Chip Scale Review®

ChipScaleReview.com

The Future of Semiconductor Packaging

Volume 26, Number 1

January • February 2022

3D interconnect inspection for heterogeneous chip packaging page 16

- Heterogeneous integration
  - Bridges for chiplet design
  - AI applications: status and future needs
  - A 2.2D die-last integrated substrate
- Managing trade-offs in the chiplet era
- FOWLP and Si-interposer for high-speed photonics
- Semiconductor packaging trends: an OSAT perspective

Subscribe

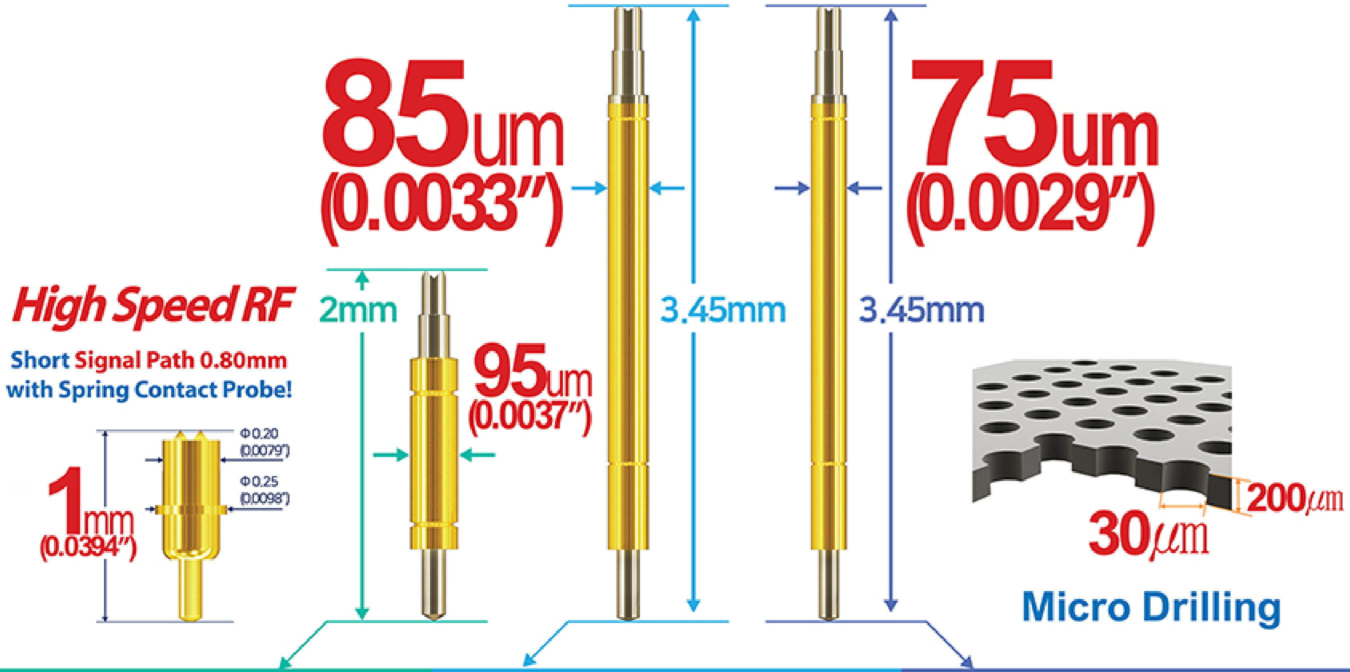
LEENO

www.leeno.com

Continuous Non-stop Innovation!

Fine Pitch Probe & Probe Head

Proven Mass Production Capability



High Speed RF

Short Signal Path 0.80mm
with Spring Contact Probe!

Micro Drilling

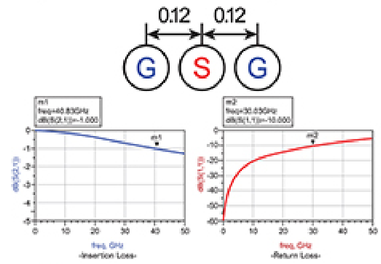
120㎛ Pitch

Mechanical Spec.

- Spring Force : 0.281oz (8.0g) @ .0098 (0.25mm)
- Recommended Travel : .0098 (0.25mm)
- Full Travel : .0118 (0.30mm)
- Material : Terminal – Pd Alloy / No plated
  Plunger – Pd Alloy / No plated
  Barrel – Ni-Au Alloy / Au plated
  Spring – Music Wire / Au plated

Electrical Spec. (Simulation Data)

- Current Rating : 1.0A
- Propagation Delay : 20.80ps
- Capacitance : 0.21pF
- Inductance : 0.38nH
- Insertion Loss : 40.83GHz @ -1.000dB
- Return Loss : 30.03GHz @ -10.000dB
  (Dielectric material : CERAMIC)



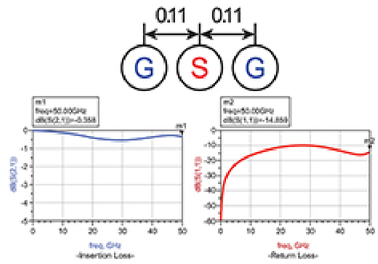
110㎛ Pitch

Mechanical Spec.

- Spring Force : 0.212oz (6.0g) @ .0118 (0.30mm)
- Recommended Travel : .0118 (0.30mm)
- Full Travel : .0138 (0.35mm)
- Material : Terminal – Pd Alloy / No plated
  Plunger – Pd Alloy / No plated
  Barrel – Ni-Au Alloy / Au plated
  Spring – Music Wire / Au plated

Electrical Spec. (Simulation Data)

- Current Rating : 0.9A
- Propagation Delay : 38.25ps
- Capacitance : 0.47pF
- Inductance : 0.63nH
- Insertion Loss : > 50.00GHz @ -1.000dB
- Return Loss : > 50.00GHz @ -10.000dB
  (Dielectric material : CERAMIC)



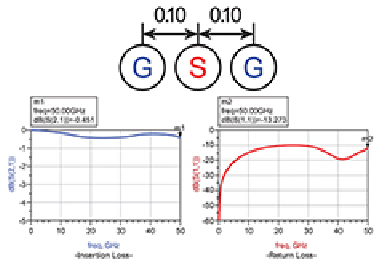
100㎛ Pitch

Mechanical Spec.

- Spring Force : 0.247oz (7.0g) @ .0118 (0.30mm)
- Recommended Travel : .0118 (0.30mm)
- Full Travel : .0138 (0.35mm)
- Material : Terminal – Pd Alloy / No plated
  Plunger – Pd Alloy / No plated
  Barrel – Ni-Au Alloy / Au plated
  Spring – Music Wire / Au plated

Electrical Spec. (Simulation Data)

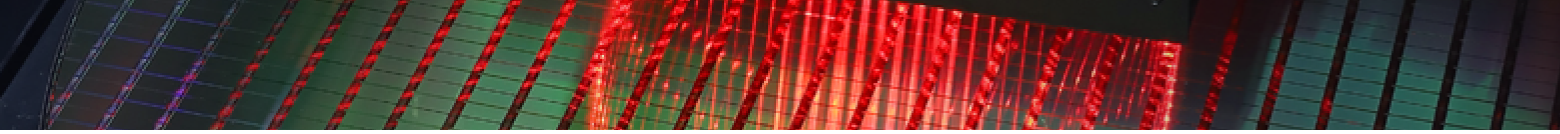
- Current Rating : 0.8A
- Propagation Delay : 35.55ps
- Capacitance : 0.44pF
- Inductance : 0.66nH
- Insertion Loss : > 50.00GHz @ -1.000dB
- Return Loss : > 50.00GHz @ -10.000dB
  (Dielectric material : CERAMIC)



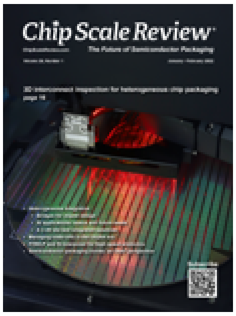
GLOBAL LEADER LEENO

HEAD OFFICE
10 105beon-gil MieumSandan-ro Gangseo-gu, Busan, Korea

CONTACT
USA: hskang@leeno.co.kr / +1 408 313 2964 / +82 10 8737 6561
Korea: sales-leeno@leeno.co.kr / +82 51 792 5639



January • February 2022
Volume 26, Number 1



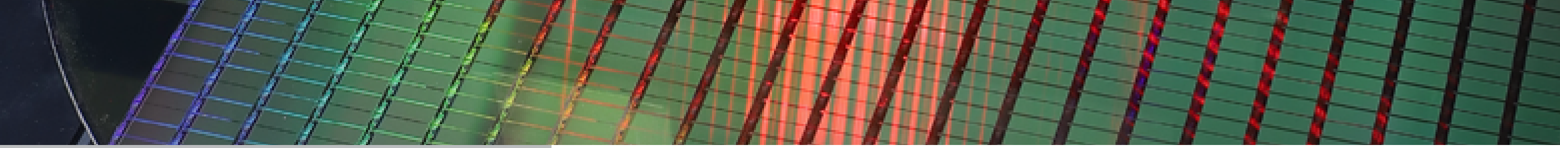
White-light scanning interferometry (WSI) applied for 100% fine-pitch interconnect inspection during wafer-level packaging. The WSI system includes scanning interferometry for large FOV, high-speed, and multi-reflectance surface 3D measurement. Interference patterns can be detected to calculate RDL dielectric layer thickness during panel/wafer-level 3D interconnect inspection.

Cover image courtesy of INTEKPLUS CO.,LTD.

CONTENTS

DEPARTMENTS

FEATURE ARTICLES



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

Volume 26, Number 1
January • February 2022

FEATURE ARTICLES *(continued)*

STAFF

Kim Newman
Publisher
knewman@chipscalereview.com

Lawrence Michaels
Managing Director/Editor
lmichaels@chipscalereview.com

Debra Vogler
Senior Technical Editor
dvogler@chipscalereview.com

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.
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 Gilroy, Calif., and additional offices.

POSTMASTER: Send address changes to Chip Scale Review magazine
P.O. Box 2165
Morgan Hill, CA 95038
Tel: +1-408-846-8580
E-Mail: subs@chipscalereview.com

Printed in the United States

Attention All Print Subscribers:

Because of current economic supply chain constraints and other associated reasons, *Chip Scale Review* will suspend the mailing of the print magazine starting with this issue. Print copies will continue to be available at select domestic (U.S.) and international conferences.

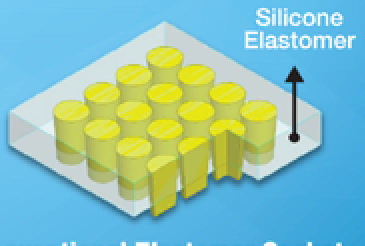
Subscribe now to receive the latest technology advances in the 6 issues coming up in 2022. *Chip Scale Review* is FREE to qualified individuals in the semiconductor advanced packaging industry.

chipscalereview.com/subscribe

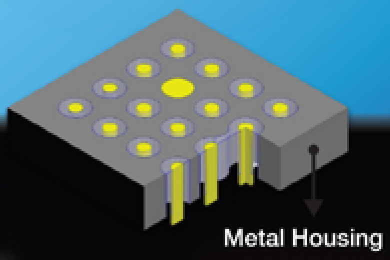
TSE
64Gbps
ULTRA HIGH SPEED
TEST SOLUTION
5G
HSIO LC Component
TL4C2
TSE
ELTUNE
(Low CTE and Dk Elastomer Socket)
MRC
(MEMS Rubber Contact)
FlexTUNE
(Coaxial Spring Pin Socket)
Load Board
TOP
A1
ELTUNE
FBGA-0.5X0.5-02A
TSE

COAXIAL ELASTOMER SOCKET

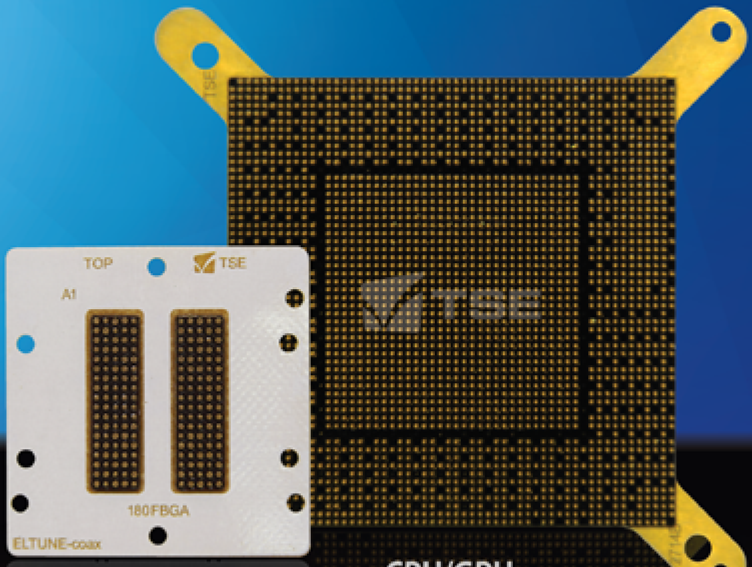
ELTUNE-coax™



Conventional Elastomer Socket



ELTUNE-coax



GDDR6
(18Gbps)

CPU/GPU
(32/64Gbps HSIO,112Gbps PAM4)

- **Mechanical Specification**

(unit: mm)

0.80mm pitch	Spring Pin	Elastomer	ELTUNE-coax
Over-drive	0.40	0.25	**0.25**
Test height	3.50	0.60	**0.60**

- **Electrical Specification**

(unit: GHz)

42.5Ω, 0.80mm pitch		Spring Pin	Elastomer	ELTUNE-coax
Insertion Loss S_{21} @-1dB	Single Ended (G-S-G)	13.93	27.81	**>100**
	Differential (G-S-S-G)	25.39	28.94	**>100**
Return Loss S_{11} @-10dB	Single Ended (G-S-G)	10.98	15.89	**>100**
	Differential (G-S-S-G)	20.77	25.20	**48.21**
Crosstalk S_{31}, @-20dB	G-S-S-G	8.50	9.43	**>100**

Contact sales@tse21.com

Korea (Headquarters)
+82-41-581-9955

USA +01-408-731-0030
China +86-512-67617299

Taiwan +886-3-657-6616
Singapore +65-9623-2650

Semiconductor packaging trends: an OSAT perspective

By David Clark [Amkor Technology, Inc.]

While supply chain issues, including severe shortages, occupied much of the visibility for semiconductors in 2021, semiconductor manufacturers and their outsourced semiconductor assembly and test suppliers (OSATS) have continued to make technical progress in many areas. These technology improvements address the advanced packaging requirements of leading-edge applications in key market segments. Before going into specifics, let's look at the overall market outlook.

Market outlook

The semiconductor packaging market continues to show a prosperous outlook and is forecast to grow to $96B by 2026 (3.8% compound annual growth rate (CAGR) from 21-26) (**Figure 1**). This market is typically divided into mainstream and advanced packaging segments with the latter being expected to exceed the mainstream segment for the first time by 2026.

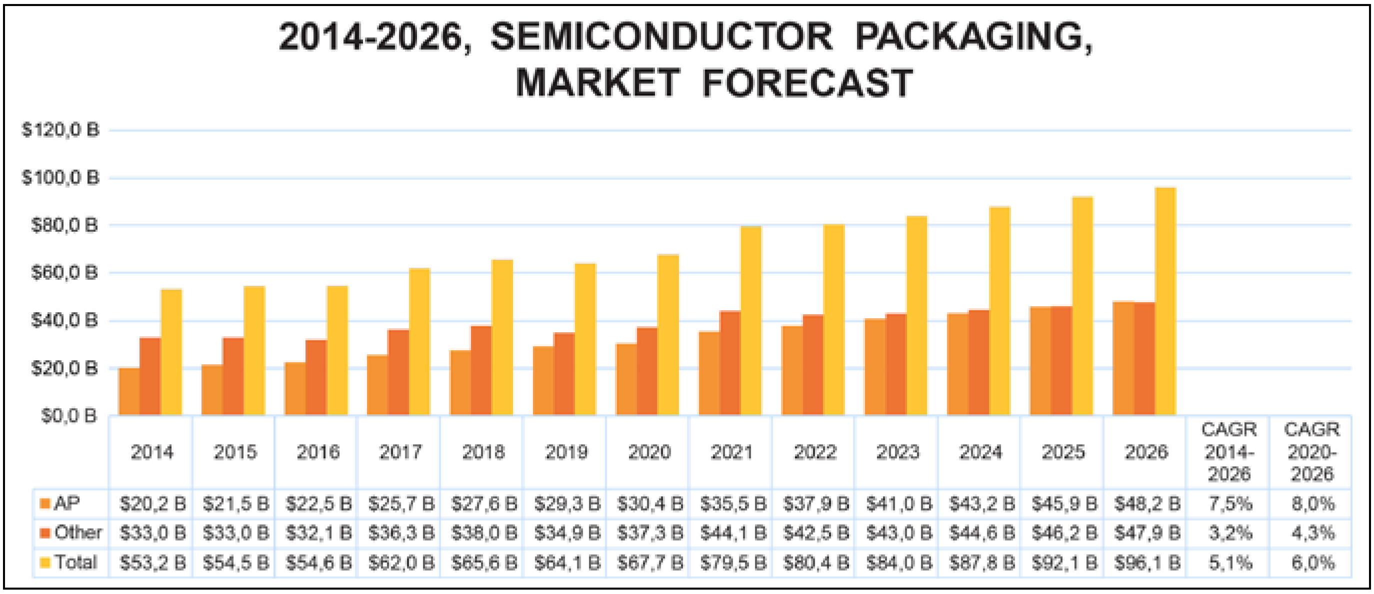
The market is underpinned by general trends in increasing manufacturing outsourcing, functionality and semiconductor content. Notable growth drivers come from multiple market segments such as 5G connectivity, automotive, data center, artificial intelligence (AI) and networking. 5G forms the backbone of many connected devices and services. While 5G is primarily a wireless connectivity growth opportunity, it also is an enabler for many adjacent markets generating further semiconductor content growth.

Despite industry-wide supply chain constraints since 2020 and expected to continue into 2022, many OSATS were still able to generate record revenues. Well reported shortages in IC foundry capacity, together with a constrained substrate supply chain, made for a challenging 2021. With newly publicized investments in these areas for capacity expansion, it is hoped that lead times will slowly reduce, and the industry will stabilize in 2022, however, substrate challenges will persist through 2023.

Mobile packaging trends

Many of the market growth drivers require increasing levels of system integration to meet the ever-increasing demand on performance, power and cost. As the OSAT supplier becomes an increasingly integral part of the overall system solution, it is in the advanced packaging segment where continued innovation in the areas of system in package (SiP), 2.5 and 3D packaging architectures are most apparent.

Cellular connectivity continues to drive advancements in radio frequency (RF) SiP technologies. With the rise of 5G, cellular frequency bands have increased considerably, requiring innovative solutions for the packaging of RF front-end modules for smartphones and other 5G-enabled devices. Amkor's double-sided molded



	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	CAGR 2014-2026	CAGR 2020-2026
■ AP	$20,2 B	$21,5 B	$22,5 B	$25,7 B	$27,6 B	$29,3 B	$30,4 B	$35,5 B	$37,9 B	$41,0 B	$43,2 B	$45,9 B	$48,2 B	7,5%	8,0%
■ Other	$33,0 B	$33,0 B	$32,1 B	$36,3 B	$38,0 B	$34,9 B	$37,3 B	$44,1 B	$42,5 B	$43,0 B	$44,6 B	$46,2 B	$47,9 B	3,2%	4,3%
■ Total	$53,2 B	$54,5 B	$54,6 B	$62,0 B	$65,6 B	$64,1 B	$67,7 B	$79,5 B	$80,4 B	$84,0 B	$87,8 B	$92,1 B	$96,1 B	5,1%	6,0%

Figure 1: Advanced packaging vs. traditional packaging market forecast (2014-2026). SOURCE: [1]

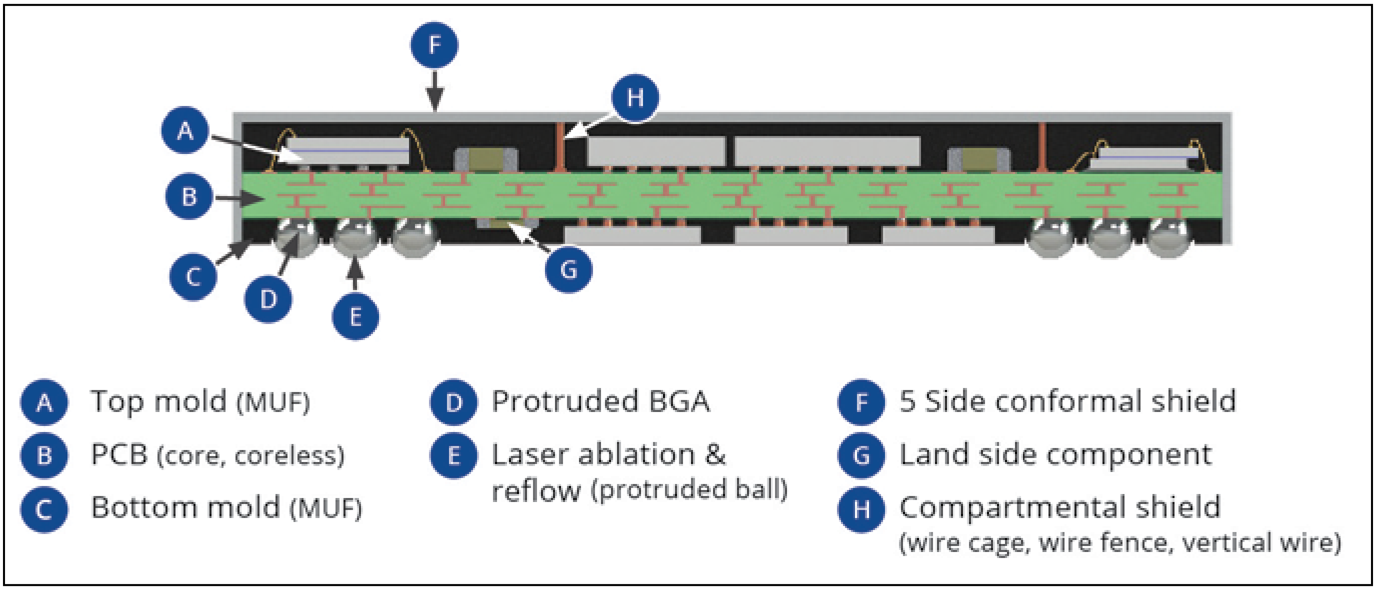

F
H
A
B
C
D
E
G
A Top mold (MUF)
B PCB (core, coreless)
C Bottom mold (MUF)
D Protruded BGA
E Laser ablation & reflow (protruded ball)
F 5 Side conformal shield
G Land side component
H Compartmental shield (wire cage, wire fence, vertical wire)

Figure 2: A double-sided molded ball grid array (DSMBGA) package.

ball grid array (DSMBGA) is the leading example of such solutions (**Figure 2**).

With the arrival of 5G networking, there has been a change in frequencies, adding frequency bands above 3GHz in FR1 and millimeter wave (mmWave) range in FR2. This growing number of new frequencies combined with the variety of multiplexing methods significantly increases the complexity of the RF front end. Integration using SiP allows customers to design, tune and test RF subsystems allowing for a reduction in design iterations and an accelerated time-to-market. Our double-sided packaging technology has vastly increased the level of integration for RF front-end modules used in smartphones and other mobile devices.

For 5G smartphones and other mmWave applications, antenna integration, either through antenna in package (AiP) or antenna on package (AoP) technologies, simplifies the challenges associated with designing products that operate at these high frequencies. A variety of AiP/AoP design methodologies provide the required form, fit and function for these applications and can include more than one antenna or antenna array. Today's AiP/AoP technologies can be implemented through standard, as well as custom, SiP modules to achieve a complete RF front-end (RFFE) subsystem.

In addition to a reduced size required for handheld and other small mmWave devices, AiP/AoP provides improved signal integrity with reduced signal attenuation and addresses the range and propagation challenges that occur at higher frequencies.

Automotive packaging trends

In the automotive area, advanced driver assistance systems (ADAS), electrification, and concepts such as the virtual cockpit, offer significant new opportunities for advanced packaging and innovation. These areas of new innovation are contributing to positive automotive semiconductor growth whereby the market is projected to grow from $38.7 billion in 2020 to almost US$82.6 billion in 2025 [2]. Implementing safety and comfort levels within ADAS-enabled vehicles result in an expectation for increased sensor deployment with the number of sensors increasing by 9.2% CAGR from 2020 to 2025 [2]. By 2026, the majority of high and some mid-range vehicles will integrate camera and radar, as well as light detection and ranging (LIDAR) sensors.

Multiple sensing techniques are being deployed to cover a vast array of range, environmental and accuracy requirements. Integration is a key focus to reduce the form factor and improve sensitivity levels. Sensor packaging platform development and re-use of mature assembly processes are key to controlling cost. For example, in

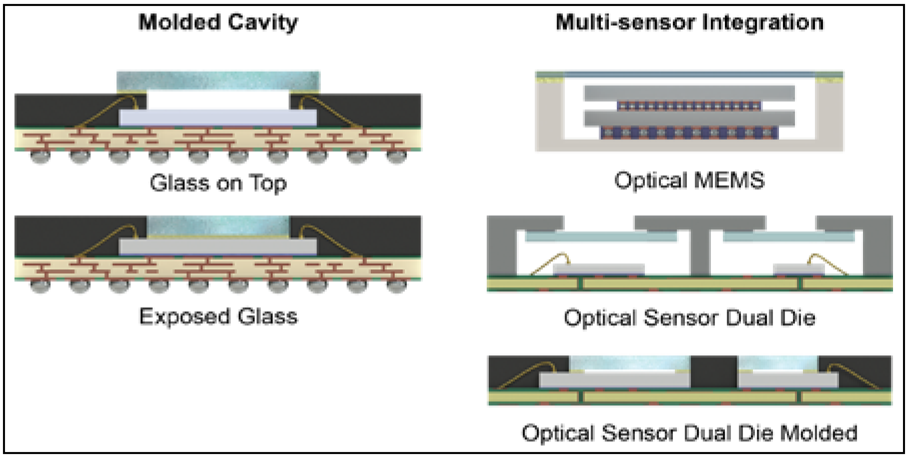


Figure 3: Molded cavity and multi-sensor integration optical sensor packages.

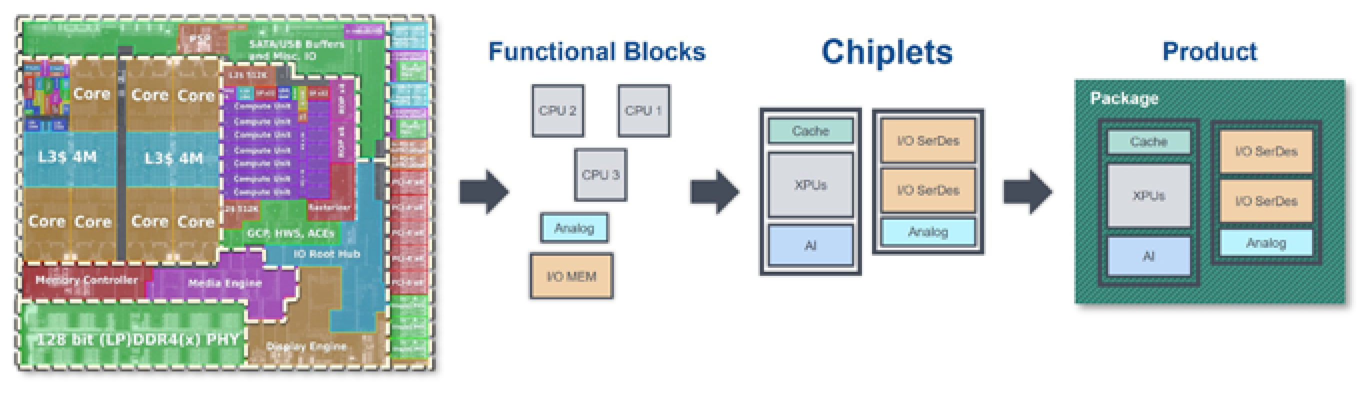


Figure 4: A heterogeneous integration platform for chiplets.

the area of optical sensing, such as the time of flight (ToF) and contact image sensor (CIS), molded single and multi-cavity microelectromechanical systems (MEMS) packaging solutions are now being deployed and qualified for these optical sensing applications (**Figure 3**).

The ADAS system-level augmentation of the sensor functions noted above will drive the need for higher levels of in-vehicle compute capability. In this area, OSATS are leveraging many years of experience emanating from the high-performance compute and network sectors. With further development of specific automotive-rated material systems, these single and multi-chip central processing units (CPUs) can be qualified to the automotive AEC Q-100 grade requirements.

