Chip Scale Review®

ChipScaleReview.com

The Future of Semiconductor Packaging

Volume 24, Number 3

May • June 2020

Maskless lithography and the shift to 3D integration

page 15

- The true cost of undetected defects
- Chiplets: a new era of smart packaging
- Over the air (OTA) testing of AiP modules for 5G apps
- AI's impact on 3D packaging: heterogeneous integration
- Embedded trace and RDL copper plating for PLP applications

XPERI®

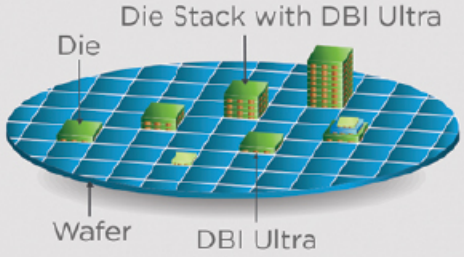
Semiconductor Technologies | invensas

DBI® Ultra Die to Wafer Hybrid Bonding

The Ultimate 2.5D & 3D Integration Technology for High Performance Computing

High Bandwidth, High Capacity, Thin Profile, Low Power, Low Cost

Data Centers | AI, Machine Learning & Deep Learning Hardware | Automotive | Gaming | Consumer Electronics | Industrial & Scientific

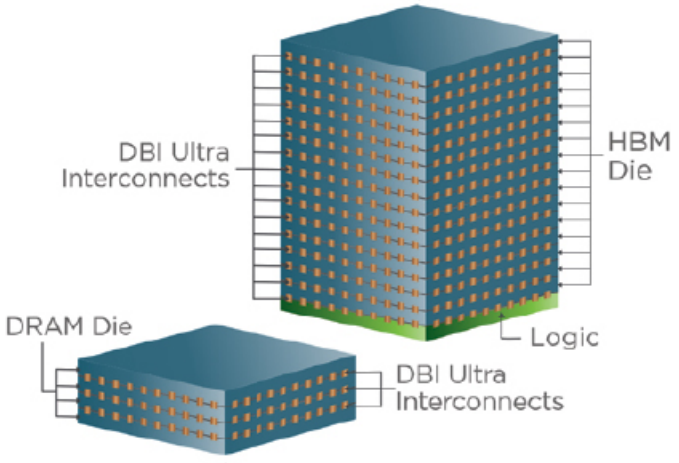


Ultimate Integration Flexibility

Accommodates various die sizes, wafer sizes, process technology nodes, etc.

Enabling 3D Stacked Memory Solutions

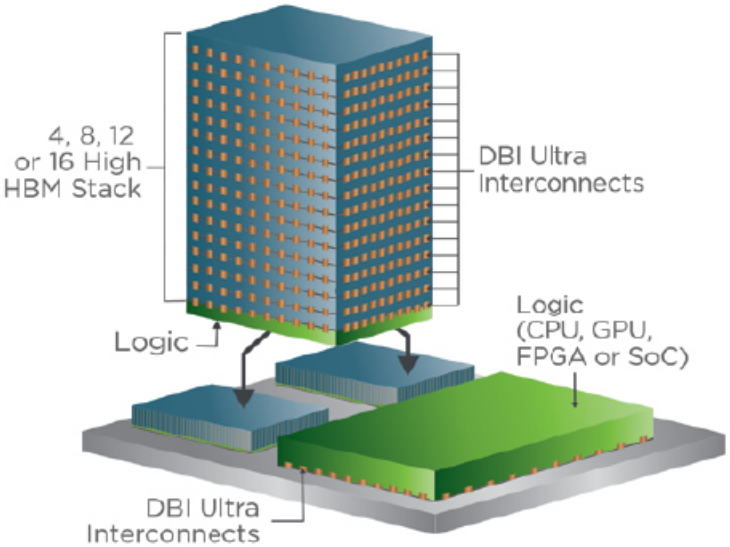
4, 8, 12, 16 or more high HBM2, HBM3 & beyond



4 High 3DS DRAM

Enabling Next Generation High Permanence Computing

2.5D Integration with DBI Ultra



invensas@xperi.com | www.invensas.com

3025 Orchard Parkway San Jose, CA 95134 +1 408.321.6000

CONTENTS

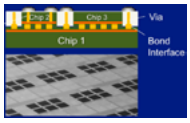
May • June 2020
Volume 24, Number 3

The requirements for back-end lithography continue to grow as heterogeneous integration becomes an increasing driving force in semiconductor development and innovation. The increased importance of design flexibility, minimizing pattern distortion and die shifts in advanced packaging, reducing consumable costs, supporting both thick and thin resists, and shortening development cycles between R&D and production, are just a few factors driving the need for maskless exposure (MLE) technology.

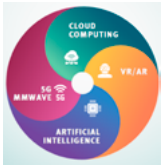
Photo courtesy of EV Group

DEPARTMENTS

5 Technology Trends
A new era of smart packaging: the introduction of chiplets
By E. Jan Vardaman
[TechSearch International, Inc.]



8 Market Update
The real third wave of semiconductor market growth and some packaging challenges
By Asif Chowdhury
[UTAC Group]



FEATURE ARTICLES

15 Maskless lithography addresses shift to 3D integration
By Bozena Matuskova and Thomas Uhrmann
[EV Group]

EVG®

invent
innovate
implement

www.EVGroup.com

ACCELERATING HETEROGENEOUS INTEGRATION

- EV Group establishes Heterogeneous Integration Competence Center™ to accelerate new product development fueled by heterogeneous integration and advanced packaging
- Open access innovation incubator for EVG customers and partners across the microelectronics supply chain, guaranteeing the highest IP protection standards
- Combining EVG's world-class wafer bonding, thin-wafer handling and lithography products and expertise, as well as pilot-line production facilities and services
- Leveraging the state-of-the-art cleanroom facilities at EVG's headquarters in Austria, supported by EVG's worldwide network of process technology teams

GET IN TOUCH to discuss your manufacturing needs
HeterogeneousIntegration@EVGroup.com

Chip Scale Review®
ChipScaleReview.com
The Future of Semiconductor Packaging

Volume 24, Number 3
May • June 2020

SUBSCRIPTION—INQUIRIES
Chip Scale Review
All subscription changes, additions, deletions to any and all subscriptions should be made by email only to
subs@chipscalereview.com

Advertising Production Inquiries:
Lawrence Michaels
lmichaels@chipscalereview.com

Subscriptions in the U.S. are available without charge to qualified individuals in the electronics industry. In the U.S. subscriptions by first class mail are $125 per year. Subscriptions outside of the United States are $225 per year to other countries.

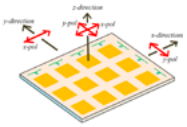
Chip Scale Review, (ISSN 1526-1344), is published six times a year with issues in January-February, March-April, May-June, July-August, September-October and November-December. Periodical postage paid at Los Angeles, Calif., and additional offices.

POSTMASTER: Send address changes to Chip Scale Review magazine
P.O. Box 2165, Morgan Hill, CA 95038

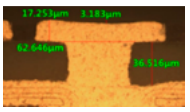
Printed in the United States

FEATURE ARTICLES

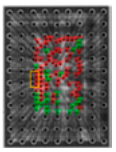
20 **Testing AiP modules in high-volume production for 5G applications**
By Jose Moreira
[Advantest]



26 **Embedded trace and 2-in-1 RDL for fan-out panel-level packaging**
By Kesheng Feng and Kwangsuk Kim
[MacDermid Alpha Electronics Solutions]

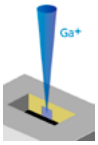


31 **The true cost of undetected defects**
By David L. Adler and Brennan Peterson
[SVXR Inc.]

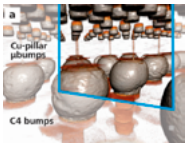


34 **AI's impact on 3D packaging: heterogeneous integration**
By Santosh Kumar
[Yole Développement, Korea]

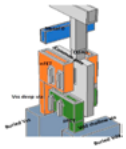
38 **Demand for 3D devices drives inspection and metrology innovation**
A Q&A with Raj Jammy
[Carl Zeiss SMT Inc., the North American subsidiary of Carl Zeiss SMT GmbH]

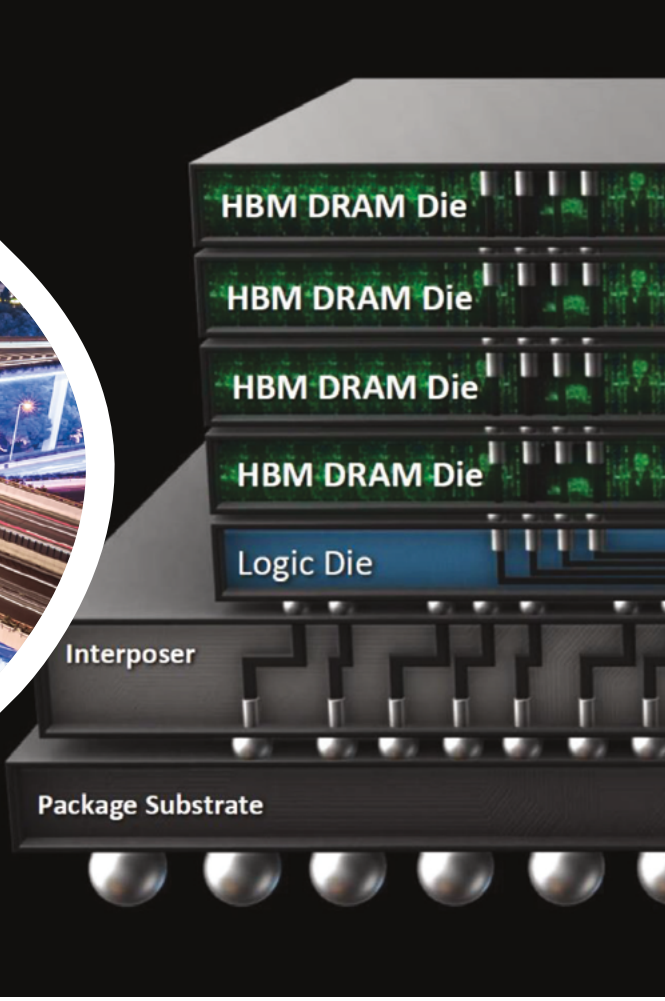


39 **Rapid analysis of buried 2.5/3D package structures**
By Cheryl Hartfield
[ZEISS Semiconductor Manufacturing Technology, Process Control Solutions]



43 **Scaling CMOS beyond FinFETs: from nanosheets and forksheets to CFETs**
By Julien Ryckaert
[imec]





When Moore's Law no longer gets your ICs where they need to go.

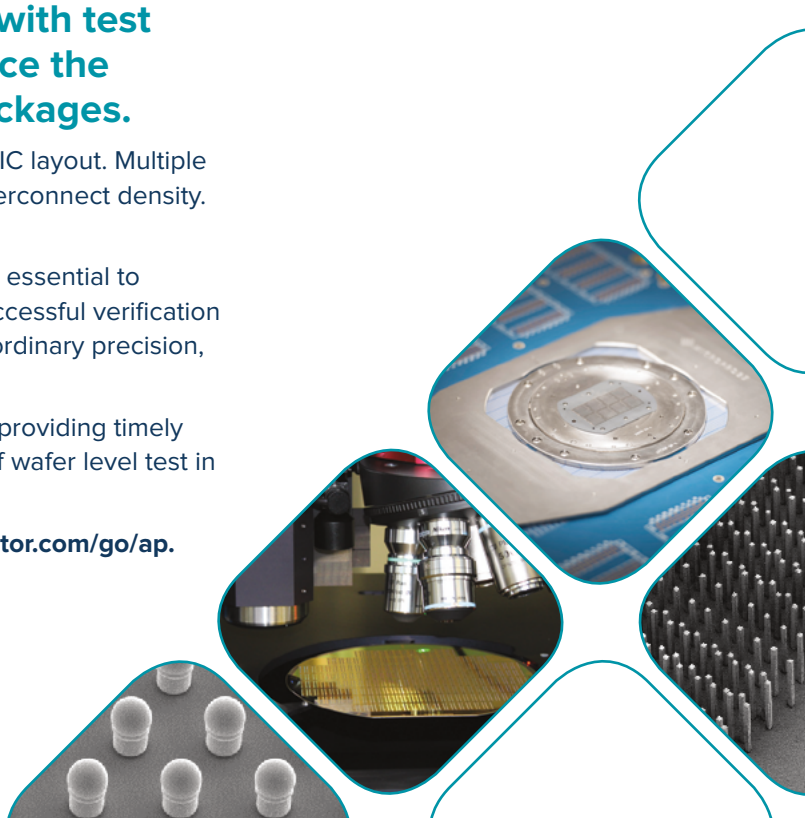
FormFactor takes on the challenge with test and measurement solutions to reduce the manufacturing cost of advanced packages.

Advanced packaging adds a new vertical dimension to IC layout. Multiple dies merge into single systems with unprecedented interconnect density. Performance goes up. Power consumption goes down.

And wafer-level test and measurement becomes nearly essential to guarantee cost-effective fabrication and packaging. Successful verification at this level requires probing and measuring with extraordinary precision, optically, electrically and mechanically.

FormFactor not only understands the problem, it's also providing timely solutions to advance yield knowledge at every phase of wafer level test in advanced IC packaging.

Let us help you rise to the occasion. **Visit www.formfactor.com/go/ap.**



FormFactor™
Test insight from lab to fab.

A new era of smart packaging: the introduction of chiplets

By E. Jan Vardaman [TechSearch International, Inc.]

The use of chiplets will be one of the most important developments for the next 10 to 20 years, according to Dr. Douglas Yu, TSMC's Vice President, Research & Development/Integrated Interconnect Packaging. Continued monolithic integration is expensive and suffers from the defect density yield loss associated with large die. An increasing number of companies are turning to new architectures using chiplets to achieve the economic advantages lost with expensive monolithic scaling, ushering in a new era of smart packaging [1]. The role of heterogeneous integration, especially chiplets, is pivotal in this new era. A chiplet is a functional circuit block and includes reusable intellectual property (IP) blocks. A chiplet can be created by partitioning a die into functions and is typically attached to a silicon interposer or organic substrate. Simply using a logic die plus high-bandwidth memory (HBM) on an interposer would not necessarily be counted as a chiplet solution.

Drivers for chiplets

The demand for chiplets is driven by the need for a more cost-effective solution, the reuse of IP, and new test flows. A cost reduction is possible because die functions, such as I/O control, are split out. These die functions are more difficult to scale, requiring more time and money. Cost reduction is achieved by chip integration. With chips divided in two, the split chips can be stacked to provide a 20% cost improvement. Chip partitioning also improves time-to-market. The chiplet is not a new idea, but electronic design automation (EDA) tool improvements are making new architectures possible. The following sections highlight chiplet technology by different entities.

AMD

AMD's current chiplet solution uses a laminate substrate. The company has shipped multiple versions of its server processor using chiplets. Up to eight 8-core processors are tightly coupled together on an organic substrate. The chiplets can be binned and speed-sorted before assembly in the package. At the system level, the multi-die package architecturally acts as if it were a monolithic die [2]. AMD's new architecture allows the chiplets to communicate with each other as well as externally, including power and ground management. This new architecture and future approaches will allow performance improvements that cannot be met with monolithic integration, including power consumption. Future versions are expected using other organic solutions.

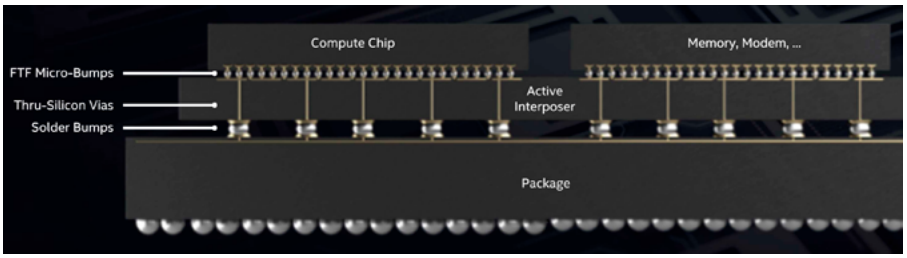


Figure 1: Foveros technology with 3D face-to-face stacking. SOURCE: Intel

Intel

Intel's chiplet solution is Foveros. It is a 3D integration technology as a form of heterogeneous system integration. Heterogeneous integration is defined as the integration of separately manufactured components into a higher-level assembly or system-in-package (SiP) that, in aggregate, provides enhanced functionality and improved operating characteristics. The technology uses a 3D face-to-face stacking process. In the process, logic die are bumped and mounted on an active interposer next to memory or die with communication functions such as a modem. The active interposer can contain active parts of the system, such as the platform controller hub (PCH) that manages I/O for the system. The active interposer is mounted on a package substrate with solder bumps (see **Figure 1**). While the first demonstration of the technology used micro bumps. Future versions are expected to use hybrid or direct interconnect bonding without a bump.

The Foveros technology will give designers greater flexibility to mix and match IP blocks with various memory and I/O elements into new form factors. Mounting memory on an active interposer removes the bottleneck of memory proximity. Intel is expected to leverage the technology across many product lines.

TSMC

TSMC and ARM demonstrated a 7nm chiplet solution on a chip-on-wafer-on-substrate (CoWoS). Two quad-core Cortex-72-based 7nm chiplets were mounted on the silicon interposer and connected using TSMC's Low-voltage-IN-Package INterCONnect (LIPINCON) interface. Each chiplet featured four ARM Cortex A72 processors and an on-die interconnect mesh bus. The die-to-die inter-chiplet connection features scalable 0.56pJ/bit power efficiency, 1.6Tb/s/mm^2 bandwidth density, and 0.3V LIPINCON low-voltage interface achieving 8GT/s and 320GB/s bandwidth. It was produced in April, 2019.

Not sure where to get started with ***Panel-Level Packaging?***

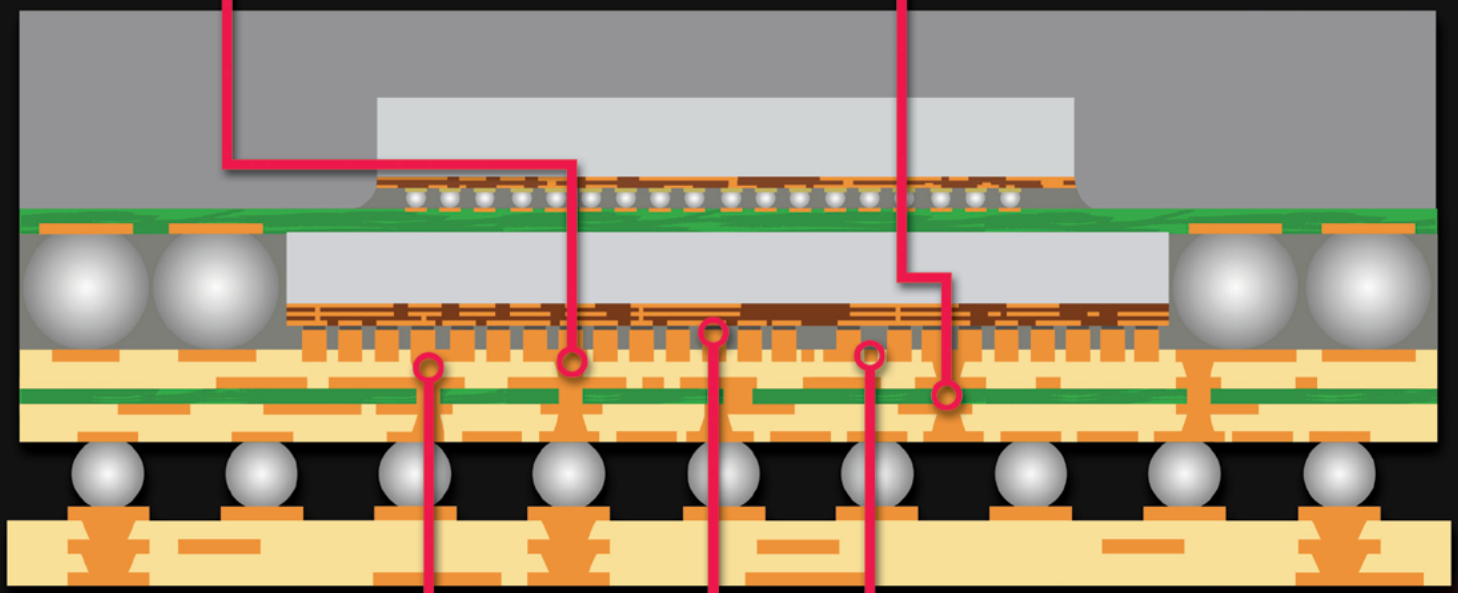
WE HAVE YOUR SOLUTIONS.

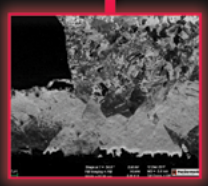
Systek UVF 100 2-in-1 Plating for RDL

Ultra via filling process for IC substrate RDL that fills micro vias and copper traces with excellent co-planarity and feature uniformity

MacuSpec THF 100

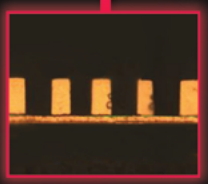
Single step copper through hole filling for core layer that provides enhanced thermal and mechanical design options





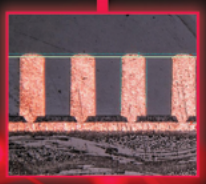
Systek SAP

High-performance semi-additive build-up process allows enhancements for high speed signals and very fine line routing



Systek ETS

Embedded trace substrate plating 5/5 micron line/space enabling next generation miniaturization of IC substrate RDL



Systek BMP

Panel-side copper pillar plating of highly co-planar copper megastructures, reducing costs and enabling new chip packaging designs

MacDermid Alpha ELECTRONICS SOLUTIONS

macdermidalpha.com

Contact us today to learn more.

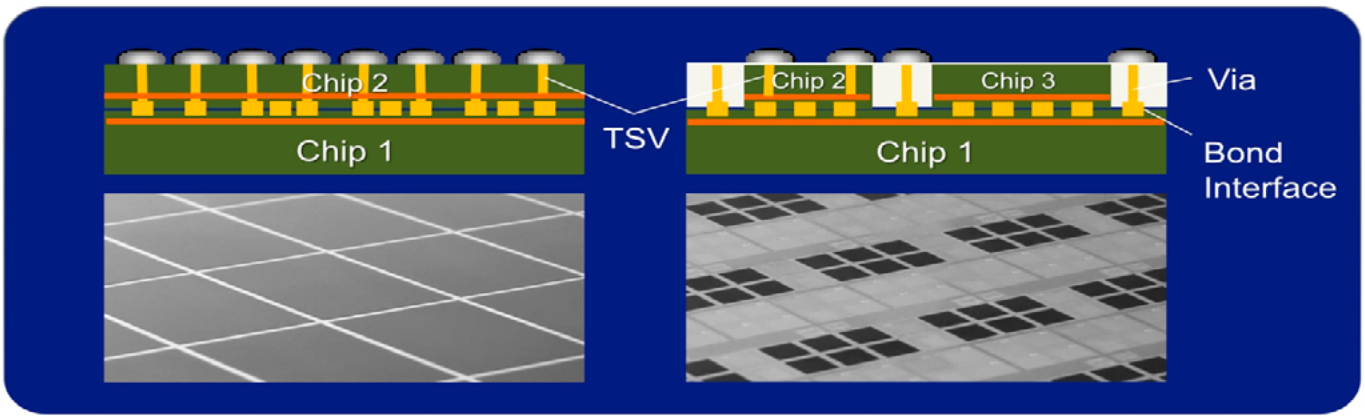


Figure 2: TSMC SoIC™ technology.

TSMC proposes its bumpless System on Integrated Chip (SoIC™) as one chiplet solution (**Figure 2**). The SoIC™ is a 3D structure formed by stacking logic, memory, or both chip types on an active interposer with through-silicon vias (TSVs). A chip-on wafer (CoW) process is used and the process can handle a <10µm bond pad pitch between chips. The chips, with or without TSVs, are bonded onto the wafer containing the active interposers using a hybrid bonding process. The 3D chip with TSV SoIC™ also uses a through-dielectric via and CoW bond. TSMC reports that the SoIC™ structure with its higher density bonding provides better signal integrity, power integrity, and lower communication latency with greater bandwidth than a conventional 3D IC using TSVs and 40µm pitch micro bumps. Lower insertion loss, important for 5G applications, is reported. Lower parasitics and low IR drop is reported. Finer interconnect pitch is possible and there is less concern with chip-package interaction (CPI) because there are no bumps [3].

zGlue

California start-up zGlue offers an online tool, called ChipBuilder, to build chiplets to connect to its smart silicon interposer that allows the use of third-party chiplets. The chiplets, fabricated as wafer-level packages (WLPs), can be connected to an active silicon interposer called Smart Fabric. The company has created a library of more than 250 off-the-shelf chiplets and plans to extend it to more than 1,500 in the near future.

The active silicon interposer or smart fabric has a fine-pitch Cu pillar micro-bump array on the surface that can conform to the chiplet ball map using a built-in programmable connectivity and power array. The smart fabric provides field-programmable connectivity and connects I/Os of various chiplets based on a program bit stream. Additionally, the smart fabric has built-in voltage regulators, power management circuits, a programmable memory, level shifters, programmable I/O to the outside, and other peripherals. The smart fabric is programmable at manufacturing time through zGlue's firmware application program interface. Software programmability allows rapid configuration for chiplet-based systems on the order of days and weeks. Connection to RF and sensitive analog signals are handled in the redistribution layer (RDL). The zGlue integration platform is called (ZiP). A unique method of surface tension is used to connect the chiplets to the active silicon interposer or smart fabric. Connections with pitch as small as 50µm are possible. Programmability of zGlue bumps also enables some repair and reconfigurability after manufacturing [4].

