We will go through the flow for W2W and D2W methods separately.

W2W bonding

What makes W2W bond yield higher is that the alignment and bonding steps are separated. First the alignment step. There are multiple techniques for W2W alignment. In the past, an IR scanner would be used to check alignment between 2 wafers. A limitation is that one wafer had to be transparent to infrared. This doesn't work for CMOS wafers as IR can’t see through the metal layers.

EVG, who are dominant in W2W bonding, have their patented SmartView alignment technology. There are 2 cameras that are calibrated with each other, one placed above the subject wafers, one below. The chuck holding the top wafer is moved so the bottom camera can identify the alignment mark, and the system registers the location of the alignment mark. The top wafer is retracted and then the bottom wafer moves between the cameras until the top camera can recognize the alignment mark. The aligner can now align the 2 wafers from calculating the relative locations of the 2 alignment marks. To help maintain accuracy and control, the wafers are very close to each other (within 50 microns) and the chucks only move on the X and Y planes with no Z-axis (vertical) movement until the pre-bond.

After alignment the wafers are moved into the bond chamber where they pressed together with a small pressure applied for around 20 minutes to form the initial bond.

Post-bond inspection can be completed in-situ via acoustics and if alignment is insufficient then the bond can also be re-worked.

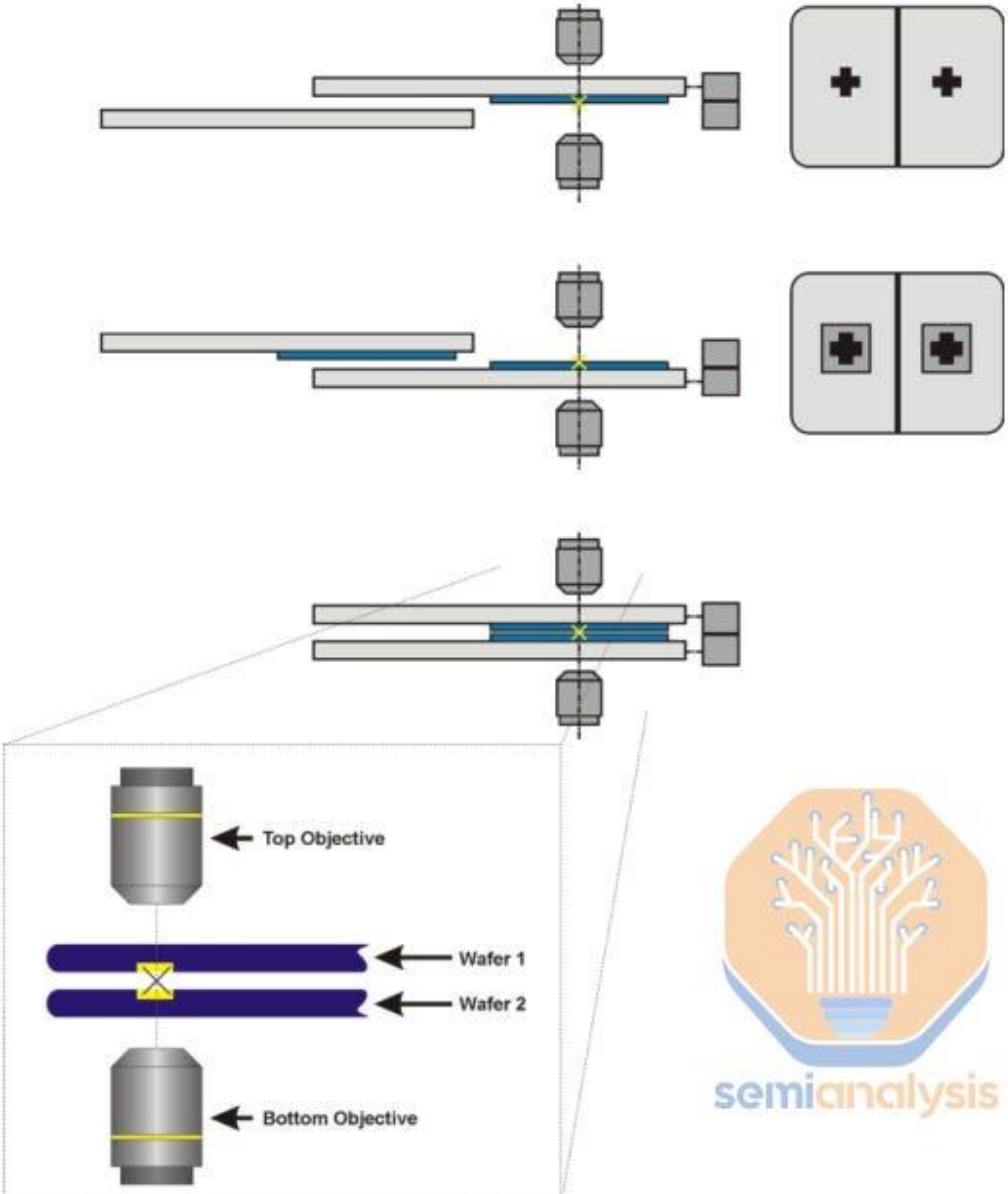


Figure 1. Working principle of SmartView® alignment

In W2W tools, there is a separate chamber that performs the alignment. Once the top and bottom wafers are aligned, they are moved into the bonding chamber (which is in a vacuum) where they are pressed together with a bit of force and after a period of around 20 minutes the initial pre-bond is formed. Key to W2W is that it is a much cleaner process in that there are fewer steps. Before the alignment and bonding, the wafers can be cleaned to remove most of the particles. Die singulation, a source of particle contamination, only happens after the bond.

More time is also afforded for the alignment step as it is a wafer-level process so longer alignment time does not hurt throughput as much as it would for a chip-level process. There is not much movement that happens in the chambers either so there are fewer contaminants generated in the chamber itself. Currently, W2W bonders can achieve **alignment accuracy of sub-50nm**. W2W bonding is already a mature process and not particularly expensive. The proof is that we see it widely adopted in mass-market products like image sensors by Sony, Omnivison, and Samsung and in NAND by YMTC, Western Digital, and Kioxia.

D2W bonding

D2W bonding is completed via pick and place tools.

The bottom target wafer sits on a wafer chuck. The dies to be bonded are placed on a tape frame face up. A flipper arm collects the individual die and flips it so that the die has its backside facing up on the flipper. There is an overhead bond arm which picks up the flipped die using vacuum suction at the bond-head.

Below we will talk about D2W bonding, annealing, self assembly, and all of the equipment companies in detail.

https://www.semianalysis.com/p/hybrid-bonding-process-flow-advanced