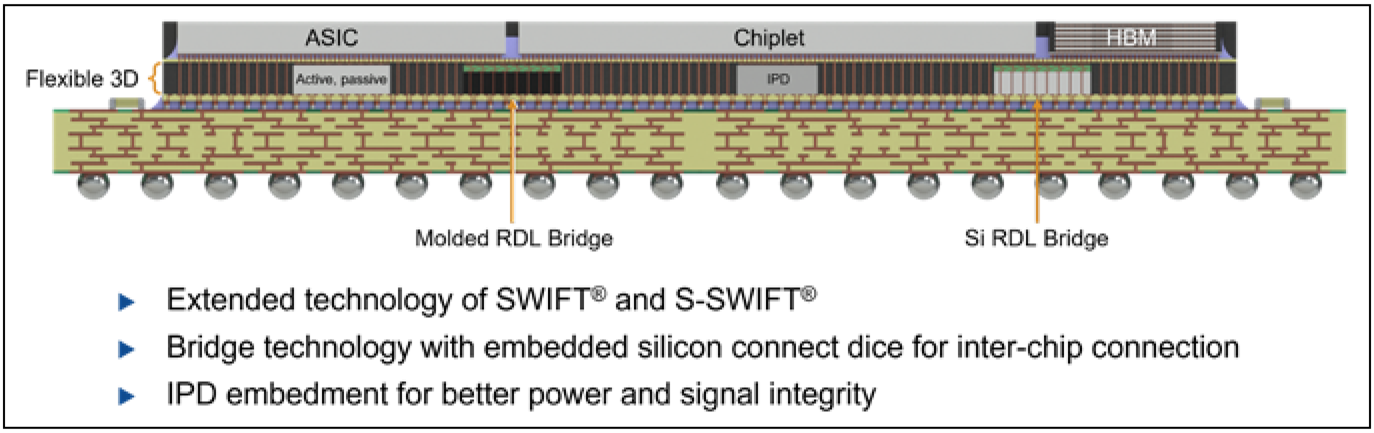
We anticipate the accelerated adoption of advanced silicon technology nodes with 5nm designs being introduced by automotive original equipment manufacturers (OEMs) later in this decade. Furthermore, SiP technology then offers automotive customers a platform to integrate these advanced CPU chips with complementary functions such as Serializer/Deserializer (SerDes), power management integrated circuits (PMICs), memory and more.

Data center and networking packaging trends

Cloud and edge computing, storage and networking form the backbone of today's connected living. The demand on voice and data traffic is driving major innovations in system architectures and fueling the partitioned chiplet trend (package-level integration) to find the ultimate, optimized balance in power, performance and cost (**Figure 4**). As these processing demands increase, transistor densities are increasingly challenging. Combined with effects like heat and noise, they are forcing designers to leverage heterogeneous architectures with specialized accelerators and memories, either on a single die or in an advanced package.

2.5 and 3D packaging solutions offer a heterogeneous integration platform for chiplets. Consequently, foundries are expanding their 3D packaging portfolios. To date, OSATS have offered complementary heterogeneous packaging and supply chain solutions, such as Amkor's SWIFT® and S-Connect technologies (**Figure 5**). Many of these approaches, whether a foundry or OSAT, die-first or die last, with or without interposer and other options, aim to quench the desire to extend Moore's Law and provide more effective package-level alternatives.

The technical challenges extend beyond the ability to co-package chiplets, so chip-package co-design is critical. When partitioning a floor plan, one needs to think carefully about where to place components within the package. Some components need to be placed very close together physically to maintain signal and power integrity. Key questions are



- Extended technology of SWIFT® and S-SWIFT®
- Bridge technology with embedded silicon connect dice for inter-chip connection
- IPD embedment for better power and signal integrity

Figure 5: Amkor's S-Connect technology.

where and what to partition, what is the workload and what silicon nodes are optimal in terms of cost and yield for each function. With this added system design freedom, the OSATS's role is increasingly important in the system-level design, chip-chip I/O routing, power distribution, thermal optimization, and more.

Today, the chiplet era is in its infancy. The way systems are designed to date has been based upon historic approaches to moving data. A more pioneering approach to the movement of data to support a metaverse future will redefine how next-generation systems are configured. Concepts such as co-packaged optics (CPOs) that are currently in the research phase are among the future package design possibilities.

Summary

To satisfy the application needs in leading markets and meet growth projections, several different advanced packaging technologies are required. For continued OSATS', as well as semiconductor manufacturers' success, a few key criteria must be satisfied. Semiconductor original equipment manufacturers (OEMs) and OSAT suppliers must continue to improve their collaboration during the design phase to make sure the right problems are being solved early in the innovation process. To minimize footprints, effectively manage power and continuously improve performance, the technology investments by OSATS must occur with financial stability as a goal. With the right packaging concepts, success is demonstrated through capability to scale up to satisfy volume requirements in these growing markets. This is essential to avoid future supply chain issues.

Acknowledgments

SWIFT is a registered trademark of Amkor Technology, Inc.

References

1. Yole Développement, Status of the Advanced Packaging Industry 2021, p. 123.
2. Gartner/Semiconductor Forecast Database, Worldwide, 3Q21 Update – Published October 4, 2021.

Biography

David Clark, Sr. Director, Amkor Technology, Inc. is responsible for Product Marketing and strategic business development in Europe. Prior to joining Amkor, David held various business development and engineering positions at FlipChip International (FCI), Leica Microsystems and Agilent Technologies. He has been granted 5 patents in Optoelectronics and Device Packaging and holds a BEng Honors Degree in Electronic, Electrical and Optoelectronic Engineering from the U. of Glasgow. Email: david.clark@amkor.com

Shrinking RADAR and LiDAR sensor packages – an introduction to TINKER

By Leo Schranzhofer [PROFACTOR], Martin Eibelhuber [EV Group], Martina Chopart [AMIRES]

Autonomous driving and self-driving cars are prominent use cases for microelectronics and sensors—most importantly, radio detection and ranging (RADAR) and light detection and ranging (LiDAR) sensors (**Figure 1**). The RADAR and LiDAR markets have enormous potential, with the market size of LiDAR sensors in automotive and industrial applications estimated to reach 26% compound annual growth rate (CAGR) in the 2020-2026 period. The market segment of advanced driver-assistance systems (ADAS) is expecting an even more impressive growth of 111% [1].

Public awareness and the industrial need for further miniaturization of RADAR and LiDAR sensor packages are the main drivers of ongoing efforts in the automotive sector to integrate these sensors into the body of a vehicle, such as in the bumpers, grilles and exterior lamps, instead of attaching them to the exterior of cars. Safety for both the driver and others is the most important consideration in the automotive sector. Therefore, high-value and high-performance RADAR and LiDAR systems are required for ADAS as well as for autonomous cars. Current bottlenecks are the relatively large size, weight and power consumption of such sensor devices. Because these factors are highly limited within cars, further miniaturization and improving functionality and efficient use of resources are highly demanded.

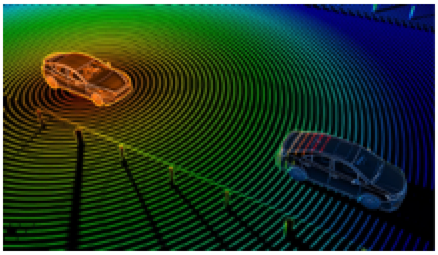


Figure 1: Autonomous cars are prominent use cases for RADAR and LiDAR sensors.

For a duration of three years beginning on October 1, 2020, the European Union's Horizon 2020 funded TINKER project [official project name is "Fabrication of Sensor Packages Enabled by Additive Manufacturing"] is set to develop a new reliable, accurate, functional, cost- and resource-efficient pathway for RADAR and LiDAR sensor package fabrication, following two main objectives. The first objective is to establish a platform based on additive manufacturing. The second objective is to fabricate RADAR and LiDAR sensor packages as use cases. TINKER's approach is to use key enabling technologies, especially inkjet printing and nanoimprint lithography (NIL), as disruptive and flexible manufacturing techniques in micro-part assembling. The proposed TINKER pilot line will offer a high degree of flexibility and reliability due to its modular character. Additional key components are inline feedback control mechanisms, which will be directly integrated. **Figure 2** shows the basic concept and components of TINKER. Starting from bare dies for LiDAR and RADAR chips, pick-and-place processes and inline inspection for feedback control will be complemented by functional inkjet printing and nanoimprinting to fabricate sensor packages for the RADAR and LiDAR use cases.

The TINKER project aims to decrease production time, measurably increase automation level, achieve a higher or similar precision level as compared to the state of the art in manufacturing of these device types, and reduce rejection

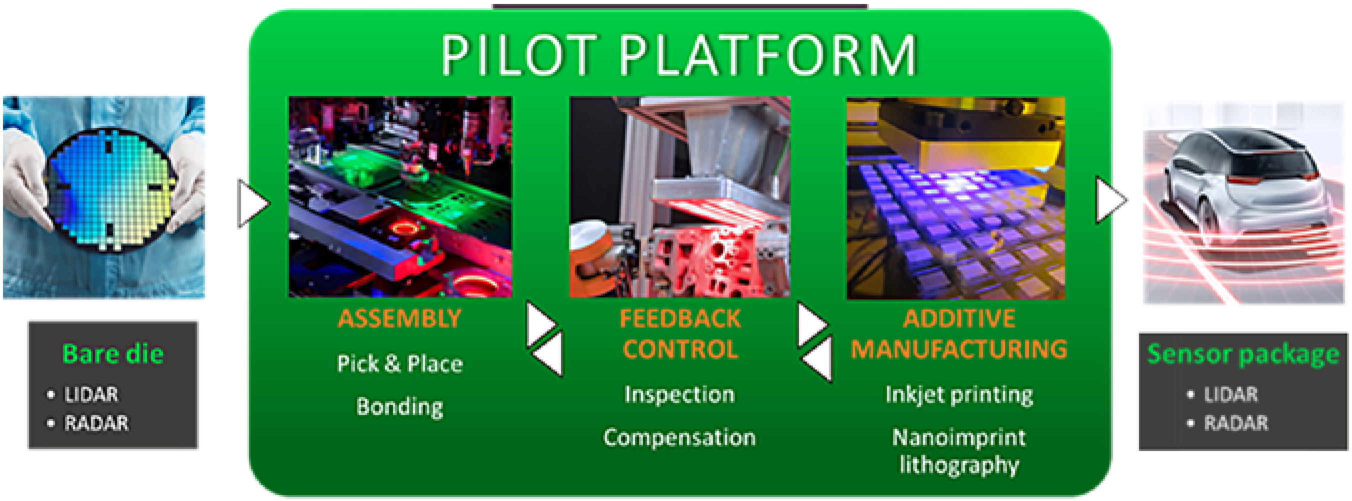


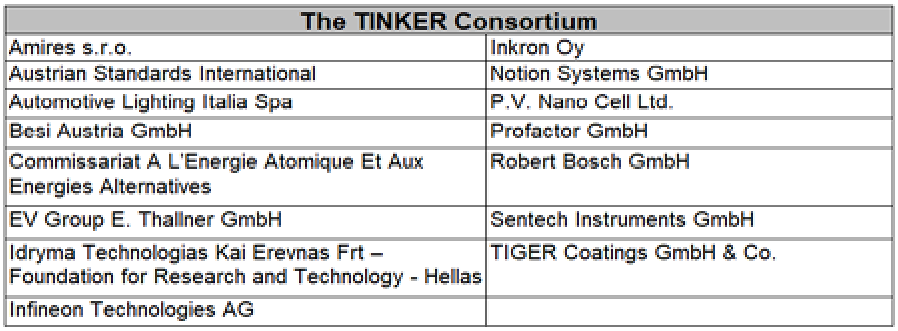
Figure 2: The TINKER platform provides new additive manufacturing concepts in a pilot line for use cases such as RADAR and LiDAR sensors.

rates during the production process. The main purpose of this project is to widen the range of available miniaturization and microelectronic fabrication possibilities, including novel approaches in assembly processes added directly into production steps. TINKER will also train PhD and MSc students and aim to publish scientific papers and protect intellectual property.

The TINKER consortium comprises 10 companies, three research specialists, one consultancy and a service association, who are major players in the field of semiconductor and microelectronic manufacturing, as well as in the fields of material and process development and industrial fields applying, or interested in applying, additive manufacturing for their needs (**Table 1**). All the partners have a track record of nationally- and internationally-funded projects in their special research fields.

Within the first year, the partners in the consortium have made significant progress in creating the proposed TINKER pilot line and the innovative techniques it aims to employ, especially pick and place, inkjet printing and NIL. Partners are working on developing and perfecting designs, processes and materials required for the two selected use cases. They have been able to overcome the obstacles that the COVID-19 pandemic posed and collaborated successfully on developing equipment and tools and shared samples, technologies, and processes. Examples of such fruitful collaborations are the completion of the pick-and-place equipment prototype, creation of the initial setup for improved inline monitoring and feedback, and the infrastructure and tools set up for NIL. The consortium has made significant progress in equipment and tools development for the TINKER pilot line, and a solid base for achieving the project's ambitious goals is now created. One year after the project's launch, it is well on track to reach its aim of widening the range of available miniaturization and microelectronic fabrication possibilities, and introducing novel approaches to assembly processes, directly in production.

The first public results can be accessed via the project webpage (https://www.project-tinker.eu/downloads/#public_deliverables) and are summarized in three reports. The first report provides a high-level overview about the activities centered on the pick-and-place equipment and shows an early prototype of the RADAR demonstrator. It also touches on the metrology systems that are developed by BESI and SENTECH.



The TINKER Consortium	
Amires s.r.o.	Inkron Oy
Austrian Standards International	Notion Systems GmbH
Automotive Lighting Italia Spa	P.V. Nano Cell Ltd.
Besi Austria GmbH	Profactor GmbH
Commissariat A L'Energie Atomique Et Aux Energies Alternatives	Robert Bosch GmbH
EV Group E. Thallner GmbH	Sentech Instruments GmbH
Idryma Technologias Kai Erevnas Frt – Foundation for Research and Technology - Hellas	TIGER Coatings GmbH & Co.
Infineon Technologies AG	

Table 1: The TINKER consortium.

Another report presents the NIL infrastructure and available tools, as well as those developed specifically for the TINKER project. For the realization of LiDAR sensors, the TINKER consortium is developing a NIL-based fabrication method for photonic integrated circuits. The photonic integrated circuit is the very precise core element necessary for the manipulation of light for each LiDAR sensor. The aim of TINKER is to replace certain fabrication steps in an optical phased array (OPA) process flow with NIL in order to achieve fast and reliable direct manufacturing of the needed passive optical elements, such as the waveguides and optical coupling structures. This will reduce the size of the photonic IC device and its fabrication costs, as well as increase the throughput of its production. With the completion of this deliverable, the TINKER consortium has now completed the equipment and tools development for the use case in TINKER.

An important goal is to integrate inline process monitoring for different production processes: pick and place, inkjet printing, and NIL. These processes pose multiple requirements on inspection and inline monitoring. Sensor technology deployed in TINKER to meet these requirements comprise optical inspection, curing sensor, and thermographic imaging. This is covered in the third report.

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement nº 958472 with an overall budget of € 10,241,526.25. The project is coordinated by scientist Leo Schranzhofer at PROFACTOR GmbH. Discover more about the project on its website (www.project-tinker.eu) and on its social media platforms linkedin.com/in/tinker-eu-ba27b01b9 and twitter.com/project_tinker.

Reference

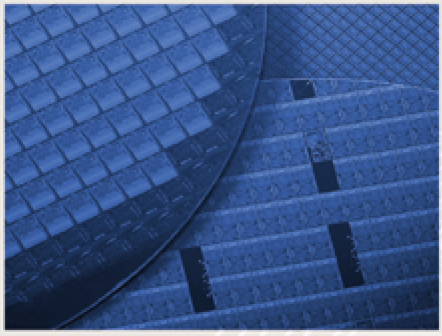
1. Yole Développement (2021, September 2). LiDAR adoption: Technology choices and supply chain management are the key enablers [Press release]. www.yole.fr/iso_upload/News/2021/PR_LIDAR_AUTOMOTIVE and INDUSTRIAL_MarketUpdate_DesignWins_Yole_September2021.pdf

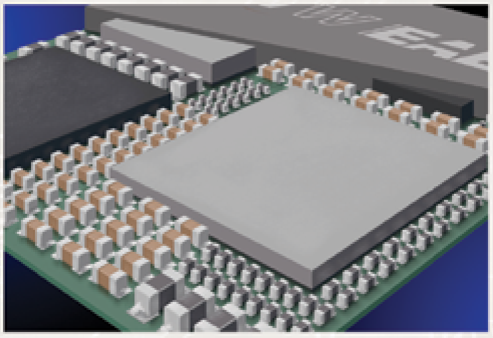
Biographies

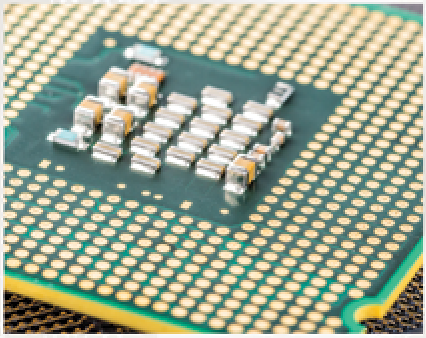
Leo Schranzhofer is Lead Scientist and Head of the "Functional Surfaces and Nanostructures" team at PROFACTOR GmbH in Steyr, Austria and project coordinator of the TINKER project. He has a doctorate in chemistry, specialized in analytical chemistry and chemical sensors, from the U. of Vienna. Email: leo.schranzhofer@profactor.at

Martin Eibelhuber is Deputy Head of Business Development at EV Group, Austria, focusing on compound semiconductors, nanotechnology, and photonics applications. He has a doctorate in Technical Physics from the Johannes Kepler U. Linz, specializing in nanoscience and semiconductor physics.

Martina Chopart, MSc., is an EU Project Manager at AMIRES s.r.o. in Prague, where she is responsible for management and dissemination of projects within the DeepTech program.







Vitronics Soltec
EDISON
MPM
Camalot

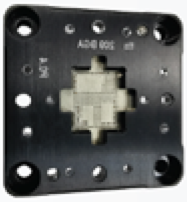
tts

www.tts-grp.com

Being in the business since 1979, TTS Group dedicates itself to bring the best possible solutions to help our customers solve some of the most challenging issues in semiconductor test and tooling today.

We understand your needs and focus on lowering the total cost of final test through relentless pursuit of innovation, yield, quality, reliability and on-time delivery for our customers.

PROBE HEAD



TEST SOCKET

PROBE PIN

3D interconnect inspection for heterogeneous chip packaging using WSI

By Shahab Chitchian [INTEKPLUS CO., LTD.]

Heterogeneous integration through the use of chiplet packaging, including fan-out wafer-level packaging (WLP) and panel-level packaging (PLP) architecture, has become a key technology to continue Moore's Law by improving yield and reducing total product cost [1,2]. Another important driver for heterogeneous integration is to improve performance and power efficiency, which can be achieved by decreasing interconnect pitch and increasing interconnect density. Therefore, capable 3D interconnect/bump inspection is necessary to enable heterogeneous packaging.

In this article, white-light scanning interferometry (WSI) technology is presented for the latest heterogeneous chip 3D bump inspection. First, 3D automated optical inspection (AOI) based on oblique and coaxial methods is introduced. As a coaxial method, scanning interferometry is described to accurately measure 3D interconnects from a few micrometers to a few hundred micrometers with various reflectances. Key features of WSI technology including multi-reflectance surface 3D measurement, high-speed camera, custom-made frame grabber, and large field of view (FOV) are introduced in the second section, followed by the 3D inspection algorithm. Furthermore, inspection results and 3D interconnect parameters of bump height, bump coplanarity, and chip area warpage (CAW) are explained to overcome challenges of large form factor multi-chip packaging. One main challenge is how to assure successful die bonding conditions before silicon die attach to the expensive multi-chip package. Last, but not the least, a case study for multi-reflectance detection is presented to measure dielectric layer thickness during redistributed layer (RDL) processing for fan-out WLP applications.

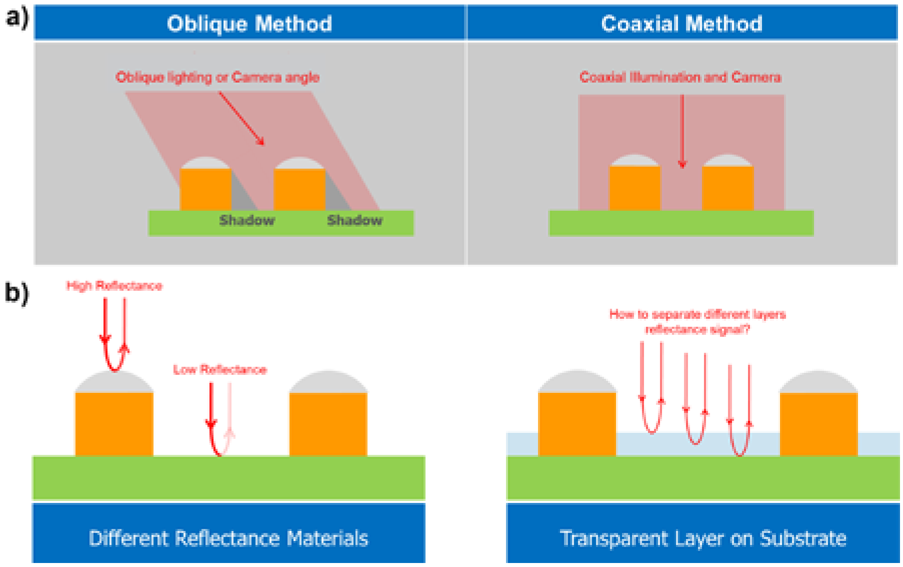

a)
Oblique Method
Coaxial Method
Oblique lighting or Camera angle
Shadow
Shadow
Coaxial Illumination and Camera
b)
High Reflectance
Low Reflectance
How to separate different layers reflectance signal?
Different Reflectance Materials
Transparent Layer on Substrate

Figure 1: a) Oblique and coaxial methods for 3D interconnect inspection; b) multi-reflectance detection for layer separation during interconnect 3D measurement.

Interconnect inspection process

100% 3D interconnect inspection applied in the semiconductor mounting process requires both high accuracy and high speed. As shown in **Figure 1a**, oblique and coaxial methods are currently being used widely in the industry. The oblique method includes technologies like moiré and laser scanning. Oblique angle illumination or camera angle causes a nonsensitive (shadow) zone that may impact measurement results. Therefore, these technologies provide a good estimate of interconnect/bump height variation—but not an accurate measurement. Another measurement error when using the oblique method is caused by a shiny surface that has a high surface reflectance. To overcome the accuracy limitation of the oblique method, coaxial illumination and a coaxial camera are applied, which result in the shadow-free and accurate measurement of interconnect/bump height. Lower throughput (in units per hour [UPH]), is a disadvantage of the coaxial method compared to the oblique method. This can be overcome by enlarging the FOV and using a high-speed camera as two solutions. The coaxial method consists of confocal and WSI.

Within coaxial technologies, WSI has the key advantage of being able to distinguish transparent layers related to RDLs, e.g., a polyimide (PI) layer, illustrated in **Figure 1b**. In this article, an interferometry system is discussed for 3D interconnect inspection and metrology of microelectronic devices; preliminary results are presented. The outlines are: 1) WSI system; 2) 3D inspection algorithm; 3) measurement results; 4) 3D interconnect inspection parameters; and 5) multi-reflectance detection for transparent layer measurement.

WSI system

The WSI system concept is shown in **Figure 2**. A 25M pixel high-speed camera, operating at 150 frames per second (FPS), is used to capture interferometry images. Because of

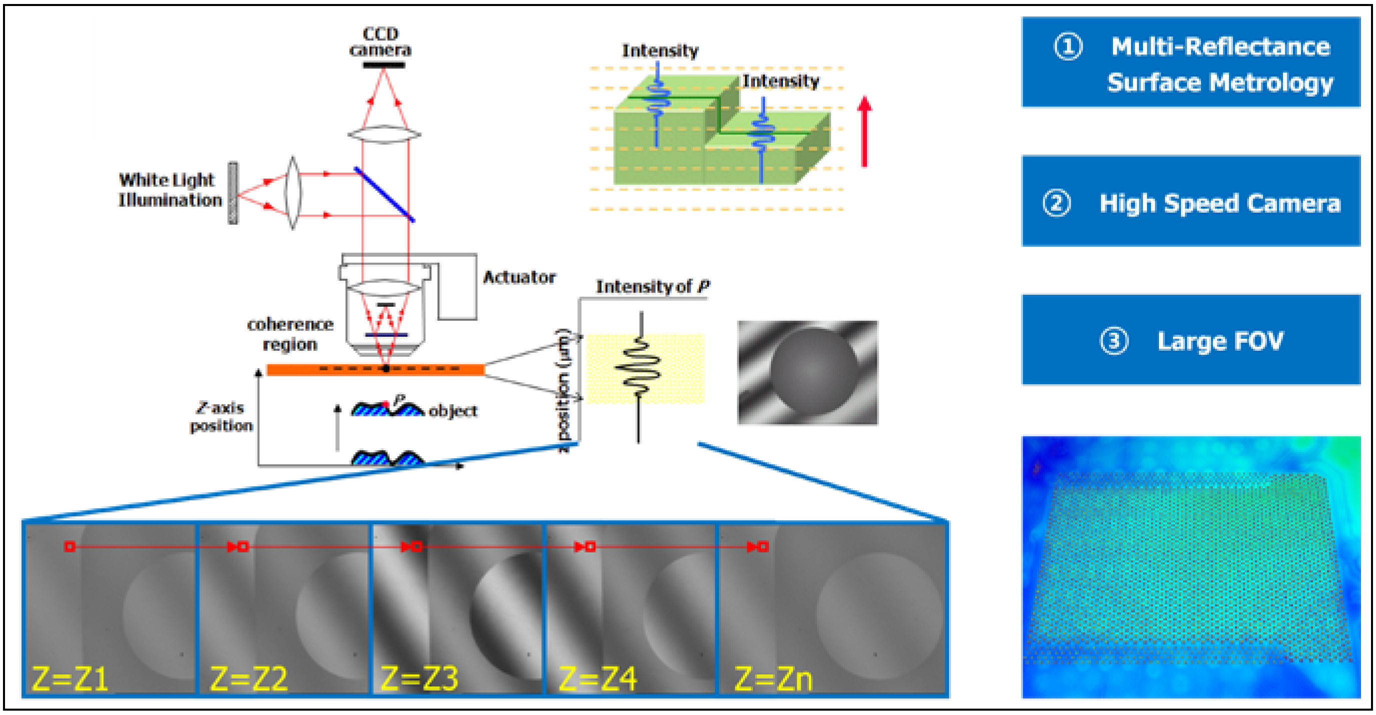


Figure 2: WSI system concept: scanning interferometry for large FOV, high speed, and multi-reflectance surface 3D measurement.

the small size of the bumps and the requirement for a fast measurement speed, a high numerical aperture (NA) imaging lens was developed so that a large FOV can be measured with high resolution.

A WSI system is described in **Figure 3**. A piezo actuator, which can move the range of a few nanometers to a few microns, is applied as the scanning method. In our system, the beam splitter (B/S) is lighter than the optical reference mirror, so we decided to use the B/S moving method in order to achieve high-speed scanning. Reference mirror scanning is also an option when we use cube B/S. A general-purpose computing on graphics processing units (GPGPU) method was implemented to process hundreds of acquired images. A light source with narrow enough bandwidth is used to expand coherence length. Therefore, scanning is possible for a few hundred um range. A light-emitting diode (LED) is used as the light source projected by a Koehler illumination system. The LED power is 24W with a center wavelength of 628nm and a 30nm bandwidth.