DARPA

The Defense Advanced Research Projects Agency (DARPA) Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) program is focused on creating an open interface, including a switch fabric interconnect to connect logic blocks inside devices, to created multi-chip modules (MCMs) using chiplets instead of traditional system on chip (SoC) devices. Breaking the SoC into separate chiplets allows proprietary operations to be fabricated separately from the licensed and tested commodity IP chips [5]. Benefits include security and potential cost reduction.

Summary

Chiplets are a heterogeneous integration solution that can move us into the next semiconductor era. While it is technologically possible to continue scaling monolithic die, the economics do not favor this approach. The use of chiplets will become a game changer in the new era for the companies that can master the design of this new approach. New players are emerging that will help others participate in the bold new chiplet era.

References

1. J. Vardaman, "Chiplets: the new era begins," The 3D InCites Yearbook, Mar. 2020, pp.13, 55.
2. M. Bhagavat, "Packaging renaissance with chiplets," IMAPS Keynote, IMAPS Conf., Oct. 2019.
3. F. Chen, et al., "System on integrated chips (SoIC) for 3D heterogeneous integration," Elec. Comp. and Tech. Conf. (ECTC), May 2019, pp. 594-599.
4. J. Nasrullah, et al., "Designing software configurable chips and SiPs using chiplets and zGlue," IMAPS Inter., Sept. 30–Oct. 3, 2019.
5. T. Coughlin, "Chiplets for All," *Forbes*, May 11, 2019.

Biography

E. Jan Vardaman is President of TechSearch International, Inc., Austin, TX. She received her BA from Mercer U., and MA in Economics from the U. of Texas. Email jan@techsearchinc.com

The real third wave of semiconductor market growth and some packaging challenges

By Asif Chowdhury [UTAC Group]

At about 0.5%, the semiconductor industry is a small percentage of global GDP. However, semiconductor devices play a very large role in advancement and growth of almost all other market sectors and industries. Semiconductors are an integral part of our everyday lives. In the last 50 years, this niche market has grown from almost nothing to about a $500B market ushering in new technologies and innovations that have not only significantly altered and improved almost all aspects of our lives, but also propelled the growth of many other markets. [Note: Forecasts in this article were made before the full potential market downside of the COVID-19 pandemic were known.]

Figure 1 shows 5-year compound annual growth rates (CAGRs) for the semiconductor market from 1995 through 2019. While the semiconductor industry has steadily grown since its inception, there were a couple of periods during which the industry saw waves of double-digit growth. The first wave was in the nineties driven by the significant growth in the personal computer (PC) and laptop markets. **Figure 1** shows 5-year CAGRs north of 20% during this growth phase. Then the market went through a down cycle due to the Dot-com bust in 2000. The next growth spurt, or the second wave of growth, was driven by the advent and proliferation of mobile handsets and tablets from the early 2000s to about 2008, just before the last financial crash (2008). Since then, the market has rarely seen any double-digit growth year-over-year. In the last decade, there wasn't a single 5-year CAGR with double-digit growth.

As the industry has matured, year-over-year growth has naturally become moderate. Still, there has been continued discussion among the semiconductor market pundits as to what will drive the next double-digit growth spurt for the semiconductor market, if anything. There were a good few years during the last decade when it was widely believed that the Internet of Things (IoT) would be the next double-digit growth driver for the semiconductor market. However, despite all the hype around IoT, that expected growth didn't take place for various reasons. Perhaps one of the key reasons is because the price points for fundamental technologies remained relatively high causing prices of many of the novel IoT end products to remain outside the reach of mass consumers.

The latest hype in the semiconductor world is 5G. There is another round of consensus building around the potential of 5G technology as the next big growth driver for the semiconductor market. This time though, the hype is likely to become reality, not only because of the potential of proliferation of 5G, but significant improvement of some other key fundamental technologies such as artificial intelligence (AI), virtual reality (VR), augmented reality (AR), and cloud computing. As shown in **Figure 2**, along with the growth of 5G, the improvement of these key technologies will likely create a convergence that will usher in the next-generation of electronics products from true smart homes to full autonomous vehicles.

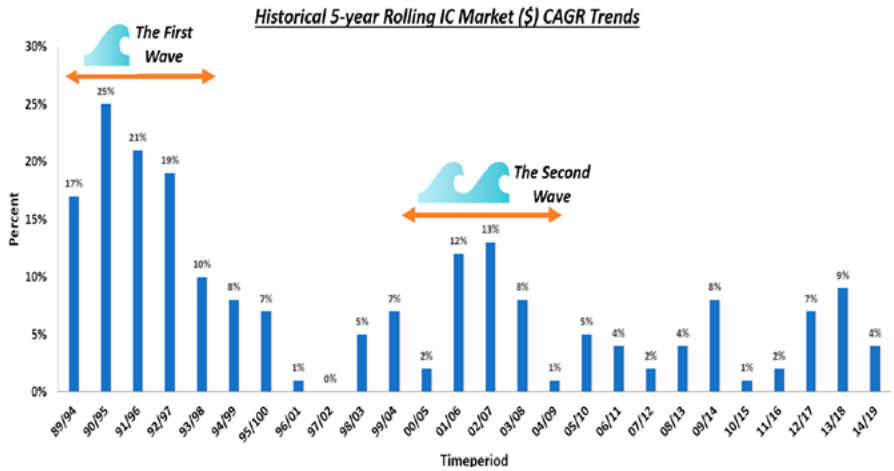


Figure 1: 5-year rolling CAGR for IC markets from 1994 through 2019. SOURCE: IC Insights

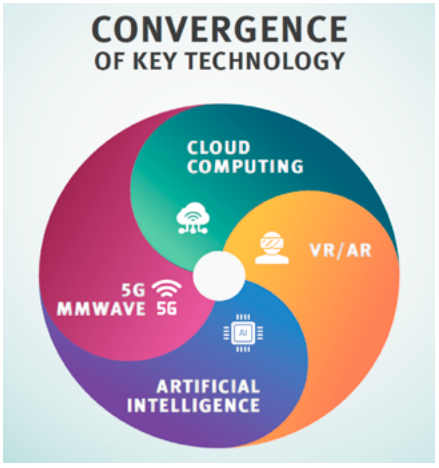


Figure 2: Development and convergence of fundamental technologies likely to fuel growth for the semiconductor market.

In addition to the significant improvement in the fundamental technologies discussed above, another important phenomenon is also likely to enable this convergence to create and grow next-generation products.

The unit price of all these technologies has come down over the last decade making "things" cheaper and hence more affordable. **Figure 3** shows significant decline in some of the key fundamental technologies such as power, various sensor products, and radio devices from 2012 through 2020. Sensor products, such as microelectromechanical systems (MEMS), are key to the success and growth of IoT products. **Figure 4** shows how the average sensor unit cost has declined by half from $1.30 in 2004, to $0.60 in 2014 and estimated to go down to $0.38 in 2020. MEMS technology allows us to "digitize" our analog world. **Figure 5** depicts the unit price reduction of MEMS sensors from over $3.00 in the early 2000s, to about $0.50 in 2018. **Figure 5** rightfully predicts two possible outcomes for MEMS unit prices for the first half of this decade. It is likely that both will happen depending on MEMS integration and application. Unit prices of advanced integrated MEMS products will likely rise because of the complexity of such products, while prices for standard MEMS products will fall driven by high volume in IoT application. However, the average unit price of MEMS will continue to decline.

The most important element for the technology convergence (as shown in **Figure 2**) is the development and deployment of 5G mmWave technology. While there is a lot of buzz going around about the potential for deployment of 5G in the next two to three years, the real 5G will not be widely available until the second half of this decade. It is this real 5G that will be the very basis for the next generations of electronic applications.

The best way to look at 5G is to put this technology deployment into two phases. Phase 1 is up to the 6GHz frequency range, and Phase 2 is in the >24GHz range as depicted in **Figure 6**. The current 5G deployment that is scheduled to start by the end of 2020 is Phase 1. In this phase, most of the 4G infrastructure can be used with 5G component upgrades while consumers will have to buy new 5G handsets. In this <6GHz frequency range, while transmission will be faster than current 4G, it will not have the step function improvement as the 5G Phase 2, which is called 5G mmWave. It is in this frequency spectrum where the true benefits of 5G lie. As shown in **Figure 7**, when compared to the performance of

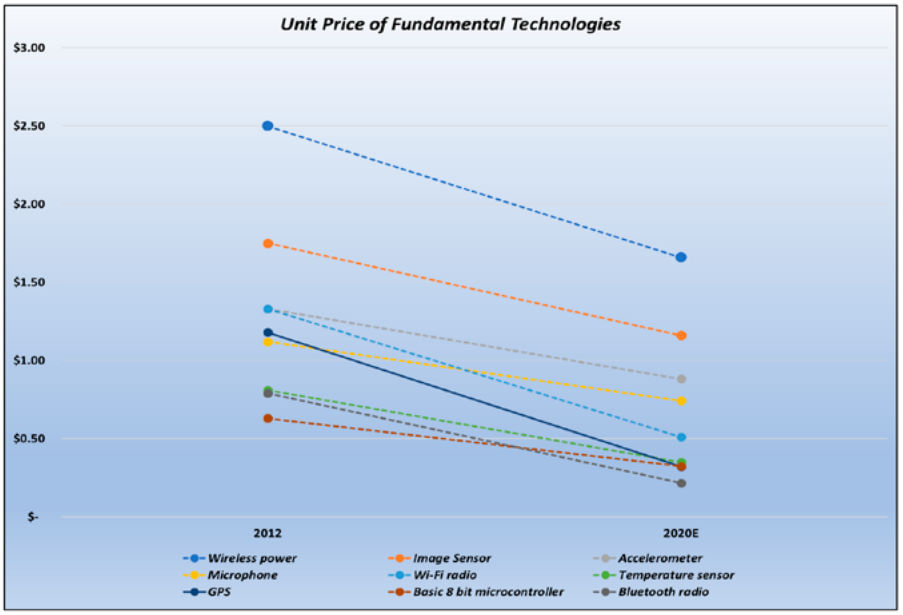


Figure 3: Unit price for fundamental technology has steadily declined in the last decade [1].

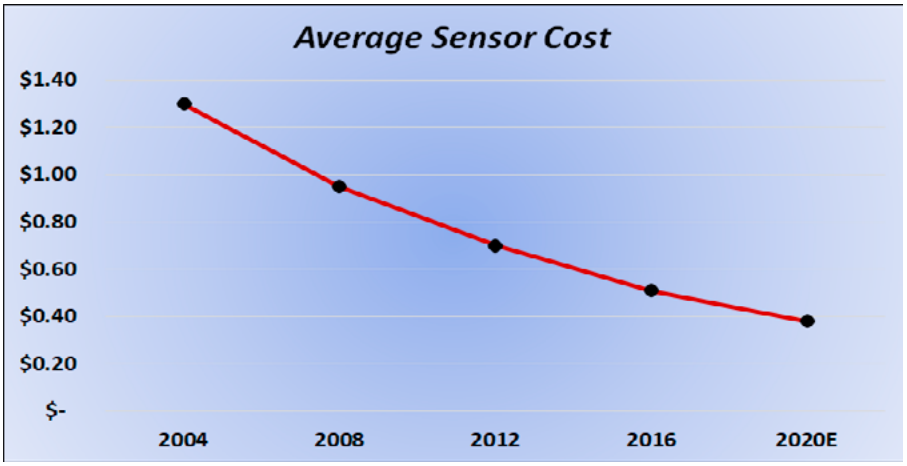


Figure 4: Average price for sensor products, key enablers for IoT, have come down significantly. SOURCES: Goldman Sachs and [2]

4G, 5G mmWave can have up to a 100X improvement in data rate, a 10X reduction in latency, and a 100X traffic capacity allowing a significantly higher number of connected devices.

5G mmWave will require brand new infrastructures as it works only with "line of sight" transmission. In other words, the cell towers will have to be much closer to each other. This will take some time to implement and most likely will not be ready in wider areas until 2024, or even later.

The significant performance improvement with the deployment of 5G mmWave technology will open new frontiers of capability and usher in the introduction and growth of advanced technologies. For example, the up to 100X transmission speed will allow seamless transmission of high-resolution 8K video enabling people to watch such video on their handsets without any interruption. This higher transmission speed along with the 10X reduction in latency is extremely important in the proliferation of autonomous vehicles. Driven by improved safety standards, a higher level of connectivity and improved

infotainment, automotive electronics and the semiconductor market have been consistently growing at 5%~8%CAGR. With the proliferation of autonomous vehicles, the amount of electronics in a car is predicted to increase significantly. **Figure 8** shows the actual growth of global vehicle production and electronic cost as a percent of total car cost from 2000, to estimated numbers in 2030. In 2030, 50% of the total cost of a vehicle is estimated to come from electronics components. As the number of electronics components grows per vehicle, so will the semiconductor content.

The technology convergence will usher in other new products and product improvements such as, 5G handsets and infrastructures (cell towers), factory and industrial automation, advanced medical technology, autonomous vehicles, gaming, security, and of course, IoT (see **Figure 9**). Indeed, we are very likely to witness the explosion of IoT during the second half of this decade.

The explosive growth of these next-generation electronic products and devices will in turn drive the next growth spurt or The Third Wave of semiconductor market growth. This Third Wave of growth is predicted to come after 2024 and the semiconductor industry could once again see years of double-digit growth in the second half of this decade as shown in **Figure 10**. Based on the above estimate, by 2030, the overall semiconductor market could go over a trillion dollars with outsourced semiconductor assembly and test (OSAT) value going over a hundred billion dollars.

The 5G revolution will also bring significant challenges to packaging and testing technologies and will drive growth for advanced new packages. However, the growth of IoT, automotive semiconductor, gaming, etc., will also raise the demand for more common semiconductor package types. **Table 1** shows actual units produced for common package types from 2016 through 2018, and estimated units produced in 2019. The table also shows the estimated forecast of these packages for 5G Phase 1 for 2024, and 5G Phase 2 for 2030. The lead frame-based quad flat no-leads (QFN) package continues to be the most cost-effective, smaller footprint solution with generally good thermal and electrical performance. QFN and molded interconnect substrate (MIS) package demand is poised to increase significantly with an estimated CAGR of 11.3% from 2019 to 2030 because of IoT proliferation as the demand will surge for analog and sensor products. Wafer-level chip-scale packaging (WLCSP) is estimated to see the second highest demand after QFN with a 9.6% CAGR from 2019 to 2030. This includes both fan-in for standard products, and fan-out for advanced integrated, multi-functional ICs. Some of the standard surface mount device (SMD) packages will also see higher growth, such as the very popular 8-lead small outline integrated circuit (SOIC). The quad flat package (QFP) will also likely see higher demand from the automotive market. Laminate packages such as ball grid array (BGA) and land grid

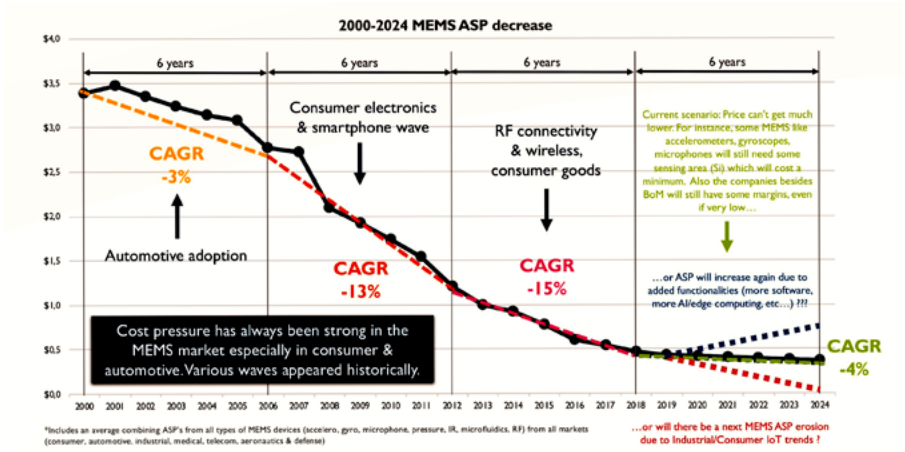


Figure 5: MEMS average selling price (ASP) evolution. SOURCE: Status of the MEMS Industry report, Yole Développment, March 2020

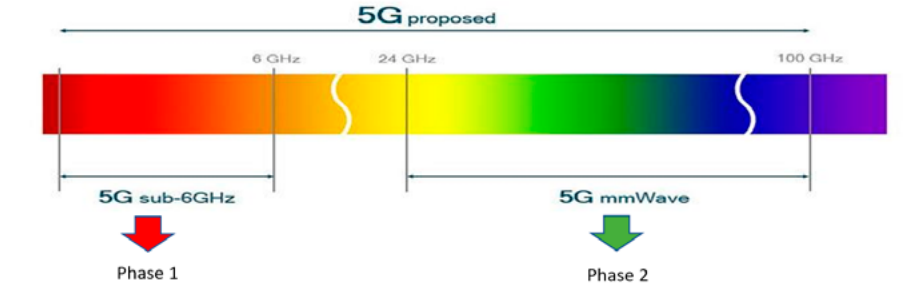


Figure 6: The two phases of 5G deployment: <6GHz in phase 1, and higher frequency, or mmWAVE in phase 2.

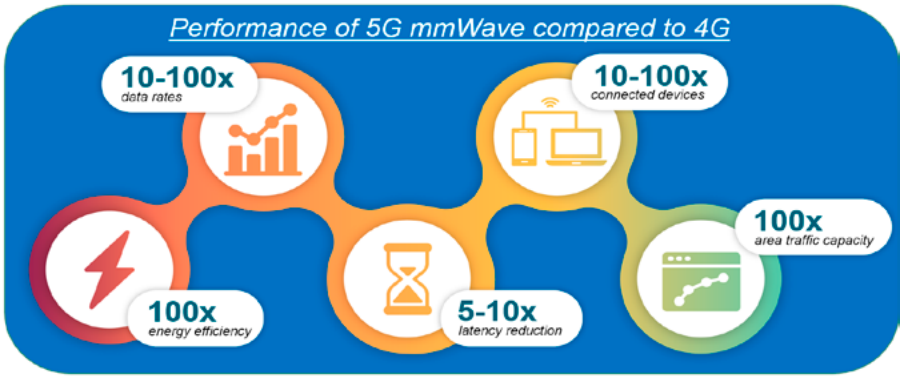


Figure 7: 5G mmWave performance summary compared to that of 4G.

SUSS MICROTEC - SHAPING THE FUTURE

SUSS MicroTec is a leading supplier of equipment and process solutions for microstructuring applications from R&D to high-volume production. We provide excellent process and maintenance support for you at any stage of your development and manufacturing cycle.

Our task is to contribute to the advancement of next-generation technologies such as 3D integration, imprint lithography and key processes for MEMS and LED manufacturing.

+ Photomask Equipment
+ Coater and Developer
+ Mask Aligner
+ Metrology
+ Imprint Lithography
+ Wafer-Bonder
+ Micro-Optics

SUSS MicroTec
info@suss.com
www.suss.com

SÜSS MicroTec

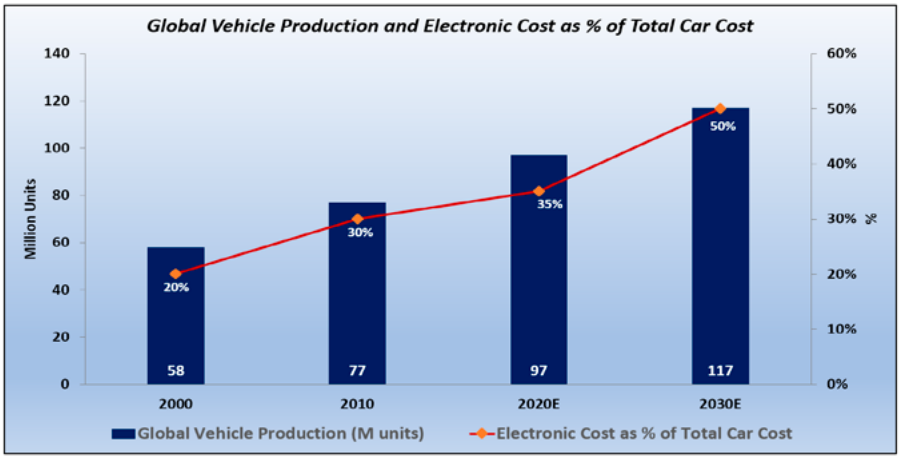


Figure 8: 2000 to 2031 global production of vehicles and the increase of electronic cost as a percent of vehicle cost [3-5].

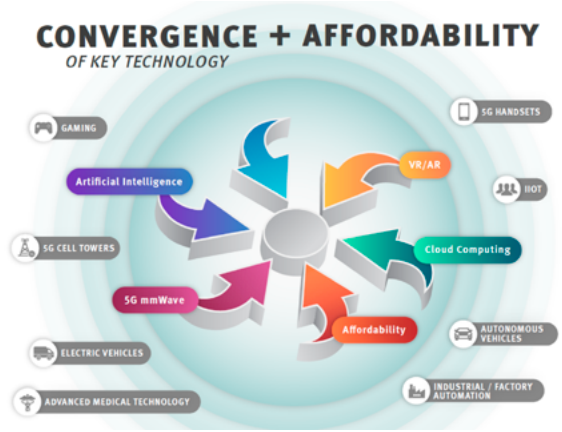


Figure 9: The convergence of key technologies, along with the significant unit price reduction of these fundamental technologies, will drive market growth in this decade.

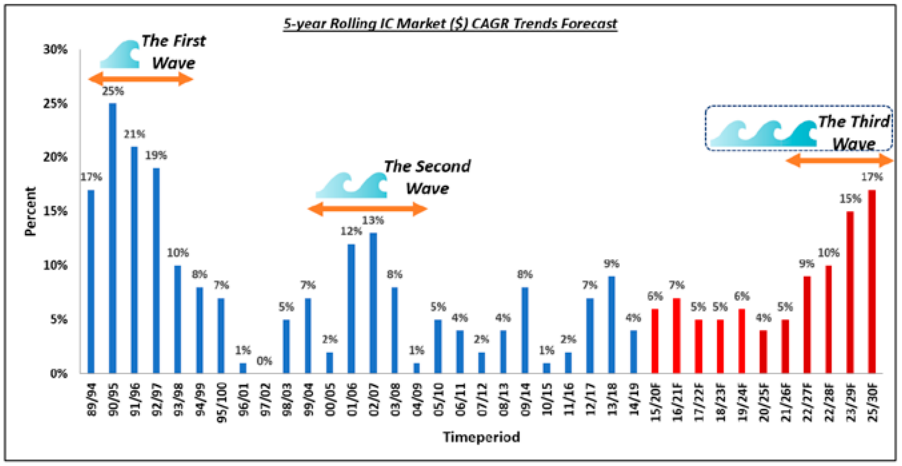


Figure 10: Driven by key technology convergence and lower unit cost of these technologies, the semiconductor industry is likely to experience double-digit growth in the second half of this decade. SOURCES: 89/94 through 14/19 data from IC Insights; 15/20F through 25/30F by Asif Chowdhury

array (LGA) will also continue to grow, especially using flip-chip interconnect technology. System in package (SIP) is expected to grow significantly because of the need for heterogeneous integration driven in turn by the need for increased functionality within the given real estate, such as in mobile phone applications.

Packaging and testing for 5G mmWave devices will require innovative thinking and perhaps even disruptive technology innovation. Historically, most IC packaging and testing had to deal with frequencies up to the 6GHz range. For 4G solutions, the frequency range was from 450MHZ to 3.7GHz. However, 5G mmWave (Phase 2) will be using a frequency range from 24GHz to 100GHz (though the industry seems to be converging towards an ~28GHz solution). These frequencies are significantly higher than what are used for the current 4G or even Phase 1 5G (at <6GHz). **Figure 11** shows the progression of RF front-end modules from 2G and provides a summary of packaging solutions required for 5G sub-6GHz and 5G mmWave. In 5G Phase 1, an extension of current packaging solutions will work. But 5G mmWave or Phase 2 will require new technologies, such as phased-array antennas and antenna-in-package (AiP). Because of the shorter wavelength of 5G mmWave technology, it has large losses at interconnects. The antenna, therefore, will need to be designed in the IC chip, or at least integrated with the RF front-end module. Testing them cost effectively will be even more challenging.

New material development will be key to produce a suitable and low-cost packaging solution for high-frequency devices. While over-molded plastic packaging is the workhorse of the industry and has the largest capacity and infrastructure in the industry, this will not be suitable for 5G mmWave RF devices. For example, the mold compound used for plastic packages has a relatively high dielectric constant and is not suitable for RF devices above 20GHz. So, unless the material supplier comes up with such a mold compound, a more expensive cavity package solution will have to be used.

Finally, there remains a lack of expertise in the assembly and test sector to effectively handle the high-frequency RF solutions discussed above. The semiconductor back-end industry would need to enlist RF engineers who understand these high-frequency requirements and can fill this knowledge gap for package design and final test.

The technology convergence because of the 5G technology revolution, improvement in artificial intelligence, virtual and augmented reality, and cloud computing, along with unit cost reduction of these key technologies will create a perfect storm that will provide the strong backwind for the next growth phase of the semiconductor industry. While 5G Phase 1 is available in a few locations and imminent in the rest of

Unit: Bn Units	2016	2017	2018	2019E	2020F	2024F	2030F	CAAGR 2016-2019	CAAGR 2019-2024	CAAGR 2019-2030
SO/TSOP/SOT/DIP	76.0	83.0	88.5	83.0	81.0	91.0	140.0	3.0%	1.9%	4.9%
QFP/LCC	13.2	14.5	15.5	14.5	14.0	15.5	18.6	3.2%	1.3%	2.3%
QFN/FC QFN/MIS	41.6	45.8	50.4	48.9	48.4	70.2	158.4	5.5%	7.5%	11.3%
WB BGA/FC BGA/PGA/LGA	1.7	1.7	1.9	1.8	1.9	2.2	3.8	1.9%	4.1%	7.0%
WB CSP/Stacked CSP/FC CSP/FC CSP for DRAM	23.4	28.2	30.7	28.6	29.5	40.5	65.3	6.9%	7.2%	7.8%
WLCSP	26.0	30.5	32.5	30.6	29.2	42.0	84.0	5.6%	6.5%	9.6%
SIP (FC and WB)	4.0	4.2	4.5	4.1	4.0	4.9	8.4	0.8%	3.6%	6.7%

Table 1: Actual and forecasted units of typical high-volume package types. SOURCE: 2016 through 2024F from Prismark; 2030F by Asif Chowdhury.

the world, the proliferation of 5G mmWave Phase 2 will occur mostly in the second half of this decade. Because this is a fundamental piece of technology for the convergence, the double-digit semiconductor market growth will likely happen in the second half of this decade as well. Cost-effective assembly and test of 5G mmWave devices will be one of the key challenges for success to ensure this third wave of semiconductor growth.

Acknowledgments

The author would like to thank Carol Chiang from UTAC Singapore, Brandon Prior from Prismark, and Sandrine Leroy from Yole Développment for their help with providing some of the updated graphs and tables.

References

1. 2020 unit price estimation by Asif Chowdhury. The 2012 unit price from the following source: https://docplayer.net/16415414-Internet-of-things-iot-and-its-impact-on-semiconductor-packaging.html
2. https://www.businessinsider.com/four-elements-driving-iot-2014-10?IR=T
3. https://www.statista.com/statistics/277931/automotive-electronics-cost-as-a-share-of-total-car-cost-worldwide/
4. https://www.statista.com/statistics/266813/growth-of-the-global-vehicle-production-since-2009/
5. https://www.statista.com/statistics/262747/worldwide-automobile-production-since-2000/

Biography

Asif Chowdhury is the SVP of Marketing & Corporate Business Development and Head of Japan Sales at UTAC Group, Singapore. He has over 25 years of experience in the semiconductor industry. Before joining UTAC, he held senior positions at Amkor Technology, Chandler, AZ, and Analog Devices, Wilmington, MA. He holds a BS in Mechanical Engineering from U. of Texas at Arlington, an MS in Mechanical Engineering from Southern Methodist U., and an MS in Finance and an MBA from Northeastern U. Email asif_chowdhury@utacgroup.com

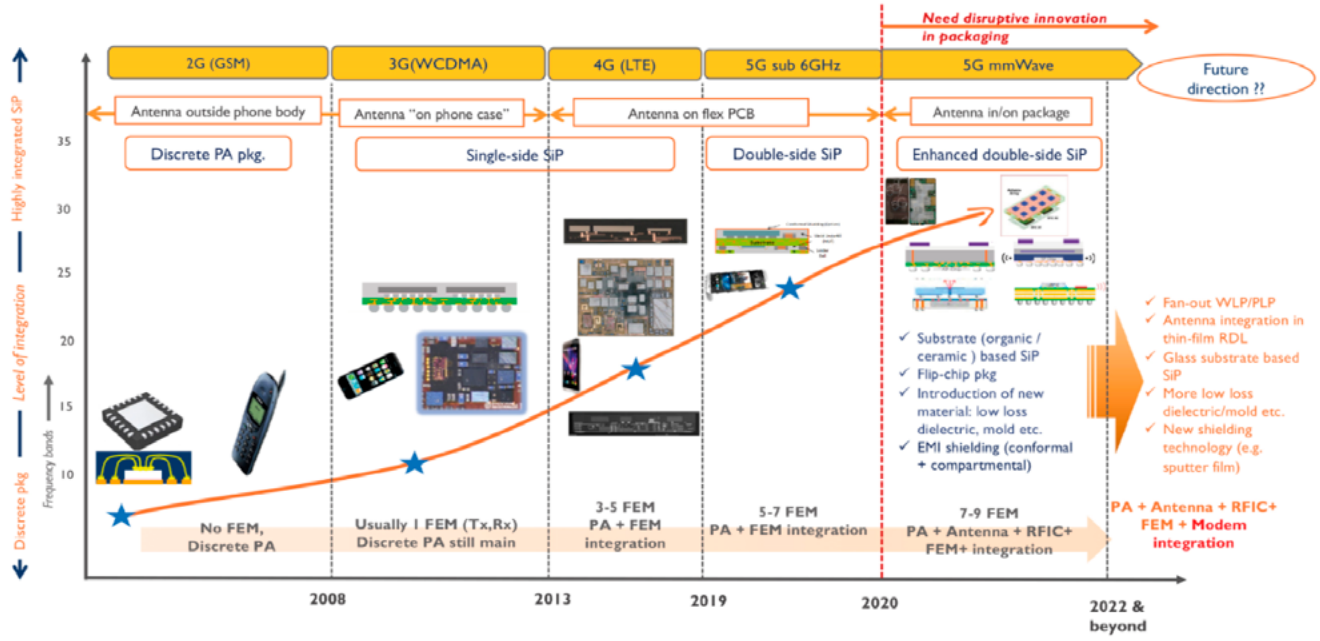


Figure 11: History and roadmap of RF front-end packaging. SOURCE: Yole Développment

PROCESS SOLUTIONS, EQUIPMENT AND SERVICES

From Lab-to-Fab, YES is the Answer!

Cure.

PB Series
for manual polyimide vacuum cure

VertaCure
for automated polyimide vacuum cure

Clean.

G Series
for plasma cleaning

RFS Series
for R&D plasma strip/descum

ÉcoClean
for automated plasma resist strip/descum

Coat.

TA Series
for HMDS vapor prime and image reversal

LabCoat
for R&D silane molecular vapor phase deposition

ÉcoCoat
for silane molecular vapor phase deposition

Yield Engineering Systems, Inc.

Call: **1-925-373-8353** (worldwide) or **1-888-YES-3637** (US toll free)

www.yieldengineering.com/csr

Maskless lithography addresses shift to 3D integration

By Bozena Matuskova, Boris Považay, Frank Bögelsack, Roman Holly, Tobias Zenger, Thomas Uhrmann, Bernhard Thallner [EV Group]

The continuous development and capability of lithographic patterning equipment for semiconductor manufacturing is driven by several megatrends that are shaping our digital society. As 2D-IC density scaling is reaching its cost limits, especially for the patterning processes, advances in miniaturization and device performance are currently being extended toward 3D integration and heterogeneous integration (HI) within advanced packaging. This novel approach is seen as a key enabler for next-generation devices, wherein mobile processors have triggered a first growth cycle in 3D/HI. This growth cycle is expected to continue as high-performance applications such as artificial intelligence (AI) and 5G gain traction in mobile devices, but also due to other megatrends such as autonomous driving, which necessitates high reliability and data network capacity, as well as the Internet of Things (IoT).

Advanced packaging technologies have both increased in complexity, as well as in the number of options over the years. Transformation from single- to multi-die packaging, enabled by 3D integration, is one example, which addresses the challenge of handling big data generated by high-performance computing. The continuous innovation of chiplet design and variety of integration schemes (on silicon, embedded or in package) can finally include several patterning levels. The increased importance of design flexibility and the ability to adopt both die- and wafer-level designs simultaneously in back-end lithographic processing have to be addressed because of the need to shorten development cycles and cover the wide variety of advanced packaging platforms at the same time. In addition, any advanced product design mix adds to multiple masking levels, and correspondingly masks and mask inventory/cleanroom storage represent a high portion of overall production costs. Additionally, replacement costs for conventional laser sources or Hg lamps can add up to significant levels. The wait time for new physical mask sets as well as overall proof of new design concepts for high product-mix designs intrinsically lead to prolonged development cycles for conventional mask-based production environments.

These requirements triggered the development of our maskless exposure (MLE™) technology to resolve critical needs in semiconductor packaging. MLE technology directly tackles this crucial demand for design flexibility while enabling scalability in both development and production facilities – thereby shortening development cycles between R&D and production phases – by eliminating mask-related difficulties and consumables costs. The technology features high-resolution (<2µm L/S), stitch-free, dynamically addressable exposure of the entire substrate surface, which enables agile processing and low cost of ownership (CoO).

Limitations of traditional exposure methods

Fundamentally, the resolution of any optical imaging system is determined by the ratio of exposure wavelength and its numerical aperture (NA). By definition, NA defines the light-gathering and light-emitting ability through the lens, and is characterized by the angle of aperture and thereby highly dependent on focal length. Technically, alteration of NA in exposure systems tends to be more cost effective than shortening the ultraviolet (UV) wavelength of the light source. Imaging exposure systems (such as steppers) are typically driven toward higher NA to enable structuring of finer critical dimensions. In contrast, higher NA reduces the depth of focus (DoF). Finding a compromise between resolution and focal depth is often a decisive parameter for the interconnect circuit design. This is especially true in advanced packaging as reconstitution of wafers is a central element in integrating die from various wafer manufacturers in a multi-die solution. Apart from those physical limitations, inaccuracies from die placement and die shift variations caused by over-molding add an additional layer of difficulty with which current lithography steppers and other mask-based systems struggle to cope. In addition, the given reticle size and optics dimensions of static exposure systems limit the exposure area. This situation can become particularly challenging for larger die interposer fabrication, where stitch lines or mismatched overlap regions of reticle exposure fields can affect electrical properties within the redistribution layer (RDL). The ability to generate a stitch-less pattern for interposers exceeding current reticle size is increasingly important for advanced devices needed for complex layouts, such as in advanced graphics processing, AI and high-performance computing (HPC).

MLE addresses these needs through a combination of sub-nanometer-range stage motion accuracy, distortion-free, high-intensity optics and real-time patterning of a vector-based mask file. For reference, other approaches that need to rasterize the mask image prior to patterning, generate in the range of 141GB of data for every 300mm wafer. Finally, the digital mask pattern is projected with sub-µm accuracy onto the substrate surface. Like most modern lenses, the MLE imaging system is diffraction-limited, and it supports a DoF of ±12µm. A measurement of the DoF process

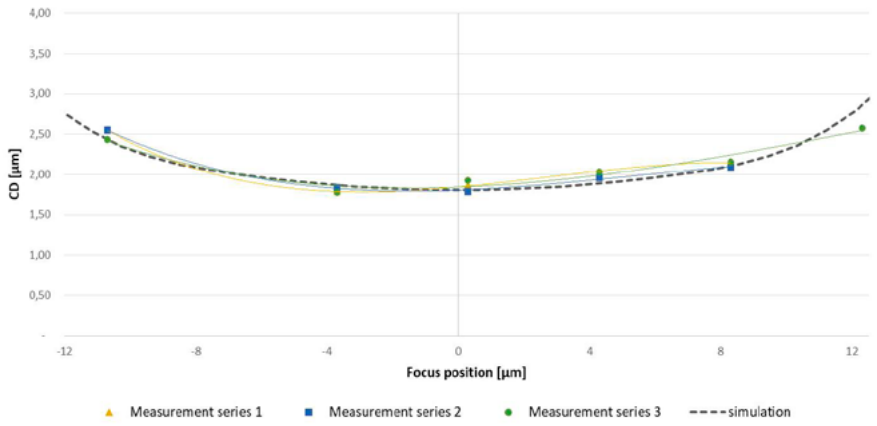


Figure 1: Critical dimension versus focus position for DoF process window evaluation.

window performance at 2µm L/S together with the simulated curve is displayed in **Figure 1**.

On top of the exposure window, the relatively small exposure field and the <1µm precise autofocus extends the usable dynamic focused range to more than 100µm. The ability to control the focus position on a larger scale of the wafer position via chuck positioning and wafer clamping enables compensation for bowed and warped substrates.

Dynamic exposure methods and active die-shift compensation

Current back-end lithographic systems have no control over distortions smaller than the exposure field and therefore face difficulties with nonlinear, high-order substrate distortions and die-shift-related issues, especially after die reconstitution on the wafer. MLE employs dynamic alignment modes with an automatic focus, in order to adapt to the substrate material and surface variations. The advanced distortion functionality relates and analyzes real-time data from synchronized visible or near-infrared topside and backside alignment. It accomplishes this by actively compensating for mechanical die placement, stress-induced inaccuracies such as rotation, displacement, expansion and high-order distortions of the substrate. The process flow of the advanced distortion correction function and dynamic alignment modes are visualized in **Figure 2**. Dynamic alignment includes both global as well as multi-point wafer alignment options, where typically up to 16 alignment marks (marked with blue and yellow), can be placed randomly in the layout in order to comprehensively cover the most critical areas on the substrate and compensate for global distortions.

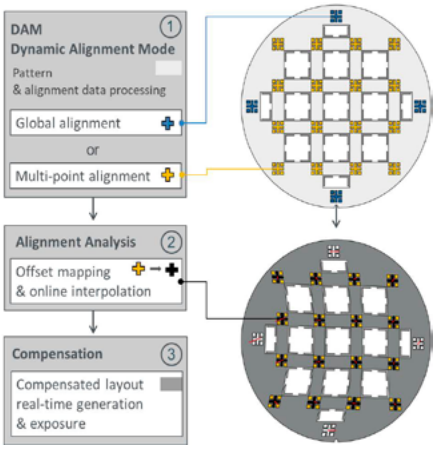


Figure 2: Advanced distortion compensation schematic process flow.

After misalignment measurement, displacement vectors are further compiled in parallel before the design is interpolated and rendered in real-time. The exposed patterns are therefore fully error-compensated without inducing overlapping or noncovered regions – delivering minimum misalignment with no impact to the throughput of the patterning process. A visualization of a compensated layout (dark grey) after an extreme atypical misalignment (indicated with red arrows) example is shown as a result after compensating the actual position of 16 marks (yellow) of multi-point alignment through the dynamic alignment mode.

Backside alignment plays a substantial role when it comes to bonded, opaque substrates or non-transparent materials used in multi-layer and multi-die processing. Referencing to the same structure for multiple exposures also helps to minimize misalignment as current pattern design density increases. Overlay errors or misalignment of any kind impact the electrical properties of contacts and insulation, and might create connection failures that significantly affect fab yield, overall productivity and cost of ownership (CoO). Systems equipped with MLE technology integrate full wafer backside alignment (see **Figure 3**) utilizing dedicated objectives with near infrared (IR) capability and proprietary chuck design accommodating wafer sizes up to 300mm.

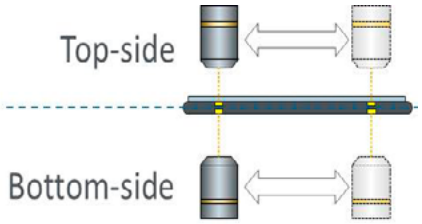
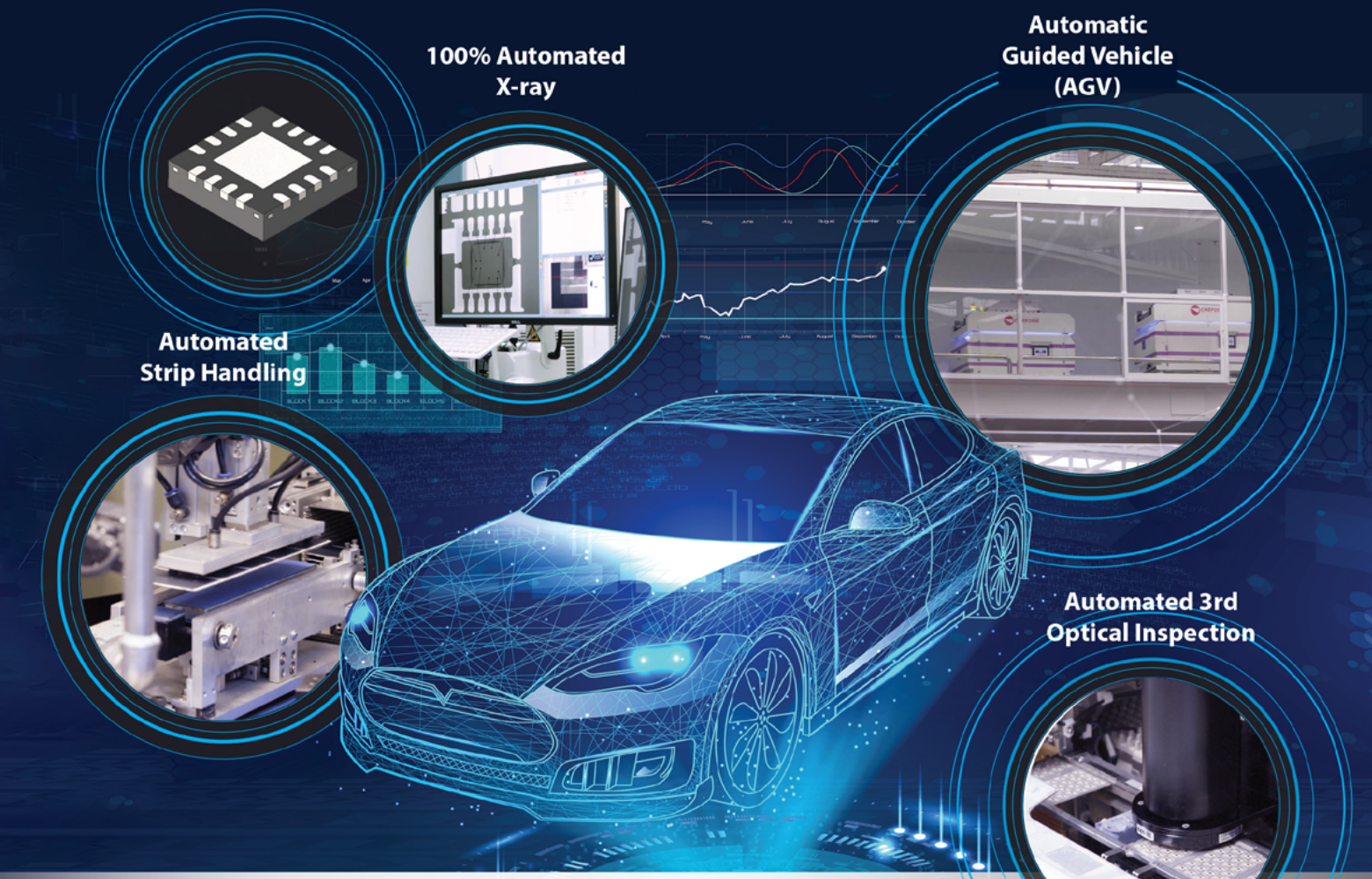


Figure 3: Schematic drawing of top- and bottom-side alignment objectives.

When considering the die distortion errors induced after reconstitution on the wafer, as is typical for fan-out wafer-level packaging (FOWLP), the advanced distortion functionality should also be applied at the die level, where active compensation and rerouting results strictly rely on external metrology data. Distortion compensation algorithms include mathematical correction of rotation, scale, shear and translation (shift). For die-placement-error compensation, the model limits distortions within the dies to the rigid body of the die, which typically is represented by two (external) alignment points per die. Because of the immediacy of the conversion process, the dynamic binary pattern generation complements externally acquired metrology data of each die individually per substrate just before the exposure in order to compensate for overlay/positioning errors caused by handling or pre-

MORE THAN A DECADE OF EXCELLENCE
IN AUTOMOTIVE IC PACKAGING AND TEST MANUFACTURING

• 20% CAGR of Automotive Volume Since 2014 • 50% of Automotive Business is Full Turnkey • Highly Regarded Automotive Assembly and Test Sites
• Leadership in Smart Manufacturing, Assembly Automation



LEADERSHIP IN SMART MANUFACTURING IMPROVES QUALITY AND REDUCES CYCLE TIME

- Hands-free Process Enhancements
- Automated Defect Detection and Screening
- Automated 3rd Optical Inspection
- Automated X-ray Inspection
- Enhanced Material Handling using Automated Guided Vehicle (AGV)
- Customized Software to Enhance Existing Manufacturing Systems
- Barcode System
- Die Traceability
- Office Automation

We can Zero Defects & Accidents

"A" RATING BY AUDITORS FROM AUTOMOTIVE TIER 1 CUSTOMERS

Customer Complaint
< 3.7 Case per Billion Units

IATF 16949 ISO26262

UTAC
www.utacgroup.com

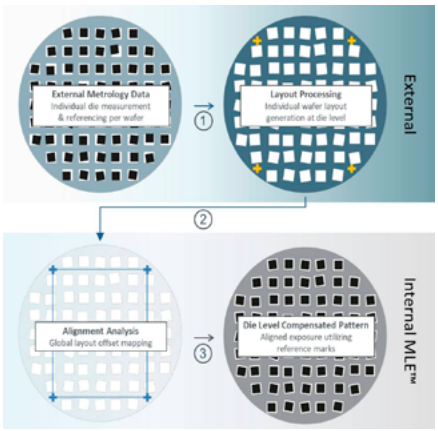


Figure 4: Die-level compensation schematic process flow.

processing excluding potential thermal influences. A simplified data integrity flow of die-level compensation is visualized in **Figure 4**.

In parallel, MLE patterning enables real-time individualized wafer-level layouts as well as simultaneous structuring of individual die layouts. In particular, the ability to implement ad hoc die annotations, serial numbers, functional and directly readable encryption codes or active patterning of fuse maps to optimize device binning for process or device tracking and documentation will lead to improved overall yield. Additionally, programmable modulations of UV dosage at the sub-pixel level enable exposure gradients that lead to controlled resist thickness-level variations after the development process, which is also known as greyscale lithography. This facilitates the fabrication of complex 3D multi-level resist patterns, which are applicable in multiple patterning processes, dual damascene, microelectromechanical systems (MEMS) devices and micro-optical elements (e.g., refractive, diffractive). The digital programmable die/wafer layout can be stored in numerous industry layout design standard file formats (e.g., GDSII, Gerber, Oasis, ODB++, or BMP). Because the vector layout is directly processed within a few seconds on the system under recipe control, neither resist type (positive/negative), exposure dose level, nor design layout complexity have any impact on the speed of the patterning process.

Patterning modularity and vital positioning in sub-grid

The resolution of MLE is aimed at typical back end of line (BEOL) resolutions with fine control of the irradiated lines, as well as their gaps (L/S <2µm), while maintaining critical dimension (CD) uniformity (CDU<10% CD) and positional accuracy of any arbitrary structures considerably within the patterning grid scale. This precision is matched by the distortion-free optics and the stage placement accuracy, which ensures seamless projection across the entire substrate. The exposure can be performed flexibly with a very high degree of freedom in intensity control, as well as precise light source spectrum tuning to achieve optimal absorption and reliable processing for a wide range of commercially established, as well as novel photoresists. The exposure light source operates at a wavelength spectrum of 375nm and 405nm, allowing for a mix and match of wavelengths to mimic known good process recipes (i.e., to follow the traditional mercury lamp spectrum) or to tailor the exposure towards specific customer demands. Both wavelengths can be simultaneously applied in any arbitrary mixture, and therefore enable thin-resist patterning, including positive, negative, polyimide, patternable dielectrics, dry

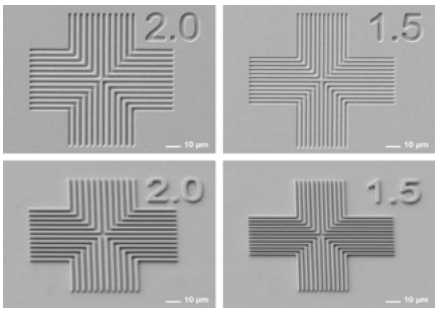


Figure 5: SEM results collage: Line space resolution tests on 1µm thick positive AZ MIR 701 (top); and line space resolution tests on 2µm thick negative AZ nLOF (bottom).

film or even printed circuit board (PCB) materials and also support thick-resist exposures at high aspect ratios typically encountered in wafer-level packaging, 3D MEMS patterning, microfluidics and integrated photonics applications. **Figure 5** displays a series of scanning electron microscopy (SEM) images of standard line-space resolution targets on 1µm thick positive AZ® MIR™ 701 resist on top, while results on the bottom show line-space resolution tests on 2µm-thick negative tone resist AZ® nLOF™. In both cases, a 1.5µm L/S result was achieved through further process optimization involving the reduction of surface reflection effects, which can be achieved by applying anti-reflective coatings or modifying substrate material properties.

The MLE technology can also finely control DoF in order to achieve steep sidewalls, thereby keeping the desired 3D contour of the resist, or preventing edge topping and footing. Large working distance and automatic adaptive focus ensures patterning uniformity across the exposure surface. Commonly used TOK P-W1000T resist for fine-line and core-line RDL creation was chosen to demonstrate various lines and spacing patterning performance, as well as sidewall patterning quality. **Figure 6** shows examples of SEM images of baseline evaluation, demonstrating: a) 2µm L/S resolution targeted on 8µm film thickness; b) 5µm L/S resolution with meander pattern; c) spacing variation of 1:2 ratio; and d) L/S variation in both horizontal and vertical directions with ratios of 1:1, 1:2, 1:3, and 1:4.

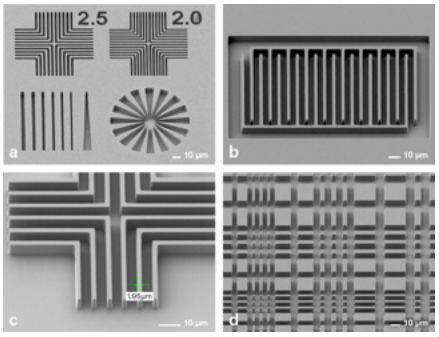


Figure 6: Baseline evaluation on a) 8µm thick TOK P-W1000T, b) meander with 5µm L/S; c) 1:2 spacing variation; d) L/S variation in both horizontal and vertical directions with ratios of 1:1, 1:2, 1:3, and 1:4.