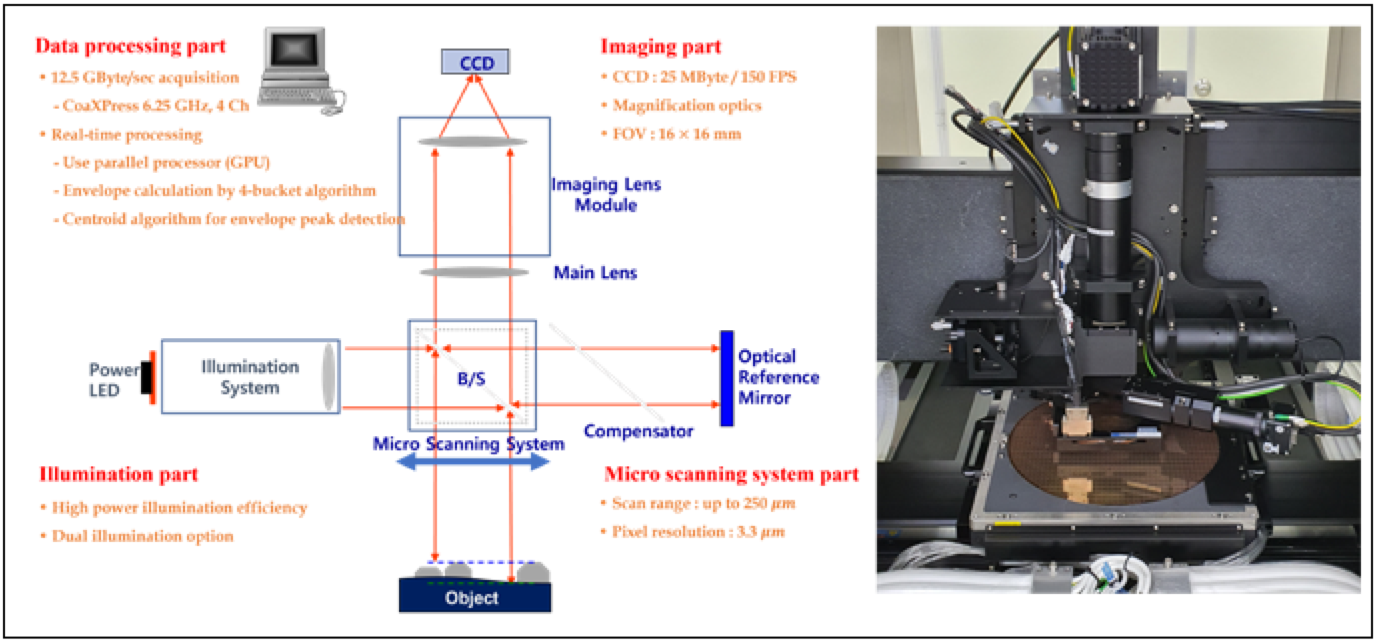

Data processing part
• 12.5 GByte/sec acquisition
- CoaXPress 6.25 GHz, 4 Ch
• Real-time processing
- Use parallel processor (GPU)
- Envelope calculation by 4-bucket algorithm
- Centroid algorithm for envelope peak detection
CCD
Imaging part
• CCD : 25 MByte / 150 FPS
• Magnification optics
• FOV : 16 × 16 mm
Imaging Lens Module
Main Lens
Power LED
Illumination System
B/S
Optical Reference Mirror
Compensator
Micro Scanning System
Illumination part
• High power illumination efficiency
• Dual illumination option
Micro scanning system part
• Scan range : up to 250 µm
• Pixel resolution : 3.3 µm
Object

Figure 3: WSI system specification and setup.

3D inspection algorithm

An envelope of interference signal acquired from the interferometer is calculated using a four-bucket algorithm [3]. The general expression of the interference pattern acquired by interferometry is shown in **Figure 4**.

The light intensity of the shifted interference pattern is defined as follows:

$$I_k = D\left[1+\gamma\cos\left(\frac{2\pi h}{\Lambda}+(k-1)\delta_0\right)\right], k = 1,2,\ldots,n. n \geq 3$$

where D is the mean intensity, γ is the visibility, $\frac{2\pi h}{\Lambda}+(k-1)\delta_0$ is the phase of the interferogram, and δ_0 is the phase shift.

The light intensity of each phase in the case of a 90 degree phase shift is given by:

$$I_1 = D\left[1+\gamma\cos\left(\frac{2\pi h}{\Lambda}+0\right)\right]; I_2 = D\left[1+\gamma\cos\left(\frac{2\pi h}{\Lambda}+\frac{\pi}{2}\right)\right]; I_3 = D\left[1+\gamma\cos\left(\frac{2\pi h}{\Lambda}+\pi\right)\right]; I_4 = D\left[1+\gamma\cos\left(\frac{2\pi h}{\Lambda}+\frac{3\pi}{2}\right)\right].$$

Using a four-bucket algorithm, the amplitude is calculated as follows, $u = \frac{1}{2}\sqrt{(I_4-I_2)^2+(I_3-I_1)^2}$ where u is the amplitude.

Finally, the envelope peak of the interference signal is detected by a centroid algorithm [4].

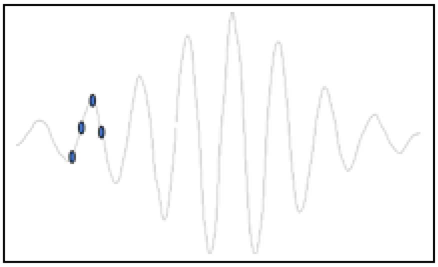


Figure 4: Interference pattern calculation using a four-bucket algorithm.

Measurement results

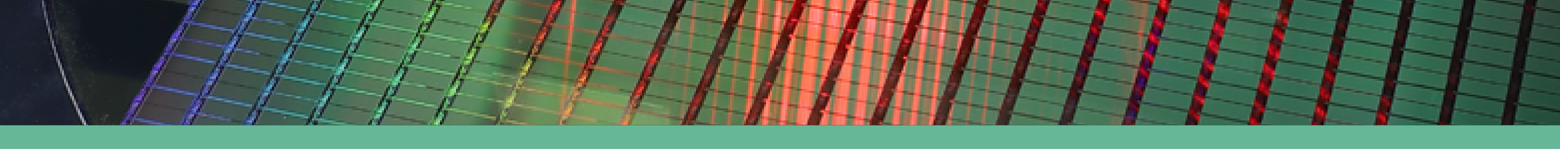
In order to evaluate the WSI system, the bump height repeatability based on 30 measurements is calculated for every bump. Because each chip comprises tens of thousands of interconnects, we focus on the distribution of standard deviations for every bump measured. Following this principle, σ_{RPT} is represented by the statistical upper three-sigma limit of the sigma distribution. The repeatability sigma is assessed based on $\sigma_{RPT} \equiv \sigma_{worst} = ave_{\sigma i} + 3\sigma_{\sigma i}$ where σ_{worst} is the worst-case standard deviation based on a 99.73% probability that the standard deviation for any given bump height result will be less than σ_{worst}. The results of σ_{RPT} for three chips are 0.714µm, 0.708µm, and 0.724µm. An example of 3σ for one sample is shown in **Figure 5a**.

We have also compared results of three samples measured by WSI and a reference system using the confocal principle [5]. A comparison of the bump height measurement using two systems is shown in **Figure 5b**. The bump height process specification is equal to 45µm±4.5µm. Correlation of measured data has the bump height R^2=0.89 with a mean difference less than 1µm, which shows measurement matching of the two systems.

WSI is sensitive to vibration, so it is necessary to correct the scan steps by using an anti-vibration system during scanning. To overcome this issue, we are now working on dispersive white-light interferometry (DWI), which uses spectral imaging to extract high vertical resolution without vertical scanning. For a DWI system, the Z accuracy depends on the spectrometer sensitivity and the fast Fourier transform (FFT) method. Therefore, DWI will provide higher speed and accuracy for in-line inspection as applied to the semiconductor mounting process.

3D interconnect inspection parameters

In this section, inspection parameters of interconnect/bump height, coplanarity, and CAW are explained to overcome challenges of large form factor multi-chip packaging. One main challenge is how to assure successful die bonding conditions before silicon die attach to the expensive multi-chip package. Three key parameters are defined as follows:



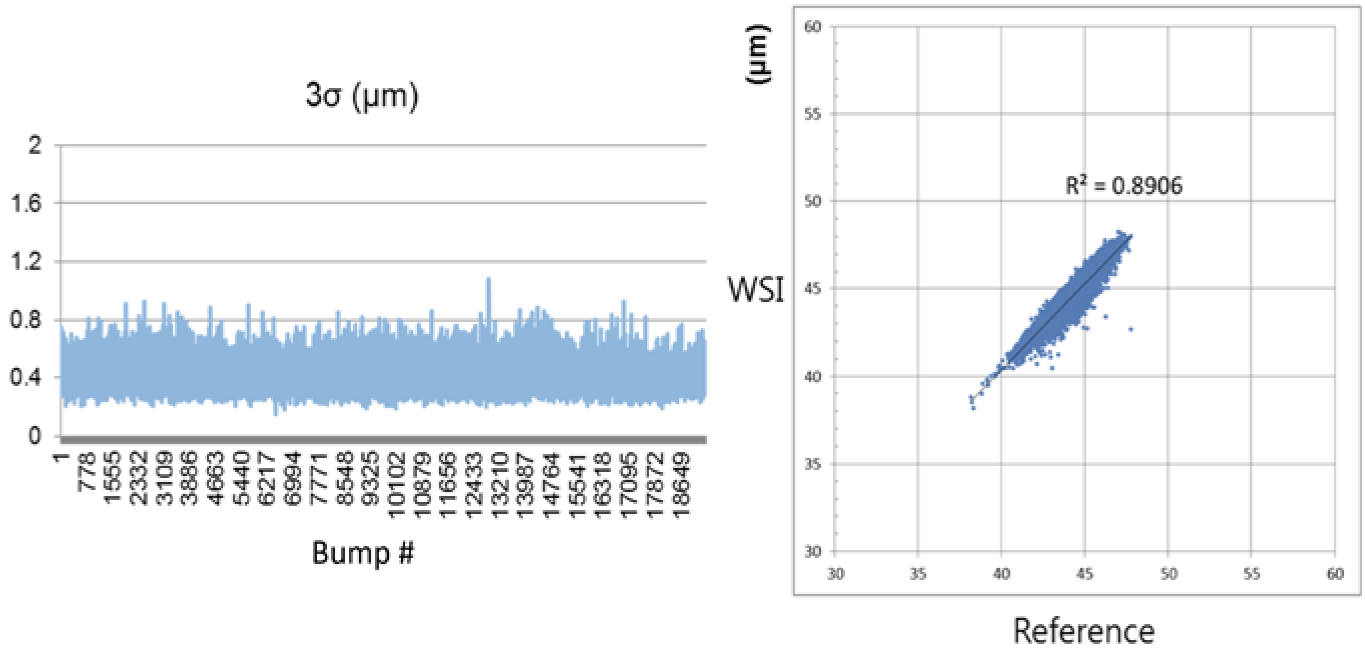


Figure 5: Results of a) (left) one sample bump height 3σ; and b) (right) WSI and reference systems bump-level bump height correlation using three samples.

1) Interconnect/bump height: the height difference for each bump relative to the surface surrounding bump depicted in **Figure 6a**.

2,3) Interconnect/bump coplanarity and CAW: **Figure 6a** illustrates chip interconnects under the free-state condition. At first, the bumps' tops, as well as the surface surrounding the bumps' (called chip area's) tops are measured in any coordinate system. Second, we find the least squares (LSQ) planes separately

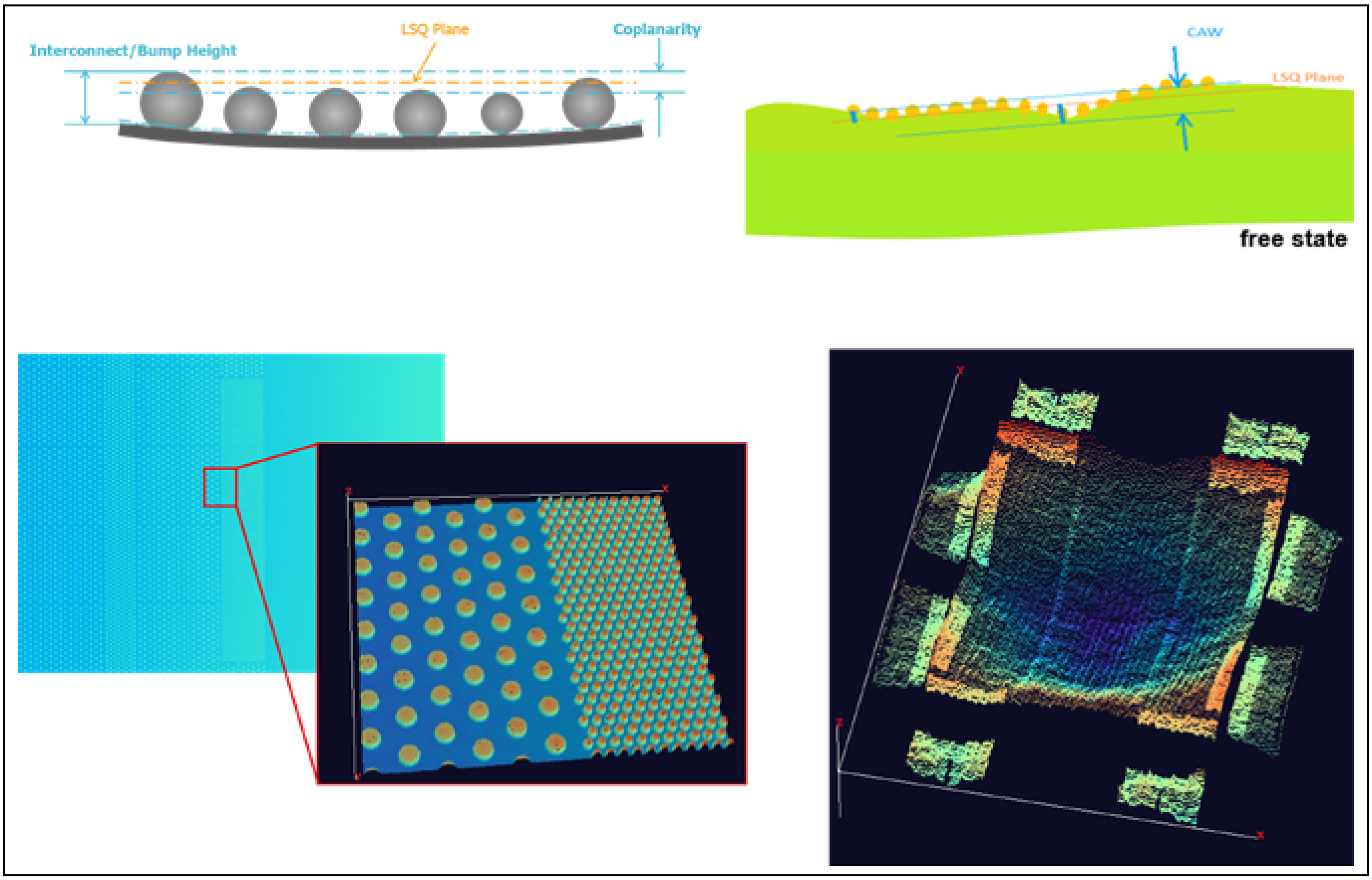


Figure 6: a) (top) Interconnect/bump height, coplanarity, and CAW calculations; and b) (bottom) fine- and coarse-pitch bump height and CAW measurement.

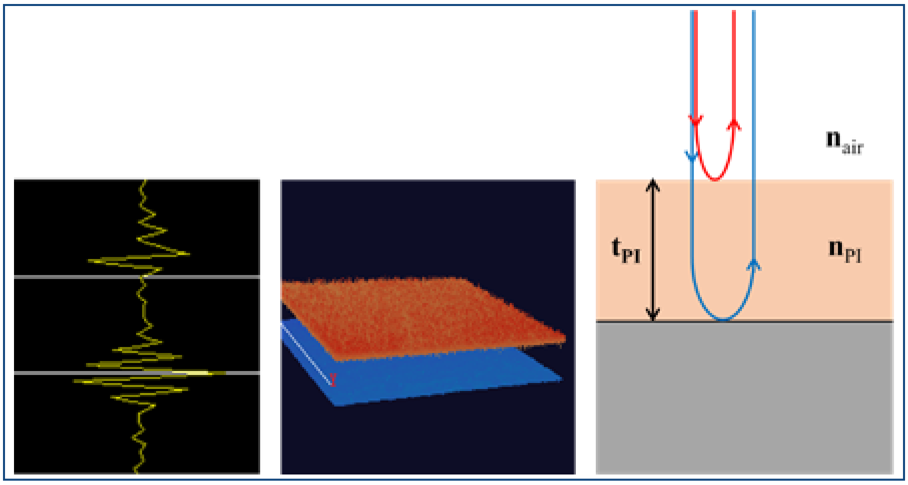


Figure 7: Interference patterns detected to calculate PI layer thickness during wafer-level 3D bump measurement.

over the bumps' tops and chip area's tops. Then, the highest and lowest points relative to the LSQ planes are calculated for the bumps' tops and chip area's tops. Finally, the distances of the highest and lowest points along the LSQ planes direction are separately measured as the bump coplanarity and CAW. According to the measurement procedure explained above, **Figure 6a** depicts interconnect/bump coplanarity and CAW calculations in the free-state condition. **Figure 6b** shows examples of bump height and CAW measurement by the WSI system.

CAW and bump coplanarity are critical parameters for the thermal compression bonding (TCB) process. TCB has been widely used for silicon chip attachment during IC packaging. **Figure 6a** illustrates interconnect/bump shape under the free-state condition, vs. the vacuum-state condition, which is related to the TCB process condition. The main challenge is how to assure successful bonding conditions before silicon chip attachment to the expensive multi-chip package. By measuring the described parameters and defining the correct specification limits, we will be able to avoid chip bonding yield loss caused by interconnect conditions that may result in non-wet or short bonding.

Multi-reflectance detection for transparent layer measurement

Another feature of our WSI system is multi-reflectance detection, which is demonstrated in **Figures 1** and **2**. Measuring interconnect/bump height with respect to the RDLs underneath, e.g., the PI layer, is crucial for chip bonding process control. PI, which is a transparent material, and other transparent dielectric materials, have been widely used during recent advanced package development. WSI can distinguish each layer of interest based on process requirements, as well as measure interconnect/bump reference to the selected layer.

Moreover, WSI has the capability to measure not only interconnect height, but also the thickness of RDLs, beneath which are fully or partially transparent layers so, therefore, we can detect light signals back from those layers. **Figure 7** shows an example of the PI layer thickness measurement during wafer-level 3D bump inspection. Light reflected from the bottom of the PI layer has a different thickness compared to the real PI thickness, t_{PI}, because of the PI material's refractive index, n_{PI}. The WSI-measured thickness is equal to the optical path, $t_{PI} \times n_{PI}$.

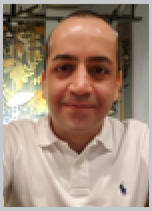
Summary

Heterogeneous package integration is going to be the main driver for semiconductor packaging in a variety of applications from system in package (SiP) to chiplet packaging. The main challenge for these expensive multi-chip packages remains how to keep assembly yield as high as possible so we can still enable the economic validity of Moore's Law—even 50-plus years after the invention of integrated circuits. The chip bonding process is the center of IC packaging, including heterogeneous integration, where multi-chips are connected to each other for chiplet applications. Therefore, keeping chip attachment yield loss as low as possible is necessary to enable heterogeneous packaging, thereby driving the industry forward.

Our WSI technology shows promising results with respect to process control and the highest yield die bonding process—both of which are necessary to enable these expensive multi-chip advanced packages to be successful products in the marketplace. In addition, the feature of transparent layer detection and thickness measurement of RDLs while inspecting interconnect height using WSI brings another dimension to advanced packaging process control.

References

1. P. Wesling, et al., "Heterogeneous integration roadmap, Chapter 8: Single chip and multi-chip integration," IEEE 2019 Ed.
2. S. Chitchian, "A deep-learning solution for heterogeneous package inspection," *Chip Scale Review* 24 (5), 2020.
3. D. Malacara, et al., "Interferogram analysis for optical testing," CRC, 2005.
4. K.G. Larkin, "Efficient nonlinear algorithm for envelope detection in white light interferometry," J. of the Optical Soc. of America A 13 (4), 1996.
5. M. Ishihara, et al., "High-speed surface measurement using a non-scanning multi-beam confocal microscope," Optical Eng. 38, 1999.



Biography

Shahab Chitchian is Chief Strategy Officer and R&D Corporate VP at INTEKPLUS CO., LTD., Daejeon, South Korea. Previously, he was Senior Staff Process Engineer at Samsung Electronics, Semiconductor Test and System Package division where he worked on advanced FOWLP development and high-bandwidth memory packaging. He also was a Senior Process Engineer at Intel Co., Assembly and Test Technology Development (ATTD), where he worked on EMIB package development. He received his MSc and PhD degrees in Electrical and Optical Engineering, from the U. of North Carolina at Charlotte. Email shahab@intekplus.com

Bridges for chiplet design and heterogeneous integration packaging

By John H. Lau [Unimicron Technology Corporation]

There are many advanced packaging technologies [1] listed in Figure 1 along with their performance and density rankings. Figure 2 shows the groups of packaging. The focus of this paper is on chiplets and heterogeneous integration.

The major difference between chiplet and heterogeneous integration is that a "chiplet" is a chip design method, while a "heterogeneous integration" is a chip packaging method [1]. The advantages of chiplet design and heterogeneous integration packaging are: a) yield improvement (lower cost) during semiconductor manufacturing; b) fast time-to-market; c) cost reduction during design; d) better thermal performance; e) reusable intellectual property (IP); and f) modularization. The disadvantages (challenges) are: a) additional area for interfaces; b) higher packaging costs; c) more packaging complexity and design effort; and d) past methodologies are less suitable. Therefore, one of the major focuses of packaging technologists is to reduce the area for interfaces of the lateral communication between chiplets and the packaging cost. In this brief article, various kinds of bridges for the lateral communications of chiplet design and heterogeneous integration packaging such as a) those embedded on top of an organic package substrate, b) those embedded in fan-out epoxy molding compound (EMC), and c) flexible bridge, will be systematically presented and discussed. Some challenges and recommendations will also be provided.

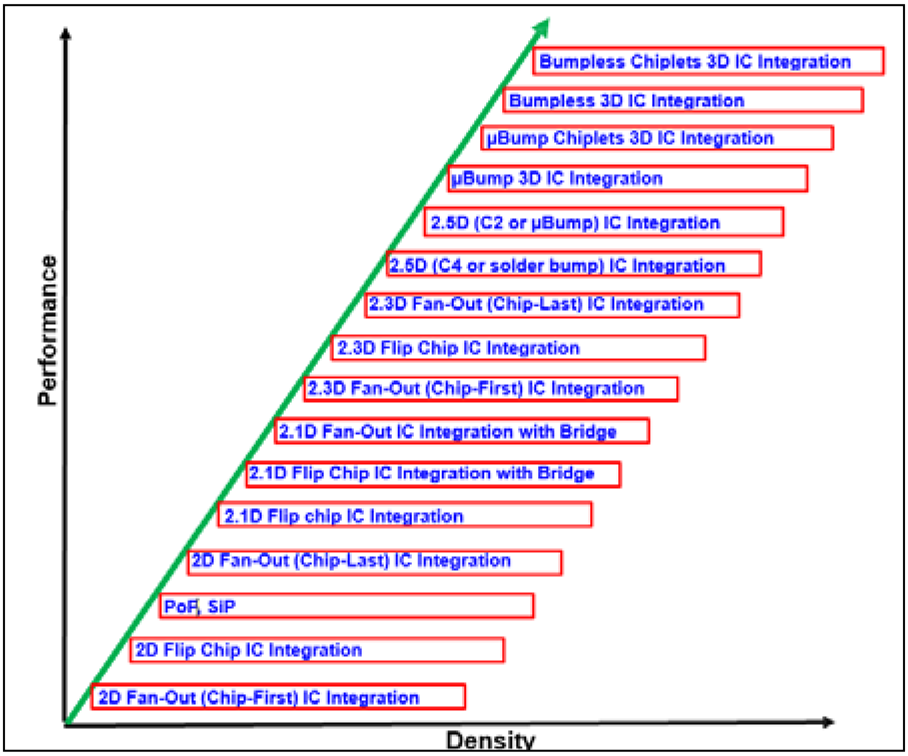


Figure 1: Advanced packaging ranking in density and performance. SOURCE: Unimicron

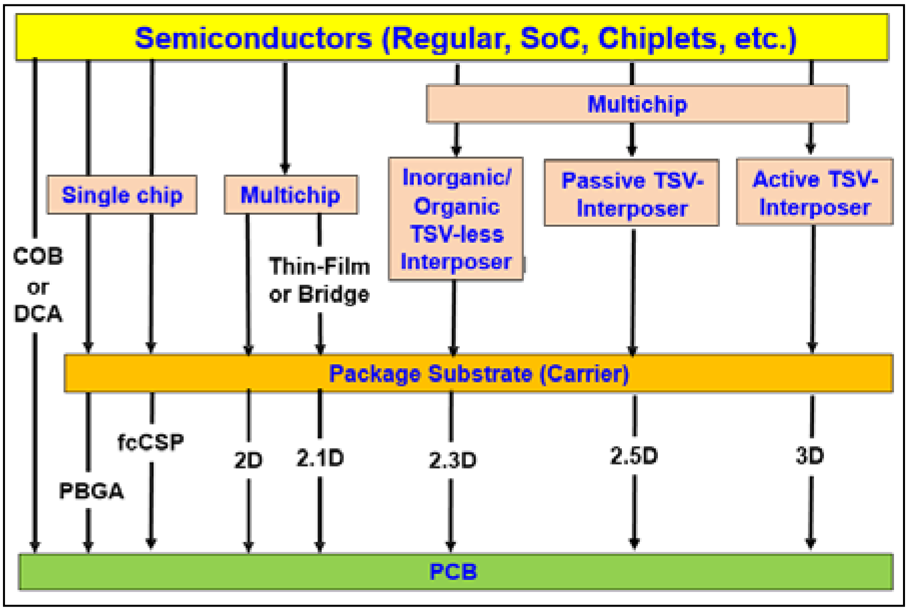
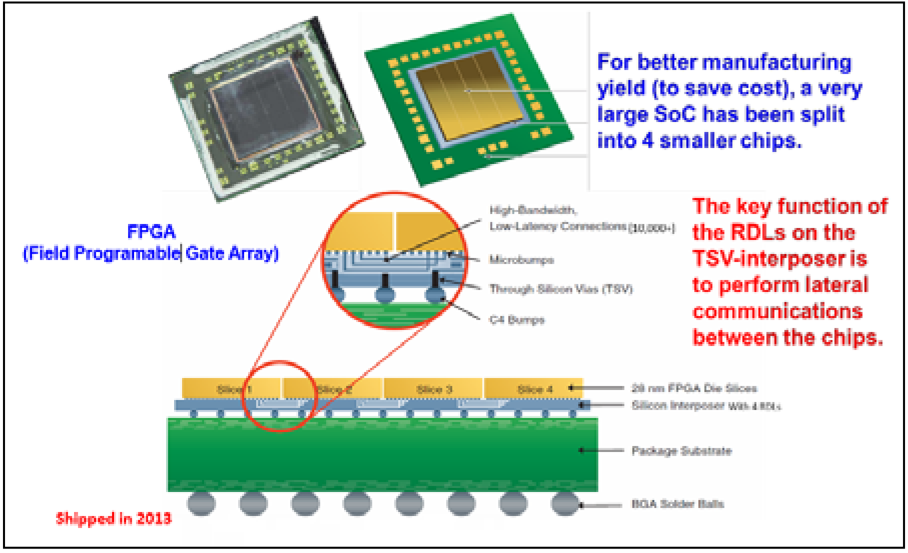


Figure 2: Groups of advanced packaging: 2D, 2.1D, 2.3D, 2.5D, and 3D IC integration. SOURCE: Unimicron

Background

In the past, the lateral communication of chiplet design and heterogeneous integration packaging is by fine metal line width and spacing (L/S) through-silicon via (TSV)-interposer and build-up high-density organic substrate. Figure 3 shows the Virtex-7 HT family shipped by Xilinx in 2013. In 2011, Xilinx asked TSMC to fabricate its field-programable gate array (FPGA) system-on-chip (SoC) with 28nm process technology [2,3]. Because of the large chip size, the yield was very poor. Then, Xilinx redesigned and split the large FPGA into four smaller



For better manufacturing yield (to save cost), a very large SoC has been split into 4 smaller chips.

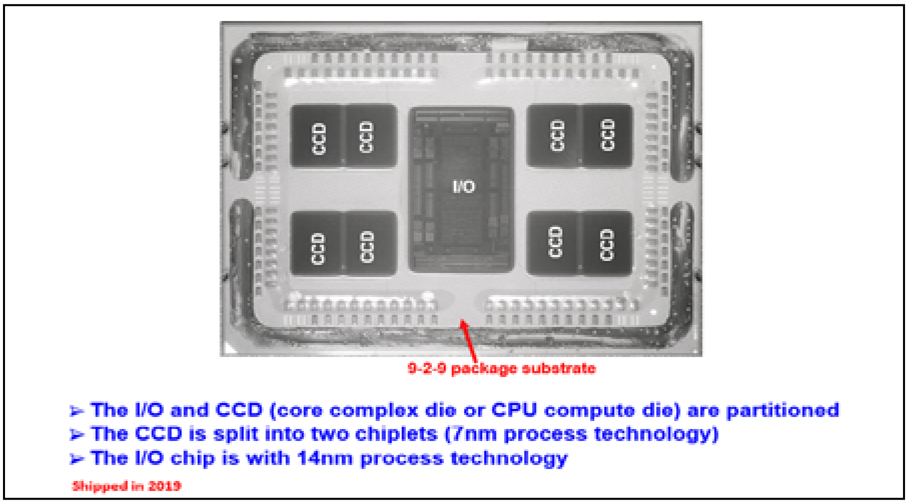
FPGA (Field Programable Gate Array)

The key function of the RDLs on the TSV-interposer is to perform lateral communications between the chips.

Shipped in 2013

High-Bandwidth, Low-Latency Connections (10,000+)
Microbumps
Through Silicon Vias (TSV)
C4 Bumps
Slice 1
Slice 2
Slice 3
Slice 4
28 nm FPGA Die Slices
Silicon Interposer With 4 RDLs
Package Substrate
BGA Solder Balls

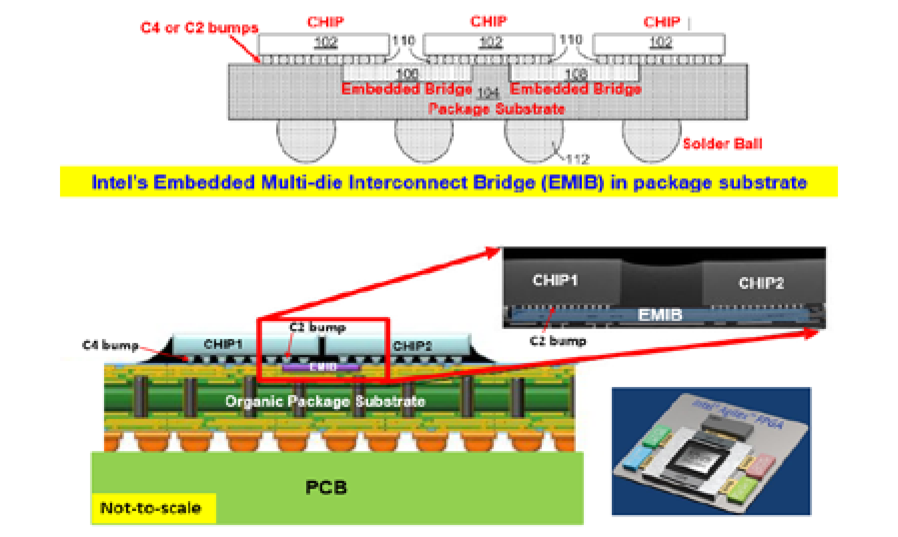
Figure 3: Xilinx's chiplet design and heterogeneous integration packaging. SOURCE: Xilinx [3]



- The I/O and CCD (core complex die or CPU compute die) are partitioned
- The CCD is split into two chiplets (7nm process technology)
- The I/O chip is with 14nm process technology

Shipped in 2019

Figure 4: AMD's chiplet design and heterogeneous integration packaging. SOURCE: AMD [4]



Intel's Embedded Multi-die Interconnect Bridge (EMIB) in package substrate

Figure 5: Intel's EMIB patent and FPGA module. SOURCE: Intel [7]

chiplets as shown in Figure 3 and TSMC manufactured the chiplets at high yield (with the 28nm process technology) and packaged them on their chip-on-wafer-on-substrate (CoWoS) technology. CoWoS is a 2.5D IC integration, which is the key structure (substrate) to let those 4 chiplets do lateral communications. The minimum pitch of the four redistribution layers (RDLs) on the TSV-interposer is 0.4µm. The TSV-interposer is known to have a very high cost.

Figure 4 shows AMD's 2nd-generation extreme-performance yield computing (EPYC) server processors [4,5], the 7002-series, shipped in mid-2019. One of AMD's solutions is to partition the SoC into chiplets, reserving the expensive leading-edge silicon for the central processing unit (CPU) core while leaving the I/Os and memory interfaces in n-1 generation silicon. Another solution is to split the CPU core into smaller chiplets. In this case, each core complex die (CCD), or CPU compute die, is split into two smaller chiplets. AMD used the expensive 7nm process technology fabricated by TSMC (in early 2019) for the core CCD chiplets and moved the dynamic random access memory (DRAM) and logic to a mature 14nm I/O die fabricated by GlobalFoundries. The 2nd-generation EPYC is a 2D chiplets IC integration technology, i.e., all the chiplets are side-by-side on a 9-2-9 build-up package substrate. The 20-layer fine metal L/S organic substrate is not cheap.

It should be noted that the requirement of lateral communications (RDLs) between chipets is fine-metal L/S/H (thickness) and at a very small and local area of the chiplets. There is no reason to use the whole TSV-interposer or the whole organic substrate to support the lateral communication between chiplets. Therefore, the concept of using small area and a fine-metal L/S/H RDLs bridge to connect the chiplets to perform lateral communication (to reduce cost) for chiplet design and heterogeneous integration packaging has been proposed and is a very hot topic today. There are at least two different groups of bridge, namely rigid bridge and flexible bridge.

Rigid bridge

The RDLs of most rigid bridges are fabricated on a silicon wafer. The most famous rigid bridge is Intel's EMIB (embedded multi-die interconnect bridge) [6-8]. For EMIB, there are at least three important tasks, namely: a) wafer bumping of two different kinds of bumps on the chiplets wafer (but there are not bumps on the bridge); b) embedding the bridge in the cavity of a build-up substrate and then laminating the top surface of the substrate;

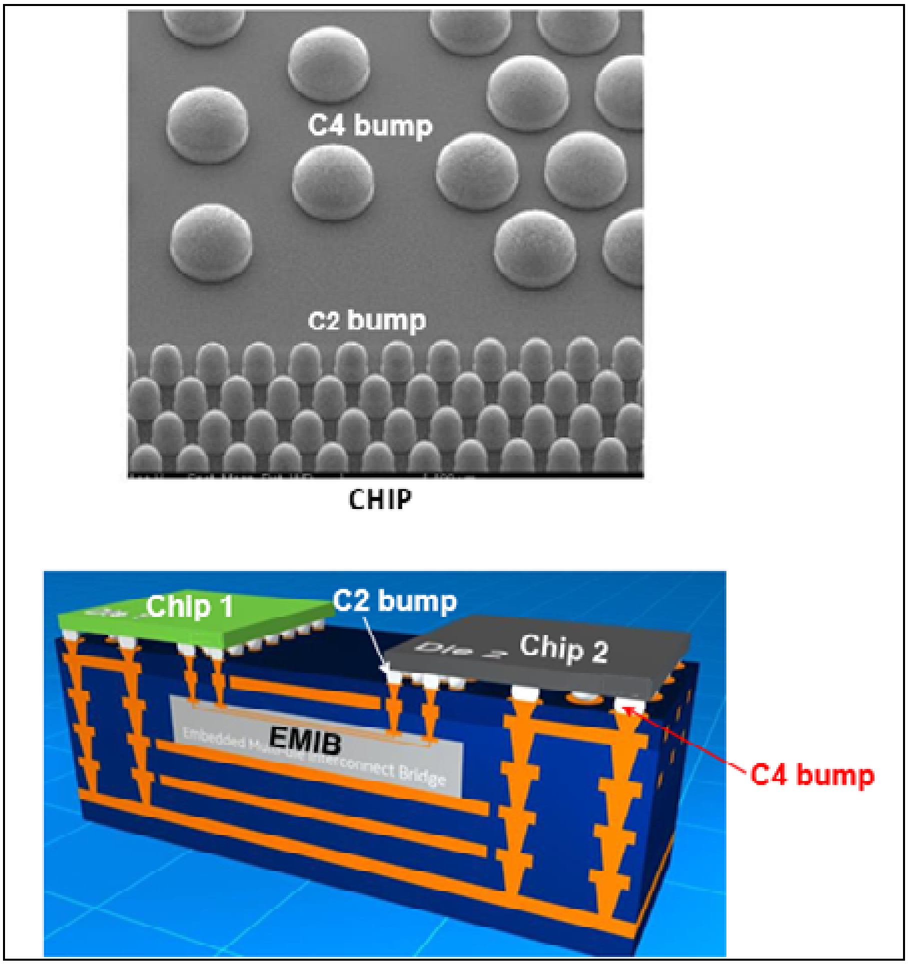


Figure 6: Intel's EMIB wafer bumping and substrate. SOURCE: Intel [7]

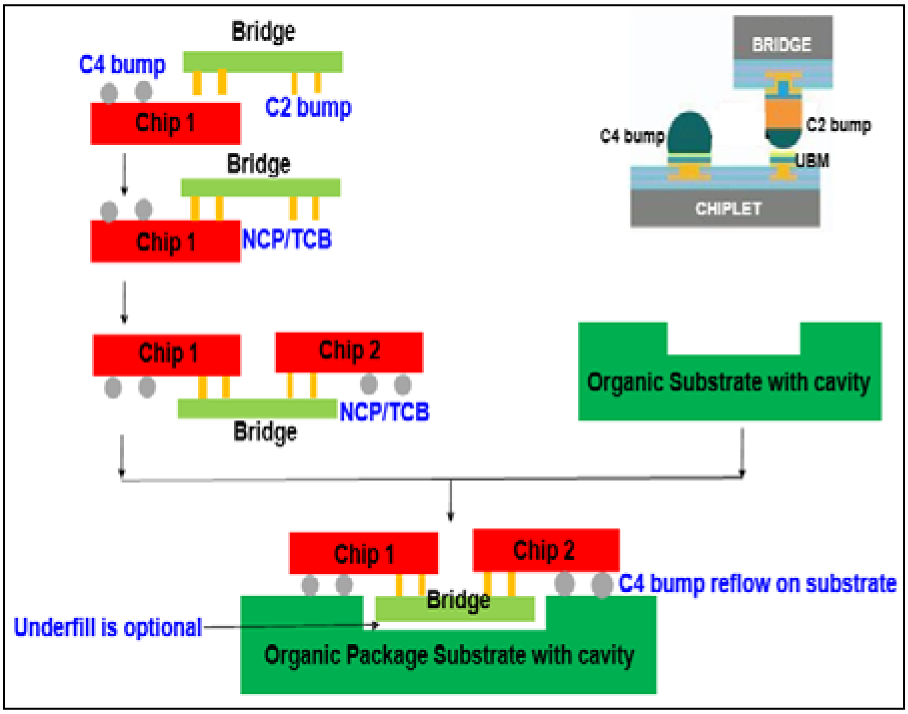


Figure 7: IBM's DHBi wafer bumping and process. SOURCE: IBM [9]

and c) bonding the chiplets on the substrate with the embedded bridge.

Figures 5 and 6 show one of Intel's patents [6], the EMIB substrate [7,8], bumps, and the Agilex FPGA module. It can be seen that there are two kinds of bumps on the chiplet, namely the C4 (controlled collapse chip connection) bumps and the C2 (chip connection or copper-pillar with solder-cap micro) bumps. Thus, wafer bumping of the chiplets wafer poses a challenge, but Intel has already taken care of this issue.

It can also be seen from Figures 5 and 6 that the FPGA and other chips are attached on top of a build-up package substrate with EMIB with fine-metal L/S/H RDLs. Today, the minimum metal L/S/H is 2µm/2µm/2µm and the bridge size is from 2mm x 2mm to 8mm x 8mm [6], but most are less than 5mm x 5mm [7]. The dielectric layer thickness is 2µm. Usually, there are ≤4 RDLs. One of the challenges of the EMIB technology is to fabricate the organic build-up package substrate with cavities for the silicon bridges and then laminate (with pressure and temperature) another build-up layer on top (to meet the substrate surface flatness requirement) for chiplets (with both C2 and C4 bumps) bonding. Intel and its suppliers are working toward high-yield manufacturing of the substrate.

A few months ago, Intel published a paper at IEEE/ECTC 2021 [8] that pointed out the bonding challenges of chiplets:

- Die bonding process;
- Manufacturing throughput;
- Die warpage;
- Interface quality;
- Die attach film material design;
- Die shift;
- Via-to-die-pad overlay alignment; and
- Integrated process considerations.

During IEEE/ECTC2021, IBM presented a paper on "Direct Bonded Heterogeneous Integration (DBHi) Si Bridge" [9]. The major differences between Intel's EMIB and IBM's DBHi are as follows:

- For Intel's EMIB, there are two different (C4 and C2) bumps on the chiplets (and there are no bumps on the bridge), while for IBM's DBHi, there are C4 bumps on the chiplets and C2 bumps on the bridge (Figure 7).
- For Intel's EMIB, the bridge is embedded in the cavity of a build-up substrate with a die-attach material and then laminated with another build-up layer on top. Therefore, the substrate fabrication is very complicated. For IBM's DBHi, the

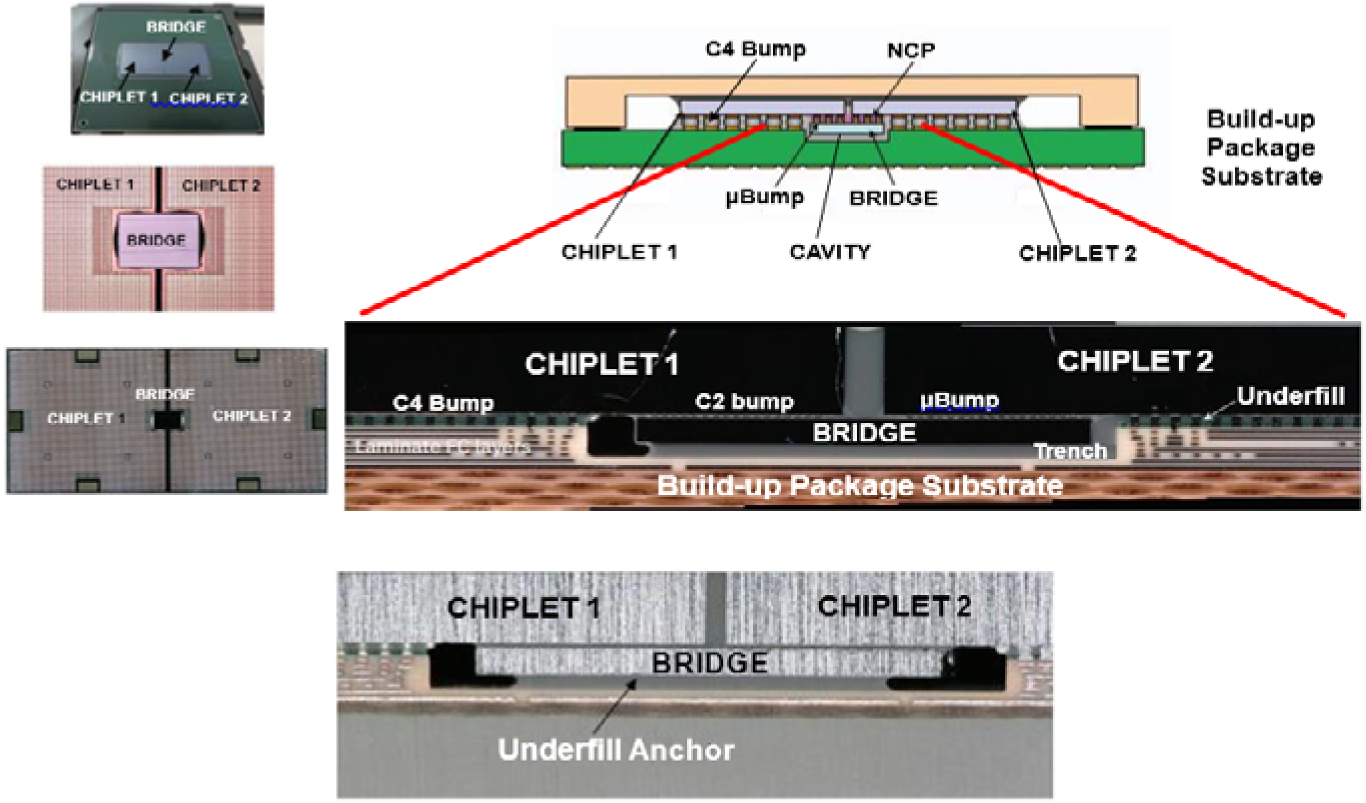


Figure 8: IBM's test vehicle and demonstration. SOURCE: IBM [9]

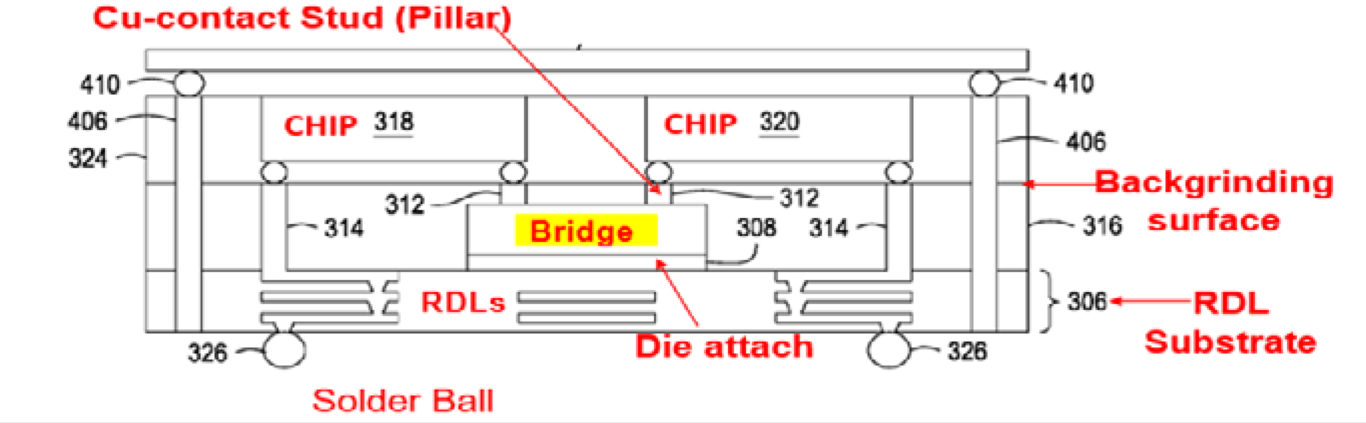


Figure 9: Applied Materials patent: fan-out chip (bridge) first and face-up process. SOURCE: US PATENT 10,651,126

substrate is just a regular build-up substrate with a cavity on top.

The bonding assembly process of DBHi is very simple (Figure 7). First, bond the Chip 1 and the bridge with nonconductive paste (NCP) and thermal compression bonding (TCB). Then, bond Chip 2 and the bridge with NCP and TCB. Those steps are followed by placing the module (Chip 1 + bridge + Chip 2) on the organic substrate with a cavity and then going through the standard flip-chip reflow assembly process. The underfill under the bridge is optional. Figure 8 shows the demonstration by IBM [9].

The challenges in IBM's DBHi are:

1. Handling and bonding of a portion of the tiny rigid bridge on a portion of the large chiplet with very fine-pitch pads;
2. Dealing with a situation in which there are more than one rigid bridge on a chiplet; and
3. Dealing with a situation in which there are more than two chiplets on a package substrate.

Intel's and IBM's rigid bridges are either embedded in, or are on an organic package substrate. There is another class of rigid bridge, which is embedded in the fan-out EMC. On May 12, 2020, Applied Materials obtained US patent 10,651,126 (filed on December 8, 2017). The company's design embedded the bridge in EMC by the fan-out chip (bridge) first and die face-up process (Figure 9). This could be the very first patent of a rigid bridge embedded in fan-out EMC.

On August 25, 2020, during TSMC's Annual Technology Symposium, the company announced its integrated fan-out

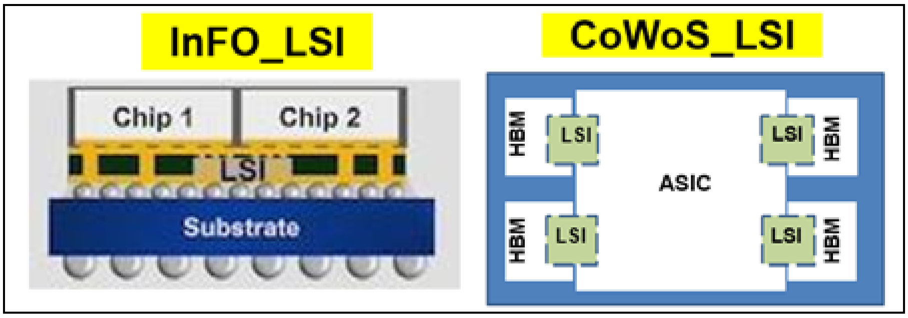


Figure 10: TSMC's InFO_LSI and CoWoS®_LSI. SOURCE: "Highlights of the TSMC Technology Symposium – Part 2," Tom Dillinger, SemiWiki, 9/7/2020

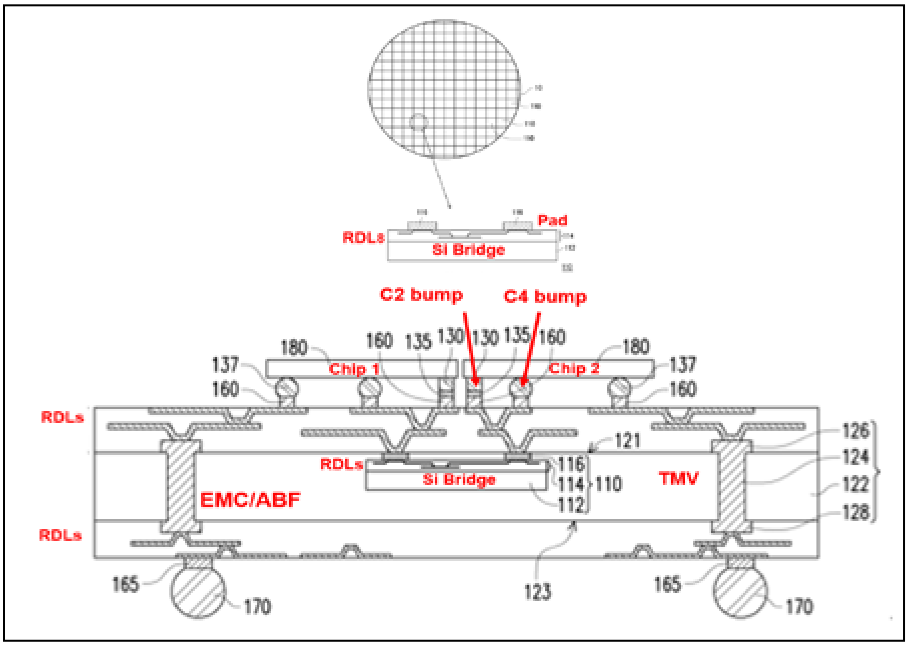


Figure 11: Unimicron patent application: fan-out chip-(bridge) first and face-down process.

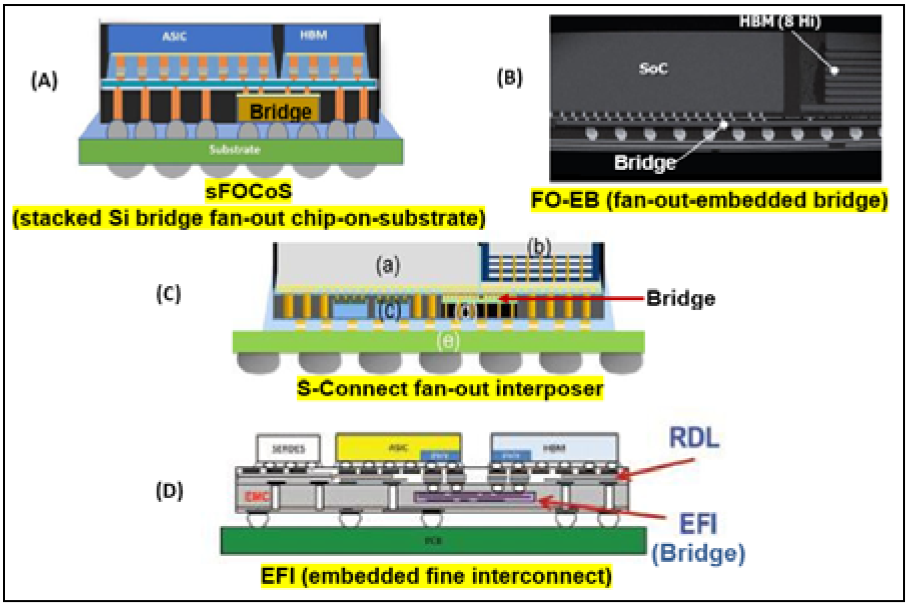

(A)
ASIC
HBM
Bridge
Substrate
sFOCoS
(stacked Si bridge fan-out chip-on-substrate)
(B)
HBM (8 Hi)
SoC
Bridge
FO-EB (fan-out-embedded bridge)
(C)
(a)
(b)
(c)
(e)
Bridge
S-Connect fan-out interposer
(D)
SERDES
ASIC
HBM
EMC
PCB
RDL
EFI
(Bridge)
EFI (embedded fine interconnect)

Figure 12: a) ASE sFOCoS; b) SPIL FO-EB; c) Amkor S-connect fan-out interposer; and d) IME EFI. SOURCES: ASE [10], SPIL [11], Amkor [12], and IME [13]

local silicon interconnect (InFO_LSI) and Chip-on-Wafer-on-Substrate local silicon interconnect (CoWoS®_LSI) (Figure 10). On May 7, 2021, Unimicron applied for a U.S. patent in which the bridge is embedded in the fan-out EMC by the chip (bridge) first and die face-down process (Figure 11).

During IEEE/ECTC (June 2021), there were at least four papers published regarding the application of fan-out packaging technology to embed the rigid bridge in the EMC for the chiplets to perform lateral communications. All four of these papers discuss very similar technologies. In [10], ASE embedded the bridge in the EMC using the fan-out packaging method and called it stacked Si bridge fan-out chip-on-substrate (sFOCoS) (Figure 12a). In [11], SPIL called its similar technology fan-out-embedded bridge (FO-EB) (Figure 12b). In [12], Amkor referred to its comparable technology as S-Connect fan-out interposer (Figure 12c). In [13], IME presented its bridge and called it embedded fine interconnect (EFI) (Figure 12d).

Flexible bridge

In addition to the rigid bridges embedded in build-up organic substrate (e.g., EMIB and DBHi) and fan-out EMC (e.g., Applied Materials, TSMC, Unimicron, ASE, Amkor, SPIL, and IME), there is the flexible bridge, which is the RDL itself. The flexible bridge consists of the fine-metal L/S/H conductors in a dielectric polymer, such as polyimide film [14]. The very first flexible bridge patent application US 2006/0095639 A1 was filed by SUN Microsystems on November 2, 2004 (Figure 13). For high-speed and high-frequency applications such as millimeter wave frequencies, the dielectric layer can also be a liquid crystal polymer and is called LCP-flexible bridge.

The assembly process of flexible bridge is very simple and very similar to IBM's DBHi as shown in Figure 14. However, both the C4 bumps and C2 bumps should be on the chiplet (just like Intel's EMIB case). This is because it is very difficult to do wafer bumping on a flexible bridge. The biggest challenge of the flexible bridge is handling the chiplets and flexible bridges during bonding. Also, there are other challenges if there are more than one flexible bridge on a chiplet and there are more than one chiplet with multiple flexible bridges.