Maskless operation scalability

Previously, back-end lithography results achieved during R&D using direct imaging equipment lacked the technological resilience for high-volume manufacturing (HVM) lines equipped with steppers. Today, the industry sees increasing product mix, such as chiplets and segmented dies, as a driver for continued performance scaling, as well as variability of applications. This triggers the need for dynamic patterning at various resist thicknesses and dose levels. MLE provides a high DoF at 2µm production resolution leveraging the physical diffraction limit established by the optics. At the same time, the scalability of MLE is broad in scope. The modular system scales

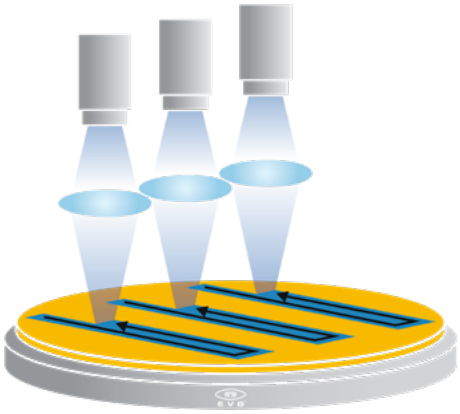


Figure 7: MLE's unique clustered exposure configuration enables exposure heads to be easily added to adjust for different throughput needs and substrate sizes.

according to user needs by adding UV exposure heads as shown in **Figure 7** for reasons such as: 1) facilitating rapid transition from R&D to HVM mode; 2) throughput optimization; or 3) adaptation to different substrate sizes and materials. It is also ideal for processing a range of substrates from small silicon or compound semiconductor wafers up to panel sizes.

Summary

MLE provides a new approach for patterning through smart and agile digital processing while delivering maskless scalability in throughput and format with a consumables-free infrastructure. It achieves the same patterning performance regardless of photoresist thanks to a flexible, reliable and scalable high-power UV laser source combination featuring multiple wavelength options. The platform allows for patterning of a wide variety of materials such as silicon, mold, glass, polymers and laminates, using the same optics. The wafer chuck and autofocus system compensate for substrate bow and warp, which is especially important for applications like FOWLP. The design flexibility that the technology brings to the current conservative environment opens up room for new innovations, helps to shorten development cycles, and at the same time bridges the gap between R&D and HVM.

Biographies

Bozena Matuskova is Business Development Manager at EV Group, St. Florian am Inn, Austria. She holds a Master of Engineering degree in Microelectronics from the Slovak U. of Technology and has 8+ years of professional experience in engineering, business development and technical product marketing activities, in both the semiconductor and automation engineering industries. Email B.Matuskova@evgroup.com

Thomas Uhrmann is Business Development Director at EV Group, St. Florian am Inn, Austria. He holds an Engineering degree in Mechatronics from the U. of Applied Sciences in Regensburg, and a PhD in Semiconductor Physics from Vienna U. of Technology.

Testing AiP modules in high-volume production for 5G applications

By Jose Moreira [Advantest]

The arrival of 5G promises enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable low-latency communication (URLLC). eMBB is already available to consumers, who can purchase 5G-capable handsets that can operate on 5G networks deployed by carriers in many regions throughout the world. But as 5G rolls out, the test community faces challenges and opportunities. That's particularly true regarding the antenna arrays that will connect handsets to base stations.

Most initial 5G deployments will take place only using the sub-6GHz frequencies used for previous generations of cellular technology, but they will be enhanced later with the possibility of short-range high data rate connections using mmWave frequencies. 5G new radio (5G NR) defines two ranges, frequency range 1 (FR1) and frequency range 2 (FR2). FR1 includes the sub-6GHz frequencies, but FR2 opens up mmWave frequencies above 24GHz for 5G deployment. 5G NR leverages the FR2 frequencies to achieve larger modulation bandwidths (for example, 800MHz). But because of the high transmission loss at these frequencies, it is necessary to use antenna arrays for multiple-input and multiple-output (MIMO) functionality and to focus the transmission beam (beam forming) in both the base station and the consumer's handset. These arrays come in the form of antenna-in-package (AiP) modules, which are a critical part of the current 5G wireless communication wave.

For the handset, these AiP modules will usually have an array of dual polarized patch antennas for top firing and, in some instances, also an array of dipole antennas for side firing as shown in **Figure 1**. To keep RF losses to the antenna radiators to a minimum, the AiP module includes an RF integrated circuit that provides the modulated mmWave signals to the AiP antenna array with the needed gain and phase to each radiating element. The module would then usually only require power, digital control signals, and modulated intermediate frequency (IF) signals.

AiP modules for 5G handsets need to be extremely small to fit into the modern cellphone form factor, and a multiple of them need to be used in a single cellphone because the user's hand position has a significant impact on the transmitted beam loss. Also, the AiP modules in a cellphone might not be all equal, but in fact have different antenna configurations depending on the handset design.

The 3GPP standard defines three methods for the over-the-air (OTA) standard compliance testing of AiP modules: direct far field, indirect far field (e.g., compact antenna test range, or CATR), and near-field to far-field transformation. Each of these methods have advantages and disadvantages, but they all require relatively large test chambers and a complex manipulator to rotate the AiP device under test (DUT) or the measurement antenna.

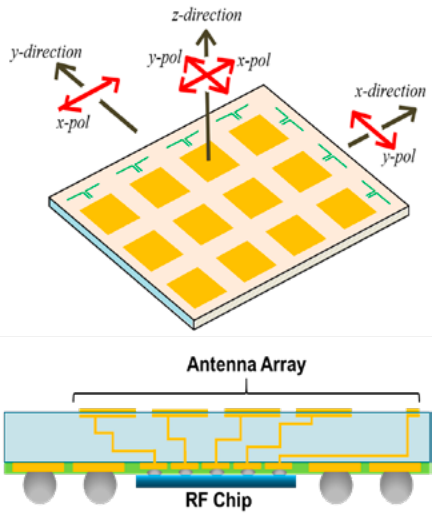


Figure 1: Drawing of an example of a generic antenna array module, comprising 12 dual-polarized patch antenna elements and seven dipole antenna elements.

For high-volume production testing, the objective is to check functionality of the DUT and not its compliance to the standard. Low cost of test is critical because most of the end applications are consumer-oriented. Also, to keep costs down it is important to be able to reuse as much as possible the test cell infrastructure already used for testing RF integrated circuits.

From a test engineering point of view there are multiple possible steps in testing an AiP module. First, the RF chip (e.g., a wafer-level chip-scale package [WLCSP] part), is tested at wafer level. This test can be either very simple and low cost (e.g., mainly DC or even a wire loopback on the mmWave ports), or it can also include full mmWave parametric measurements using an appropriate probe card and automatic test equipment (ATE) system. After the AiP module is assembled, the same question on the types of tests to be performed on the AiP module can be evaluated. It might consist of a simple low-cost DC test or even some kind OTA loopback test, to a full parametric OTA test with a measurement antenna. Finally, after integration in the end product (e.g., a 5G handset) a system-level type of test might also be performed. The test strategy at each stage depends on the overall test strategy for the AiP module.

Another important point is the AiP module calibration. For proper beam steering of the AiP module, it is critical that one is able to accurately set the gain and phase at each antenna element. If this accuracy cannot be guaranteed by design or a built-in self-test (BIST) calibration technique, then this calibration step needs to be performed at one of the test stages of the AiP module testing, or worst case at system level when the handset is assembled.

We will now concentrate on the case that a parametric OTA test is required for an AiP module, but because of

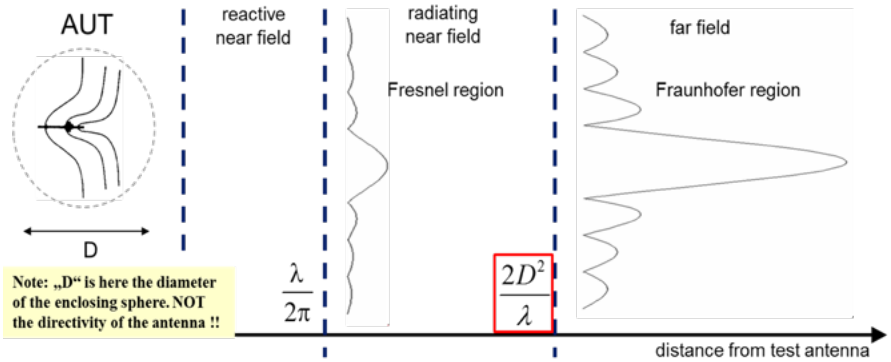


Figure 2: Measurement region definition for an antenna under test (AUT).

cost reasons, a methodology different from the defined 3GPP compliance test methods is needed. But first it is important to understand how in the field of antenna measurement one defines the measurement regions of an antenna under test (AUT) as shown in Figure 2 [1].

The start of the far field is usually defined by the Fraunhofer formula. Note that this is not a hard boundary, but a continuous transition where the radiated waves become locally more planar. From an antenna measurement point of view, the far-field region is the best because the radiated waves are locally planar, and also the measurement antenna is too far away to have an impact on the AUT. But the far-field distances also imply large dimensions for the measurement setup.

In this paper we will discuss three possible options available for the production OTA testing of AiP modules with automated test equipment: far-field testing, radiating near-field testing, and reactive near-field testing.

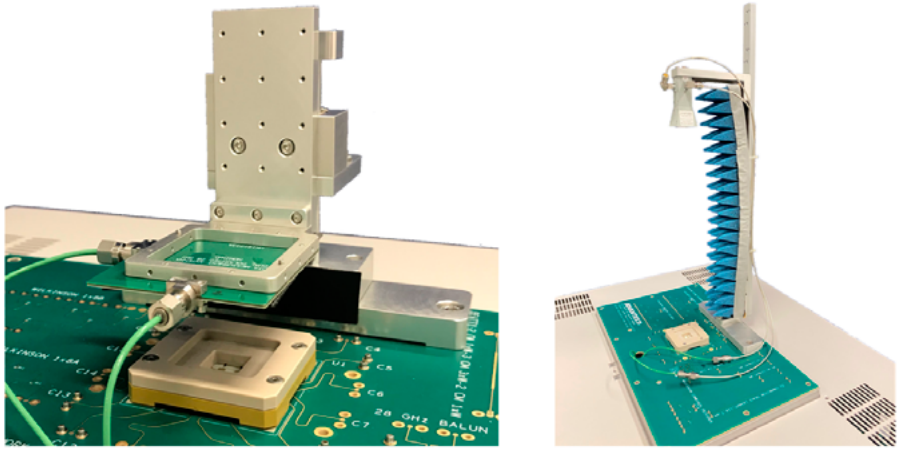


Figure 3: Examples of an OTA ATE far-field measurement setup: a) (left) motorized linear stage; and b) (right) static setup.

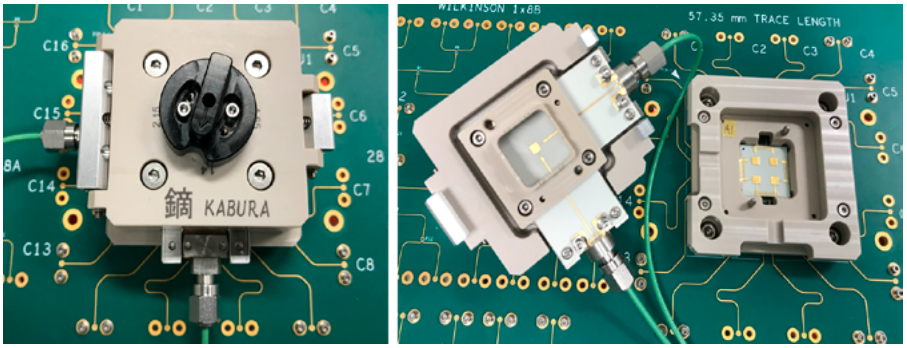


Figure 4: Example of a low-cost radiating near-field OTA socket for manual ATE-based OTA testing.

OTA far-field testing

OTA far-field testing on ATE usually means that although the measurement antenna is on the far-field, neither the AiP DUT, nor the measurement antenna move—they stay static at a certain predefined distance. This means that no traditional beam forming measurements are possible. Figure 3 shows two examples of a simple far-field measurement setup on an ATE system. This approach is excellent for an initial start with OTA testing on ATE because one can start in the safety of the far-field measurement range while doing correlation and debugging of the AiP DUT using the ATE system. Calibration on a far-field setup is also trivial using standard antenna measurement calibration procedures [1]. The problem arises when considering high-volume production by integrating a far-field OTA methodology on a standard ATE test cell.

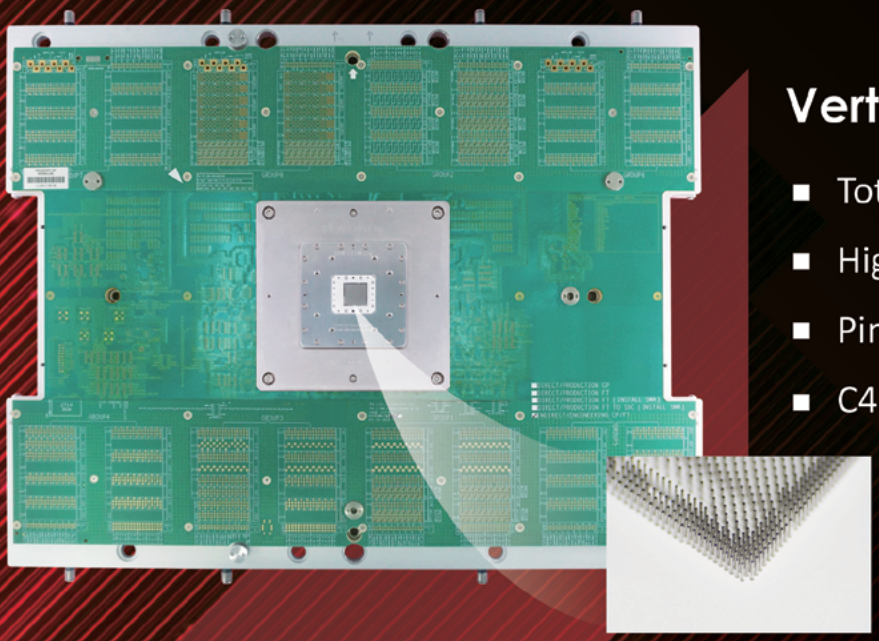
The mechanical dimensions required for a far-field OTA test solution prevent the usage of standard ATE test-cell commercial handlers, thereby requiring the usage of custom robotic handlers. This creates additional costs and hurdles especially for outsourced semiconductor and test suppliers (OSATS). Cost reduction through multisite implementation on ATE is also far from trivial with a far-field OTA ATE implementation.

OTA radiated near-field testing

To address the mechanical challenge of integrating an OTA measurement setup into a standard ATE test cell it is necessary to shrink the physical dimensions between the AiP DUT and the measurement antenna. One straightforward approach is to move the antenna into the radiating near-field region. Figure 4 shows an example of a low-cost radiating near-field test socket for a patch type antenna array AiP. In this example, the measurement antenna is 11mm from the DUT AiP antenna array. A radiating near-field antenna test has the advantages of easy integration within a standard ATE test cell along with easy multisite implementation, which results in a low cost of test.

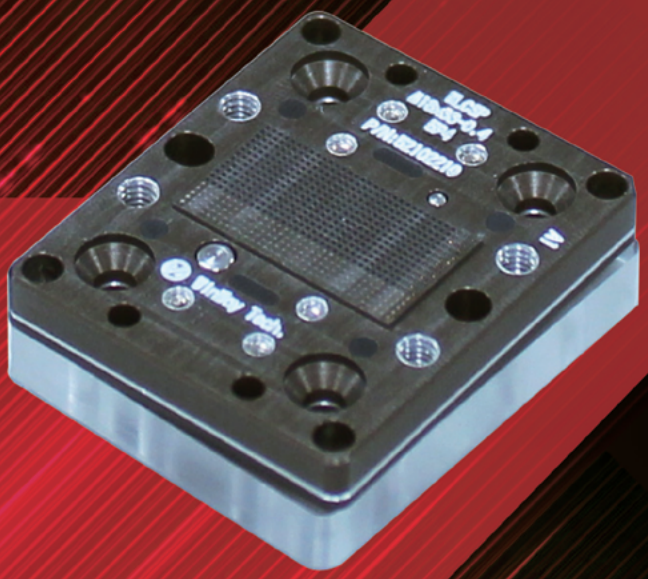
Because in a production test environment the objective is to identify failed AiP modules and not to characterize them, one could

WinWay Technology



Vertical Probe Card

- Total solution (Turnkey Service)
- High CCC Solution
- Pin count : Up to 30,000pins
- C4 pad pitch : Down to 75um



WLCSP Probe Card

- Maximum pin count up to 8000
- Maximum multi-site up to 32
- Minimum pitch down to 120um

  New Coaxial probe head down to 350um (Patented)

sales@winwayglobal.com

www.winwayglobal.com

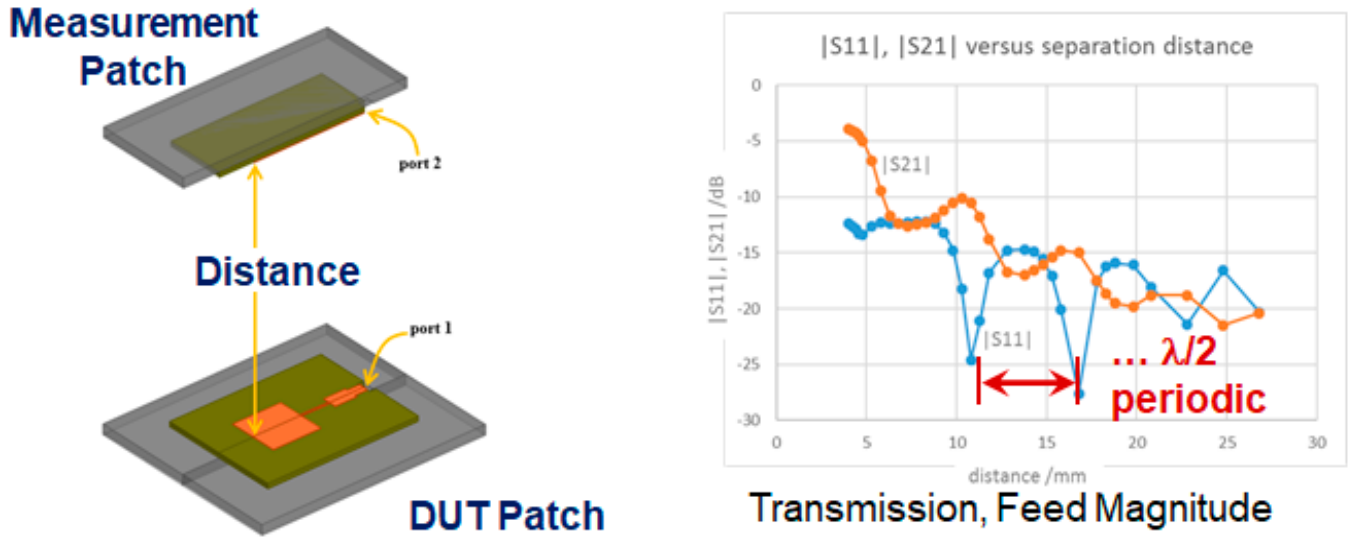


Figure 5: Impact of the measurement antenna on the DUT AiP antenna elements in a radiating near-field measurement.

assume that there would be some easy correlation between good AiP modules tested in a far-field setup with failing AiP modules tested on a near-field setup, assuming of course a comprehensive list of performed tests. This is a valid thinking, but one needs to be aware of two important drawbacks on a radiating near-field measurement setup. The first is that the measurement antenna is now so close to the AiP DUT antenna array that it will have an impact on the DUT AiP antenna elements (antenna detuning) and can even result in a standing-wave effect. This is shown with a simple simulation in **Figure 5** where two patch antennas are simulated with varying distances between them. This effect is also easy to see on a real measurement on an AiP module [2,3].

The second drawback is shown in **Figure 6**. Because only one measurement antenna is used, depending on the DUT AiP antenna array geometry, the distance of each DUT antenna array element to the measurement antenna will be different. This can have a significant impact on a worst-case scenario as show in [2,3]. Finally, calibration in the radiating near-field is not trivial. If golden-device calibration is used, results are critically dependent on the golden device's performance, and absolute measurements are not possible.

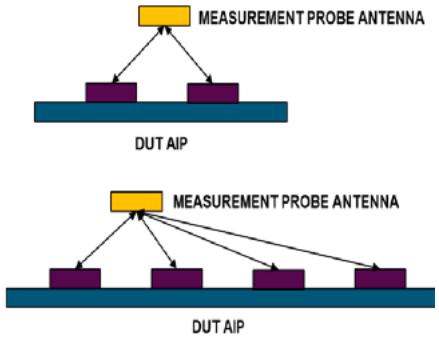


Figure 6: Impact of the single measurement antenna position on the distance to the different antenna array elements on an AiP DUT.

OTA reactive near-field testing

Another alternative approach for OTA testing with ATE is to measure the DUT AiP antenna array in reactive near field. In this case, a classical measurement antenna cannot be used because in the reactive near-field range it would have a dramatic effect on the DUT AiP antenna elements. To measure on the reactive near field, the antenna or probing element needs to be very small. **Figure 7** shows one possible reactive near-field probing concept for OTA ATE that has been patented by Advantest using two very thin parallel needles to probe the electric or magnetic field on the DUT AiP reactive near field. The main advantages are that each element of the DUT AiP array is individually measured (power and phase) and that the probe size is very small to minimize the disturbance of each radiating element. This concept is explained in more detail in [4,5]. **Figure 8** shows an example of a prototype reactive near-field socket [3]. Note that in the example of **Figure 8**, a dual-polarized 2x2 AiP array is measured resulting in eight individual signals. To keep ATE resources to a minimum (for cost of test reasons), a solid-state relay switches each of the antenna/polarization signals

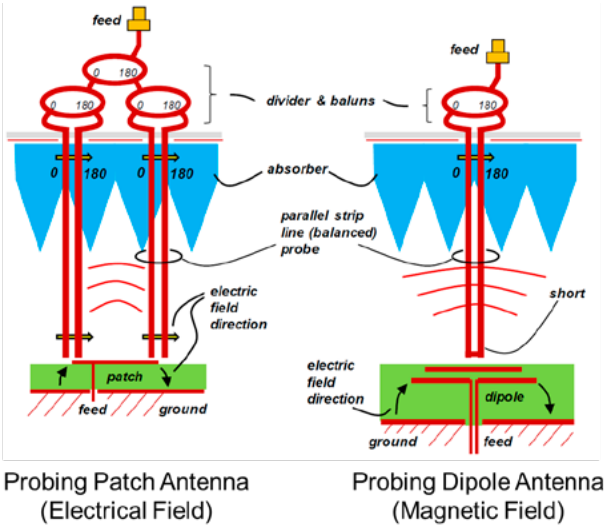


Figure 7: Reactive near-field probing concept for patch and dipole antennas.

in series to the ATE measurement instrument. A parallel measurement approach is also possible, but requires eight ATE measurement instruments. The optimal setup will depend on a detailed cost of test analysis.

Summary

For OTA testing with ATE of AiP modules, there is no right or wrong answer. Depending on the testing requirements and on the testing stage (for example, initial ramp up or mature high-volume manufacturing), the OTA test strategy might be different. Table 1 shows a high-level comparison of the different OTA test strategies presented on this paper.

In a future follow-up paper we will use a custom-designed 28GHz 2x2 path antenna array in a 0.4mm ball grid array (BGA) pitch package to compare the different approaches in terms of OTA measurement results with the Advantest V93000 Wavescale Millimeter CardCage ATE system.

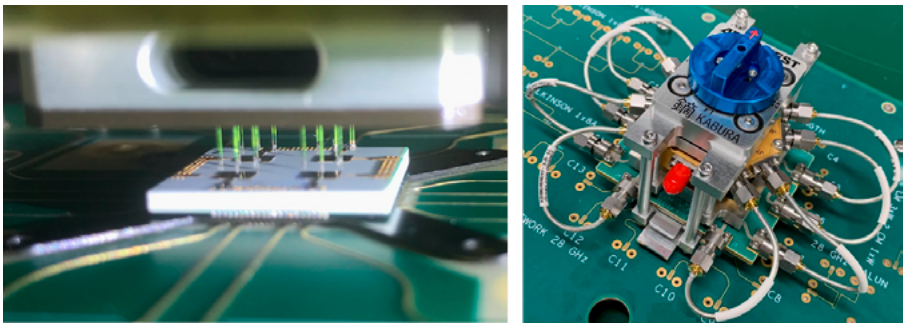


Figure 8: Example of a reactive near-field probing prototype socket for demonstration of the concept on a 2x2 patch antenna array.

OTA Test Strategy	Advantages	Disadvantages
Far-Field Antenna	• Far-field measurement • DUT antenna is not impacted by the measurement antenna • Easiest setup to correlate with measured data using 3GPP compliant methods	• Integration in standard ATE test cell difficult due to mechanical dimensions • Multisite implementation complex • High cost of test for volume production
Radiating Near-Field Antenna	• Easy integration on standard ATE test cell • Easy multisite implementation • Low cost	• Measurement antenna (e.g., patch) will have an impact on the DUT antenna performance (standing wave effect) • Possible different distances between the measurement antenna and DUT antenna array elements • If "golden device" calibration used, results are critically dependent on "golden device" performance. No absolute measurements possible
Reactive Near-Field Probe	• Probing measurement elements have a minimal impact on the DUT antenna • Each DUT radiating element is individually measured • Easy integration on standard ATE test cell • Multisite implementation possible	• Higher NRE and manufacturing costs • Long lead-time • Higher loss due to weak coupling of the probes to the DUT radiating elements

Table 1: Comparison of the different ATE OTA testing methodologies.