Summary

Some important results and recommendations are summarized as follows:

WinWay Technology

Your trusted partner in IC testing

E-Flux

Extremely High Heat-Flux Thermal Controller

- Support high power chips testing in various temperature conditions
- All in one system benefits convenient operation
- Long-lasting operation & reliable performance

Temperature Range	Performance
-40°C to 150°C	1500W (60°C ~150°C)

Ramping Rate	
Heating > 0.7 °C/sec	Cooling > 0.45 °C/sec

✉ sales@winwayglobal.com

www.winwayglobal.com

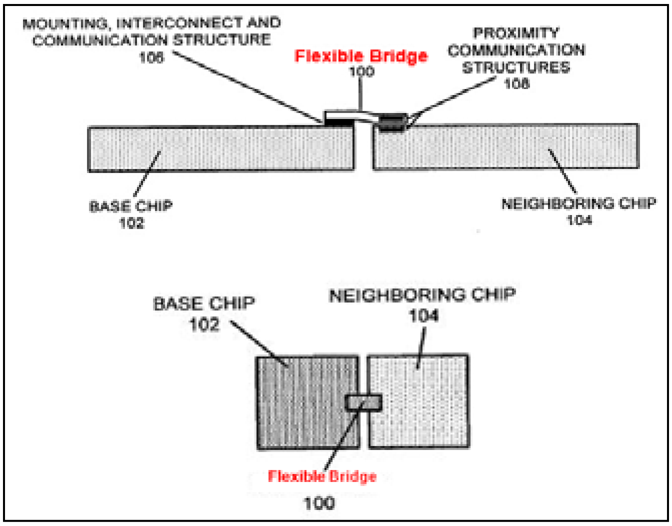


Figure 13: SUN Microsystems patent application: flexible bridge. SOURCE: US PATENT Application 2006/0095639 A1

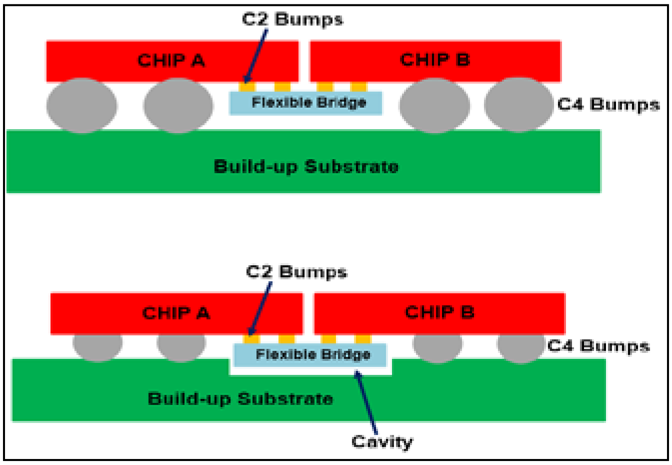


Figure 14: Flexible bridge for chiplet design and heterogeneous integration packaging. SOURCE: Unimicron

- Advanced packaging technologies have been ranked according to their density and performance and grouped into 2D, 2.1D, 2.3D, 2.5D, and 3D IC integration.
- Chiplet is a chip design method, while heterogeneous integration is a chip packaging method.
- The key advantages of chiplet design and heterogeneous integration packaging are: 1) yield improvement (lower cost) during manufacturing, 2) fast time-to-market, 3) cost reduction during design, 4) better thermal performance, 5) reuse of IP, and 6) modularization. The key disadvantages are: 1) additional area for interfaces, 2) higher packaging costs, 3) more complexity and design effort, and 4) past methodologies are less suitable for chiplets.
- There are two groups of bridges: rigid bridge and flexible bridge.
- For rigid bridges, the RDLs are fabricated on a silicon wafer substrate. Today, the rigid bridges are embedded on an organic package substrate such as the EMIB and DBHi, and embedded in fan-out EMC, such as those by Applied Materials, TSMC, Unimicron, ASE, Amkor, SPIL, and IME.
- For a flexible bridge, the RDL comprises the conductor layer and the polyimide dielectric layer. For 5G millimeter wave high-frequency applications, it is recommended to replace the polyimide with the liquid crystal polymer (LCP), i.e., a LCP-flexible bridge.
- The challenges of various bridges have been provided.

References

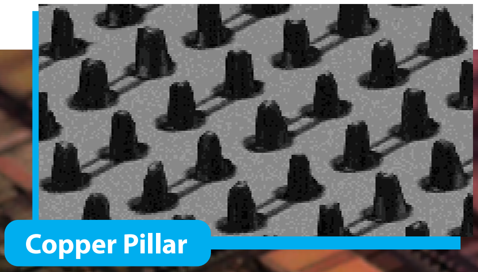
1. J. H. Lau, *Semiconductor Advanced Packaging*, Springer, New York, 2021.
2. http://press.xilinx.com/2013-10-20-Xilinx-and-TSMCReach-Volume-Production-on-all-28nm-CoWoS-based-All-Programmable-3D-IC-Families
3. B. Banijamali, C. Chiu, C. Hsieh, T. Lin, C. Hu, S. Hou, et al., "Reliability evaluation of a CoWoS-enabled 3D IC package," IEEE/ECTC Proc., May 2013, pp. 35-40.
4. S. Naffziger, K. Lepak, M. Paraschour, M. Subramony, "AMD chiplet architecture for high-performance server and desktop products," IEEE/ISSCC Proc., Feb. 2020, pp. 44-45.
5. S. Naffziger, "Chiplet meets the real world: benefits and limits of chiplet designs," Symp. on VLSI Tech. and Circuits, June 2020, pp. 1-39.
6. C. Chiu, Z. Qian, M. Manusharow, "Bridge interconnect with air gap in package assembly," US Patent No. 8,872,349, 2014.
7. R. Mahajan, R. Sankman, N. Patel, D. Kim, K. Aygun, Z. Qian, et al., "Embedded multi-die interconnect bridge (EMIB) – a high-density, high-bandwidth packaging interconnect," IEEE/ECTC Proc., May 2016, pp. 557-565.
8. G. Duan, Y. Knaoka, R. McRee, "Die embedded challenges for EMIB advanced packaging technology," IEEE/ECTC Proc., May 2021, pp. 1-7.
9. K. Sikka, R. Bonam, Y. Liu, P. Andry, D. Parekh, A. Jain, M. Bergendahl, et al., "Direct bonded heterogeneous integration (DBHi) Si bridge," IEEE/ECTC Proc., June 2021, pp. 136-147.
10. J. You, J. Li, D. Ho, J. Li, M. Zhuang, D. Lai, et al., "Electrical performances of fan-out embedded bridge," IEEE/ECTC Proc., June 2021, pp. 2030-2034.
11. J. Lee, G. Yong, M. Jeong, J. Jeon, D. Han, M. Lee, et al., "S-connect fan-out interposer for next-gen heterogeneous integration," IEEE/ECTC Proc., June 2021, pp. 96-100.
12. L. Lee, Y. Chang, S. Huang, J. On, E. Lin, O. Yang, "Advanced HDFO packaging solutions for chiplets integration in HPC application," IEEE/ECTC Proc., June 2021, pp. 8-13.
13. C. Chong, T. Lim, D. Ho, H. Yong, C. Choong, S. Lim, et al., "Heterogeneous integration with embedded fine interconnect," IEEE/ECTC Proc., June 2021, pp. 2216-2221.
14. J. H. Lau, *Handbook of Tape Automated Bonding*, McGraw-Hill Book Company, New York, 1992.

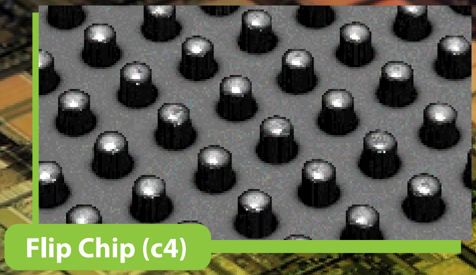
Biography

John H. Lau is a senior special project assistant at Unimicron Technology Corporation, Taoyuan City, Taiwan (ROC). He has more than 40 years of R&D and manufacturing experience in semiconductor packaging, 511 peer-reviewed papers, 40 issued and pending US patents, and 22 textbooks. He is an ASME Fellow, IEEE Fellow, and IMAPS Fellow. He earned a PhD degree from the U. of Illinois at Urbana-Champaign. Email John_Lau@unimicron.com

Seriously Fast.

WX3000™ Metrology and Inspection Systems for Wafer-Level and Advanced Packaging





2-3X Faster with High Resolution and High Accuracy

WX3000 3D+2D metrology and inspection system provides the ultimate combination of high speed, high resolution and high accuracy for wafer-level and advanced packaging applications to improve yields and processes.

Powered by Multi-Reflection Suppression™ (MRS™) Sensor Technology

The 3-micron NanoResolution (X/Y resolution of 3 micron, Z resolution of 50 nanometer) MRS sensor enables metrology grade accuracy with superior 100% 3D and 2D measurement performance for features as small as 25-micron. 100% 3D and 2D metrology and inspection can be completed simultaneously at high speed (25 300mm wafers/hour and 55 200mm wafers/hour) as compared to a slow method that requires two separate scans for 2D and 3D, and only a sampling process.

CYBEROPTICS®

www.cyberoptics.com

Managing trade-offs in the chiplet era

By Rob Munoz [Intel]

In his famous 1965 paper [1] Gordon Moore foreshadowed that a chiplet-like approach would become attractive. By chiplets, we mean die that have been optimized to connect to other die within the same packaged device. Intel has used the term "tiles" to describe chiplets that are integrated using high-density, high-bandwidth interconnects enabled by advanced packaging technologies such as Intel's 2.5D embedded multi-die interconnect bridge (EMIB) [2], 3D Foveros [3], and combined EMIB-Foveros (Co-EMIB) [4]. **Figure 1** is an illustration of chiplet packaging and physical connectivity taxonomy. In this figure, AIB refers to the advanced interface bus chiplet interface standard [5] and HBMIO refers to the JESD235-specified high-bandwidth memory (HBM) interface standard [6].

Chiplets are poised to become "the new normal," especially when building chips (both merchant and custom) for data center and edge deployments. The IEEE, in collaboration with SEMI and ASME, has sponsored a Heterogeneous Integration Roadmap [7] effort to foster collaboration and technology preparedness. As is outlined in an earlier article [8], there are several key adoption and scaling challenges that must be addressed and key trade-offs that must be understood and properly managed to optimally take advantage of a chiplet approach. In this article, we'll delve more deeply into the trade-offs that were previously outlined.

Potential chiplet benefits

Key potential chiplet benefits include reducing portfolio (both product/solution and project) costs, helping scale innovation and delivery capabilities, and improving time to solution. Product/solution cost reduction is the most frequently mentioned potential chiplet benefit. This cost benefit is typically most pronounced in usages that require a large amount of silicon area, especially in the early years of manufacturing at a leading-edge process node. Smaller die will have fewer defects than larger die and can help achieve higher mask/lithographic field utilization (which improves manufacturing velocity and efficiency).

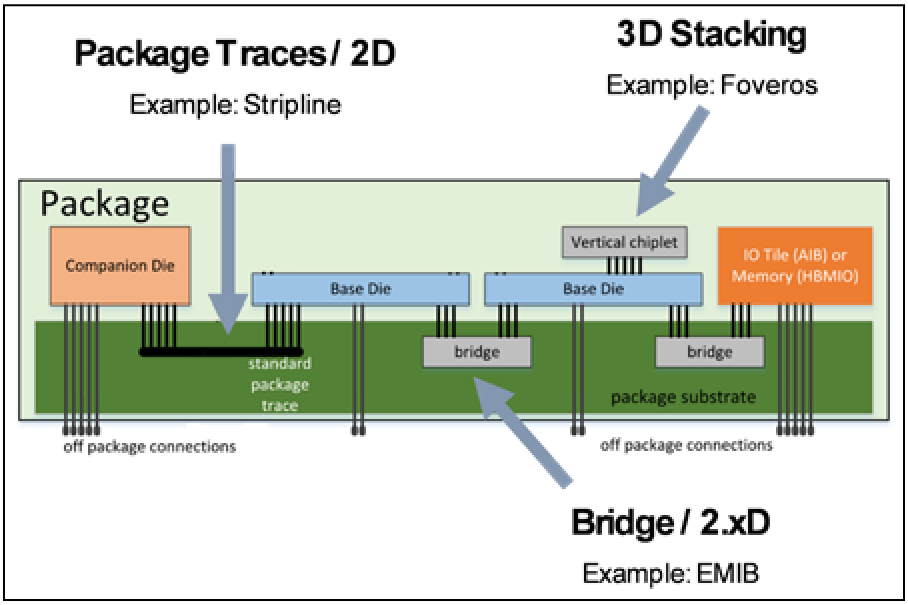


Figure 1: Chiplet packaging and physical connectivity taxonomy with examples.

Likewise, the fundamental value proposition of heterogeneous integration is that each chiplet can be built in (or ideally, reused from) a process node that is well targeted for its intended usage. For example, it is often preferable to keep analog intellectual property (IP) in a more mature process node because porting it can be difficult and time consuming, it often scales very poorly, and it typically has much more limited options for repair and redundancy to tolerate defects. Optical, radio frequency (RF), and other specialized analog IP may require process attributes that are not available in leading edge process nodes optimized for digital IP. Likewise, dynamic random access memory (DRAM) and embedded DRAM typically require specialized process nodes. Similarly, some process nodes may be heavily optimized for low-power operation. While such nodes may be ideal for implementing energy efficient accelerator IP that can be configured to go "slow and wide" when higher throughput is needed, these process nodes are typically unattractive for implementing high-performance general-purpose processors, high-speed memories, and high-speed I/O.

Furthermore, the number and type of chiplets populated in a chip can more closely match the configuration that the customer is nominally purchasing (e.g., for a processor, this might include core count and I/O capabilities). This reduces the amount of disabled or "dark" silicon that needs to be manufactured to fulfill customer orders. When manufacturing capacity is tight, the opportunity cost of wasting silicon area is much higher than the nominal accounting cost of this silicon. The above factors in combination can significantly improve efficiency and cost effectiveness per silicon wafer when using a chiplet approach.

Large chips are often very desirable when addressing the highest performance data center and high-performance computing (HPC) processing usages. We can use a chiplet approach to build chips with a silicon area that is substantially larger than the nominal reticle lithography exposure limit (currently 33mm x 26mm = 858mm^2; the IEEE IRDS chapter on lithography [9] has projected trends). However, the maximum practical size of a commercial volume chip will still be limited by thermal, mechanical warpage, etc., considerations.

Chiplets can also help reduce project costs, scale innovation and delivery capabilities, and reduce time to solution. Ideally, chiplets would largely be reusable within and between product/solution generations and product lines while supporting per segment and potentially per customer feature tailoring where it is desired. For example, a chiplet approach would help a hyperscalar deploy a particular and potentially proprietary type of machine learning accelerator uniformly (albeit perhaps at different performance scaling levels) in their cloud and edge solutions. Within a given product/solution generation, chiplets can be combined in different arrangements to create many combinations of useful chip and solution configurations. Customized solutions can be created that use or reuse existing chiplets created internally, by customers, and/or by third parties, each of whom can ideally develop and evolve chiplets they create asynchronously and independently. Given the rapidly rising cost of designing chips targeted for leading-edge process nodes (see Figure 2 for a rough estimate based on averaging previous IBS [10] and Gartner [11] estimates), it will be increasingly important to amortize these costs over a broader market opportunity of products/solutions to make new products/solutions economical.

Ideally, future high-volume standard product offerings will support one or more standardized chiplet attach "slots" to enable easy customization and high agility to adapt to rapidly changing customer and market needs. If these "slots" support protocols like Compute Express Link (CXL) and PCIe they can address a broad variety of transactional use cases (load/store data transfer via the PCIe or CXL.io protocols, memory access via the CXL.mem protocol, and cache coherent accelerator and I/O access via the CXL.cache protocol). Such an approach would provide a proven interoperability model for connecting processors, accelerators, memory, and external input/output interfaces together. Using CXL/PCIe also helps address system-on-chip (SoC) construction issues (e.g., address space configuration, reset, initialization, register access, etc.) and security considerations that might otherwise hinder interoperability and early adoption. Furthermore, this approach enables a uniform software model across a solution portfolio that blurs the distinctions between whether functions are integrated on die, in package, or at the board or system level. A chiplet approach can also significantly shrink the required printed circuit board (PCB) area needed for solutions. Space reduction benefits can be even more substantial when employing 3D packaging technologies like Intel Foveros and Foveros Omni.

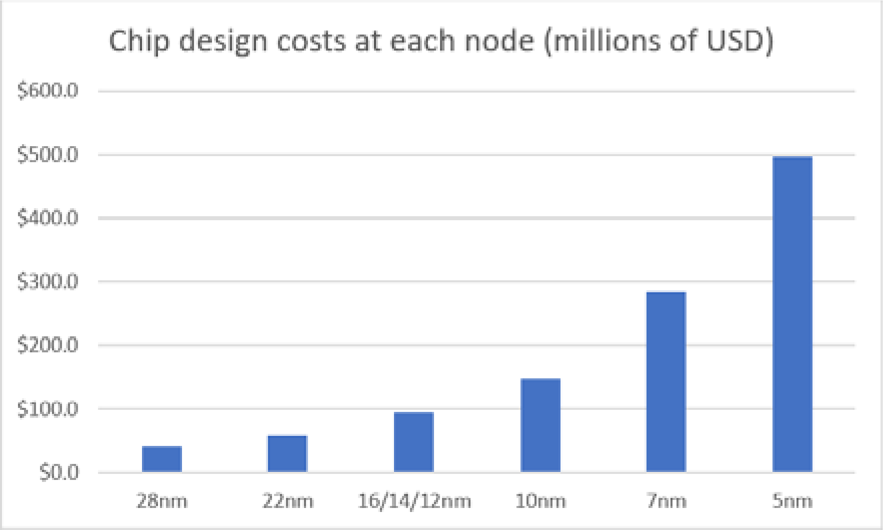


Figure 2: Rough estimate of conventional chip design costs (average of IBS [10] and Gartner [11] estimates).

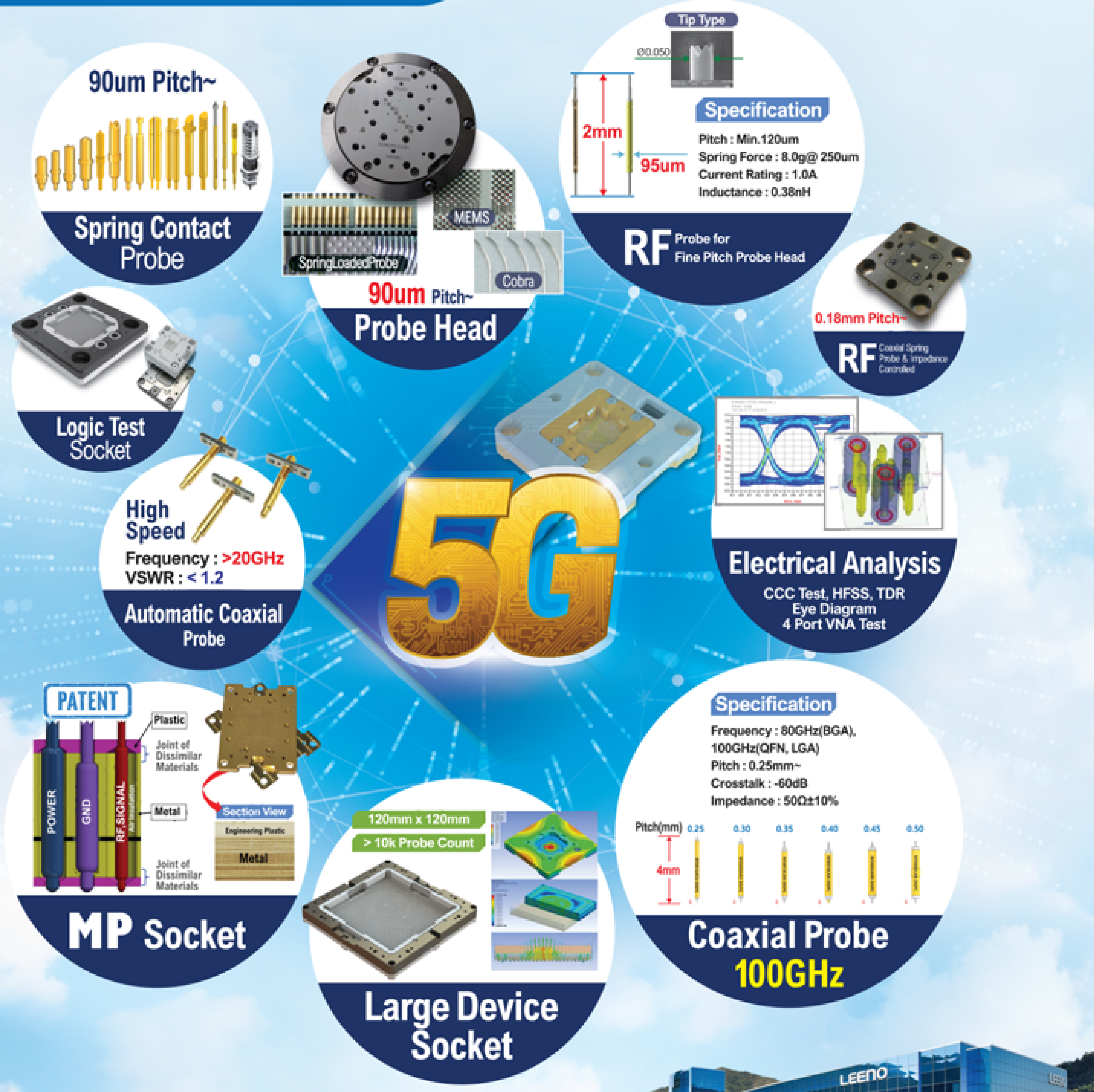
Potential chiplet disadvantages and associated mitigations

Unfortunately, there is no "free lunch" with chiplets. A key initial hurdle to achieve the full benefits of industry-scale systematic chiplet reuse is broad adoption of fully-specified interface standards (see Figure 3 for details). Intel is collaborating on a cross-industry effort to standardize a widely applicable chiplet interface supportable at all major foundries and outsourced assembly and test (OSAT) suppliers that supports CXL-based protocol connectivity [12] to help tackle this hurdle. While not all chiplets used in all solutions need to have a fully-standard interface, using industry or internal standards where possible will often reduce development and verification efforts and improve quality. The industry must also continue to collaborate on additional associated work in several important areas including design automation tooling, test [13], reliability [14], modeling and simulation [15], etc., as described in the IEEE Heterogeneous Integration Roadmap [7] and other places. Assuming this fundamental work on interoperability, modeling, tooling, etc., will be adequately addressed, when planning a chiplet-based solution portfolio there are still associated trade-offs around tiling overheads, thermal and input/output (I/O) escape constraints, and associated supply chain and economic considerations that must be managed.

The tiling overheads on power, performance, area, and cost are often the most visible potential disadvantages of using chiplets. The die-to-die interface that connects chiplets together in a package will typically consume more area, power, and performance overhead than hypothetical on-die connectivity in a monolithically-integrated alternative would require. While an efficient 2D/2.xD die-to-die chiplet interface can provide a 10x or better improvement in bandwidth area and shoreline density (as measured in GB/s/mm^2 and GB/s/mm, respectively) and energy efficiency (measured in pJ/bit) compared to current board-level PCIe and similar SerDes-based interconnects, tiling overhead is still higher than comparable figures for on-die interfaces. However, it is not necessarily always "fair" to directly compare a chiplet interconnect with an on-die interconnect because it may be impractical or impossible to even build

LEENO

www.leeno.com



90um Pitch~

Spring Contact Probe

SpringLoadedProbe

MEMS

Cobra

90um Pitch~

Probe Head

Tip Type

Ø0.050

2mm

95um

Specification

Pitch : Min.120um
Spring Force : 8.0g@ 250um
Current Rating : 1.0A
Inductance : 0.38nH

RF Probe for Fine Pitch Probe Head

0.18mm Pitch~

RF Coaxial Spring Probe & Impedance Controlled

Logic Test Socket

High Speed

Frequency : >20GHz
VSWR : < 1.2

Automatic Coaxial Probe

5G

Electrical Analysis

CCC Test, HFSS, TDR
Eye Diagram
4 Port VNA Test

PATENT

Plastic

Joint of Dissimilar Materials

Metal

Joint of Dissimilar Materials

POWER

GND

RF,SIGNAL

Section View

Engineering Plastic

Metal

MP Socket

120mm x 120mm

> 10k Probe Count

Large Device Socket

Specification

Frequency : 80GHz(BGA), 100GHz(QFN, LGA)
Pitch : 0.25mm~
Crosstalk : -60dB
Impedance : 50Ω±10%

Pitch(mm)	0.25	0.30	0.35	0.40	0.45	0.50

4mm

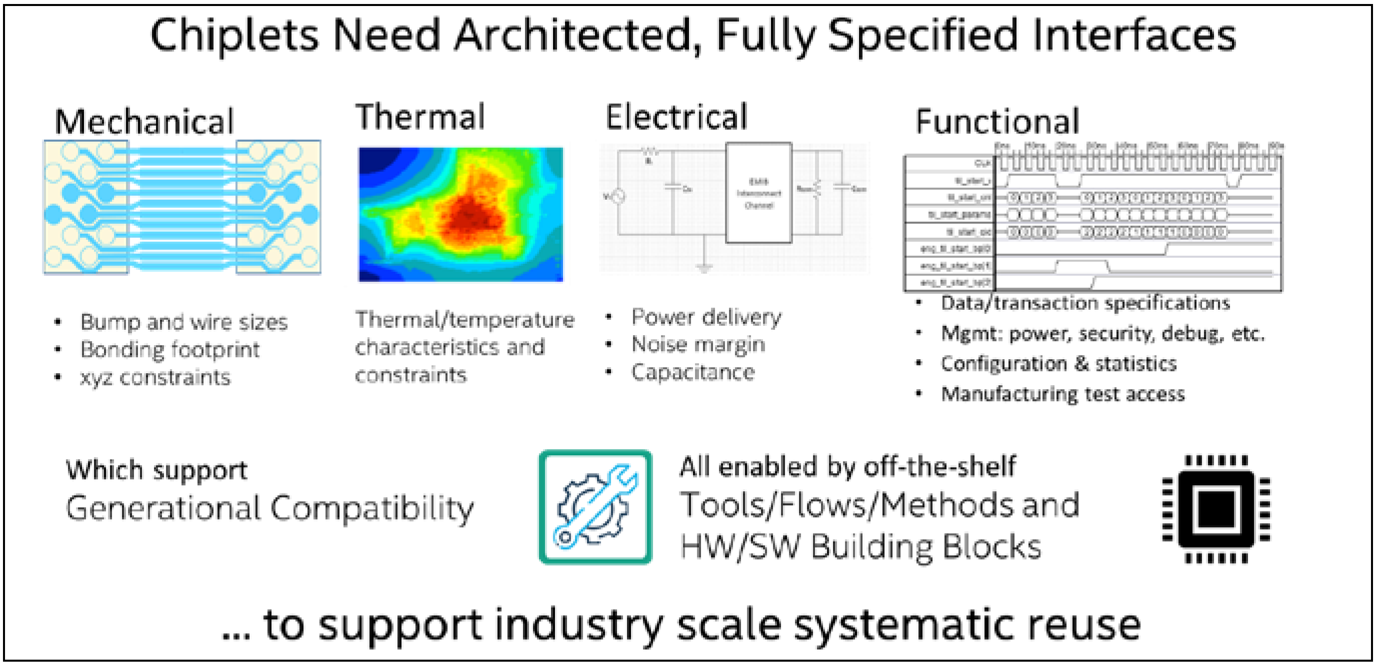
Coaxial Probe 100GHz

LEENO LEENO INDUSTRIAL INC.

GLOBAL LEADER LEENO

HEAD OFFICE
10 105beon-gil MieumSandan-ro Gangseo-gu, Busan, Korea

CONTACT
USA: hskang@leeno.co.kr / +1 408 313 2964 / +82 10 8737 6561
Korea: sales-leeno@leeno.co.kr / +82 51 792 5639



Chiplets Need Architected, Fully Specified Interfaces

Mechanical

- Bump and wire sizes
- Bonding footprint
- xyz constraints

Thermal

Thermal/temperature characteristics and constraints

Electrical

- Power delivery
- Noise margin
- Capacitance

Functional

- Data/transaction specifications
- Mgmt: power, security, debug, etc.
- Configuration & statistics
- Manufacturing test access

Which support Generational Compatibility

All enabled by off-the-shelf Tools/Flows/Methods and HW/SW Building Blocks

... to support industry scale systematic reuse

Figure 3: Chiplet interface specification requirements for industry scale adoption.

a monolithically-integrated alternative (due to die size-based yield or reticle size limitations, intellectual property portability/suitability considerations, development cost challenges to achieve all the required configuration variations, etc.). Still, chiplets are not necessarily going to be economically optimal in cases where a single "sweet spot" monolithic alternative is feasible and attractive.

Continuing advances in packaging and interconnect technology will help reduce tiling overheads over time. For example, Intel's Foveros Omni is projected to support bump pitches down to 25µm and Intel's Foveros Direct is projected to support bump pitches of 10µm or less. Each N-fold improvement in bump pitch and wire density scaling for advanced packaging potentially enables an N^2 improvement in bandwidth density. Likewise, as process technology improves and connection distance shrinks, we can reduce the voltages chiplet interfaces use to further improve energy efficiency. Chiplet interconnects should therefore, ideally, use a "many wires, simple I/O" philosophy rather than a SerDes-like "few wires and complex I/O" approach to take better advantage of upcoming improvements in connectivity technology.

A more subtle disadvantage of using chiplets is that packaging, assembly, and test costs and durations will generally be higher for multi-die chips than monolithic chips. Using chiplets that can be individually, comprehensively, and efficiently tested prior to assembly together with a chiplet interconnect and packaging integration approach that enables efficient and comprehensive post-assembly testing with an adequate degree of redundancy and repair can help mitigate this disadvantage. Likewise, not all chiplet solutions will necessarily require the high bandwidth densities that advanced packaging can provide, so ideally, the chiplet interconnect selected can support an option to use conventional low-cost 2D "standard package trace" packaging when that is economically optimal.