Acknowledgements

We would like to thank Natsuki Shiota, Aritomo Kikuchi, Hiromitsu Takasu, Hiroyuki Mineo, Sui-Xia Yang, and Frank Goh from Advantest for their support and collaboration on the OTA project development. We would like also to thank Prof. Jan Hesselbarth from the University of Stuttgart.

References

1. C. Parini, et al., *Theory and Practice of Modern Antenna Range Measurement*, IET, 2014.
2. J. Moreira, J. Hesselbarth, K. Dabrowiecki, "Challenges of over the air (OTA) testing with ATE," TestConX China, Shanghai, Oct. 29, 2019.
3. N. Shiota, A. Kikuchi, H. Mineo, J. Moreira, H. Takasu, "Socket design and handler integration challenges in over the air testing for 5G applications," To be presented at TestConX 2020, May 2020.
4. J. Hesselbarth, G. Sterzl, J. Moreira, "Probing millimeter-wave antennas and arrays in their reactive near field," 49th Euro. Microwave Conf., 2019.
5. U. Dey, J. Hesselbarth, J. Moreira, K. Dabrowiecki, "Over-the-air test of dipole and patch antenna arrays at 28GHz by probing them in the reactive near field," To be presented at 95th ARFTG Microwave Meas. Conf., 2020.

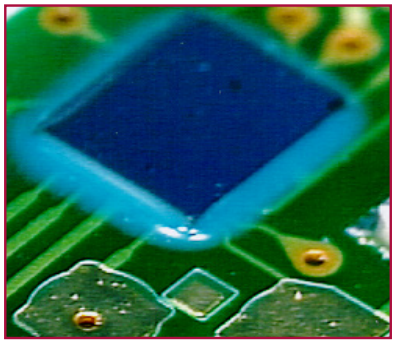
Biography

Jose Moreira is a Senior R&D Staff Engineer at Advantest in Böblingen, Germany. He focuses on testing high-speed digital, silicon photonics, and 5G mmWave devices and holds an MS degree in Electrical and Computer Engineering from the Instituto Superior Técnico, Lisbon U., Portugal. He is a senior member of the IEEE, technical member of the IEEE P370 standard and co-author of the book *An Engineer's Guide to Automated Testing of High-Speed Digital Interfaces*. Email jose.moreira@advantest.com

Laser Assisted Bonding

PacTech
member of nagase group

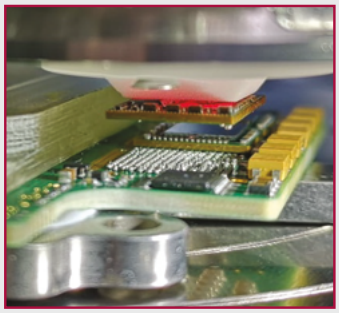
The LAPLACE-system provides an integrated solution for flip chip assembly, the laser reflow will be done during the placement - all in one step. The laser assisted assembly can be applied for soldering, ACF and NCP interconnections. The optional dispensing unit in the flip chip assembly platform allows a maximal flexibility for flux, solder paste and/or ACF, NCP dispensation.

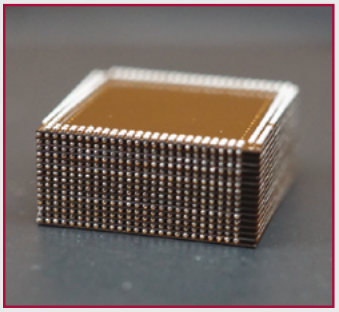


New Applications

Selective Chip Rework

- Selective removal of bad die
- In-situ replacement & reflow
- Cost saving for high density fan-out packaging



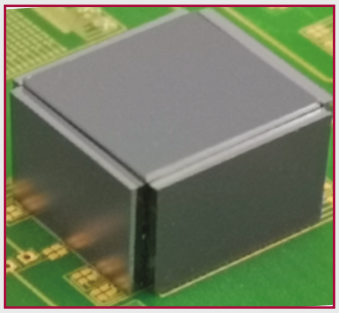


3D Packaging

- In-situ die placement & reflow
- Lowest thermal & mechanical stress
- High chip stack uniformity

3.5D Multilayer Die Stacking

- Horizontal & vertical chip assembly
- No requirement for TSV structures
- Simplification of complex package design



Embedded trace and 2-in-1 RDL for fan-out panel-level packaging

By Kesheng Feng, Kwangsuk Kim, Saminda Dharmarathna, William Bowerman, Jim Watkowski, Johnny Lee, Jordan Kologe
[MacDermid Alpha Electronics Solutions]

With respect to fan-out panel-level packaging (FOPLP), fabricators have struggled to justify the upfront costs of installing the process because of challenges with copper plating performance and package cost reduction compared to fan-out wafer-level packaging (FOWLP). A key challenge is obtaining a combination of high line resolution and height uniformity from the copper plating process that forms the circuitry of the redistribution layer (RDL). This is especially challenging in simultaneous plating of traces and filling of vias, in what is called 2-in-1 (RDL) plating. Coplanarity between the surface of the plated copper traces, pads and/or filled vias is a critically important qualifying feature for these copper plating processes. A surface that is not coplanar could result in signal transmission loss and distortion of the circuitry layers after lamination. These defects can cause short circuiting and catastrophic failure of the device. Because of the scale of the circuitry involved, planarization processes that can restore coplanarity can impart registration errors into the layer. Therefore, copper plating solutions providing a uniform surface profile without requiring additional post treatment are key to successful implementation of RDL plating for panel-level packaging. In this article, we present a background on the manufacturing technologies that create copper metallization for FOPLP, discuss the chemical and equipment influences on copper electroplating for this application, and examine the performance of a commercial electrolyte system in both embedded trace and 2-in-1 RDL plating.

SAP, mSAP, and ETS technologies

The technologies currently utilized for manufacturing the RDL for FOPLP include semi-additive processing (SAP), modified semi-additive processing (mSAP), and embedded trace plating. SAP has been used for making fine lines on organic substrates for decades. The process begins with electroless copper plating to form an ultra-thin conductive seed layer, followed by photolithography to pattern a photo resist on the surface. Electrolytic copper plating is then used to form metallization structures between the photo resist patterns. This is then followed by removal of photo resist and flash etching away the copper seed layer to complete the patterning. SAP has advanced enough to allow wiring dimensions down to 9µm, but due to the small amount of side etching that occurs to the plated copper traces during the flash etching step, there has been a challenge in reducing linewidth scales further. There are also other challenges as line/space trends approach 9µm/9µm, including adhesion to the organic substrate, process tool accuracy across uneven substrate surfaces on panels that can be in the range of 500mm x 400mm in size, plating thickness uniformity, and the high costs of specialized equipment, such as advanced photolithography tools.

mSAP is a newer process that has been widely adapted in reducing circuitry dimensions for mobile electronics by effectively fulfilling the utility of printed circuit board (PCB) and integrated circuit (IC) substrate in the device. The typical mSAP flow begins with an organic substrate clad with a very thin copper foil of approximately 1-5µm in thickness. Microvias are then laser drilled and the panels desmeared, either through a plasma or chemical desmear or a combination of both. This process cleans any resin residues from the target pad and imparts topography to the via walls for adhesion of subsequent copper deposits. The panels are then processed through a primary metallization process such as, electroless copper, carbon-based direct metallization, or conductive polymer, to make the via walls conductive for electrolytic copper plating. The panels are subsequently imaged and pattern-plated with electrolytic copper to both fill the microvias completely with copper, and build the copper traces to the required height in a single step. After plating, the resist is stripped and a differential or flash etch is done to form the final circuitry. The fine-line resolution from this technology is typically limited to 13µm dimensions.

Embedded trace substrate (ETS) plating technology provides additional cost reduction and higher resolution advantages without a flash etching step, allowing fine-line resolution approaching 5µm. ETS technology uses a photolithographic process to create pattern-plated copper metallization structures onto a conductive

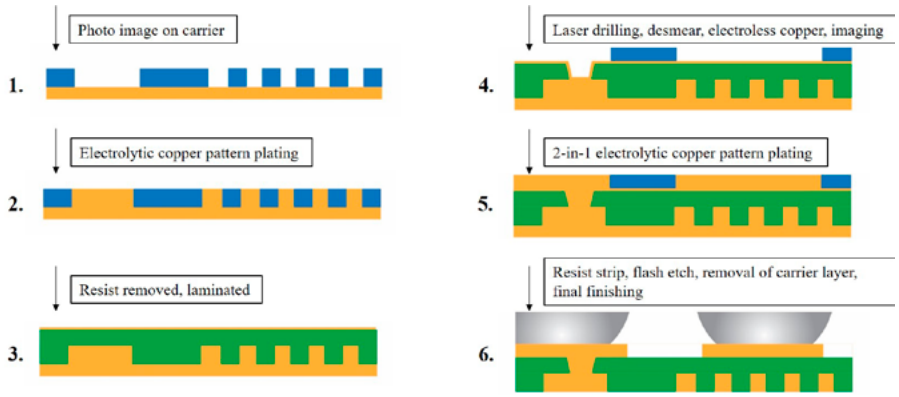


Figure 1: Combining embedded trace substrate technology (steps 1, 2, 3, and 6) and mSAP technology (steps 4, 5, and 6) in a single process flow.

carrier. These structures are laminated into dielectric to embed the circuits, and then built up with an mSAP or SAP type process.

Figure 1 describes the process for the fabrication of a panel-level substrate over 6 steps by using a combination of ETS and mSAP processes. ETS technology involves using a photolithographic process to create pattern-plated metallization structures onto a conductive carrier after photo imaging as seen in steps 1 and 2. The fine-line traces created are then laminated onto the dielectric to embed the circuits, forming the M1 layer, in step 3. The conductive carrier is removed in step 6 after the substrate is processed with the M2 layer and M3 layers using mSAP technology. The mSAP process starts in step 3 with a thin copper foil of approximately 1 to 5µm in thickness on the organic substrate, which allows for good adhesion. This foil, along with electroless copper, are plated with electrolytic plating processes to create the M2 or M3 layer, which can contain both fine-line RDL and vias, as seen in steps 4 and 5. This is the previously mentioned 2-in-1 plating step. The copper carrier that covers the embedded trace is removed in step 6.

Factors influencing copper plating quality

Typical acid copper electrolytes contain copper sulfate, sulfuric acid, chloride ions, and organic additives. These additives play a crucial role in controlling the deposit distribution as well as the physical properties of the copper deposit. To meet the specific objectives of the plating process these additives must be monitored and controlled properly. The additives work in combination when they are controlled within a given range to improve plating uniformity. Namely, these additives are the wetter, brightener and leveler. The wetter works in the presence of chloride ion to adsorb onto the cathode and increase the effective thickness of the diffusion layer. As a result, the plating current increases at the cathode and the deposit becomes more uniform, so a densely-packed copper deposit can be obtained without burning. This modified diffusion layer improves the distribution of the deposit in fine-line plating. The brightener reduces suppression and acts as a grain refiner to deposit copper with a fine grain structure in random orientation. Because of its strong effects on overall grain structure, the brightener has the greatest influence on physical properties of the deposit, such as tensile strength and elongation.

The leveler is a mild suppressor that adsorbs onto specific locations such as corners and peaks of base materials, aiding in evening out the thickness of copper deposit. Within the microprofile at the surface of the panel, the diffusion layer tends to be thin at the peaks and thick at the valleys. Without a leveler, the copper plating will exaggerate the microprofile resulting in higher peaks. On the other hand, the plating on the peaks will be

Systek ETS	Chemistry	Concentration
VMS	$CuSO_4$	100 g/L
	H_2SO_4	200 g/L
	Cl^-	60 ppm
Additives	Wetter	5 ml/L
	Brightener	2.5 ml/L
	Leveler	10 ml/L
Parameter	CD	1.0-4.0 ASD
	Temperature	20-27 °C

Table 1: ETS electrolyte and plating parameters.

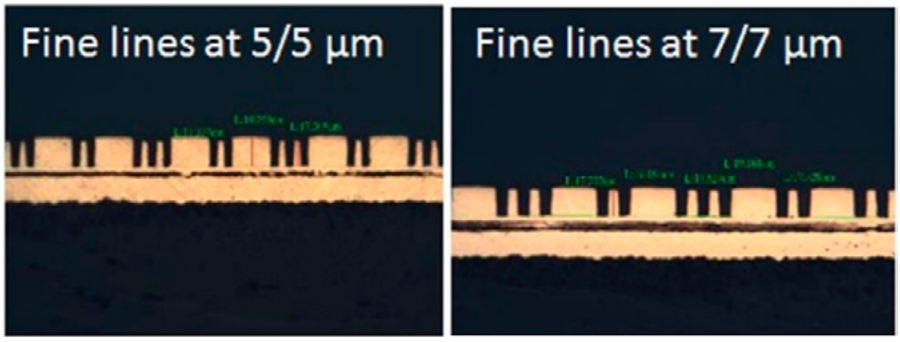


Figure 2: Cross sections of fine lines from the plating process with soluble anodes.

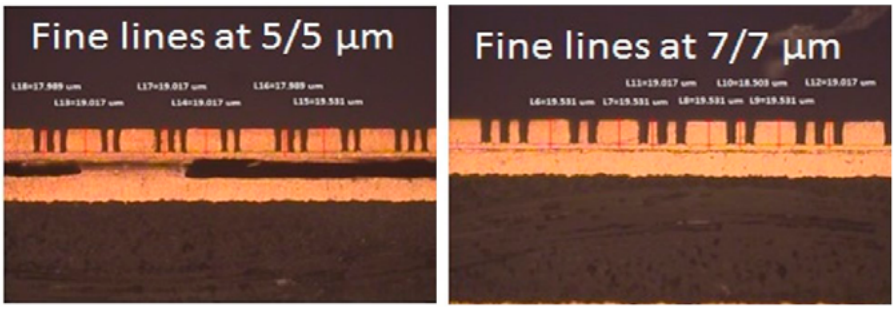


Figure 3: Cross sections of fine lines from the plating process with insoluble anodes.

suppressed and the microprofile will be diminished if a leveler is present. Proper additive selection and control are crucial to obtain plating uniformity and desirable physical properties of electroplated copper. In addition to selection and optimization of additives, anode type, virgin makeup solution (VMS), and plating current density also have to be taken into consideration for their effect on plating performance.

Evaluation of commercial electroplating for FOPLP

In the following sections, we examine the plating capability of a commercial embedded trace plating electrolyte system. The plating processes we describe were designed for panel-level packaging boards that are up to 2 to 3 layers thick. We also show the results of adapting this same electrolyte system for 2-in-1 RDL plating by varying the concentration of the copper and acid in the VMS. Performance was evaluated by measuring plating uniformity and coplanarity of both fine lines, pads, and filling of microvias on panel-level type substrate. We compare the effect of soluble versus insoluble anodes on plating uniformity and examine the grain structure of the deposit by XRD and FIB-SEM imaging. The plating for the samples shown was conducted in vertical continuous plating (VCP) equipment in high-volume production conditions. Tight control was kept over all additive components through cyclic voltammetry stripping (CVS) analysis.

Embedded trace plating

For embedded fine-line plating we used an electrolyte branded as Systek ETS as described in Table 1. The test panels used were carriers with photo imaged dry-film patterns on them with a thickness of 25µm. Each test panel went through a pre-clean cycle of 1min acid cleaner, 1min rinse, and 1min 10% sulfuric acid before plating in the acid copper containing the electrolyte.

Table 2 describes an example of the influence of anode type on plating uniformity for fine lines and vias in VCP equipment using the electrolyte system previously described. Panels were plated at 1.5 ASD, for 60min to obtain a copper thickness of approximately 20µm. The plated height variation between fine lines of 5µm or 7µm width to the larger pads was measured from cross sections as shown in Figures 2 and 3. When the equipment was operated with soluble anodes, this variation between fine lines and pads was 1.4µm for the panel that had 5µm linewidth, and 2.19µm for the panel that had a 7µm linewidth. When the equipment was operated with insoluble anodes, the plated height variations between fine lines and pads were all below 1.0µm for both 5µm

Systek ETS Soluble anode	Line thickness(µm)				Pad thickness (µm)				Thickness variation R (µm)
Fine line (µm)	1	2	3	avg.	1	2	3	avg.	(line-pad)
7/7	17.82	18.03	17.73	17.86	19.45	19.16	19.16	19.26	1.40
5/5	17.14	17.86	17.37	17.46	20.14	19.55	19.26	19.65	2.19
Systek ETS Insoluble anode	**Line thickness(µm)**				**Pad thickness (µm)**				**Thickness variation R (µm)**
Fine line (µm)	1	2	3	avg.	1	2	3	avg.	(line-pad)
7/7	19.02	18.5	19.02	18.85	19.53	19.53	19.53	19.53	0.68
5/5	17.99	19.02	17.99	18.33	19.02	19.02	19.53	19.19	0.86

Table 2: Anode type influence on the performance in fine-line and pad variation for embedded trace plating.

	Pad (µm)					Trace (µm)					
Trial#	1	2	3	4	Avg.	1	2	3	4	Avg.	R Value (µm)
Top	19.43	19.54	19.49	19.43	19.47	18.76	18.98	19.15	19.32	19.05	0.42
Center	19.43	19.26	19.43	18.93	19.26	18.98	18.99	19.04	18.65	18.91	0.35
Bottom	19.82	20.10	19.77	19.88	19.89	19.71	19.60	19.71	19.48	19.63	0.26

Table 3: Plated height variations between fine lines and pads within unit, as well as within panel.

Copper Sulfate	Sulfuric Acid	Chloride	Via Size/Dimple (µm)	Fine Lines(µm)	Pads(µm)	Height Variation
220 g/L	50 g/L	60 ppm	60x40 µm /3.1 µm	18.22/18.09/18.26	17.60/17.41/17.25	Average @1.01 (µm)

Table 4: Plated height variations between fine lines and pads when vias were filled.

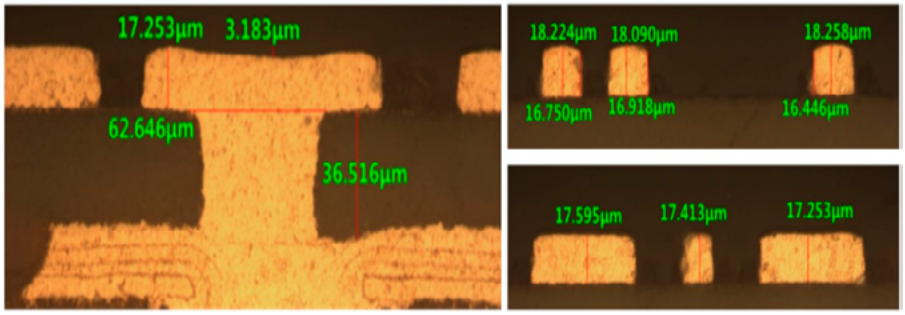


Figure 4: Cross sections and measured data of fine lines and filled vias.

and 7µm linewidth panels. This data is shown in Table 2.

In addition to the uniformity control benefit, insoluble anodes are also easy to maintain and allow the plater to apply higher plating current density. For larger linewidths, the process has even tighter control. For fine lines of 10µm width, the plated copper height variation is below 0.5µm when plating at a current density of 1.5 ASD. At a higher current density of 4 ASD, the thickness variation was approximately 1µm. It is common knowledge that plating uniformity can be improved when plating current density is reduced.

The variation of plated thickness across a panel-level substrate that is 410mm x 510mm in size was below 0.5µm. Measurements taken on the panel at the top, center, and bottom allowed across-panel uniformity to be assessed. This data is shown in Table 3.

2-in-1 simultaneous fine-line plating and via filling

For RDL applications containing both vias and fine lines, it is a challenge for an electrolyte to maintain great via filling capability while also obtaining good uniformity and coplanarity on fine lines. We found that by adjusting the electrolyte VMS to have copper sulfate at 220g/L and sulfuric acid at 50g/L, it enabled the process to fill vias of 60 x 40µm with a dimple of less than 5µm (Table 4). Under these conditions, the plated height variation between fine lines of 15µm width to the larger pads was approximately 1.0µm (Figure 4).

Various Data	Lit. Value Cu	Systek ETS 1200
111	100	100
200	46	11.4
220	20	10.7
311	17	5.5
Lattice constant a [A]	3.615	3.609
Density [g/cm³]	8.92	8.98
Stress [MPa]	----	-6.7 ± 4.7

Table 5: Copper deposit XRD data.

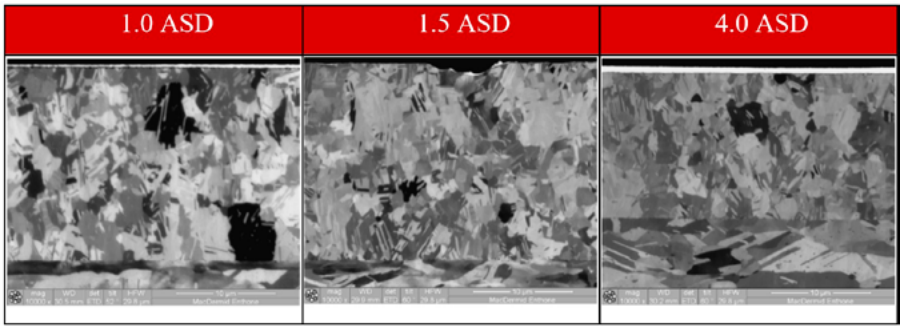


Figure 5: Copper deposit FIB/SEM pictures under different plating current densities.

Deposited copper physical properties

Physical properties of the plated copper deposit are essential for the reliability of the substrate. A few of the most important physical properties are tensile strength, elongation percentage, and internal stress. These properties show the tolerance of the deposit for thermal stress and warpage. Standard testing equipment was used to measure the tensile strength, elongation percentage, and internal stress of the deposit. The copper plated from the process has a greater than 36,000psi tensile strength and more than 18% elongation. Because of the extraordinarily thin dimensions required by modern packaging RDL, internal stress of the copper metal is an important parameter. With high internal stress, the deposit may warp, and warpage may get worse with time or with temperature. The plated deposits from the process at various current densities all exhibit low stress—below 1.0Kg/mm^2.

Deposited copper grain structure

An X-ray diffraction (XRD) study was performed for the plated deposits at a current density of 1.5 ASD to identify the crystal phase and different planes. The diffraction pattern of the copper grains obtained was the same as the standard reported for copper in the literature. In addition to relative intensities of the crystal orientations, the crystallographic density and the lattice constant is also of interest in determining whether there is a preferred orientation. The data indicates that the deposits from the bath have preferred [111] planes as shown in Table 5.

Focused ion beam-scanning electron microscopy (FIB-SEM) photos showed that the plated copper deposit had equiaxial grain structure, which does not have much variation under different plating current densities (Figure 5).

Summary

Panel-level packaging presents unique challenges for electrolytic copper metallization systems. With proper selection of VMS electrolyte and equipment, advanced packaging fabricators can achieve the required level of coplanarity for embedded trace and 2-in-1 plating for RDL build-up. Commercial electrolyte systems for embedded trace plating are available that can provide good performance and, with adjustment, are capable of 2-in-1 fine-line plating and via

filling with minor adjustments of the VMS. The physical properties, tensile strength and elongation of these copper-plated deposits pass IPC class III specification and have low internal stress as-plated and after annealing, thereby providing a reliable deposit that will withstand the stresses of assembly and device usage.

References

1. H. Akahoshi, M. Kawamoto, T. Itabashi, O. Miura, A. Takahashi, S. Kobayashi, "Fine-line circuit manufacturing technology with electroless copper plating," IEEE Trans. on Components, Packaging, and Manufacturing Tech.: Part A, vol. 18, pp. 127-135, 1995.
2. Y-H. Chen, S-L. Cheng, D-C. Hu, T-J. Tseng, "L/S ≤5/5µm line embedded organic substrate manufacturing for 2.1D/2.5D SiP application," IMAPS, Nov. 2015.
3. C. Chen, M. Lin, G. Liao, Y. Ding, W. Cheng, "Balanced embedded trace substrate design for warpage control," IEEE Elec. and Comp. Tech. Conf., May 26–29, 2015, pp. 193–199.
4. Y. Li, D. Goyal, *3D Microelectronic Packaging From Fundamentals to Applications*, Springer, Jan. 20, 2017 - Technology.
5. S. Dharmarathna, et al., "High throw DC acid copper formulation for vertical continuous electroplating processes," IPC APEX Expo 2017, San Diego, CA, Feb. 14-16.
6. Y. Zhang, G. Ding, P. Cheng, H. Wang, "Numerical simulation and experimental verification of additive distribution in through-silicon via during copper filling process," Jour. of The Electrochemical Soc., 162 (1) D62-D67 (2015).

Biographies

Kesheng Feng is Director of Research, Metallization Technology at MacDermid Alpha Electronics Solutions, Waterbury, CT, USA. He has 22 years of experience in the PCB industry with MacDermid Alpha with a primary focus on IC substrates, HDI, and primary metallization techniques. He holds 12 US patents, has authored more than 25 technical articles, and holds a PhD in Organic Chemistry from Nankai U. in China. Email Kesheng.Feng@MacDermidAlpha.com

Kwangsuk Kim (Leslie Kim) is Business Manager, IC Substrates and Electronics Materials at MacDermid Alpha Electronics Solutions, Bundang-Gu Sungnam-City, Kyungki-Do, Korea. In his 21 years with MacDermid Alpha he has dedicated himself to the development and application of new processes and technologies in high technology PCB, IC substrate, LED and molded interconnects. He holds a Bachelor's degree in Applied Chemistry and a Master's in Business Administration, both from Anjou U. in Korea.

The true cost of undetected defects

By David L. Adler and Brennan Peterson [SVXR Inc.]

The cost of an undetected defect can be much higher than the cost of the defective part: an electronic component that fails after installation into a phone, car or medical device can cost hundreds, thousands, or potentially millions of times the cost of the part. Eliminating defects early in the supply chain saves money and possibly lives. As advanced packaging makes its way into automotive applications, finding latent defects that affect reliability becomes more important. In this paper, we explain the benefits of 100% inline process control using X-ray technology. Lastly, we present a new technology, using advanced X-ray imaging and artificial intelligence (AI)-based defect detection, that addresses the ever-increasing complexity of highly integrated post-Moore's Law devices.

Background

In 1985, Gordon Moore noted the following: "It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected. The availability of large functions, combined with functional design and construction, should allow the manufacturer of large systems to design and construct a considerable variety of equipment both rapidly and economically." A Walkman®, phone, boombox, and camera became a phone. And within that phone, the separate central processor unit/graphics processing unit (CPU/GPU)/Northbridge/Southbridge/cache all merged onto a system on chip (SoC) with in-package dynamic random access memory (DRAM). The driving force behind semiconductor advances through today was in integration, and the benefits of adding both performance and capability. Hidden away in this integration effort was a subtle benefit: making more devices together is both functionally better and more reliable. For example, one good chip was more reliable than 10 put together, thereby giving profound benefits in multiple domains, from healthcare to transportation.

Now, however, the issues of integration become varied in nature when the different manufacturing technologies no longer scale as usefully together. It is also increasingly true that the right DRAM controller technology is no longer the same as the static random access memory (SRAM), or the GPU, or the machine learning (ML) subsystem, or perhaps even some of the logic. These features work better (and are cheaper to boot) when separated. Integration, in the more modern parlance, is taken to be integrated circuit (IC) integration on high-density substrates. Figure 1 shows an idealized future assembly, with multiple levels of integration. There is, however, one lurking issue: integration carries with it high costs for failures. The $1 IC can cause a fail in a $50 board, which necessitates a $1000 repair, that in turn might necessitate a $50,000 engineering change, or a $5,000,000 recall.

The scenario described above was solved with single chips because the causes of errors on chips (process variation and error) were fully correlated for an integrated chip, so long as you can continue to co-optimize the different

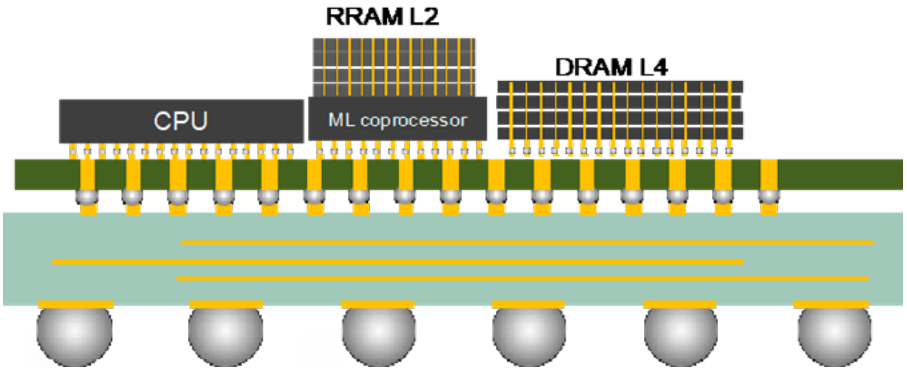


Figure 1: Example of a heterogenous, multilayer integration. There are (from bottom to top) ball grid array (BGA) bumps, µbumps/pillars, then µpillars, then through-silicon vias (TSVs) to connect memory. The darker green is a Si interposer, the light green is an organic substrate.

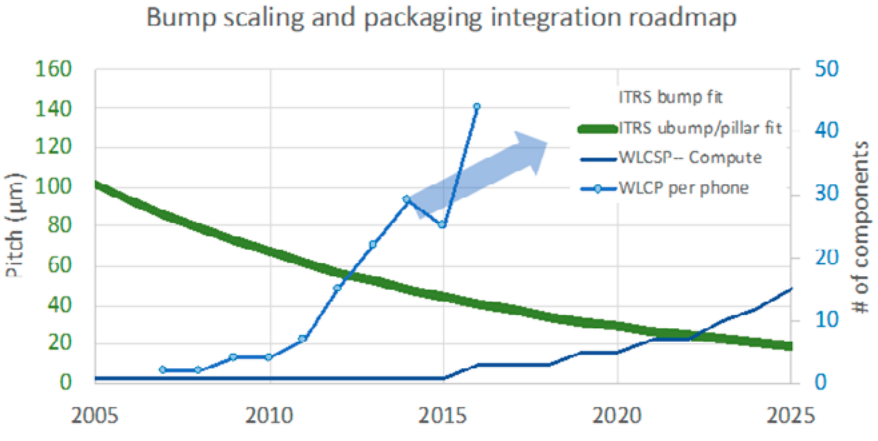


Figure 2: Bump scaling and packaging integration roadmap.

features. It is no longer possible to co-optimize all the different parts at <10nm geometries, so the parts are splitting. In the case of multiple components, errors add—or more precisely, success rates multiply—as in 10 components with 99% success have a 0.99^{10} or 90.4% success rate).

Keeping the above discussion in mind, there is an additional wrinkle: the increase in chip count also increases the demands on the bandwidth between the different components. With more bandwidth, there is a greater need for connection density. So even with a slowdown in the typical Moore's Law implementation, package components will add connection density, and will therefore continue to reduce pitch, likely at an accelerated rate. An extended roadmap is shown in Figure 2. Pitch continues to decrease, and the supported bandwidth continues to double roughly every two years. The International Technology Roadmap for Semiconductors (ITRS) data is fit to the roadmap, but linearized. This works out to an ~8% per year pitch decrease for the small scale, and ~4% for the larger bumps. In addition, the number of integrated components will increase significantly.

Case studies

To put the above discussion in practical terms, let's look at a pair of simple cases.

A phone. Consider a simple device on a phone, with an ~100M per year rate (~20% market share) and a $5 cost. So we can assume the following components: a sound chip, an amplifier, some modem components, and some Bluetooth device. The board for this is approximately $200, from TechInsights and IHS Markit reports. Table 1 shows the cost of fails, where we look at failure rates in parts per million (ppm). The failure rate is the rate for a single component.

A rate of 100ppm is a .01% fail rate. And it costs about $10M per year in returns at the assembly (phone) level. I calculate a logic board cost for the phone of ~$260, and I rounded up the entire assembly to $1000. For a basic 100ppm failure rate, that is $2.6M per year in costs just to replace the board, if the design allows that. And that is ignoring return costs and failure analysis (FA) costs. An improvement from 500 to 100ppm defect rate would save $10M per year at the integration level.

There are two notable aspects to consider in light of the above: first, that the cost to the end device manufacturer is much, much higher than a .01% failure of a $5 part would indicate. And that is a core lesson of reliability: the cost is borne by the final manufacturer or ultimately, the consumer. Second is that at high volumes, extensive inspection even for fairly low cost items pays for itself.

These costs also ignore the cost of failure analysis on the producer side, as well as goodwill and reputation. These may well outweigh the direct costs. On the production side, the critical benefit of inspection is typically in product qualification and test. For example, months could be saved from qualification time, saving (in this "toy" model) ~$10M per week.

An automotive electronic control unit (ECU). An automotive ECU is a more complex device compared to the simple phone discussed above. In this case, we have a more expensive unit, but lower volumes. There are about 70M cars sold worldwide, and if we assume 1/7th use these electronics, then we can make the calculations shown in Table 2.

Similar to the phone "case," the cost to the IC manufacturer of low yield is fairly trivial (keep in mind that the manufacturer will make about $1B in revenue). The cost of repair in this case may well be higher at the assembly level—an ECU replacement is perhaps a $1000 fix, and maybe more. That means that for a 500ppm failure rate, the replacement cost is $5M annually. If the car fails, this is a $20-100M cost. Across the industry, a 100ppm defect rate on critical components still gives rise to an ~$10-$150M in costs. Considering the vast increase in electronic components projected in the next decade because of new automated driver assistance programs, major improvements are needed.

Input data			Annual cost		
Failure Rate (PPM)	Device cost ($)	Annual production (M)	IC	Board	Assembly
10	5	100	$5,000.00	$260,000.00	$1,000,000.00
100	5	100	$50,000.00	$2,600,000.00	$10,000,000.00
500	5	100	$250,000.00	$13,000,000.00	$50,000,000.00

Table 1: Cost of failures for a phone component.

Input data			Annual cost		
Failure Rate (PPM)	Device cost ($)	Annual production	IC	Assembly	System
10	100	1.00E+07	$10,000.00	$100,000.00	$2,000,000.00
100	100	1.00E+07	$100,000.00	$1,000,000.00	$20,000,000.00
500	100	1.00E+07	$500,000.00	$5,000,000.00	$100,000,000.00
1000	100	1.00E+07	$1,000,000.00	$10,000,000.00	$200,000,000.00

Table 2: Cost of failures for an automotive controller.

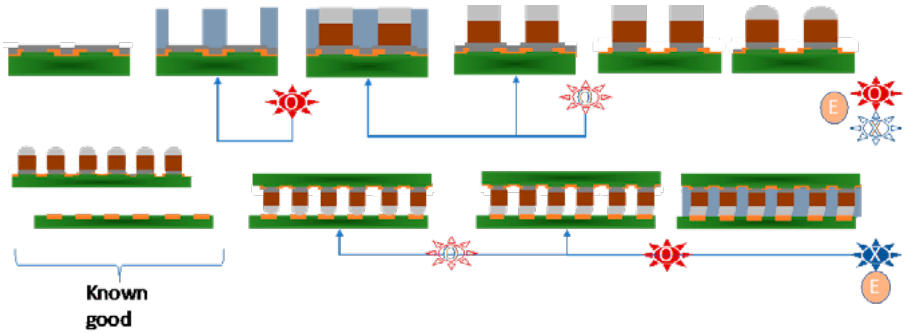


Figure 3: Simplified process flow for 100% inspection, optimized etest, and 100% post-attach inspection.

Other costs of quality

A more practical example for yield monitoring would be a case where a few months term run at a factory led to partial voiding in one of the device interfaces in a run of a few thousand devices. Electrically, these devices pass, and some will pass a stress test. After longer periods of time, however, these devices fail. Because it was unmonitored, and concentrated, a number of devices will need to be pulled, repaired, and a recall issued for items at risk. Even a 100ppm failure rate on average can lead to an expenditure of $20M or more, depending on the cost of repair or recall.

The costs of quality go beyond the simple calculation of recalls and returns. There are also development cost, reputation, product delays, returns (and expedition), and time delays on future programs. Quality costs are paid for a very long time, and the cost of quality can easily be as much as 3-5X the simply calculated estimates. Again, fast inspection would save both money on the device side, reduce qualification time, and reduce lab and other costs. Inspection at the 100% level on all, fully-assembled parts is critical.

Circling back to integration—the method of integration is adding many more interconnect pins, thereby reducing pitch. Figure 3 shows a simplified process flow in both the part level, and global metrology and inspection. There are multiple cases for electrical, optical, and X-ray testing. X-ray testing provides the ability to see through, and increasingly, to add resolution on critical features where optical resolution is insufficient. Each part is tested before final assembly, and the final assembly is inspected for variance.

Existing X-ray tools are capable of finding most of these defects, but the cost of ownership is extremely high. As previously discussed, it is not possible in complex heterogeneous assemblies to bring defectivity down to less than 100ppm with process control alone: 100% X-ray inspection is needed to find random defects, which are relatively common in assembled parts. Additionally, 100% X-ray inspection requires a very low cost per part: typically a few cents per part. This low cost can be achieved only by a very high-speed, high-resolution fully-automated X-ray inspection system—a system that is at least 100X faster than most X-ray tools available today with the required resolution. In the past, it has been impossible to achieve the required throughput (at least 3000mm^2/min) at the required resolution (less than three microns). A third requirement is that the X-ray inspection is safe for sensitive ICs, like CMOS and DRAM memory chips.

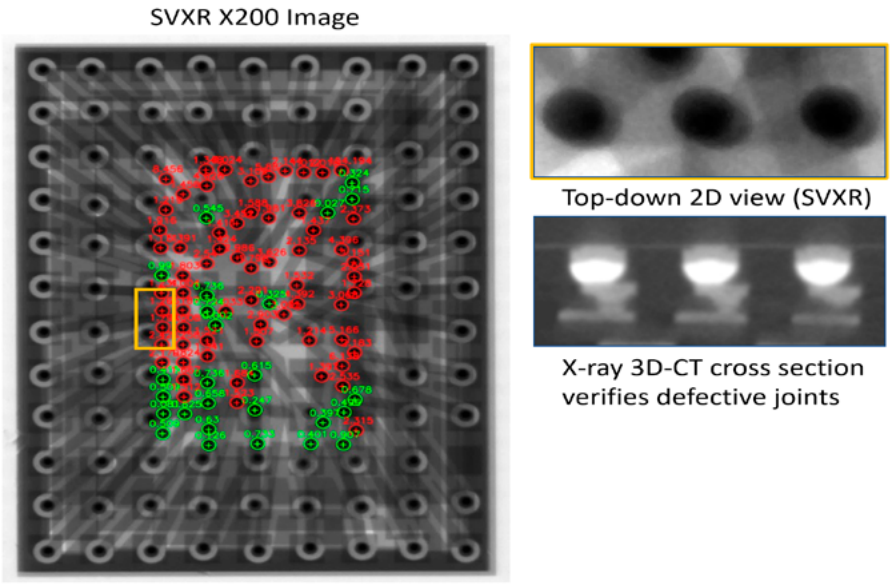


Figure 4: High-resolution X-ray image from a SVXR X200. Potential fails are indicated by the orange square. At in-fab inspection, these potential fails can be reviewed with higher resolution, slower X-ray techniques to finalize root cause.

Recently, a significant breakthrough in X-ray technology has allowed a 100x improvement in high-resolution imaging (S. Jewler, *Chip Scale Review*, Sept/Oct 2019). This new technology takes images with 2.5 micron resolution at 3000mm^2/min, and does not damage sensitive ICs (Figure 4). As such, it helps IC packaging manufacturers lower defect rates to well below 100PPM, while also improving reliability, at a low cost-of-ownership.

Summary

In the long-run, the costs for reliability are paid by the last company in the supply chain: either the automotive company, the cell-phone company, or the telecom company. By reducing the defectivity earlier in the supply "food chain," the costs of defects are drastically reduced. A new technology now exists that can reduce this defectivity earlier in the supply chain. It remains to be seen whether the end customer can push suppliers to implement this new technology. Currently, there is no real financial incentive for manufacturers at the beginning of the supply chain to improve their quality. They will not invest in technology for reliability on their own: they will only do this if their customers require it. New technology is expensive, and if the company reaping the benefits—the final customer—will not pay for it, then they will continue to pay 100X more for poor reliability.

Biography

David Adler is the CEO, President, and Board Chair of SVXR Inc., San Jose, CA. He co-founded SVXR in 2013 with the specific objective of developing an inline, automated X-ray inspection system to provide advanced process control and defect detection for 3D-IC packaging. Previously, he served as Chief Technologist at KLA. He received a PhD in Physics from Cornell U. and a BS from California Institute of Technology. Email dave@svxr.com

Brennan Peterson is a Director of Marketing at SVXR Inc., San Jose, CA. He has progressively increased responsibility and scope from experience in engineering, applications, and strategy from working at Intel, Thermo-Fisher/FEI, Nanometrics, ASML, and KMLabs.

AI's impact on 3D packaging: heterogeneous integration

By Santosh Kumar [Yole Développement, Korea]

Artificial intelligence (AI) has been in development for more than fifty years but recently it has emerged as one of the key drivers of semiconductor growth fueled by smartphones, personal assistants, social media and smart automotive. AI requires various computing hardware and high-end memories; and because of requirements of high bandwidth, low latency and low power consumption, AI has created opportunities for the advanced packaging business.

AI technology trends

AI is now widespread and has become an integral part of the technology industry. Whenever a machine mimics human cognitive function, we can say it is AI. In the AI field, some people begin to distinguish between the types of machine learning. Machine learning is the subset of AI that includes abstruse statistical techniques that enable machines to improve task performance with experience. The first goal of machine learning is to give the machine the ability to learn without being programmed. The next goal allows the machine to assess the data collected and make predictions. Besides academic research and military programs, there are machine learning flagship applications aimed at consumers. The most important applications are voice identification and language processing used for an intelligent personal assistant (e.g., Siri, Cortana, Alexa, etc.) and image recognition for autonomous driving.

There are several algorithmic approaches that enable enhancement and acceleration of machine learning—deep learning is one of them and it is gaining more and more interest. Deep learning is the subset of machine learning comprising algorithms that allow software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data. These new ways of processing heavy data, like video and photo, were made possible because of the availability of efficient data computing hardware, such as new large bandwidth memories, general processor units (GPUs), central processor units (CPUs), application specific integrated circuits (ASICs), and field-programmable gate arrays (FPGAs), etc.

Deep learning is made up of two phases: training and inference. What is called training in AI, is training a virtual machine to recognize objects and sounds. The training phase requires huge computing power and can be extremely long (hours, days, months) depending on the required precision. Currently, most of the training is done in the cloud where the computing capabilities are in-line with such operations. Nevertheless, some training can still be done on edge. An example would be for face detection systems on phones where once-off training of a couple of seconds is required to complete the neural network model to recognize the face of the phone's owner. Inference can't happen without training. Inference can occur on edge, and will give similar prediction accuracy, but simplified, compressed and optimized for runtime performance. Inference can also occur in the cloud. The act of using the trained neural network with new data on a device or server to identify something is known as inference. System on chips (SoCs) with GPUs and a CPU inside are used to do this computation on edge (on a phone for example). Inference requires less computational capabilities than training, as this was already performed in the cloud.

Hardware for AI

Training and inference have two different missions and that makes the hardware requirements different. Training requires intensive calculations, consequently large bandwidth, using CPU, GPU, FPGA and dynamic random access memories (DRAMs), and it is the first and main user of 3D interconnected devices. Inference (inference processing) workload looks like the processing of digital signal processing (DSP) algorithms. Inference can take place in two places, or in a datacenter, or locally (embedded inference), such as in a car. The requirements for inference include low latency, less expensive and much lower power consumption, especially when inference is embedded. Inference products could integrate an accelerator onto a SoC. Inference is typically conducted at the application or client endpoint (i.e., edge), rather than on the server or cloud. It requires fewer hardware resources, and depending on the application, can be performed using CPUs, FPGAs, ASICs, DSPs, etc. Inference is expected to shift locally to mobile devices. Here, precision can be sacrificed in favor of greater speed or less power consumption.

As mentioned before, the key computing hardware for training and interference of AI include CPUs, GPUs, FPGAs and ASICs. CPUs offer a great degree of programmability, however, they tend to provide less performance power than optimized and dedicated hardware chips. FPGAs are extremely flexible and have excellent performance, making them ideal for specialized applications that need a small volume of reprogrammable microchips. That said, FPGAs are quite difficult to create and expensive as well, not to mention that they still falter in terms of power and performance when compared to the likes of GPUs and ASICs. GPUs are ideal for graphics, as well as their underlying matrix operations and scientific algorithms, as they are super fast and flexible. With an ASIC, you get the best of all worlds as it is basically a customizable chip that can be designed to accomplish a very specific task at high power, efficiency, and performance. ASICs are now increasingly being developed for the purpose of supporting artificial intelligence AI and associated technologies. Google tensor processing units (TPUs) are a series of ASICs designed for machine learning, and optimized to run open source machine learning software. Baidu developed dedicated ASICs for its "Kunlun" AI accelerator for data centers.

High-bandwidth memory (HBM) is an ideal memory solution for AI training hardware. HBM2E is the latest version of HBM—its specification was announced by JEDEC in 2018 to support increased bandwidth and capacity. Samsung announced the industry's first HBM2E memory, "Flashbolt," in March 2019, which

boasts a 3.2Gbps data transfer speed per pin, which is 33% faster than the previous-generation HBM2. Flashbolt has a density of 16Gb/die, double the capacity of the previous generation. With these improvements, a single Samsung HBM2E package will offer a 410 gigabytes-per- second (GBps) data bandwidth and 16GB of memory. In August 2019, SK Hynix announced an HBM2E DRAM product with the industry's highest bandwidth. The new HBM2E boasts an approximately 50% higher bandwidth and 100% additional capacity compared to the previous HBM2. It supports an over 460 Gigabyte (GB)-per-second bandwidth based on the 3.6Gbps speed performance per pin with 1,024 data I/Os. By using through-silicon via (TSV) technology, a maximum of eight 16 gigabit chips are vertically stacked, forming a single, dense package of 16GB data capacity. Compared with traditional wire bond connections, DRAM chips that are stacked using TSV interconnects result in a shorter signal path and a high-speed performance with lower power consumption. While traditional structures package memory chips into a module form that can be connected to system boards, an HBM chip is packaged closely to logic chips, such as GPUs, leading to shorter distances between chips, which further accelerates data processing rates.

3D packaging for AI

Various 2.5D/3D packaging solutions, with or without TSV, are available today for packaging AI chips for inference and training. As shown in [1, Figure 1], TSV-based technologies include 2.5D Si interposer (e.g., chip-on-wafer-on-substrate [CoWoS]), 3D TSV stacking (e.g., for HBM stack), Foveros, 3D SoC, etc. TSV-less packaging technologies include embedded multi-die interconnect bridge (EMIB), integrated fan-out package on package (InFO-PoP), integrated fan-out on substrate (InFO-oS), fan-out chip-on-substrate (FOCoS), silicon wafer integrated fan-out SWIFT®, integrated thin-film high-density organic package (i-THOP®), RDL interposer, etc. The positioning of various 3D packaging technologies in terms of I/O and package size is shown in [1, Figure 4].

3D/2.5D TSV and heterogeneous integration technologies have emerged as the choice technology for AI, and particularly for deep learning applications, as they provide higher bandwidth, low latency, and low power consumption. For AI accelerators, it is important to keep logic and high-capacity memory as close as possible to provide low latency and lower power. When two chips or more are integrated on an interposer, the distance between logic and memory is shortened, which enables lower latency and lower power consumption. DRAM, based on a 3D TSV solution, offers an unequaled bandwidth performance because of the ability of the TSV solution to connect several layers of the device.

3D IC and 2.5D solutions based on interposer are not stand-alone products—they need to be integrated onto a final package to be functional. The interposer acts as an intermediate layer between the dies; it is the solution that has enabled die partitioning. Si interposers provide a high-density routing connection between logic and memory.

Interposers can be made out of silicon, glass and organic laminate materials. Only silicon is used in commercial applications, but glass and laminate are under development. However, silicon interposers are the only solution ≤1µm for routing today. Silicon interposer is considered costly, so many alternative technologies from outsourced semiconductor assembly and test suppliers (OSATS) and integrated device manufacturers (IDMs) are in development, and some are already in production, e.g., EMIB from Intel, InFOoS from TSMC, FOCoS from ASE, etc. Some of the recent developments in the 3D stacking technologies are listed in the sections below.

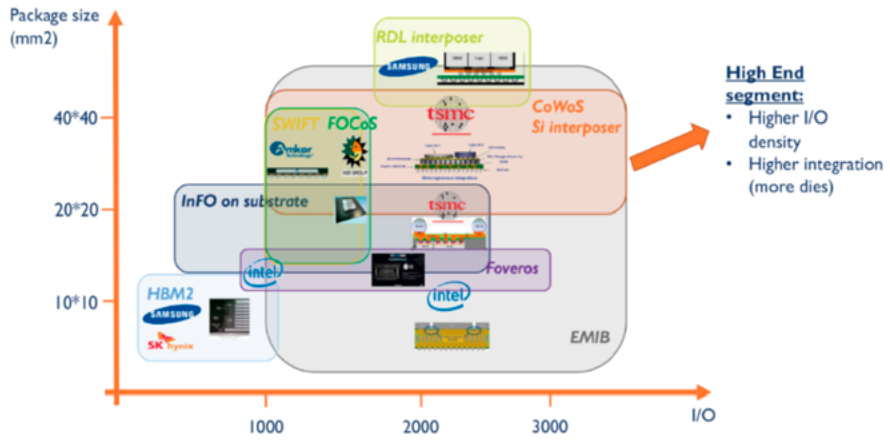


Figure 1: Suppliers involved in 3D packaging technologies. SOURCE: [2]

EMIB from Intel. EMIB technology is a replacement for TSV interposer. It consists of a silicon bridge integrated into the package substrate to ensure the connections between two or more different dies. Intel already used this technology in its Stratix 10 MX & Core I7-8809G 8th Gen. The advantages of EMIB technology as given by Intel include: 1) similar density die-package and die-die interconnections as for interposer; 2) no more Si interposer and therefore lower cost (instead, there is the need to etch a cavity into the substrate and to manufacture the bridge and attach it to the substrate using an adhesive). This silicon bridge technology can connect different dies together. For the moment, on the I7-8809G core, Intel only used it to interconnect the GPU to HBM2. Why not use it to interconnect the GPU to Intel's CPU?

3D SoIC® from TSMC. Last year TSMC announced its system-on-integrated chip (SoIC) advanced packaging technology that is slated to go into mass production in 2021. SoIC is a type of innovative multi-chip stacking technology, which can be used to carry out wafer bonding in the manufacture of chips that are 10nm and less. SoIC is a "bumpless" interconnect method between multiple die. The SoIC solution enables known good dies of different sizes, process technologies, and materials to be directly stacked together. Compared to typical 3DIC solutions with microbumps, TSMC's SoIC delivers higher bump density and speed, while consuming much less power. What's more, SoIC is a "front-end" integration solution connecting two or more dies before they are packaged. Therefore, a SoIC stack can be further integrated with other SoIC or chips in one of TSMC's "back-end" advanced packaging technologies such as InFO or CoWoS, offering a powerful "3D-by-3D" system-level solution.