Thermal and mechanical constraints can also strongly influence system-level partitioning. It is often impractical to co-package multiple "hot" die together when doing so exceeds the thermal capabilities of the resulting system. This limitation can be especially severe in those industrial or far-edge usages that cannot rely on forced-air cooling. It is likewise similarly impractical to co-package die that each require a lot of external PCB connectivity because doing so will increase package-level and PCB routing requirements and associated costs. Instead, it is generally preferable to co-package functions that require high-bandwidth connectivity between themselves to take full advantage of the previously noted 10x improvements in connection density and energy efficiency while incrementally reducing communication latencies and improving system performance.

In production systems and supply chains, any additions to cycle time (processing and/or transit times) or supply chain uncertainty will often require additional inventory to be carried. Purchasing external chiplets typically incurs margin stacking and additional inventory carrying costs. The use of a common set of chiplets in many chip configurations can mitigate the impact of these considerations by enabling inventory pooling. Likewise, consignment arrangements can help mitigate some of the margin stacking and carrying cost impacts.

While 3D stacking can provide substantial savings in board space, interface density, and power efficiency, chiplets built to be 3D stacked with other chiplets have historically needed to be carefully co-designed (e.g., due to thermal, power delivery, signal integrity, etc., considerations), hindering the ability to reuse them in other contexts. However, as HBM memory has illustrated, we can usefully specify a standard 2D/2.5D interface to a 3D stack of chiplets that is supplied as an integrated subassembly. This consideration is a key reason why it is preferable to initially focus multivendor standardization efforts on 2D and 2.xD chiplet connectivity usages before attempting to standardize 3D interface details.

Summary

To summarize, a chiplet approach has tremendous potential to reduce portfolio (both product/solution and project) costs, help scale innovation and delivery capabilities, and improve time to solution. Assuming the industry successfully collaborates to address initial adoption and scaling challenges, planning a chiplet-based solution portfolio will still require understanding and managing the associated trade-offs around tiling overheads, thermal and input/output (I/O) escape constraints, and associated supply chain and economic considerations. By doing so, industry participants can effectively leverage a rich and innovative chiplet ecosystem to economically deliver a portfolio of innovative semiconductor-based solutions. Together, we can realize a future where high-performance chiplet express lanes interconnect the separately constructed and manufactured functions that Gordon Moore foreshadowed in 1965, fundamentally reshaping how our industry collaborates to build future systems.

Acknowledgments

The perspective I have shared is based on the work of Intel's internal Chiplet Work Group. I especially want to thank Dave Kehlet, Tanay Karnik, Ramune Nagisetty, Peter Onufryk, and other collaborators in this effort over the last 5+ years.

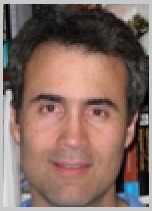
Legal disclaimer

Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Intel technologies features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and actual results may differ materially from those expressed or implied in such statements.

References

1. G. E. Moore, "Cramming more components onto integrated circuits," Electronics, Vol. 38, No. 8, Apr. 19, 1965, reprint downloadable from https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf
2. "Embedded Multi-Die Interconnect Bridge," https://www.intel.com/content/www/us/en/silicon-innovations/6-pillars/emib.html
3. "Up Close with Lakefield – Intel's Chip with Award-Winning Foveros 3D Tech," https://newsroom.intel.com/news/up-close-lakefield-intels-chip-award-winning-foveros-3d-tech
4. "The Revolutions that Led to Chips Made Like Quilts," June 30, 2021, https://www.intel.com/content/www/us/en/newsroom/news/revolutions-led-chips-made-quilts.html
5. AIB specification, https://github.com/chipsalliance/AIB-specification, retrieved Nov. 21, 2021.
6. [6] JESD235D: High Bandwidth Memory (HBM) DRAM, JEDEC, Mar. 2021, https://www.jedec.org/standards-documents/docs/jesd235a
7. Heterogeneous Integration Roadmap, IEEE Electronics Packaging Soc., https://eps.ieee.org/technology/heterogeneous-integration-roadmap.html
8. R. Munoz, "Scaling the Chiplet Adoption Wall," MEPTEC Report, Fall 2021, p. 14-18, https://issuu.com/mepcom/docs/meptec_report_fall_2021
9. Lithography 2021 update, International Roadmap for Devices and Systems, IEEE, 2021, https://irds.ieee.org/images/files/pdf/2021/2021IRDS_Litho.pdf
10. "FinFET and FD SOI: Market and Cost Analysis," International Business Strategies, Sept. 18, 2018, http://soiconsortium.eu/wp-content/uploads/2018/08/MS-FDSOI9.1818-cr.pdf
11. As quoted in Mark Lapedus, "Foundry Challenges in 2018," Semiconductor Engineering, Dec. 17, 2017, https://semiengineering.com/foundry-challenges-in-2018
12. "Intel Innovation: Event Keynote," Oct 27,2021, (replay) https://www.youtube.com/watch?v=3IZ9fpUUoqc (22:07-23:02)
13. P. Tadayon, et al., "Moore's Law and the Future of Test," *Chip Scale Review*, May – June 2021, https://www.chipscalereview.com/wp-content/uploads/2021/05/ChipScale_May-Jun_2021-Intel.pdf
14. IEEE Reliability Committee, IEEE Electronics Packaging Society, https://cmte.ieee.org/reliability
15. "Ch. 14: Modeling and Simulation," Heterogeneous Integration Roadmap, IEEE Electronics Packaging Soc. Sept. 2020, https://eps.ieee.org/images/files/HIR_2020/ch14_ms.pdf



Biography

Rob Munoz is a Principal Engineer in Intel's Design and Engineering group located in Austin, TX. His day job is architecture for custom/semi-custom/standard products targeted at wireless infrastructure. He also co-leads the cross-Intel Chiplet Working Group. He has been with Intel since November 2014, joining as part of Intel's acquisition of LSI/Avago's networking business. He started his career at Bell Labs, has an MS in Computer Science from UT Austin, and currently has 15 granted patents. Email robert.munoz@intel.com

Global No.1! Total Test Solution Provider!

GLOBAL INNOVATION

ISC

www.isc21.kr

iSC

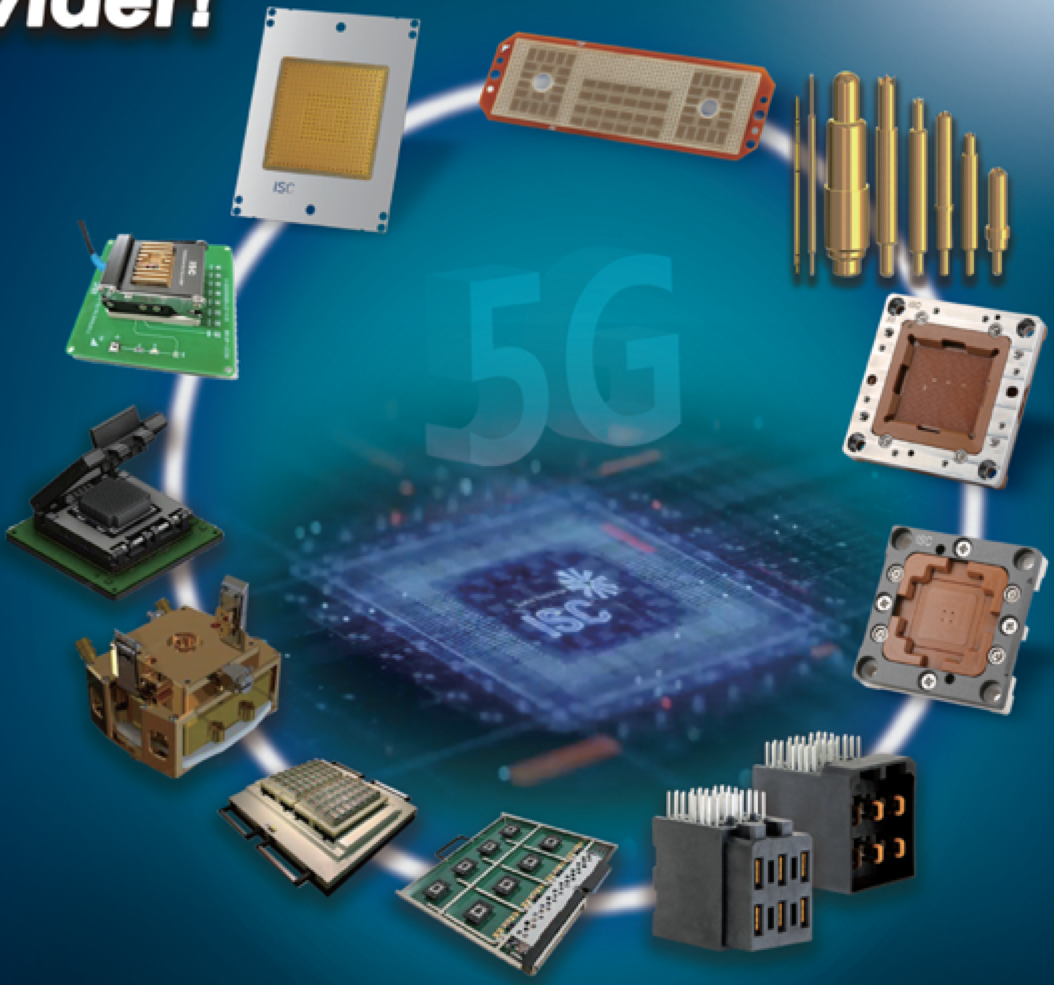
iSP

iSB

iSTIU

iSCON

iSTCU



BURN-IN SOCKET SOLUTIONS

- No ball damage
- Direct inserting on Burn In Board without soldering
- No damage on Burn In Board land
- Simple structure without sub PCB & easy maintenance

POGO SOCKET SOLUTIONS

- Various design available
- Excellent gap control & long lifespan
- High bandwidth & low contact resistance
- Adjustable Impedance
- Fully Metal Shielding

Hi-fix & Burn In Board

- High performance and competitive price
- Test fine pitch, high speed device at hot & cold temperature
- Customized design to meet individual requirements

ELASTOMER SOCKETS & INTERPOSERS

- High performance and competitive prices
- High speed & RF device capability
- No load board pad damage & no contact trace on the ball
- Customized design to meet challenging budget constraints
- Full thermal and electrical simulation

ISC Connector

- ISC Connector Solutions solve many problems across a wide range of circuit sizes, configurations, pitch, and PCB-attach terminations.
- Designed to have the strong resistance against torque forces on mating area to achieve the high reliability in the natural fall shock.

THERMAL CONTROL UNIT

- Extreme active temperature control
- Customized design to meet challenging requirements
- Price competitiveness through self designing and fabrication
- Safety auto shut down temperature monitoring of the device & thermal control unit
- Full FEA analysis for strength, deflection, air flow and any other critical requirements

GLOBAL INNOVATION
ISC

CONTACT	ISC HQ	ISC International	Tel: +82 31 777 7675 / Fax : +82 31 777 7699
ISC CO., LTD	Seong-nam, Korea	Siliconvalley, CA	Email:sales@isc21.kr / Website: www.isc21.kr

A 2.2D die-last integrated substrate for heterogeneous integration applications

By Dyi-Chung Hu [SiPlus Co.]

The semiconductor industry follows Moore's Law by shrinking dimensions from 5nm, 3nm, 2nm, and beyond. The progress in semiconductor technology drives up the need for a higher interconnection density and the requirement of reducing the interconnection length between chips.

Traditional packaging is divided into several packaging levels. However, because of more stringent performance requirements, the "level one" package (die to the substrate) is disappearing because it's more beneficial to reduce the interconnecting distance between chips by using bare dies. For example, AMD's 2016 Fiji product uses bare dies for the interconnection between a graphics processing unit (GPU) and high-bandwidth memories (HBMs). This structure has become the de facto standard for high-performance computing (HPC) packaging. Figure 1 shows that the number of HBMs on one substrate has been increasing over the years. The authors in [1] have predicted that HBMs on one substrate will double every five years based on past data. It is, therefore, expected that a larger packaging substrate with fine lines is needed to meet the demand for increased processing power in the future.

Silicon manufacturing has an excellent infrastructure to meet the acceptable line requirements of the interposer. However, the current silicon-based 2.5D structure has limitations in extending its size to accommodate more HBMs on time.

The term silicon interposer is an interesting one because silicon can only act as the mechanical support for a fine line. However, silicon is a semiconducting material that needs the isolation of through-silicon vias (TSVs). Moreover, the TSV components have resistance, capacitance, and inductance [2], as shown in circuit modeling. Therefore, efforts have been made to reduce these adverse effects by reducing the thickness of the silicon interposer, or removing it altogether.

When considering the manufacturing technologies needed to reduce the thickness of silicon interposers, however, one must take into account environmental, social and governance (ESG) compliance—a major effort by many corporations. The trend toward ESG compliance encourages products that use less energy, materials, and processes needed for manufacture. Therefore, suitable heterogeneous integration solutions have to meet both criteria of high performance and ESG compatibility.

Various TSV-less solutions have been developed in the packaging industry for the interposer. For the substrate, the "coreless" structure has been in mass

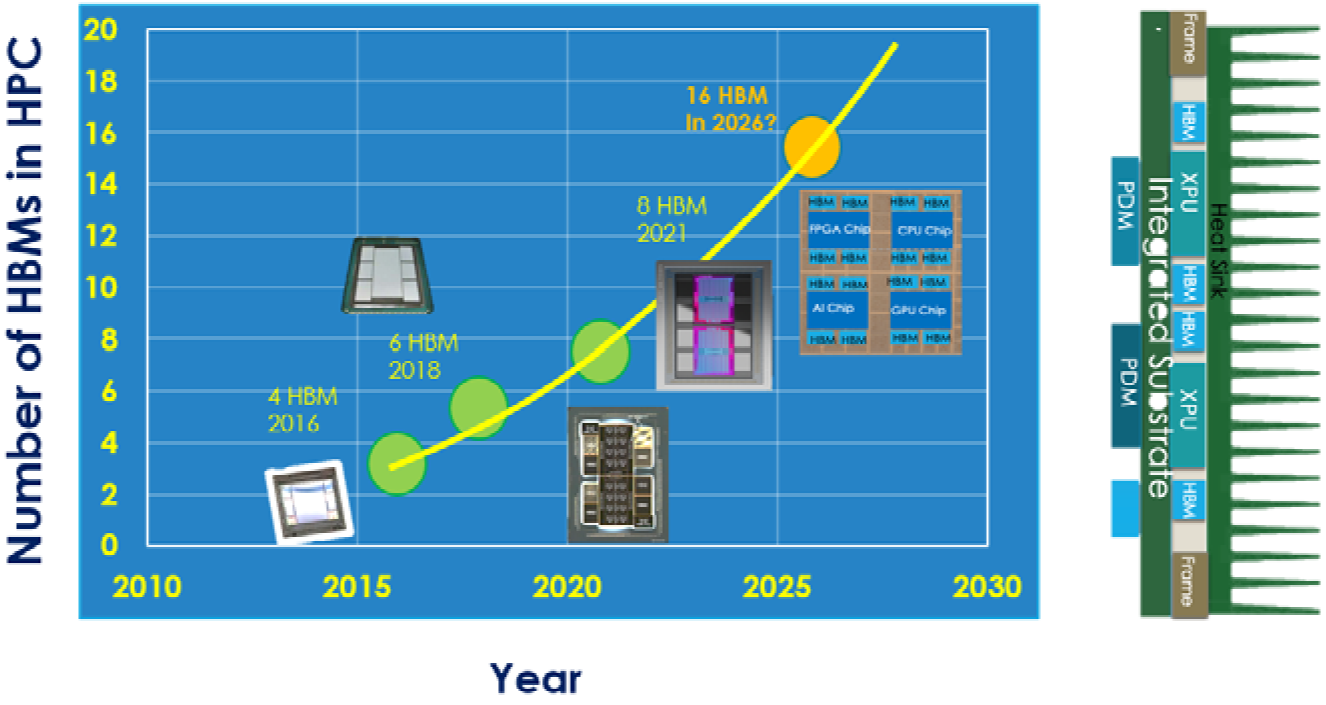
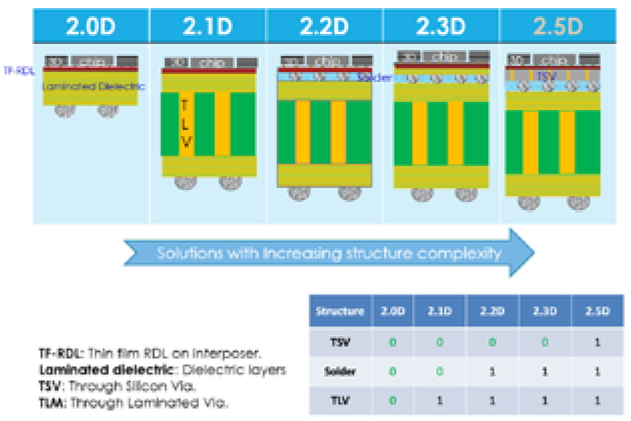


Figure 1: The number of HBMs on a substrate doubles every five years.


2.0D
2.1D
2.2D
2.3D
2.5D
TF-RDL
3D
chip
Laminated Dielectric
T L V
Solder
TSV
Solutions with increasing structure complexity
TF-RDL: Thin film RDL on interposer.
Laminated dielectric: Dielectric layers
TSV: Through Silicon Via.
TLM: Through Laminated Via.

Structure	2.0D	2.1D	2.2D	2.3D	2.5D
TSV	0	0	0	0	1
Solder	0	0	1	1	1
TLV	0	1	1	1	1

Figure 2: Classification of HPC substrates according to their complexity of packaging structure.

production for over ten years because of its benefit to the system's electrical performance. Solders are used for the interconnection of the silicon interposer to the underlying substrate. However, the solder itself also can be modeled as resistance, capacitance, and inductance. So removing the solders in the package would be beneficial to the electrical performance of the system. Therefore, solder-less solutions like fan-out wafer-level packaging (FOWLP), integrated fan-out wafer-level packaging (InFO-WLP), and embedded die have been in production for many years.

Figure 2 categorizes HPC packaging by the generic "through-x via" (TXV) and solder components needed inside their structure, thereby increasing structure complexity from 2.0D [3] to 2.5D. For the 2.5D structure, a silicon interposer with TSVs is connected to the underlayer cored substrate with solders. The cored substrate has through-laminated vias (TLVs). On the other hand, the 2.0D packaging is a TSV-less, TLV-less, and solder-less structure with the shortest interconnection distance in the Z direction for the packaging. SiPlus has verified the 2.0D test vehicle's (TV's) manufacturability and reliability with Nanya PCB Co. [4].

Recently, the TSV-less redistribution layer (RDL)-first technology has become popular among foundries and outsourced semiconductor assembly and test suppliers (OSATS). Various trade names are used for this thin-film RDL (TF-RDL)-based TSV-less interposer structure, namely, silicon-less interconnect technology (SLIT), organic interposer chip-on-wafer-on-substrate (CoWoS®-R), R-Cube, fan-out chip on substrate (FOCoS), etc. However, those are the "die-middle" solutions and not a true die-last solution. Because dies are bonded to the TF-RDL first and then molded and debonded before joining to the substrate, it is desirable to have a true die-last substrate solution based on TF-RDL. A true die-last TF-RDL substrate solution benefits flexibility, as well as enabling known good die and known good substrates before assembly, and reworkability, et al. SiPlus has proposed [1] a 2.2D die-last solution to enable TF-RDL directly on the substrate.

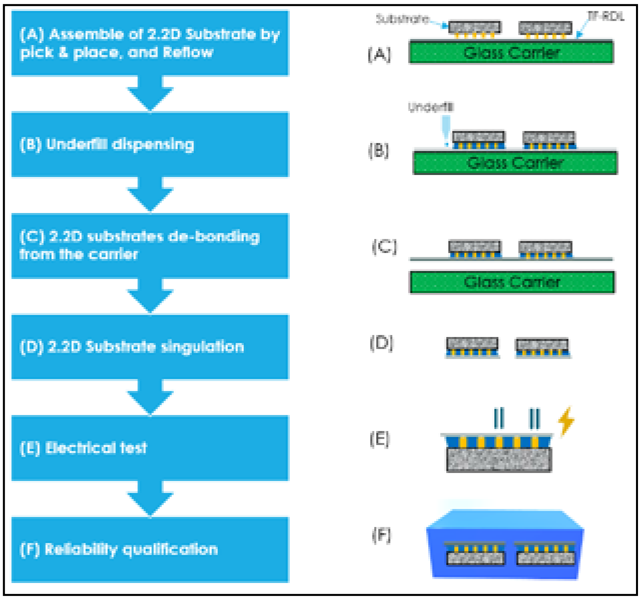
2.2D structure verification requirements

Because TF-RDL is fragile and easy to curl if one removes it from a carrier, in order to be able to use it as a die-last substrate it needs to meet the following requirements: 1) The dimensional stability of the TF-RDL needs to be maintained, otherwise, there will be misregistration between the TF-RDL and the substrate; 2) The robustness of the joints between the TF-RDL and the substrate needs to pass the substrate reliability test; and 3) It is beneficial to utilize the existing infrastructure to realize the low-cost manufacturing process.

A 2.2D TV was designed to verify the above requirements. The TV comprises two portions: the TF-RDL and the substrate. The joining segment of the TF-RDL is composed of a copper pad with solder plating on a polyimide dielectric. The substrate can be ceramic, laminated organic or glass, etc. Copper pillars with Au surface finish on AlN ceramic substrate are used for the TV demonstration. Moreover, 3x3mm daisy chain loops on the TF-RDL and ceramic substrate are designed to check the connectivity of the TF-RDL directly to the substrate. The TV also aims to build a fine-line 2.2D substrate larger than the reticle size (33x26mm). To improve ease of manufacture of the 2.2D substrate, adapting thermal reflow in a conventional solder reflow furnace is desirable.

2.2D integrated substrate TV manufacturing process

The TV used to demonstrate the 2.2D integrated substrate manufacture process flow is shown in Figure 3. The copper pillars are manufactured on AlN substrate, where the upper part of the pillars is electroplated with 5µm Ni


(A) Assemble of 2.2D Substrate by pick & place, and Reflow

(B) Underfill dispensing

(C) 2.2D substrates de-bonding from the carrier

(D) 2.2D Substrate singulation

(E) Electrical test

(F) Reliability qualification

Figure 3: The 2.2D TV integrated substrate manufacturing process flow.

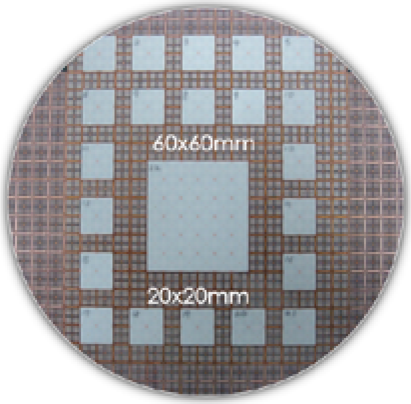
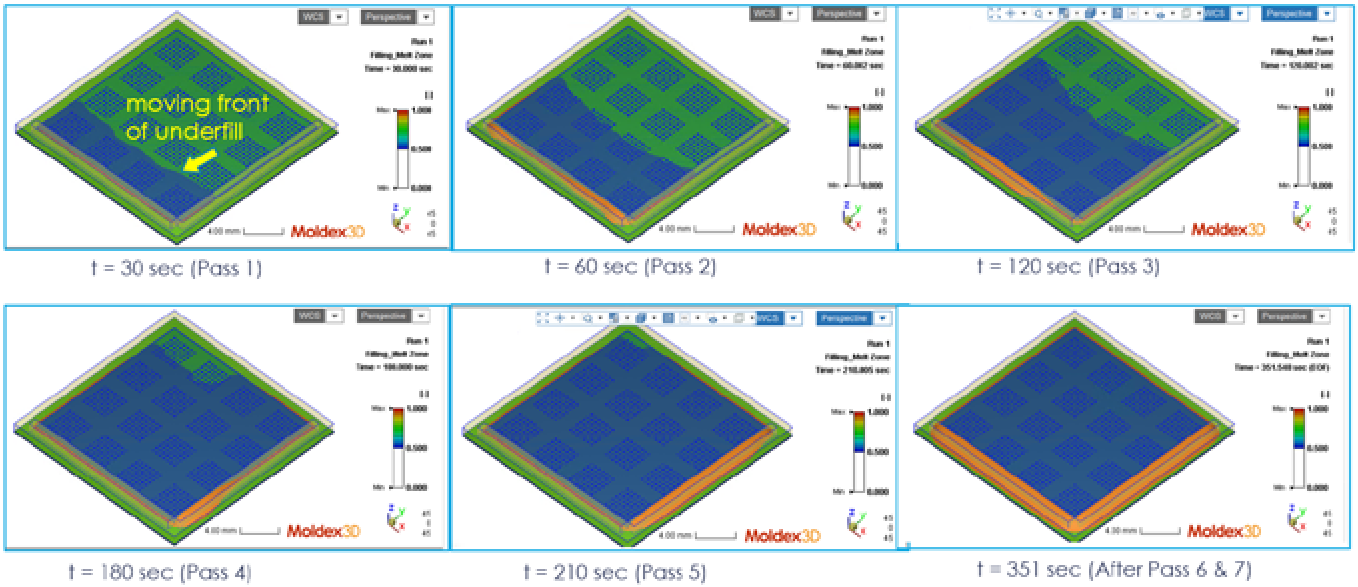


Figure 4: One 60x60mm ceramic substrate and twenty-one 20x20mm ceramic substrates are mounted on the 12" patterned glass wafer.



t = 30 sec (Pass 1)

t = 60 sec (Pass 2)

t = 120 sec (Pass 3)

t = 180 sec (Pass 4)

t = 210 sec (Pass 5)

t = 351 sec (After Pass 6 & 7)

Figure 5: Simulation of underfill flow propagation.

and 0.3µm Au. Two 5µm polyimide (PI) layers with two 3µm Cu metal layers were built on a 12" glass wafer for TF-RDL. Each 10x10mm unit has four 3x3mm daisy chain blocks that were populated on a 12" glass wafer. There is lead-free solder on top of each TF-RDL pad. The Cu pillars were built with a height of 80µm on top of the ceramic substrate

One of the designs of the 2.2D TV pattern used for this study can accommodate substrates underneath the TF-RDL with an incremental size of 10mm. Figure 4 demonstrates the substrate integration with two different sizes of the ceramic substrate that can be placed at will on the surface of the 12" TF-RDL glass wafers. For the 2.2D TV, the joining TF-RDL to the ceramic substrate with dimensions of 20x20mm, 40x40mm, and 60x60mm have been evaluated. One 60x60mm TV sample is used to check the feasibility of the surface mounting process for a sizable 2.2D substrate. The 40x40mm samples are used to get the warpage information of the 2.2D integrated substrate. The 20x20mm TVs were prepared for the reliability qualification test. Ceramic substrates were placed on a 12" glass wafer by pick and place equipment, which had a not too demanding placement accuracy of +/-25µm. Later, the assembly went through a conventional furnace with a peak temperature of 260°C for solder reflow.

After the 2.2D substrate assembly process, underfill is dispensed to protect and enhance the bonding strength of the TF-RDL to the ceramic substrate. The underfill material A is the one that is commonly used for the protection of packages on a printed circuit board (PCB)—it has a low viscosity for smooth underfilling. After underfilling, the assembly was cured in an oven at 150ºC for 15mins.

To further implement the underfill process for the 2.2D substrate, a good simulation tool is needed to optimize the selection of copper pillar height, underfill gap and underfill material. A computer aided engineering (CAE) tool from Moldex3D is used for the simulation of the underfill process. For accurate simulation, we measure the underfill material A's viscosity, curing kinetic and surface tension vs. the process temperature for the data input of the model. The simulation of underfill flow propagation by underfilling material A is shown in Figure 5. The simulation result at the end-of-filling state (351s) is in good agreement with the actual measurement. In addition, the liquid underfill moving front will finally meet at Corner B. The shape predicted by the model also agrees well with the experimental result, as indicated in Figure 6. The fillet at Corner A has a larger volume than Corner B. After the underfill dispensing process, laser debonding is used to remove the 12" glass carrier wafer. All the ceramic substrates mounted on the TF-RDL are singulated into 2.2D integrated substrates using laser cutting.