RDL interposer from Samsung. Samsung is working on RDL interposer to replace Si interposer. RDL interposer advantages include: 1) It's based on standard deposition and lithography processes (passivation and metal layers deposition and patterning), which results in low cost manufacturability, higher memory density > 4 HBM per package (6 to 8); 2) It's a good platform for high-density applications, better thermal performance and higher bandwidth using a side-by-side package vs. a package-on-package (PoP); 3) It's scalable to panel-level production; 4) It has less of a coefficient of thermal expansion (CTE) mismatch between the RDL and the printed circuit board (PCB) vs. a Si interposer and PCB, and therefore better warpage control

Fan-out based 3D packaging from TSMC. TSMC recently introduced an advanced 3D Multi-stack (MUST) system integration technology, and a 3D MUST-in-MUST (3D-MiM) fan-out package has been developed as a next-generation wafer-level fan-out package technology. 3D-MiM technology utilizes a more simplified architecture that eliminates ball grid arrays (BGAs) between packages for system-level performance, power and form factor (PPA) purposes. This technology also makes use of a modularized approach for both design and integration flow to improve design flexibility and integration efficiency. Known-good pre-stacked memory cube and/or logic-memory cubes are fabricated by leveraging the established integrated fan-out technology platform (InFO) in tools, materials, design rules, and processes to shorten development cycle time and achieve cost effectiveness. Near memory processing is attractive for high bandwidth, low latency, and power saving. 3D-MiM was proposed to be an alternative heterogeneous integration solution to FC-PoP for mobile and 3DIC stacking for high-performance computing (HPC) applications, to realize in-package near-memory computing. For mobile applications, in comparison to the FC-PoP, the 3D-MiM fan-out offers a thinner package profile (~0.5mm z-height), higher data bandwidth (2X~4X), with lower latency (0.2X) and thermal resistance to meet demands of future 5G/AI-driven edge computing. For AI, in comparison to 3DIC HBM, the 3D-MiM fan-out offers a lower cost alternative with a new memory-to-memory, and SoC-to-memory integration architecture that has promising electrical and thermal performances. With respect to manufacturing, 3D-MiM fan-out technology leverages the well-established infrastructure of wafer-level system integration (WLSI) in capacity and materials, tools, processes, design rules for yield, and competitive cost.

Foveros. Intel recently introduced the "Foveros" 3D packaging technology. Foveros is Intel's new active interposer technology designed as a step above its own EMIB designs for small form factor implementations, or those with extreme memory bandwidth requirements. For these designs, the power per bit of data transferred is super low, however the packaging technology has to deal with the decreased bump pitch, the increased bump density, and also the chip stacking technology. The interposer contains the TSVs and traces required to bring power and data to the chips on top, but the interposer also carries the plated contact hole (PCH) or I/O of the platform. It is, in effect, a fully working PCH, but with vias to allow chips to be connected on top.

3D stacking technology: who's winning

The level of complexity of TSV manufacturing and integration has made it difficult for this technology to penetrate the market. In particular, it took time to set up the supply chain. One of the challenges is that there are several wafer/die donors (processing die, memory cube and interposer). They may come from two to three different suppliers. To achieve this, semiconductor players have aligned their strengths, which has changed the usual packaging supply and value chains. 3D integration has to be taken into account at the beginning of the package and chip

designs, which gives an important role and opportunity for design and intellectual property (IP) houses. In reusing IP blocks, time can be saved on product development time. If we consider the integration of a memory cube and a processing unit, between 5 to 7 companies are involved on average, which adds to cost and complexity. That's why the supply chain is considered from the beginning of product development. **Figure 1** shows the suppliers involved in 3D packaging technologies.

Four different types of players want to claim a share of the growing $5.5 billion stacking market. "Foundries, integrated device manufacturers (IDMs), OSATS and intellectual property (IP) companies all want their share of the stacking business. Foundries, like TSMC, UMC, and GLOBALFOUNDRIES, dominate the TSV heterogeneous stacking technology market because of their ability to produce the interposer in-house. Intel is the only IDM that is trying to compete in this sector with its "Foveros" technology. In the 3D stacked-memory market, the battle is between the big 3 IDMs: Samsung, SK Hynix, and Micron. These companies will continue to reign supreme in the stacked-memory market. 3D SoC is a foundry technology. Most probably one foundry will manufacture the technology to ensure high yield and limit risks. TSMC is leading the time-to-market race ahead of GLOBALFOUNDRIES. For the TSV-less technologies, the game is a bit more intense between foundries, IDMs, OSATS and substrate makers. Samsung, Intel and TSMC are into both with- and without-TSV technology development. OSATS like ASE have introduced their FOCoS technology to the market in 2016. Other players like Amkor developed their proprietary technology but are still waiting for orders. The substrate companies, like Shinko, Unimicron, and lately, Fujitsu Interconnect, are still in R&D. Xperi will have a positive impact on the market as its hybrid bonding technology is midway between technologies with- and without-TSV. The other advantage of Xperi is that its technology is compatible with both high and mid/low-end segments. Foundries, IDMs, and IP companies have the advantage over OSATS in stacking technology, where the latter encounter difficulties in getting orders.

References

1. F. Shoo, "Stacking technologies: the road ahead," *Chip Scale Review*, Nov/Dec 2019, pp.: 4-8.
2. "2.5D/3D TSV & Wafer Level Integration Technology & Market updates report," Yole Développement, 2020.

Biography

Santosh Kumar is a Principal Analyst and Director Packaging, Assembly & Substrates, Yole Développement, Korea. He is involved in the market, technology and strategic analyses of the microelectronic assembly and packaging technologies. He is the author of several reports on fan-out/fan-in WLP, flip-chip, and 3D/2.5D packaging and received Bachelor's and Master's degrees in Engineering from the Indian Institute of Technology (IIT), Roorkee and U. of Seoul, respectively. Email Santosh.kumar@yole.fr

Demand for 3D devices drives inspection and metrology innovation

CSR asked Raj Jammy, President of the Process Control Solutions Business Unit at Carl Zeiss SMT group and President of Carl Zeiss SMT Inc., the North America subsidiary of Carl Zeiss SMT GmbH, to provide an update on how demand for 3D devices is driving inspection and metrology innovation.

CSR: 3D devices drive a good deal of semiconductor technologies today and in turn, such devices need to have 3D integrated information when it comes to 3D inspection and metrology. Why is imaging alone not sufficient for these devices?

RJ: In the pursuit of root cause of failures and process defects, structural information tells only part of the story. Something may look structurally sound, but be electrically or chemically defective. For example: often one wants to do energy dispersive X-ray spectroscopy (EDS) analysis of a focused ion beam (FIB) cross section before final transmission electron microscopy (TEM) prep and electron energy loss spectroscopy (EELS) analysis. It can be important to understand if any oxidation present at a defect site occurred before exposure to atmosphere, or is a side effect from exposure during the transfer from FIB to TEM. 3D device architectures, indeed, pose significant challenges for all three types of analyses, because of shrinking features that are buried beneath the surface and often not accessible for testing.

CSR: With respect to shrinking features and imaging, how is cross-section accuracy impacted?

RJ: Traditional package cross-section approaches have an accuracy in the range of about 15 microns, and it can take two days or longer to prepare a high-quality advanced-package cross section at exactly the right location with minimal artifact. As package interconnects continue to shrink, the ability to cross section to exactly the right location becomes less successful. We are dealing with technologies today in packaging that require nanometer resolution for micron-sized objects located within millimeters of volume. Just as shrinking semiconductor dimensions in the 1990s drove the transition from scanning electron microscopy (SEM) imaging to FIB prep and TEM imaging, today we see the cycle repeat with package fine pitches driving a transition from traditional cross-sectional methods to new cross-section approaches with the speed and accuracy for package interconnects at 50 micron and finer pitches.

CSR: What other considerations come into play with respect to cross-sectional methods for advanced packages?

RJ: Generally, advanced packages are used for the advanced Si nodes that use fragile dielectric materials in the back end of line (BEOL) Si interconnects. To achieve fine pitches, Cu-pillar solder bumps are being adopted for advanced package interconnects. Cu is a hard material, and the combination of low-modulus brittle materials and hard materials can lead to cracks and delaminations, whether from handling steps of sample preparation or from reliability failures caused by chip-

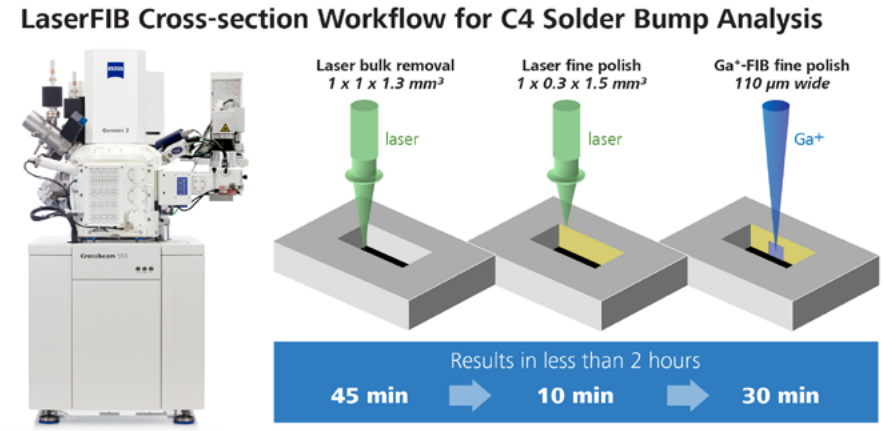


Figure 1: Crossbeam Laser FIB-SEM throughput for each step of the workflow for a customer's application.

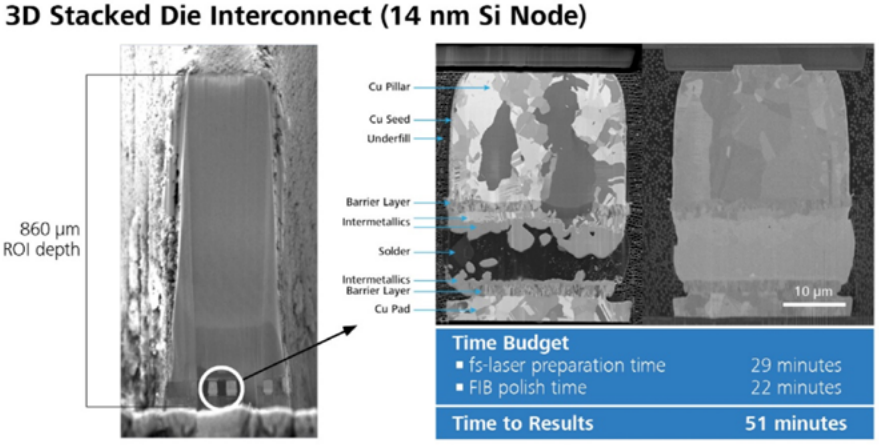


Figure 2: Crossbeam Laser FIB-SEM provides fast, high-quality cross sections of Cu-pillar microbumps buried 760µm deep with total time to results of <1 hour.

package interactions. Artifact-induced cracks from traditional cross sections look exactly like the type of real defects observed from reliability failures. For this reason, while it is always challenging to make an artifact-free cross section, it is even more so with advanced packages.

CSR: What kinds of innovations or discoveries led to the combination of hardware in a single instrument that is vital to the Crossbeam laser system?

RJ: We have implemented a patented architecture that maintains the laser ablation in a chamber that is integrated into the FIB-SEM, yet segregated from the FIB-SEM chamber. The large volumes of ablated material tend to coat surfaces. The ZEISS architecture maintains the pristine condition of the imaging chamber for high-quality imaging at highest SEM resolutions. The sample can be shuttled between the ablation and imaging chambers without exposure to atmosphere, preventing oxidation of defects prior to their analysis, and the entire workflow (**Figure 1**) can be completed rapidly because it exists within a single instrument with no requirement to manage different instrument queues or repeat the sample set-up, as would be required with a non-integrated solution. Some people think a plasma focused ion beam PFIB integration with a laser would be a better choice for processing large volumes. However, the purpose of the FIB in the workflow is for fine polishing, and the reality is that a Ga+ beam has significantly higher current densities than a PFIB at the typical low currents used for fine polishing. The fs-laser speed of large-volume material removal, combined with the Ga+ beam resolution and high current density for fast fine polishing with highest process accuracy, results in an effective solution that provides the fastest workflow for site-specific cross sections of deeply buried structures (**Figure 2**). The integrated system is versatile as it can be applied to all the traditional Ga+ FIB applications that require high resolution and accuracy for semiconductor analysis, such as TEM sample preparation.

Biography

Raj Jammy is President of the Process Control Solutions business unit at Carl Zeiss Semiconductor Manufacturing Technologies group and President of Carl Zeiss SMT Inc., the North America subsidiary of Carl Zeiss SMT GmbH. With 25 years in the semiconductor industry, Dr. Jammy has previously held executive leadership positions at IBM, SEMATECH and Intermolecular. He received a doctorate in Electrical Engineering from Northwestern U. Email raj.jammy@zeiss.com

Rapid analysis of buried 2.5/3D package structures

By Cheryl Hartfield, Marcus Kaestner, Sascha Mueller, Juan Atkinson-Mora, Ingo Schulmeyer
[ZEISS Semiconductor Manufacturing Technology, Process Control Solutions]

Semiconductor packaging plays a key role in the relentless pursuit of better electronic system performance. Diverse package technologies and strategies are advancing next-generation products for smart systems and a connected world, and package engineers have many options for designing the fullest functionality into the smallest footprints for system-in-package (SiP) and system-on-chip (SoC) packages. Recent work shows chiplets can be connected to an active interposer using 150,000 microbumps at 20µm pitch in a 40mm x 40mm 3D package [1]. Additionally, package interconnect dimensions are crossing over into the space dominated by the silicon back end of line (BEOL) dimensions. Hybrid bonding produces some of the smallest package interconnects. Submicron pitches have been demonstrated, as well as the bonding of 300mm wafers with submicron accuracies [2]. Recent announcements by multiple companies show hybrid bonding is now spreading beyond CMOS imaging sensors (CIS) into dynamic random access memory (DRAM), 2.5D logic, and SoC, with pick and place accuracy requirements ranging from 3-5µm down to 250nm, depending on application [3].

The advances noted above present challenges for package fault isolation, process characterization, and failure analysis (FA). Design for test (DfT) and FA strategies need to be paired with new analysis tools to enable fast development of reliable processes and packages. To achieve rapid analysis of buried fine-pitch package and silicon interconnects, a new approach for high-resolution cross-sectional imaging of structures in 3D packages has been developed. It leverages a femtosecond (fs) laser integrated with a focused ion beam scanning electron microscope (FIB-SEM). This article details the capability of this new system for microbump analysis of a 3D stacked-die package.

Package analysis with a laser-integrated FIB-SEM

Some of the most challenging devices for characterization are those used in high-performance computing and artificial intelligence, where packages can be 80mm in diameter or larger, and the package interconnect pitches are 40µm and driving smaller [4]. Reconstructed 3D X-ray microscope (XRM) images show the complexity of fine-pitch interconnects in these devices, which becomes evident at successively higher resolutions (**Figure 1**).

The XRM images are from a delidded, but otherwise fully intact, 55mm x

55mm 2.5D package used in an artificial intelligence application. The inset image showing 25µm-diameter microbumps was acquired using 1.8µm/voxel. 3D XRM has become standard in FA labs because of its ability to image fully intact packages with high spatial resolution [5], and state-of-the-art 3D XRM has a spatial resolution of 500nm with voxel sizes of 40nm [6]. The XRM images are helpful to guide subsequent cross-sectional scanning electron microscope (SEM) analysis.

Traditional mechanical cross-section techniques are under pressure to deliver artifact-free results at high throughput [7]. Focused ion beam (FIB) processing, while having adequate accuracy and quality for the finest-pitch interconnects, lacks efficiency for removing large volumes of packaging material to analyze buried features. To address these deficiencies, a new instrument, ZEISS Crossbeam laser, was recently developed. It extends the nanoscale imaging and process accuracy of FIB-SEM to packages by enabling site-specific removal of large volumes of packaging material. It includes a fs-laser attached to the external load lock of a gallium ion (Ga^+) FIB-SEM, delivering an improved workflow for site-specific cross-sectional imaging. Integration of a fs-laser and Ga^+ FIB into a single system ensures a streamlined "cut and look" workflow for fastest time to results, as well as a pristine sample that is not oxidized by exposure to atmosphere, thereby enabling accurate analysis. The fs-laser interaction is essentially athermal [8], producing a laser affected zone (LAZ) smaller than 1µm under optimized processing conditions. This aids fast Ga^+ FIB polishing times, and rapid results are further enabled by a streamlined single-instrument queue, rather than managing two queues of separate tools.

The laser-integrated FIB-SEM (laserFIB) represents a new class of FIB-SEM, optimized for imaging targeted features at nanoscale resolutions within SiP and 2.5/3D packages. **Table 1** shows it is well-suited for removing cubic millimeters of material, unlike the Xe^+ plasma FIB (PFIB). Using parameters for high-quality laser-processed surfaces, it takes four minutes for the fs-laser to remove a half cubic millimeter of silicon, compared to two days for a PFIB or 15 days for a Ga^+ FIB at published milling rates [9].

The time-consuming conventional cross-section steps of epoxy embedding and mechanical polishing are not used in the laserFIB workflow (**Figure 2**), and if downsizing of larger samples is required, it can be done in areas far away from the desired target location. This reduces the risk of preparation artifacts, even in high-stress packages containing advanced-node silicon die with ultra-lowK dielectrics.

The sloped walls produced by the fs-laser ensure the Ga^+ beam has a short milling path in the z-dimension of this wedge-shaped edge, enabling efficient local Ga^+ FIB polishing and high-resolution imaging across areas of 100µm to 500µm wide and equal or greater depths. High imaging quality and low maintenance is ensured by segregating the laser from the FIB-SEM chamber to avoid contaminating the columns and detectors with ablated and recast material. Efficient transfer between chambers enables the repeated cycles of laserFIB processing and imaging that may be required for new recipe set-up or for analyzing multiple sites in a sample.

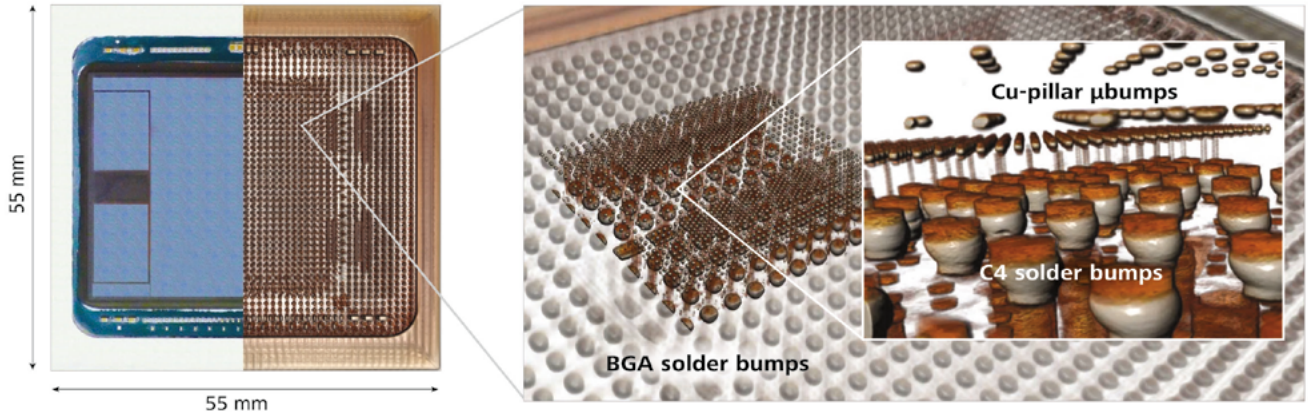


Figure 1: 3D XRM images from three scans of a 55mm x 55mm 2.5D package, showing multiple levels of interconnect, the smallest being 25µm-diameter Cu-pillar microbumps.

Si Removal Rate Comparison

Technology	fs-laser (515 nm)*	Xe^+ PFIB**	Ga^+ FIB**
Si Removal Rate	5.4 x 10^5 µm³/sec	6.7 x 10^2 µm³/sec	1.0 x 10^2 µm³/sec
Volume Processed	Calculated Process Time		
0.10 mm³	2 sec	25 min	2.8 hr
0.25 mm³	29 sec	6.5 hr	1.8 days
0.50 mm³	4 min	2 days	15 days
0.75 mm³	13 min	7 days	49 days
1.0 mm³	30 min	17 days	116 days

* Ablation rate used for a high-quality finish
** FIB milling rates based on [15]

Table 1: Technology timing comparisons for removal of up to one cubic millimeter of silicon.

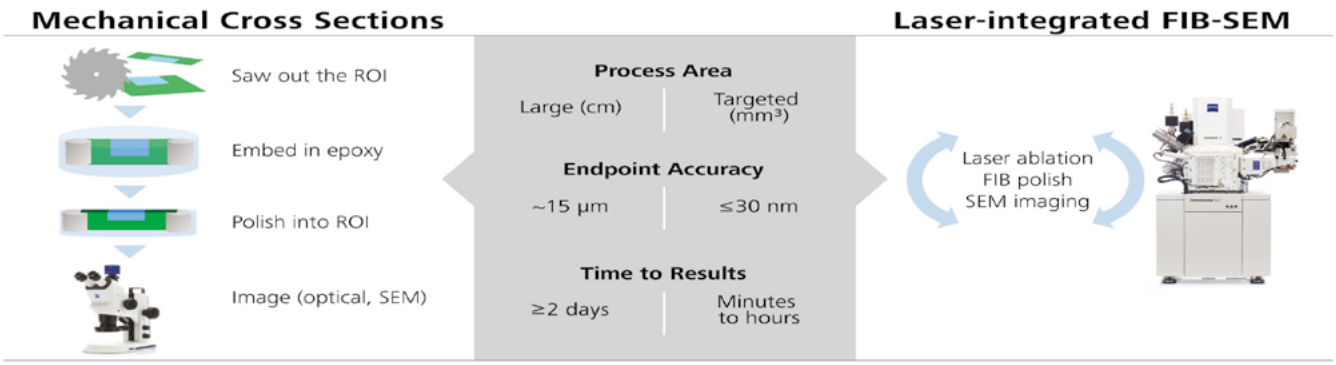


Figure 2: Cross-section workflow comparisons.

Cu-pillar microbumps in a 3D package

A laserFIB workflow for imaging 25µm-diameter Cu-pillar microbumps buried almost one millimeter deep in a 3D stacked-die package is detailed in **Figure 3a-h**: a) registration of 3D XRM virtual cross-sections to SEM images via the system's software for precise laser pattern placement [10]; b) bulk laser ablation of a large region; c) fine laser polishing; d-e) imaging after the laser polish; f) Ga^+ FIB polishing of a 300µm-wide local area; g) imaging 25µm diameter microbumps using high-resolution secondary-electrons; and h) imaging the same using backscattered electrons. The laser quality allows imaging of some structures right after laser polishing. Ga^+ beam polishing provides the highest-quality surfaces for imaging, removing the shallow LAZ and laser-induced periodic surface structures (LIPSS) to enable imaging intermetallic compounds and other features.

The combined speed and accuracy of the laserFIB has been demonstrated on a 3D package test vehicle for 14nm node silicon technology. The test vehicle has 50µm pitch Cu-pillar microbumps sandwiched beneath a 725µm thick top die and a 50µm thick bottom die. Using 1µm voxel resolution, 3D XRM scans were done at a region of interest. **Figure 4** shows a SEM image of the cross-sectioned microbump void superimposed upon the virtual plan-view XRM slice that guided the laser cuts. The void is indicated by a red arrow. The microbump containing the void is circled in the XRM image. The large-area ablation volume was 0.9 x 0.9 x 1.0mm^3, and the cumulative laserFIB process time to cross-section the void in the microbump was less than one hour, following the workflow described in **Figure 3**. Both large-area laser ablation and laser fine polishing were completed in 29 minutes. The Ga^+ FIB polishing was done over a 110µm-wide area that included the affected microbump and progressed until reaching the void. This step took 22 minutes. Therefore, the 5µm void

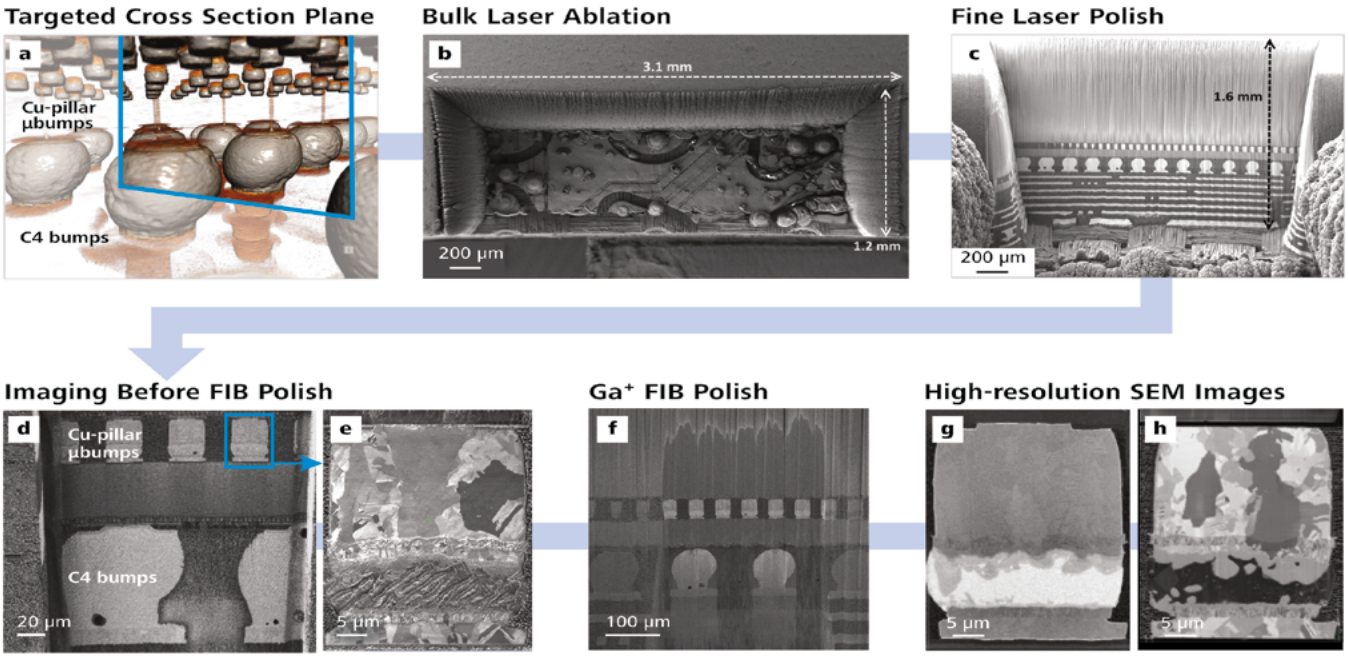


Figure 3: Steps for correlated 3D XRM and laserFIB analysis of deeply buried microbumps in a 3D package.