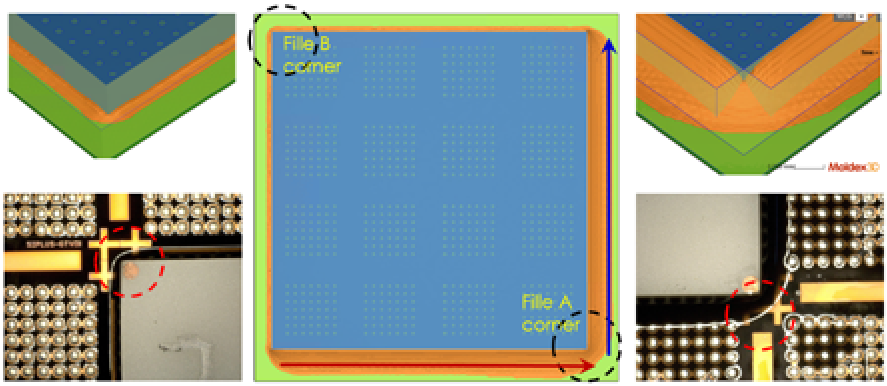


Figure 6: Comparison of the simulation of an underfill fillet formation vs. the experimental result.

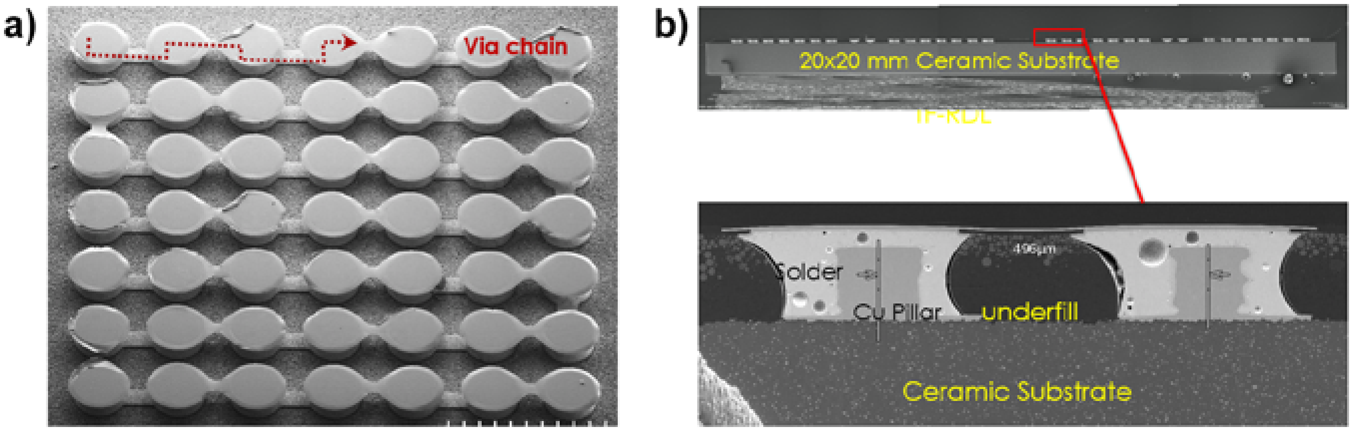


Figure 7: a) The schematic structures of a 2.2D TV integrated substrate; and b) SEM cross-section view of a 20x20mm 2.2D substrate.

2.2D substrate structure evaluation

The schematic daisy chain structures of the TF-RDL and ceramic substrate after the surface mount process are shown in Figure 7a. This 3D metal daisy chain structure on the ceramic substrate can be revealed by etching off the two layers of polyimide dielectric. The red line indicates the daisy chain electrical path from the TF-RDL pad through the ceramic's copper pillar and moves to the adjacent copper pillar connected to the top TF-RDL pad. In total, there are 7x7 TF-RDL connection pads with matching copper pillars on the ceramic substrate.

Figure 7b is a scanning electron microscope (SEM) cross-section view of a 20x20mm 2.2D substrate. The TF-RDL is not visible under this magnification. The magnified portion of two connecting pillars is shown in Figure 7c. The formation of solid solder joints is evident. The microstructure of the 2.2D substrate can also be revealed in detail by 3D X-ray as shown in Figure 8. Figure 8a shows the top view of the 3D X-ray image including cross-section views along the green line (see Figure 8b) and along the red line (see Figure 8c). The formation of solid solder joints show up as "clear" in color. Several small voids inside the solder joint can be revealed as black dots on a white background of solder under X-ray examination. The warpage of the 2.2D substrate is measured by shadow Moiré test to be within +/-20µm in a 40x40mm diagonal range from room temperature to 260ºC.

Electrical and reliability measurements

Each 10x10mm TV daisy chain unit has four daisy chain blocks. For 20x20mm and 40x40mm TVs, daisy chain blocks are measured in the two crossed diagonals. For the 60x60mm TV, except for the blocks in the two crossed diagonals, 4-side peripheral ones are also measured. The measured results are plotted in the accumulation with rather tight distributions, as shown in Figure 9a. The averaged resistances for 20x20mm, 40x40mm and 60x60mm TVs are 0.366Ω, 0.336Ω and 0.347Ω, respectively, with standard deviations of 0.035Ω, 0.027Ω and 0.021Ω, respectively. Figure 9b shows the electrical measurement after the 1000-cycle -65 to +150ºC temperature cycling test (TCT) (reliability test). The 20x20mm 2.2D TV test pattern averaged a resistance of 0.38Ω measured after 500 TCT cycles. The TV test pattern averaged a resistance of 0.386Ω after 1000 TCT cycles. The 2.2D integrated substrate passed the MSL level 4 and 1000 cycles of -65 to +150ºC TCT test.

Applications of a 2.2D integrated substrate to heterogeneous integration

The 2.2D substrate can have various applications as heterogeneous integration integrated substrates for mid-end and

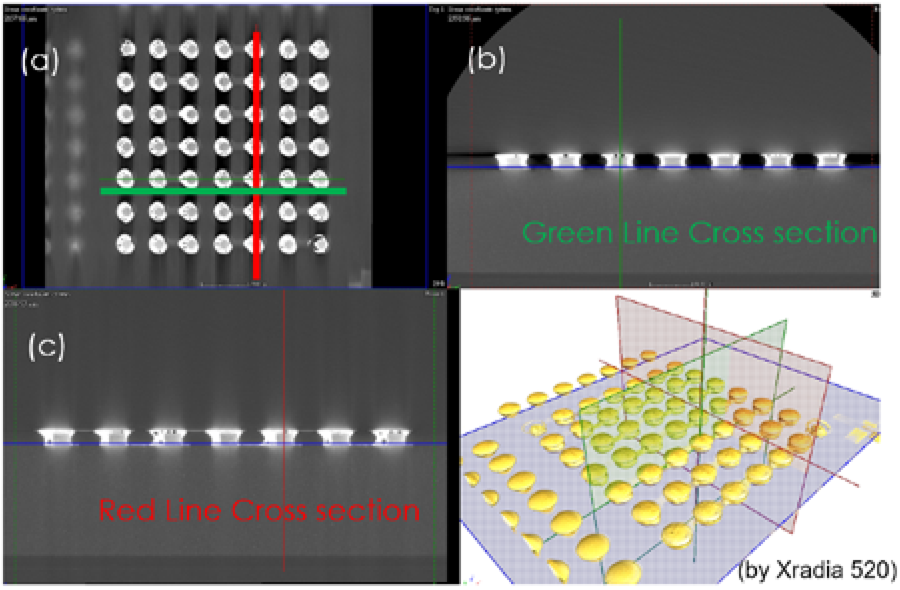


Figure 8: A 3D X-ray image of the 2.2D TV after assembly: a) Top view; b) Cross-section view of the joints along the green line; c) Cross-section view of the joints along the red line; and d) An Xradia 520 is used for the 3D X-ray image.

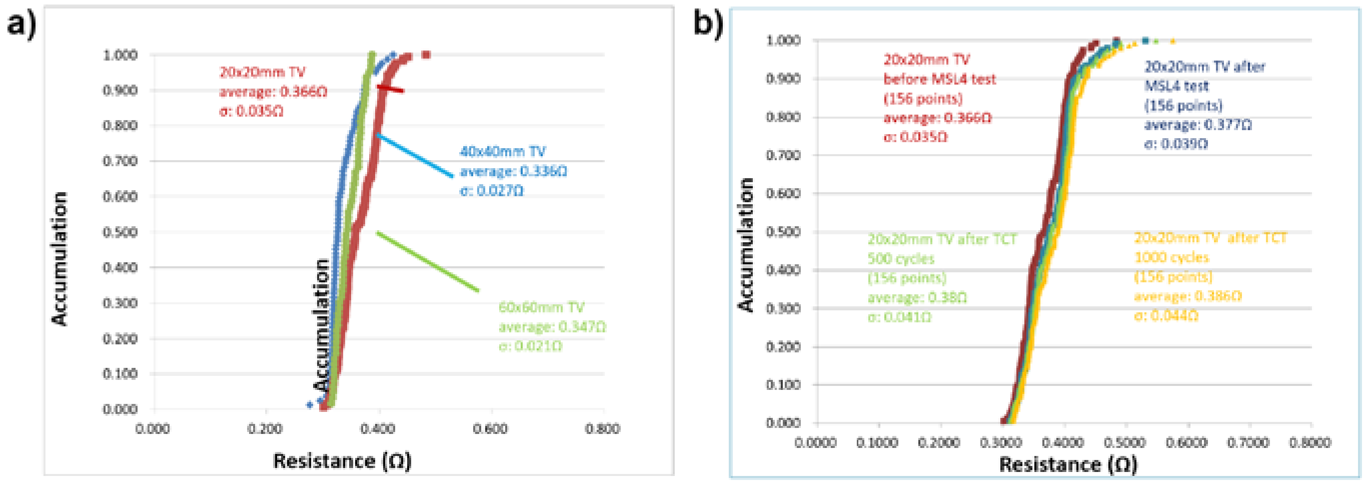


Figure 9: TCT test results for a 2.2D integrated substrate: a) Different sizes of a 2.2D TV substrate resistance distribution before the MSL 4 test; and b) A 20x20mm 2.2D TV daisy chain resistance measured after MSL4, 500, and 1000 TCT cycles—all passed.

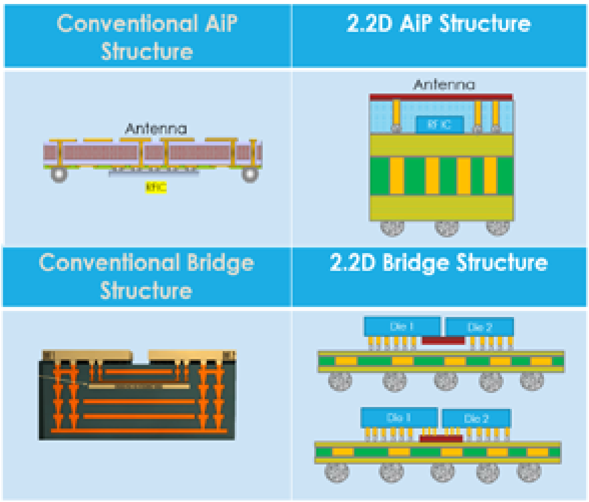


Figure 10: 2.2D integrated substrate applications: 5/6G antenna and high-density bridges.

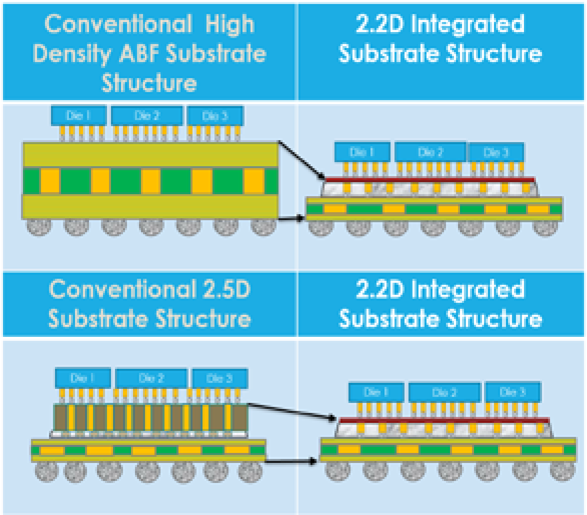

Conventional High Density ABF Substrate Structure
2.2D Integrated Substrate Structure
Die 1
Die 2
Die 3
Die 1
Die 2
Die 3
Conventional 2.5D Substrate Structure
2.2D Integrated Substrate Structure
Die 1
Die 2
Die 3
Die 1
Die 2
Die 3

Figure 11: 2.2D integrated substrate applications: ABF substrate alterative, heterogeneous integration substrate.

high-end applications. Figure 10 shows the applications of a 2.2D substrate. The sections below discuss them.

Antenna in package (AiP) for 5/6G. The TF-RDL has an excellent pattern definition and dielectric thickness control, essential for a consistent antenna quality. This TF-RDL antenna array is also thinner than that of the standard laminated PCB.

Localized high-density bridge interconnect. One can place a TF-RDL on localized areas for high-density applications similar to the bridge structure on the substrate. The bridges can also act as a direct die-to-die interconnection for side-by-side 3D die integration.

A substitute supplier to the current high-end ABF substrate. Figure 11 shows the additional applications of a 2.2D substrate. The 2.2D substrate can be used as a substitute supplier to the current high-end Ajinomoto build-up film (ABF) substrate. A TF-RDL can achieve a finer line width than the metal lines on the present laminated dielectric substrate. The increased wiring density of a 2.2D TF-RDL will help to reduce the layer counts of the existing high layer count ABF substrate. It is interesting to know that the ABF substrate is usually symmetrically positioned against the core. Therefore, the bottom portion of the ABF substrate wiring capacity can't be fully utilized. On the contrary, the 2.2D substrate has an asymmetric structure to meet the optimal wiring needs.

High-end interposer-like applications. The 2.2D substrate can be applied to high-end interposer-like applications. Higher TF-RDL layer counts and 2µm fine lines are needed for this application. We have witnessed fine line and layer count improvement over the years, especially in semiconductor foundries.

Summary

An integrated substrate structure, "2.2D," is demonstrated. The 2.2D substrate is a true die-last integrated substrate solution. The developed solution can put TF-RDL directly bonded on top of the substrate. Therefore, the 2.2D substrate simplifies the 2.5D structure, reduces the cost, and improves the product cycle time. We demonstrated the 2.2D substrate and manufacturing process by designing a 2.2D test vehicle with two metal layers of thin film bonded directly to a ceramic substrate. The connectivity of a 2.2D integrated substrate using a daisy chain run from TF-RDL to the substrate was validated. The 2.2D TV demonstrates good warpage behavior up to the solder reflow temperature. The 2.2D structure TV also passed the substrate

MSL level 4 and TCT 1000 cycle TCT electrical test. Therefore, the 2.2D structure shows good potential for a heterogeneous integration substrate for both mid-end and high-end applications. Furthermore, we also disclosed that the 2.2D integrated substrate solution can be used in a 5/6G AiP structure, as a substitute in a current advanced ABF substrate, and in localized high-density bridges.

Acknowledgment

This article is based on materials originally presented at the 2021 IEEE 71st Electronic Components and Technology Conference (ECTC), June 2021.

References

1. D. C. Hu, E. Hao Chen, J. ChangBing Lee, C. P. Sun, C. C. Hsu, "2.2D die-last integrated substrate for high-performance applications," ECTC 2021.
2. M. Swaminathan, K. J. Han, *Design and Modeling for 3D ICs and Interposers,* World Scientific Publishing Co. Pte. Ltd. 2014.
3. D. C. Hu, "An innovative system integration interconnection technology beyond BEOL," IEEE Inter. Interconnection Conf., Santa Clara, U.S.A., 2018.
4. D. C. Hu, J. Ho, "Methods to reduce the hierarchy of interconnections in electronic system," Proc. of IMAPS 2020.
5. W. H. Yang, R. Y. Chang, "Numerical simulation of mold fill in injection molding using a three-dimensional finite volume approach," Inter. Jour. for Numerical Methods in Fluids, vol. 37, 2001.

Biography

Dyi-Chung Hu is CEO of SiPlus Co., Hsinchu, Taiwan, ROC. He is the founding member of a number of high-tech companies including SiPlus Co., Raytek Semiconductor Co., Hannstar Display, and E-ink Co. He was a SVP of R&D at Unimicron Technology Co. and was the founding chairman of the SEMICON Taiwan Packaging Committee. He earned a PhD from MIT in Material Science and Engineering. Email hu@si2plus.com

Heterogeneous integration for AI applications: status and future needs (part 1)

By Madhavan Swaminathan, Siddharth Ravichandran [Georgia Institute of Technology]

The semiconductor industry has been driven by Moore's Law [1] for over five decades. With the number of transistors on a chip doubling every two years, this has led to almost exponential performance increase for microprocessors while making these chips affordable. The performance increase from one microprocessor generation to the next was supported through metal-oxide semiconductor field-effect transistor (MOSFET) scaling proposed by Robert Dennard [2], which enabled area reduction while maintaining constant power densities. Until the mid-2000s, this trend continued until transistor leakage became a major problem because of thinning of the gate oxide to a few atomic layers. This has led to reduced frequency and reduced single thread performance scaling for microprocessors since the mid-2000s.

Several innovations over the last fifteen years related to materials, transistor structure and architecture have enabled continued area scaling to continue Moore's Law. One such innovation is the use of multiple cores supported by software parallelism to increase performance. Today, microprocessors in data center applications contain more than a hundred cores with ten billion transistors and this trend of increasing core count appears to be continuing. Unfortunately, due to leakage, the power densities of microprocessors have increased since the mid-2000s. A combination of prohibitive costs associated with chip fabrication in advanced nodes, reduced area scaling, increased power densities and the general feeling that Moore's Law is slowing down, is causing the semiconductor industry to pursue non-traditional approaches to transistor scaling and computing. John Shalf [3] posits that the path forward is along three fronts, namely: 1) Relying on more efficient architectures supported by advanced packaging, 2) Developing new materials and devices that enable non-traditional transistors, and 3) Innovation using new models for computation such as quantum computing. We focus on heterogeneity using advanced packaging in the context of artificial intelligence (AI) as one path forward to continue Moore's Law in this article.

AI is gaining momentum in data science as a means for solving difficult problems that are otherwise unsolvable. The AI algorithms being developed by the computer science community to support such solutions rely on neural network architectures for training and deriving inferences. Over the last several years architectures based on feedforward neural networks (FFNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and others have emerged that rely on several layers of neurons interconnected through dense connectivity to address complex problems arising in science, computer vision, finance, robotics, and others. These computational architectures need to be mapped to computer hardware so that the neural networks can be suitably trained to derive inferences from data.

As the complexity of data to be learned increases, the number of hidden neurons in a neural network increase converting neural networks into deep neural networks (DNN), making the hardware required for computations even more complex. Unlike traditional microprocessor-based computing platforms, AI algorithms require significantly increased computations and storage needs, thereby limiting the performance gained from general purpose central processing units (CPUs). An alternative is the use of graphics processing unit (GPU)-based platforms that provide better performance than CPUs for neural network-based computations. However, GPUs consume high power and are not very energy efficient, thereby limiting their capability and applicability for many AI applications. Because neural network processing is highly parallelizable, exploitation of both data- and thread-level parallelism is required in the implementation of such computing architectures, but at low energy and power levels as compared to multi-core CPUs. In addition, the performance of neural network computations is limited by insufficient memory bandwidth and latency.

Rather than integrate all logic functions using a single process through monolithic (or homogeneous) integration as in system on chip (SoC), there is a trend towards polylithic (or heterogeneous) integration for microprocessors. This is being driven by the exponential costs associated with large dies implemented using advanced process nodes, the reduced time to market possible using smaller dies from optimized technology nodes, and the move towards heterogeneous semiconductor systems with dies connected from different process nodes. Such connectivity is being enabled by two fundamental technologies namely, 2D interposers and 3D stacking. The interposer connects dies together laterally using high-density wiring and resides between the dies and a package substrate, which can then be mounted onto a printed wiring board (PWB). As compared to interposers, 3D stacking supports much higher wiring density and shorter wires, but with drawbacks related to power delivery and heat removal. With AI applications being memory intensive, highly parallelizable, and requiring memory to be placed near the logic for reducing latency, there is a natural fit for using both interposers and 3D stacking to maximize performance. As AI applications evolve, we expect the resulting system architectures to require extreme heterogeneity further justifying the need for a heterogeneous integration platform enabled by advanced packaging.

In this paper we provide a survey and comparison of the various 2D, and 3D technologies available and in development based on the present and future needs posed by AI. This comparison is based on a set of metrics derived from data speed, energy efficiency and latency that have a direct impact on system performance.

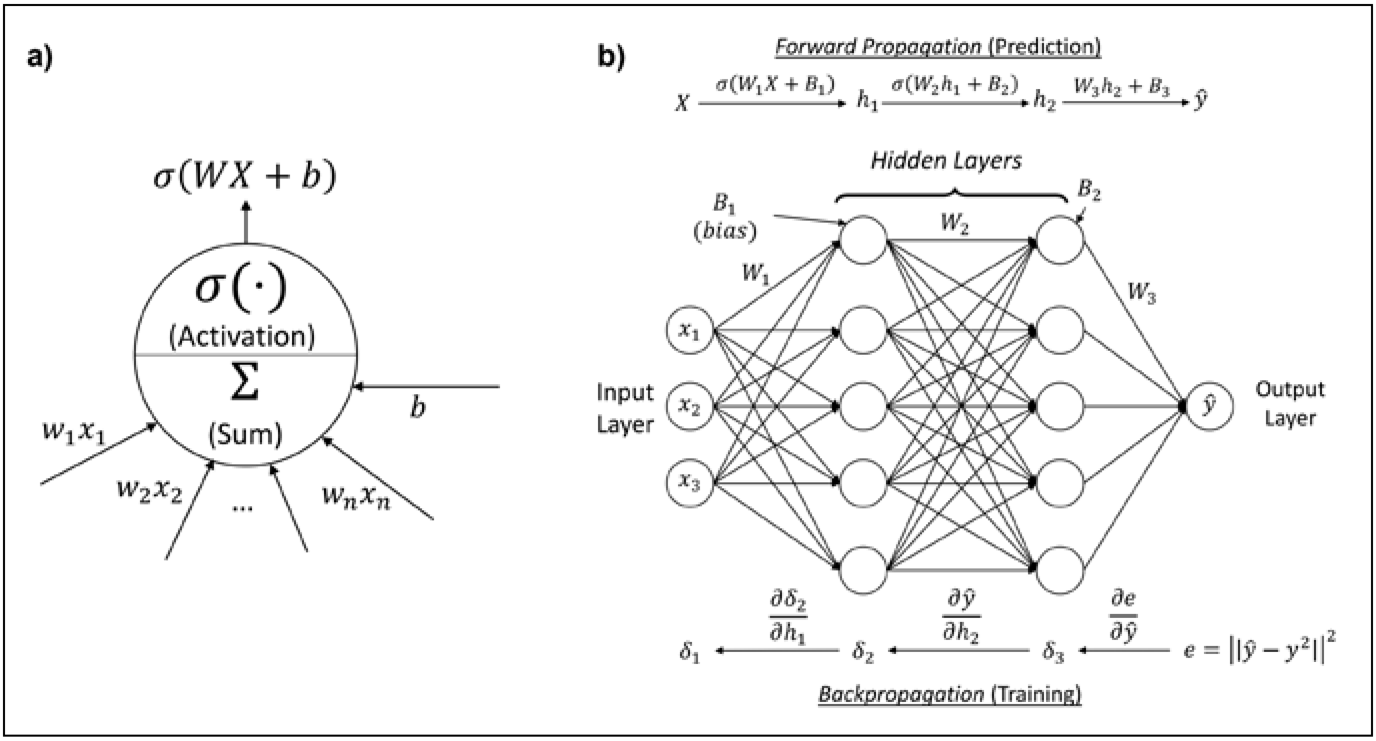

a)
$\sigma(WX+b)$
$\sigma(\cdot)$
(Activation)
Σ
(Sum)
b
w_1x_1
w_2x_2
...
w_nx_n
b)
Forward Propagation (Prediction)
$X \xrightarrow{\sigma(W_1X+B_1)} h_1 \xrightarrow{\sigma(W_2h_1+B_2)} h_2 \xrightarrow{W_3h_2+B_3} \hat{y}$
Hidden Layers
B_1 (bias)
B_2
W_1
W_2
W_3
Input Layer
x_1
x_2
x_3
$\hat{y}$
Output Layer
$\delta_1 \xleftarrow{\frac{\partial\delta_2}{\partial h_1}} \delta_2 \xleftarrow{\frac{\partial\hat{y}}{\partial h_2}} \delta_3 \xleftarrow{\frac{\partial e}{\partial\hat{y}}} e=||\hat{y}-y^2||^2$
Backpropagation (Training)

Figure 1: a) Single neuron, and b) Feedforward neural network (FFNN) [1].

System architectures for AI

We start with the basic building block for a neural network (NN), namely, the neuron as shown in **Figure 1a**. A neuron takes a vector of inputs X=(x$_1$,…,x$_n$), constructs their weighted sum WX=(w$_1$ x$_1$,…,w$_n$ x$_n$) and adds a bias (b) to generate the result $WX+b$. This is then passed through a nonlinear activation function to obtain σ($WX+b$), which represents the output of this single neuron. The activation function introduces non-linearity and bounds the output. Using the neuron, a NN can be constructed consisting of input, hidden and output layers, as shown in **Figure 1b**. The purpose of the hidden layers in the figure is to capture non-obvious interactions between the overall input-output relationships. Each hidden layer consists of multiple neurons where each neuron connects to each of the neurons in the subsequent layer, where each connection describes a different interaction pattern. This is an example of a FFNN used for inference (or prediction). As the number of hidden layers increases for capturing more complex patterns in data, the NN transforms into a DNN. For training, the error (e) generated by the NN is minimized by adjusting the weights in each layer through their gradients, which are then back propagated to the previous layer [4].

The mathematical operations to be performed in a DNN include matrix-vector multiplication (*WX*), addition (*WX*+*b*), a pointwise operation as in activation using ReLU, sigmoid and others and fully-connected layers using batch matrix multiplication. A NN may also include convolutional layers as in CNN, which involves matrix-vector multiplication using a sparse, or Toeplitz matrix [5]. In a NN, around 80-90% of the computations are related to matrix multiplications and convolution while the remaining 10-20% are used in the computation of vector functions such as ReLU, sigmoid and others. Rather than using a general-purpose CPU for such computations, accelerator chips can be used with smaller dies from advanced process nodes to reduce cost.

NN computations are memory intensive and therefore designs that only use on-chip cache (including high-density eDRAM) are not scalable to larger dimensions, as required for a DNN. Moreover, computing for AI requires high-speed communication between logic and memory at low energy per bit, and therefore, designing energy-efficient hardware represents a major challenge for implementing memory-compute systems.

The NN parameters (weights, bias, hyper-parameters, and others) need to be stored in memory, and therefore, one possible compute architecture is to integrate high-bandwidth memory (HBM) near the logic die, as shown in **Figure 2a**. The logic could be a CPU that includes a single accelerator where the NN computations are performed using processing elements (PE), where each PE constitutes a neuron that performs the multiplication, addition, and activation. The logic could also be implemented using multiple accelerators that constitute a larger NN. For such an architecture to work, the HBM and logic must be connected using short wires through an interposer as shown in the figure to reduce off-chip memory latency. For an HBM configuration that has four memory stacks, each having 8 channels that contain 128-bit data interface [6], it has been shown that the achievable total bandwidth aggregates to 1.63TB/s using a silicon interposer [7].

A method for reducing off-chip latency is by using the hybrid memory cube (HMC) based on the Neurocube architecture that integrates logic within a 3D-stacked DRAM memory. Here, each Neurocube accelerator contains one HMC containing 16 vaults, where each vault consists of a PE that performs multiplication-accumulation (MAC) and a router for transferring between the logic and DRAM dies. Multiple HMCs can be assembled on an interposer and connected as shown in **Figure 2b**. In both architectures described, the memory is near the logic, and are dictated by logic centric computations. We refer to such architectures as near-memory processor (NMP). NMP architectures are generally limited by logic-centric computations because the logic and memory are separated from each other.

A better approach is to directly perform computation inside memory, referred to as processor-in-memory (PIM) architectures where the memory array is re-purposed for computation, thereby realizing massive parallelism and almost nullifying data movement [7]. PIM architectures are emerging and use CMOS with non-volatile memory (NVM) such as resistive RAM (ReRAM) or ferroelectric FETs (FeFETs). The ReRAM crossbar structure can accelerate matrix-vector multiplications where the vector representing the input signal stored along the word lines can be multiplied with the matrix elements programmed into the conductance cells through current summing, with the output available through the bit lines. For a large matrix that does not fit into the array of the crossbar structure multiple arrays can be used where partial sums can be added to obtain the output [5].