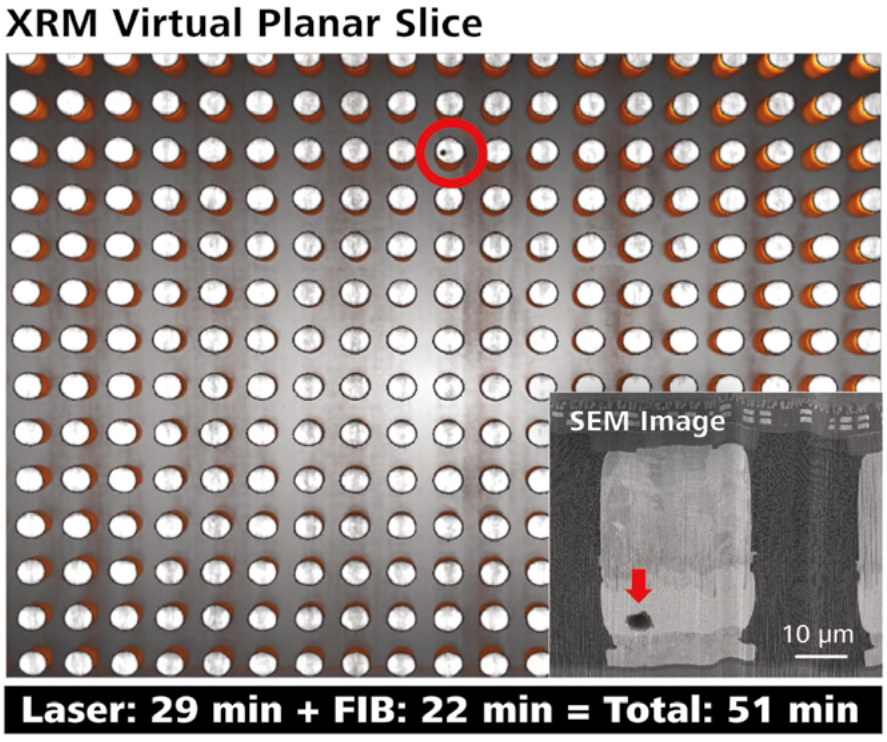


Figure 4: A 5µm void (red arrow) found in a single Cu-pillar microbump (red circle) was precisely cross sectioned in less than one hour.

in the 3D package was accurately cross-sectioned with high quality and no artifacts in less than one hour.

Summary

A laserFIB was used to produce, within one hour, a high-quality cross section of a targeted ~5µm void within a 25µm-diameter microbump buried almost 1mm deep in a 3D package. In comparison, high-quality mechanical cross sections made through a row of 100µm-diameter C4 bumps can take more than 2 days [7], and smaller structures like microbumps further reduce throughput and success rates. This work demonstrates that a laser-integrated FIB-SEM enables faster package characterization and FA by enabling rapid access to deeply buried interconnects and interfaces in 2.5/3D packages. The integration strategy enables a streamlined workflow capable of meeting the throughput and success requirements of advanced packages.

References

1. P. Coudrain, et al., "Active interposer technology for chiplet-based advanced 3D system architectures," 2019 IEEE 69th Elec. Comp. and Tech. Conf. (ECTC), 2019.
2. M. Fujino, K. Takahashi, Y. Araga, K. Kikuchi, "300mm wafer-level hybrid bonding for Cu/interlayer dielectric bonding in vacuum," Japanese Jour. of Applied Physics, vol. 59, no. SB, 2020, doi: 10.7567/1347-4065/ab4b2b.
3. L. Mirkarimi, "The proliferation of hybrid bonding for 3D integrated circuits," presented at the 3D & Systems Summit, Dresden, 2020.
4. J. Ryckaert, E. Beyne, "A 3D technology toolbox in support of system-technology co-optimization, *imec Magazine*, 2019.
5. C. Hartfield, C. Schmidt, A. Gu, S. T. Kelly, "From PCB to BEOL: 3D X-ray microscopy for advanced semiconductor packaging," Inter. Physics of Failure Analysis, Marina Bay Sands Conv. Center, Singapore, 2018.
6. ZEISS Xradia 610 and 620 Versa Product Information, version 1.1. (2019).
7. P. S. Pichumani, F. Khatkhatay, "Mechanical milling and polishing of cross sections using a micro CNC machine for failure analysis," Electronic Device Failure Analysis (EDFA), 14-19 (2020).
8. K. Sugioka, Y. Cheng, "Ultrafast lasers—reliable tools for advanced materials processing," Light: Science & Applications, vol. 3, no. 4, p. e149, 2014, doi: 10.1038/lsa.2014.30.
9. S. J. Randolph, J. Filevich, A. Botman, R. Gannon, C. Rue, M. Straw, "In situ femtosecond pulse laser ablation for large volume 3D analysis in scanning electron microscope systems," Jour. of Vacuum Science & Tech. B, vol. 36, no. 6, 2018, doi: 10.1116/1.5047806.
10. J. Favata, V. Ray, S. Shahbazmohamadi, "Correlative 3D X-ray, laser ablation, and SEM/EDS mapping establishing access point for FIB tomography of defects in multi-layer ceramic capacitors," Microscopy and Microanalysis, vol. 25, no. S2, pp. 344-345, 2019, doi: 10.1017/s1431927619002459.

Biographies

Cheryl Hartfield is Solutions Manager, ZEISS Semiconductor Manufacturing Technology, Process Control Solutions, Pleasanton, CA. She received her MA and BS in Microbiology from UT Southwestern Medical Center, and Texas A&M, respectively. She co-founded Omniprobe and was Senior Member of Technical Staff at Texas Instruments, working for 12 years in characterization and package development. Cheryl is an ASM Fellow and Past President of EDFAS. She has >13 patents and authored >70 papers. Email cheryl.hartfield@zeiss.com

Marcus Kaestner is Project Manager, ZEISS Semiconductor Manufacturing Technology, Process Control Solutions, Oberkochen, Germany. He received his PhD and MS in Micro- and Nanotechnologies from Technical U. of Ilmenau.

Scaling CMOS beyond FinFETs: from nanosheets and forksheets to CFETs

By Julien Ryckaert [imec]

Get ready! New transistor architectures below 3nm will impact interconnect structures in a chip's BEOL. CSR asked imec for input.

The FinFET transistor architecture is the workhorse of today's semiconductor industry. But as scaling continues, undesired short-channel effects require the introduction of new transistor architectures. In this article, an evolutionary path towards 2nm and beyond technology nodes is presented [1]. This evolutionary path comprises the nanosheet transistor, the forksheet device, and the complementary FET (CFET). Part of these insights have been presented at the 2019 IEEE International Electron Devices Meeting (IEDM) [2,3]. As with so many scaling transitions, changes in device architectures will impact the back-end-of-line (BEOL), and these are also discussed.

The FinFET: today's leading-edge transistor

At every new technology generation, chipmakers have been able to scale transistor specifications by 0.7x, delivering a 15% performance boost, a 50% area gain, a 40% power reduction and a 35% cost decrease at device level. Several years ago, the industry made the transition from "good old" planar metal-oxide-semiconductor field-effect transistor (MOSFET) to FinFET transistor architectures in order to maintain this scaling path. In a FinFET, the channel between source and drain terminals is in the form of a fin. The gate wraps around this 3D channel, providing control from three sides of the channel. This multi-gate structure could eliminate short-channel effects, which started to degrade the transistor's performance at reduced gate lengths. Superior short-channel control is crucial because it sets the foundations of device scaling-–allowing shorter channel lengths and lower operating voltages.

In 2012, the first commercial 22nm FinFETs were introduced. Since then, FinFET architectures were improved for enhanced performance and reduced area. For example, the 3D nature of the FinFET allowed an increased fin height to obtain a higher device drive current at the same footprint. Today, industry is ramping up production of 10nm/7nm chips with FinFETs "inside." At the cell level of the most advanced nodes, standard cells with a track height of 6T (which is a measure of the cell area) feature down to 2 fins per device.

Vertically-stacked nanosheets: an evolutionary step

As scaling is pushed beyond 5nm, the FinFET is expected to run out of steam. At reduced gate length, the FinFET structure in turn fails to provide enough electrostatic control. On top of that, the evolution to lower track height standard cells requires a transition to single-fin devices, which cannot provide enough drive current – even if fin height is further increased.

With changing technology nodes, however, the semiconductor industry is not eager to switch to other transistor architectures. Some companies might even decide to stay at certain nodes longer. But still, there are applications – such as machine learning, big data analysis and data center servers – that will require the latest "universal" CMOS solutions. With such a universal CMOS solution, one and the same transistor architecture in one and the same technology node can be used to perform all functionalities on the chip.

Vertically-stacked nanosheet transistors can come to the rescue of the challenges described above (**Figure 1**). They can be considered a natural evolution of the FinFET device. Just imagine placing a FinFET on its side and dividing it into separate horizontal sheets, which make up the channels. A gate now fully wraps around the channel. This gate-all-around nature of the nanosheet provides superior channel control compared to the multi-gate FinFET. At the same time, the more optimal distribution of the channel cross-section in the 3D volume optimizes the effective drive per footprint.

The need for scaling boosters

The migration to nanosheet devices becomes optimal at low cell track heights of 6T and 5T, where fin depopulation would degrade drive current in traditional FinFET-based cells. But reducing track heights (and hence, cell area) from 6T to 5T cannot happen without introducing structural scaling boosters such as buried power rails and wrap-around contacts. Power rails provide power to the different components of the chip and are traditionally implemented as metal lines in the chip's back-end-of-line (BEOL) (i.e., the M_{int} and M1 layers). There, however, they occupy considerable space. In a buried power rail construct, the power rails are buried in

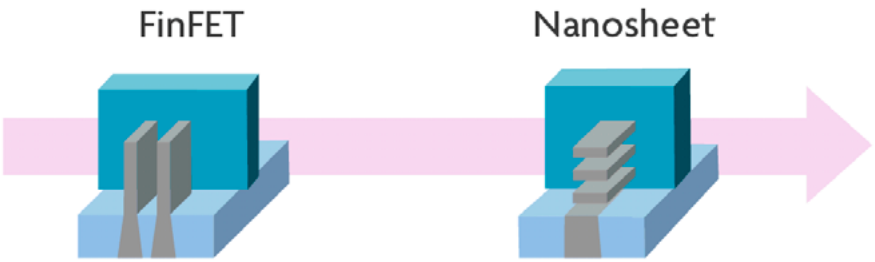


Figure 1: Natural evolution from FinFET to nanosheet.

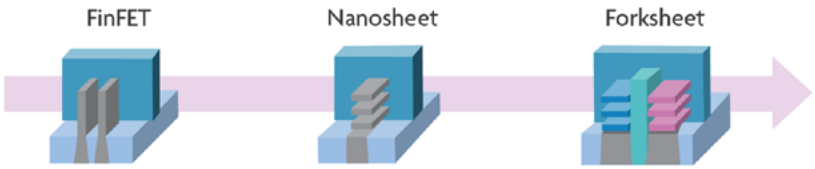


Figure 2: Natural evolution from FinFET to nanosheet, and to forksheet.

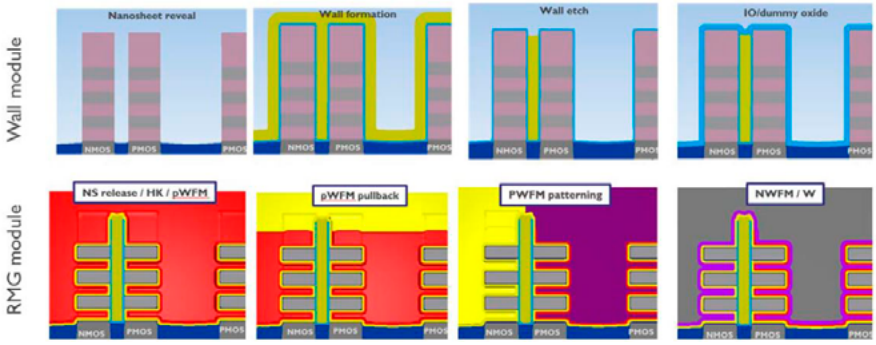


Figure 3: Key steps in the forksheet process flow, showing the wall formation step after active patterning.

the chip's front-end-of-line to help free up routing resources for the interconnects. Moreover, they provide a lower resistive local distribution of the current to a technology that suffers from increasing BEOL resistance with pitch scaling. By removing the power rails from the BEOL, the standard cell height can be further reduced from 6T to 5T.

The next step: reducing the spacing between p and n

As the journey to smaller track heights continues, a further reduction of cell height will require a much tighter spacing between nFET and pFET devices within the cell. However, for both FinFET and nanosheet devices, process limitations pose a limit to how close these n and p devices can be brought together. In FinFET architectures, for example, two dummy fin spacings are typically required between n and p devices, consuming up to 40-50% of the total available space.

To extend the scalability of these devices, imec has recently proposed an innovative alternative architecture, called the forksheet device (Figure 2). The forksheet can be considered a natural extension of the nanosheet device. Contrary to the nanosheet device, the sheets are now controlled by a forked gate structure, realized by introducing a dielectric wall in between the p- and nMOS devices before gate patterning. This wall physically isolates the p-gate trench from the n-gate trench, allowing a much tighter n-to-p spacing.

The process flow used for making the forksheet devices is similar as the one for making nanosheet devices, with only a few additional process steps (Figure 3). The dielectric isolation between n and p even holds a few process advantages, including, for example, a more simplified process for filling the work function metal. On top of this process window enhancement, the forksheet is expected to have superior area and performance scalability due to the large reduction in n-to-p separation.

Forksheet device: improved performance and area

Researchers at imec have recently used technology computer-aided design (TCAD) simulations to quantify the expected power-performance-area (PPA) potential of the forksheet device architecture. The device under study targets imec's 2nm technology node, using a contacted gate pitch of 42nm and a 5T standard cell library with a metal pitch of 16nm. The proposed design includes scaling boosters such as buried power rails and wrap around contacts.

Compared to a nanosheet device, a 10% speed gain (at constant power) and a 24% power reduction (at constant

speed) was calculated. This performance boost can be partly explained by a reduced (parasitic) Miller capacitance, resulting from a smaller gate-drain overlap. The available space can also be used to increase the sheet width, and as such, enhance the drive current. Finally, the n-to-p separation reduction can be exploited for shrinking the track height from 5T to 4.3T – resulting in a 20% cell area reduction. When implemented in a static random access memory (SRAM) design, the simulations reveal a combined cell area scaling and performance increase of 30%, for 8nm p-n spacing (Figure 4).

The forksheet can be considered a next step in the natural evolution from planar to FinFET and on to vertically-stacked nanosheets. The above characteristics demonstrate its potential as an ultimate logic "universal" CMOS device for the 2nm technology node. In further research, the process challenges to fully bring these devices into manufacturing need to be resolved.

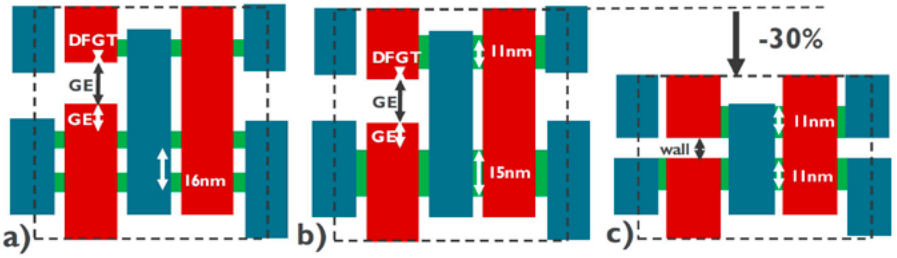


Figure 4: Layout of SRAM half cells for: a) FinFET; b) gate-all-around nanosheet; and c) forksheet. The forksheet can provide up to 30% scaling of the bit cell height as the p-n space is not governed by gate extension (GE), gate cut (GE) or dummy fin gate tuck (DFGT).

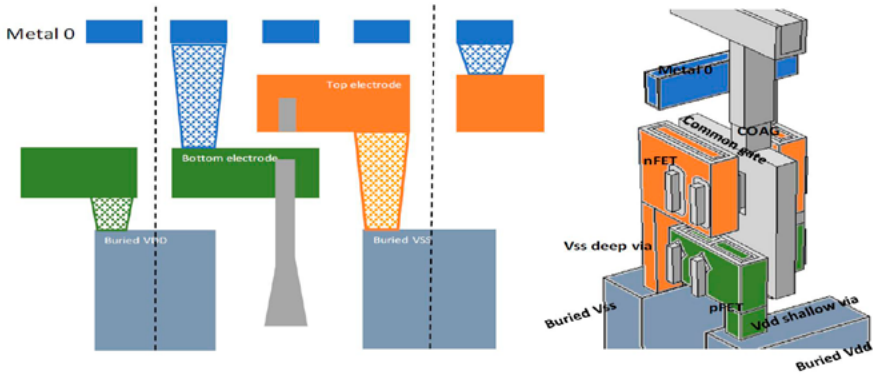


Figure 5: The CFET architecture forming a stacked p-n CMOS primitive structure with 2-level local interconnects.

CFET: the road towards 3T logic standard cells

Beyond 5T, a further reduction of the cell height is now mainly limited by routability issues, which should be evaluated at the logic block level. Optimizing routability brings us to the CFET, or complementary FET device – pushing the horizon for Moore's Law further out. The concept of CFET consists in "folding" the nFET on top of the pFET (either fin-on-fin or sheet-on-sheet) – thereby fully exploiting the possibilities of device scaling in 3D (Figure 5). By its stacked nature, the CFET exhibits two levels of local interconnects – providing more freedom for internal cell routing and for reducing cell area. Routing between cells can also be largely improved.

First assessments have shown that a FinFET-based 4T CFET can match and even surpass the standard cell power-performance metrics of a 5T "standard" FinFET device. It can also yield standard cells and SRAM cells with 25% smaller layout area. A nanosheet-based CFET could offer an extra performance boost and be necessary for scaling down to a 3T logic standard cell.

Impact on BEOL

The introduction of new transistor architectures impacts the interconnect structures in the chip's BEOL. Scaling

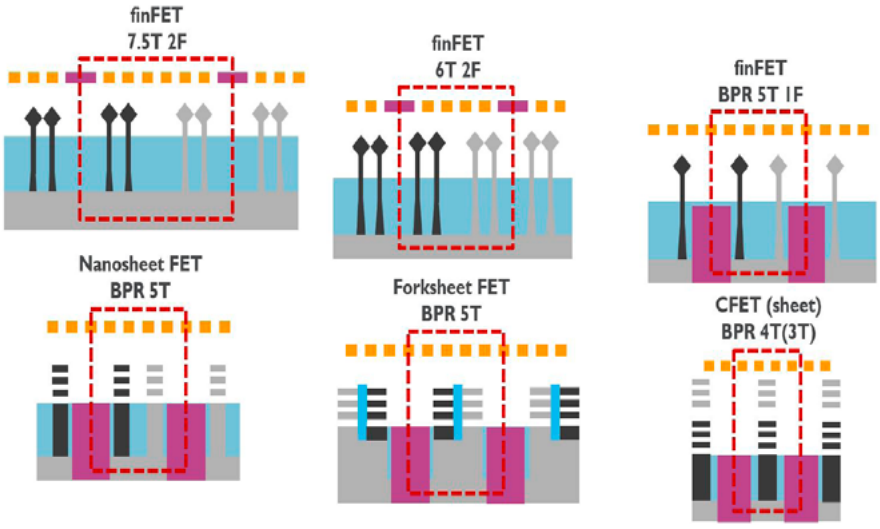


Figure 6: From FinFET to nanosheet, to forksheet, and then CFET.

is for an essential part limited by the interconnect. Therefore, building more compact logic circuits by track height scaling would require an optimization of the routability at the block level to enable the introduction of these new device architectures at reduced track heights. For example, as explained above, the transition from 6T to 5T standard cell height can only be enabled by moving the power rails – traditionally implemented in the chip's BEOL (i.e., M_{int} and M1 layers) – to the FEOL.

In the forksheet architecture, the local interconnect (active trench contact) to connect p- and nMOS source/drain within the same vertical trench can be more efficiently used in standard cell routing. This n-to-p-routing results in reduced M1 and M_{int} at cell level. And finally, CFET, because of its complementary stacked nature, exhibits two levels of local interconnect. This can significantly relieve inter-cell or intra-cell routing congestion. The large reduction of vertical routing resources inside the cell allows migrating the cell I/O pins to the first horizontal interconnect level increasing the aperture of the cell to the BEOL routing.

Summary

In this article, imec has mapped out an evolutionary path towards ultimately scaled logic devices for 2nm and beyond technology nodes (Figure 6). After today's mainstream FinFET comes the nanosheet device, offering superior channel control with limited additional process complexity. When complemented with scaling boosters, standard cells with 5T track height come within reach. As a next step, the forksheet may enter the scene, offering a path to 4.3T cells on account of a reduced n-to-p spacing. First simulations confirm its potential for the 2nm technology node. The CFET completes the roadmap as the ultimate compact CMOS structure, holding promise for 3T logic standard cells.

References

1. "Imec presents forksheet device as the ultimate solution to push scaling towards the 2nm technology node," imec press release, 12/2/19.
2. J. Ryckaert, et al., "Enabling sub-5nm CMOS technology scaling – thinner and taller!" IEDM 2019 Conference Proc. (invited paper); pp.: 29.4.1–29.4.4.
3. P. Weckx, et al., "Novel forksheet architecture as ultimate logic scaling device towards 2nm," IEDM 2019 Conference Proc.; pp.: 36.5.1–36.5.4.

Biography

Julien Ryckaert is Program Director at imec Leuven, Belgium. He received an MSc degree in Electrical Engineering from the U. of Brussels (ULB), Belgium, and a PhD degree from the Vrije U. Brussel (VUB). Since 2013, he has been in charge of imec's design-technology co-optimization (DTCO) platform for advanced CMOS technology nodes. As program director, he is focused on scaling beyond the 3nm technology node as well as the 3D scaling extensions of CMOS. Email Julien.Ryckaert@imec.be

ADVERTISER INDEX

Advantest www.advaantest.com ..... OBC

Amkor Technology www.amkor.com 1

Brewer Science
www.brewerscience.com 27

Chip Targets www.chiptargets.com 45

DL Technology www.dltechnology.com 30

E-tec Interconnect www.e-tec.com 48

EV Group www.evgroup.com 2

FormFactor www.formfactor.com/go/ap 4

Indium Corporation
www.indium.com/SiP/CSR 35

Ironwood Electronics
www.ironwoodelectronics.com 19

IWLPC www.iwlpc.com 48

Johnstech www.johnstech.com ..... IBC

MacDermid Alpha
www.macdermidalpha.com 6

PacTech www.pactech.com 25

Plasma Etch www.plasmaetch.com 46

SEMI www.semi.org/standards 47

Sonix www.sonix.com 44

SÜSS MicroTec www.suss.com 11

SVXR www.svxr.com 37

UTAC Group www.utacgroup.com 17

Winway Technology
www.winwayglobal.com 22

Xperi Invensas www.invensas.com ..... IFC

Yield Engineering Systems
www.yieldengineering.com/csr 14

July August 2020
Space Close July 3rd
Materials Close July 10th

For Advertising Inquiries
ads@chipscalereview.com

Johnstech®

Providing our customers accuracy, reliability and a solid partner to deliver today's and tomorrow's test solutions.

FOR MORE INFO, VISIT JOHNSTECH.COM

ISO 9001:2015 CERTIFIED

ADVANTEST

5G Made **Real** By Our Customers
Made **Possible** By Advantest

Advantest. Changing the World Through Our 5G-enabling Technologies and Solutions.

As a global leader in the semiconductor industry–we are working to improve the safety, security and reliability of millions of devices for applications including 5G, the Internet of Things, autonomous vehicles, artificial intelligence, machine learning, wearables, mobile electronics, and many more. Our innovative products and integrated solutions are used in the most advanced semiconductor production lines in the world.

Integrity. Innovation. And Trust.

These are just some of the core values that form the foundation of all that we do.

WWW.ADVANTEST.COM