Several possibilities arise in the implementation of the PIM architecture with two of them described in [7] namely: 1) A single PIM engine coupled with a DRAM module used to store the DNN parameters and establishes connections through an interposer similar to **Figure 2a** with the logic replaced by PIM; and 2) Multiple PIM accelerators where all the DNN parameters are stored within the PIM leading to a DRAM-free design, where the PIMs are connected through the interposer, as shown in **Figure 2c**. In the former, the DRAM stores the DNN model parameters and intermediate data—these parameters are then written into the PIM accelerator for each layer followed by the computations. This architecture could be limited by energy and latency overhead due to off-chip communication. In the latter, one possibility is to map the whole DNN into the on-chip memory of a single chip, but this can result in a large chip that would make testing complex, thereby, significantly increasing cost.

An alternative to the architecture discussed above is a multi-chip design where each die can be used for the computation of a layer, or compact layers in each die are combined using multiple dies. The input, output and intermediate layers are transported between dies in a pipelined manner. For inference, because the weight parameters are fixed, and the intermediate

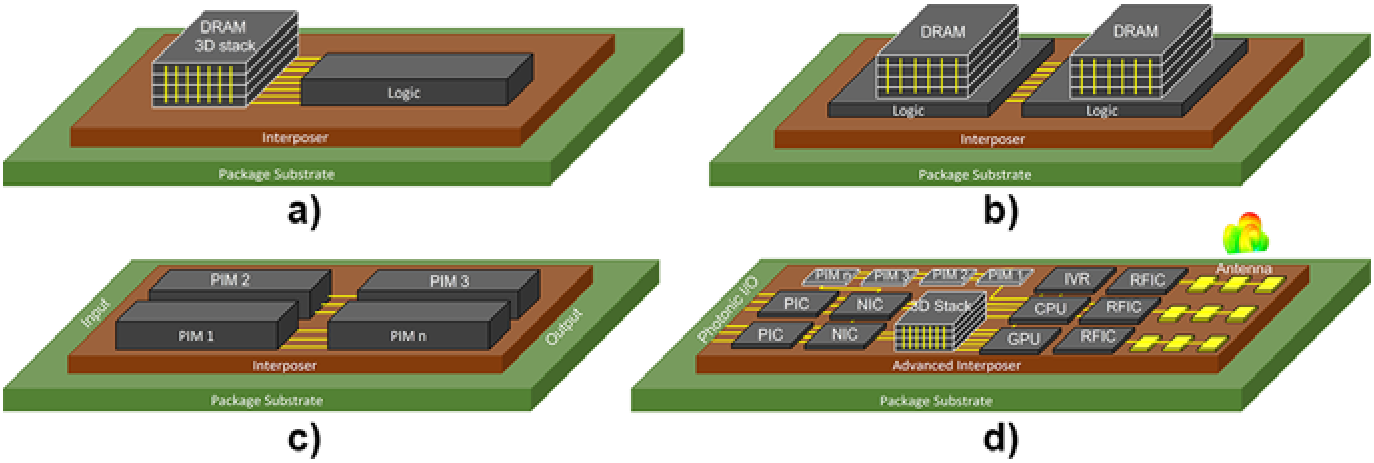


Figure 2: AI packaging architectures: a) NMP using logic and HBM; b) NMP using HMC; c) Multi-chip PIM accelerator; and d) Future extreme heterogeneity.

data size is smaller than the weight parameters, the data communication latency and energy can be reduced. In [7], it has been shown that for AlexNet, a CNN implemented using a FeFET-based PIM accelerator, the multi-chip design leads to lower energy efficiency and latency as compared to a single-chip design. In addition, the multi-chip design provides higher throughput because of pipelined execution.

In both NMP and PIM accelerator architectures, the DRAM, logic, and other dies come from different process nodes and are connected using 2D (interposer and package substrate) or 3D (stacking) leading to heterogeneous integration. As AI architectures become more prevalent there will be a need for integrating additional dies from other domains (analog, radio frequency (RF), and photonics), as well as different process nodes leading to extreme heterogeneity as shown in Figure 2d.

Comparison metrics

With 2D and 3D solutions available for connecting dies together, metrics are required to compare these technologies. In this section we describe five important metrics related to interconnect density, energy per bit, data rate, power delivery, and thermal design power (TDP) for comparing the various options.

Interconnection density. As shown in Figure 2, the implementation of NMP and PIM DNN architectures requires connectivity between adjacent dies. When dies on a package substrate or interposer are connected, the number of die input/output (I/O or IO) terminals that can escape along the die edge to connect to an adjacent die becomes an important metric. Because the number of interconnects that can be routed depends on the length of the die edge and number of redistribution or wiring layers (RDL), a better metric is the interconnect density with units of IO/mm/layer [8].

Two dies connected together are shown in Figure 3 where the circular pads of diameter D represent the positions where the dies are assembled using solder or other means. The center-to-center pad pitch is P with the pads staggered as shown in the figure, like pad arrangements in HBM. The number of interconnects that can be wired between the two dies over a distance. P_y is given by:

$$n = \frac{P_y - D - S}{L + S} \quad \textbf{Eq. 1}$$

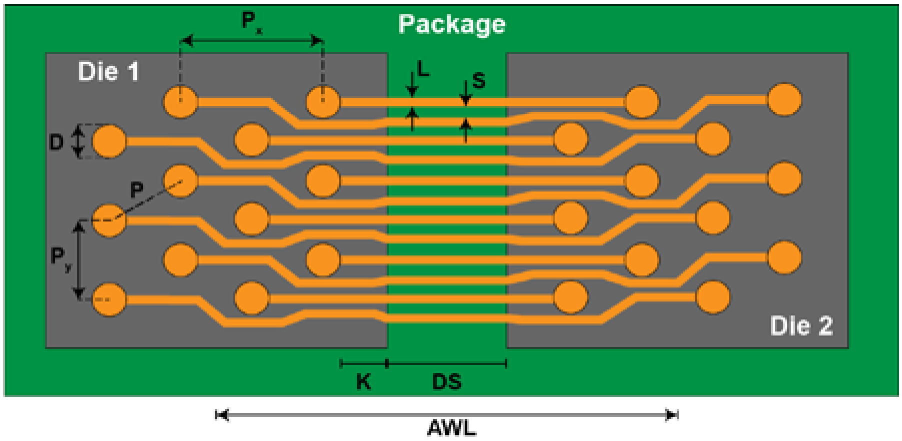


Figure 3: Interconnect density and wiring length.

where L and S are the linewidth and spacing, respectively. This leads to an interconnect density per layer:

$$W = \frac{n+1}{P_y} \quad \textbf{Eq. 2}$$

(In part 2 of this article we will present Table 2; the data in that table will show that the 10X higher interconnect density for silicon (250 IO/mm/layer) as compared to organic interposer (25 IO/mm/layer) results in the use of fewer layers in the former. For example, 7,500 wires can be routed along a die edge of length 10mm using just three wiring layers with a silicon interposer as compared to many more layers in an organic package. Because vias add additional parasitic losses and latency, increased layers will reduce performance. Layer count can be further reduced by decreasing the assembly pitch. For 3D stacking, the dies are connected vertically, and therefore, $W=(1/P)^2$. For a non-staggered pitch of P=10µm, this translates to an interconnection density of 10,000 IO/mm^2.)

Interconnect length. The length of the wire connecting adjacent dies determines the total resistance and capacitance of the interconnections, and therefore, represents an important design parameter to consider. From Figure 3, not all wire lengths are equal and therefore, an average wirelength, AWL, can be calculated as:

$$AWL = DS + (2 * K) + n \times \sqrt{2}P \quad \textbf{Eq. 3}$$

where DS is the die-to-die spacing, and K is the keep-out zone (KOZ). Most advanced packages have DS=100µm and K=50µm, and therefore, n and P are the main parameters that affect AWL. As more columns in each die are routed, AWL will increase, but compensation for the increased AWL can be achieved by using smaller pad pitch, P. For 3D stacking, the wirelength is the physical length that connects dies together, which includes the length of the through-silicon via (TSV). All wires have the same length in 3D integration. (In part 2 of this article, we will present Tables 2 and 3 that show that some of the wirelengths are estimated based on **Eq. 3**, and others are based on published data.)

Data rate. Dies communicate with each other using a driver-receiver pair through the interconnection. With short wires that are a few mm long, simple driver and receiver circuitry can be used, as compared to long interconnections where equalization and error correction schemes are required, as shown in Figure 4a [9]. In

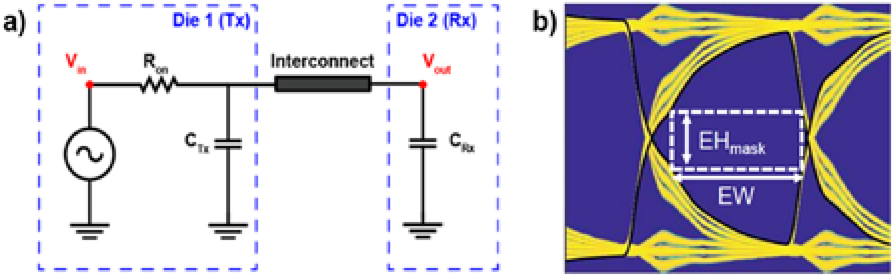


Figure 4: a) Driver-interconnect-receiver communication, and b) Eye diagram at the far end of the center line.

the figure, R_{on} is the driver resistance while C_{Tx} and C_{Rx} are the driver and receiver capacitances, respectively. The interconnections are designed with minimum dimensions, where five coupled lines carry the signals in parallel. The RC parameters of the line are extracted using an electromagnetic (EM) simulator [10] and the eye-diagram at the far end of the center line is computed using the wirelengths derived from **Eq. 3**. The maximum data rate per IO that can be supported based on a bit error rate (BER) of 10^{-12} is then computed by increasing the signal frequency until the EH_{mask} and EW shown in **Figure 4b** reach $0.1V_{in}$ (along the vertical voltage axis) and 0.1UI (UI = unit interval along the horizontal time axis), respectively. Because the linewidths and spacing are small, the interconnect response is RC dominated as can be seen from the eye-diagram. (In part 2 of this article [Tables 2 and 3], we will show that the difference in data rate/IO between the various 2D and 3D approaches is due to dielectric permittivity (scales C), L/S (scales R & C), driver/receiver parameters and linelength.)

Bandwidth density. The bandwidth density is derived as:

$$BW = IO/mm \times Datarate/IO \quad \textbf{Eq. 4}$$

(For example, in part 2 of this article [Table 2], we will show that 500Gbps/mm enables 5Tbps of data to be transmitted between dies across a 10mm edge for the silicon interposer.)

Energy per bit (EPB). The energy consumed to transmit one bit through the interconnect channel for non-return to zero (NRZ) signaling is given by:

$$EPB = P_D \times T = \frac{C_T V_{swing}^2}{2} \quad \textbf{Eq. 5}$$

where P_D is the dynamic power, T is the time period for one clock cycle, C_T is the total capacitance ($C_{Tx} + C + C_{Rx}$) to be charged, and V_{swing} is the voltage swing at the far end of the line. (In part 2, we will show that it is important to note [from Tables 2 and 3] that EPB is always lower for 3D stacking as compared to 2D approaches because of shorter interconnect lengths and smaller capacitances.)

Power delivery. The system level power efficiency is defined as:

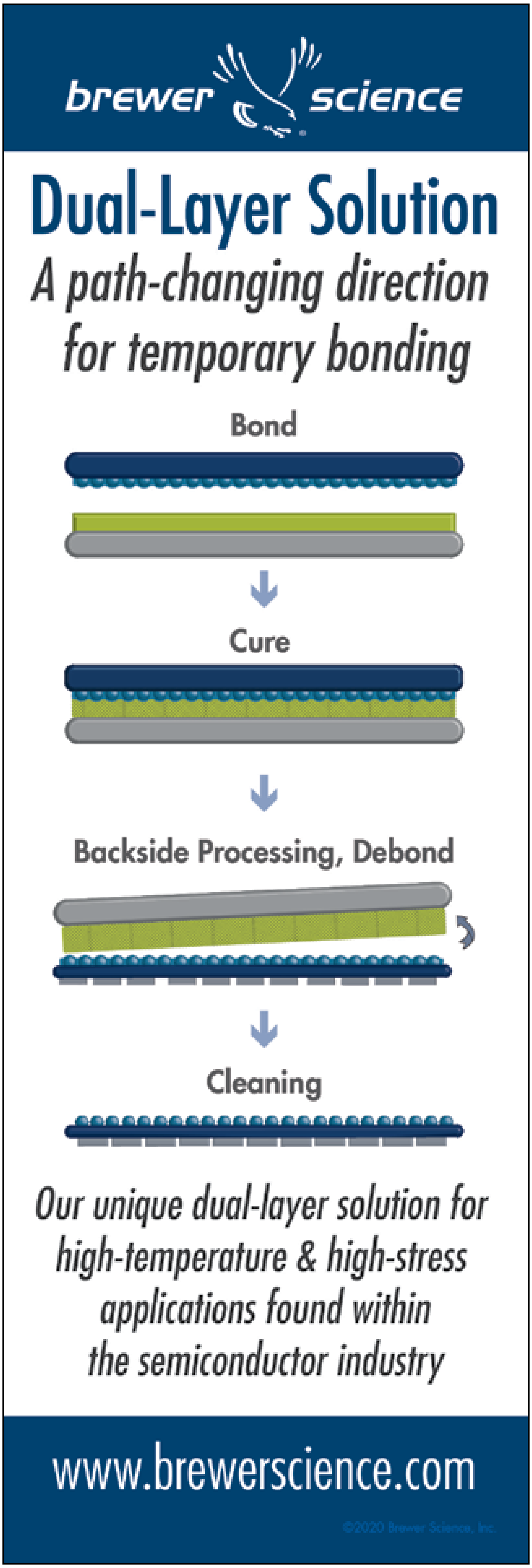
$$\eta = \frac{P_{out}}{P_{in}} \quad \textbf{Eq. 6}$$

where P_{out} is the power delivered to the die and P_{in} is power delivered to the voltage regulator (VR). The efficiencies are typically in the range 75-80% for the higher power applications. These efficiencies can be increased by either decreasing the effective resistance between the VR and die or/and integrating high-voltage conversion ratio regulators near the die on the same package [11]. (Because power delivery is an application-specific problem, we do not include it in the comparison Tables 2 and 3 in part 2.)

Thermal design power (TDP). The heat dissipated by the dies need to be removed through a thermal dissipation structure such as a heat sink, heat pipe, immersion cooling, or others using air or liquid. The TDP is defined as:

$$TDP = \frac{T_j - T_a}{R_{th}} \quad \textbf{Eq. 7}$$

where the junction temperature, T_j, for the devices varies between 85°C-130°C with an ambient temperature T_a of around 40°C. The



TDP has been increasing ~7%/year and today has a range between 200W – 300W for both the CPU and the GPU [12]. The AI chips developed so far have a TDP in the range of 75W-300W and most use engines that accelerate computations in conjunction with a CPU or GPU [13]. For example, the Tensor Processing Unit (TPU) from Google uses HBM combined with CPU for implementing training and inference engines. With average TDP of 250W and T_j=107°C, the effective thermal resistance R_{th} required translates to 0.3°*C/W* [12]. Achieving such low thermal resistances requires liquid cooling and is a metric that needs to be supported irrespective of the packaging type used.

References

1. G. E. Moore, "Cramming more components onto integrated circuits," Electronics, vol. 38, no. 8, pp. 114-117, 1965.
2. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," IEEE Jour. of Solid-State Circuits, 9(5), 256-268 (1974).
3. J. Shalf, "The future of computing beyond Moore's law," Philosophical Trans. of the Royal Soc. A, 378(2166), 20190061 (2020).
4. M. Swaminathan, H. M. Torun, H. Yu, J. Ale Hejase, W. D. Becker, "Demystifying machine learning for signal and power integrity problems in packaging," IEEE Trans. on Components, Packaging and Manufacturing Tech., V: 10, Issue: 8, pp: 1276 – 1295, 2020.
5. Y. Chen, Y. Xie, L. Song, F. Chen, T. Tang, *A Survey of Accelerator Architectures for Deep Neural Networks*, pub.: Elsevier Engineering, 6, pp. 264 – 274, 2020.
6. M. O'Connor, "Some highlights of the high-bandwidth memory (HBM) standard," Proc. Memory Forum Workshop, Jun. 14, 2014; https://www.cs.utah.edu/thememoryforum/mike.pdf.
7. S. Mukhopadhyay, Y. Long, B. Mudassar, C. S. Nair, B. H. DeProspo, H. M. Torun, M. Kathaperumal, V. Smet, D. Kim, S. Yalamanchili and M. Swaminathan, "Heterogeneous integration for artificial intelligence: challenges and opportunities," IBM Jour. Res. & Dev., Vol. 63, no. 6, paper 4, 2019.
8. R. Mahajan, et al., "Embedded multi-die interconnect bridge (EMIB)--a high-density, high-bandwidth packaging interconnect," 2016 IEEE 66th Elec. Comp. and Tech. Conf. (ECTC), 2016.
9. A. C. Durgun, Z. Qian, K. Aygun, R. Mahajan, T. T. Hoang, S. Y. Shumarayev, "Electrical performance limits of fine-pitch interconnects for heterogeneous integration," IEEE 69th ECTC, pp. 667-673, May, 2019.
10. ANSYS, HFSS. "3D full-wave electromagnetic field simulation, ANSYS.," https://www.ansys.com/products/electronics/ansys-hfss
11. S. Mueller, K. Z. Ahmed, A. Singh, A. K. Davis, S. Mukhopadyay, M. Swaminathan, et al., "Design of high-efficiency integrated voltage regulators with embedded magnetic core inductors," IEEE 66th ECTC, pp. 566-573.
12. L. T. Su, S. Naffziger, M. Papermaster, "Multi-chip technologies to unleash computing performance gains over the next decade," IEDM, 2017.

Biographies

Madhavan Swaminathan is John Pippin Chair in Microsystems Packaging and Director - 3D Systems Packaging Research Center (PRC) Georgia Institute of Technology, Atlanta, GA. Email: madhavan@ece.gatech.edu

Siddharth Ravichandran is a recent PhD graduate from the School of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA.

FOWLP and Si-interposer for high-speed photonics packaging

By Lim Teck Guan, Eva Wai Leong Ching, Jong Ming Ching, Loh Woon Leng, David Ho Soon Wee, Sajay B G, et al.
*[Institute of Microelectronics, A*STAR (Agency for Science, Technology and Research)]*

The ever-increasing demand for higher data bandwidth in the data center has led to the requirement of higher speed, smaller form factors and scalable, integrated optical engines (OE). An OE comprises various electronic integrated circuits (EICs) and photonics ICs (PICs) to enable the optical/electrical (O/E) signal conversion. With the advancement in semiconductor device technology, packaging and integration technologies are becoming the limiting factors to enable the OE to meet the demand of the data center's high data bandwidth requirement of more than 800Gbps. It is, therefore, important to develop a cost-effective PIC and EIC packaging platform to realize a high data rate OE. The integration and packaging technologies must support the scaling of the number of optical channels and provides a high-speed electrical interconnect between the PIC and EIC of each channel of more than 100Gbps.

Advanced wafer-level packaging has been successfully used in state-of-the-art field-programmable gate array (FPGA) ICs, smartphone application processors, and graphics processing units (GPUs) to provide power-performance-form factor boosts that are not obtained by conventional packaging. A key benefit of advanced wafer-level packaging is the capability to achieve heterogeneous integration whereby ICs from diverse technologies (complementary metal-oxide semiconductor [CMOS], SiGe, silicon-on-insulator [SOI], PIC, III-V semiconductors, etc.) can be independently optimized and tightly integrated in small form factor packages to achieve power-performance-form factor-cost optimization that is not otherwise possible with monolithic integration.

Data center OEs use ICs from diverse technologies such as CMOS/SiGe for drivers and amplifiers, and III-V lasers. These ICs need to be integrated into very small form factor system-in-packages (SiPs) to be able to bring in optical signals and convert them into electrical signals (and vice versa) to be processed by switches, FPGAs or other application-specific ICs (ASICs). Therefore, electronic-photonic heterogeneous integration enabled by advanced wafer-level packaging is a promising technique to realize high-speed optical engines for the data center.

For the best performance with respect to speed, the EIC is integrated directly on top of the PIC. Currently, silicon photonics is the most promising technology that can provide the necessary performance and highest functionality integration. The PIC here is an active interposer—it not only consists of the photonics circuit, but also provides the necessary physical area for the EIC and external routing and integration [1]. However, this is a costly solution because the PIC will require an additional large area to support the EIC. In addition, for high-speed integration, through-silicon vias (TSVs) will be required to be used for the PIC active interposer. The fabrication of the TSV on the PIC requires many complex process steps as described in [2]. Furthermore, the Si photonics circuit is fabricated in a 200mm wafer instead of a 300mm wafer, which makes the cost of this integration solution unattractive for most commercial applications. The alternative solution is using various complex customized multiple assembly technologies for the multi-chip module integration. However, because of the complex assembly requirement and the limited scaling capability, this solution is not preferred.

IME is currently developing two packaging platforms based on fan-out wafer-level packaging (FOWLP) and the Si-interposer to address the EIC and PIC integration requirement. The first solution is based on the established FOWLP technology, which has been demonstrated for digital and radio-frequency (RF)/mmWave applications [3] integration. The FOWLP has good RF performance and allows multi-chip heterogeneous integration to be well suited for this high-performance OE application. For the Si-interposer with a high-resistivity substrate, it can also provide very high-speed and high-bandwidth PIC and EIC integration. Additionally, it can provide a sub-micron alignment feature for the fiber to the PIC passive assembly.

Photonics FOWLP

The FOWLP integration platform we describe is for high-speed PIC and EIC integration. It is low cost and it leverages advanced FOWLP development for electronic semiconductor packaging. Currently, only the PIC with vertical coupling in FOWLP has been demonstrated [4]. The integration of the edge PIC remains a big challenge because the embedding mold compound is not transparent in the optical signal spectrum, and the optical signal cannot be coupled from the optical I/O at the edge of the PIC after molding. The optical I/O is very sensitive to contamination because the optical signal has a very short wavelength (~1.3µm to ~1.5µm), therefore, polishing and cleaning the package to reveal the PIC edge I/O is very challenging and time consuming. To overcome this problem, the PIC optical I/O at the edge will be designed with an additional silicon buffer section, as

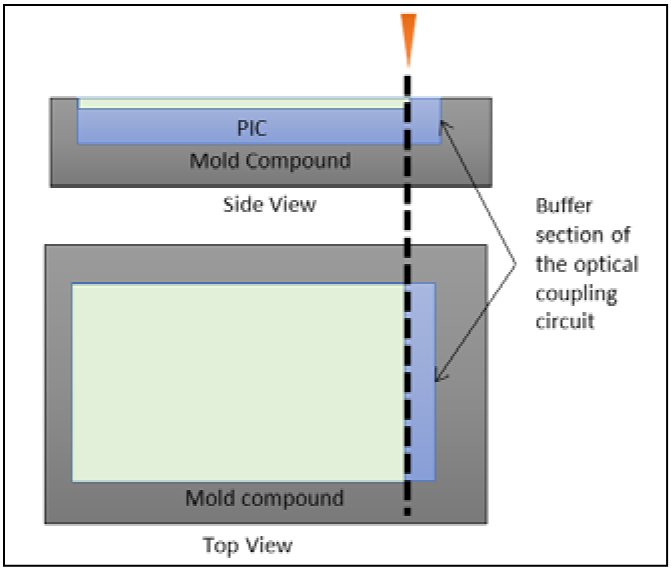


Figure 1: Schematic showing the FOWLP embedding solution of PIC with edge optically-coupled I/Os.

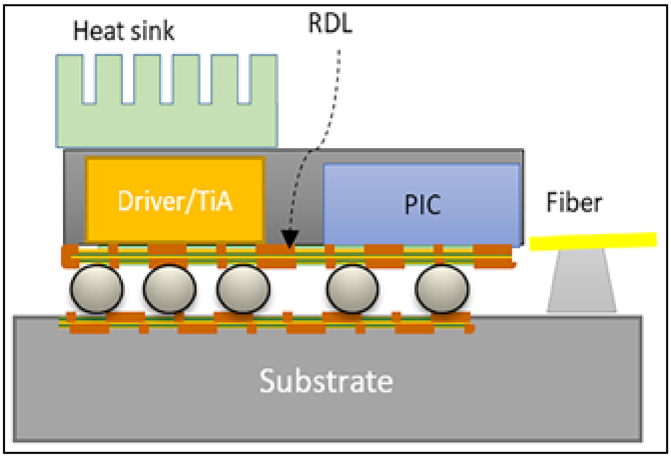


Figure 2: Schematic side view of the proposed integrated PIC and EIC FOWLP flip-chip attached to the substrate.

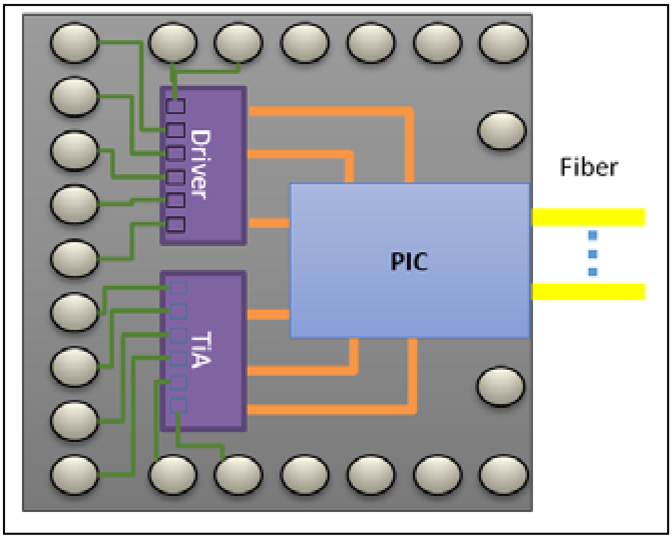


Figure 3: Schematic bottom view of the FOWLP with PIC and EIC (TiA and driver).

shown in Figure 1. The silicon buffer is just an additional section of the Si substrate of the PIC, extending beyond the edge optical coupling circuit. The PIC with this buffer section design is then embedded in the FOWLP process. The PIC buffer section protects the optical I/Os and will only be diced off during the singulation, or dicing process, to expose the PIC optical I/Os. Depending on the application requirement, the EIC can be either embedded side by side with the PIC, or integrated on top of the FOWLP PIC.

The extended mold area of the FOWLP is used to support the electrical lateral and vertical routing connections using the redistribution layer (RDL). The RDL can provide very fine line width connections and reduce the parasitic components with good impedance control for high-speed connections. The PIC and EIC can be integrated laterally with a gap of less than 500µm. In this way, it can achieve a very short electrical length and a well-matched interconnect to support the high-speed and high-density integration. Through-mold vias (TMVs) can be designed on the FOWLP and the EIC can be stacked directly on top of the PIC for 3D integration. Although this can help to reduce the package size, it increases the design and process complexity.

Figures 2 and 3 show the schematic of the proposed laterally-integrated PIC and EIC using the FOWLP. The EIC here consists of the driver for the photonics modulator and the transimpedance amplifier (TiA) for the photodetector in the EIC. The FOWLP is designed for the flip-chip assembly on the substrate. Flip-chip assembly is preferred as it can provide high-speed and high-bandwidth connection to the switch IC or ASIC on the mainboard. In addition, the flip-chip assembly allows the heat sink to be designed directly on the backside of the EIC to reduce thermal effects.

For the integration design described above, the high-speed signal lines between the PIC and EIC, and the EIC to the external circuit, are usually designed using a differential transmission line. The performance of the differential transmission lines of 1.0mm length is modeled using a 3D (electromagnetic) EM simulator as shown in Figure 4. The typical dielectric constant and the loss tangent of the mold compound are around 3.4 and 0.008, respectively. The simulated results are shown in Figure 5. The return loss is more than 30dB, and the insertion loss ranges from less than 0.15dB up

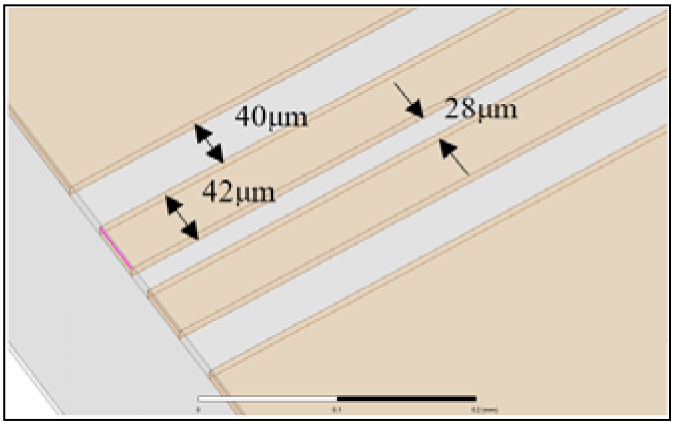


Figure 4: Simulation model of a differentiated GSSG transmission line design on FOWLP.

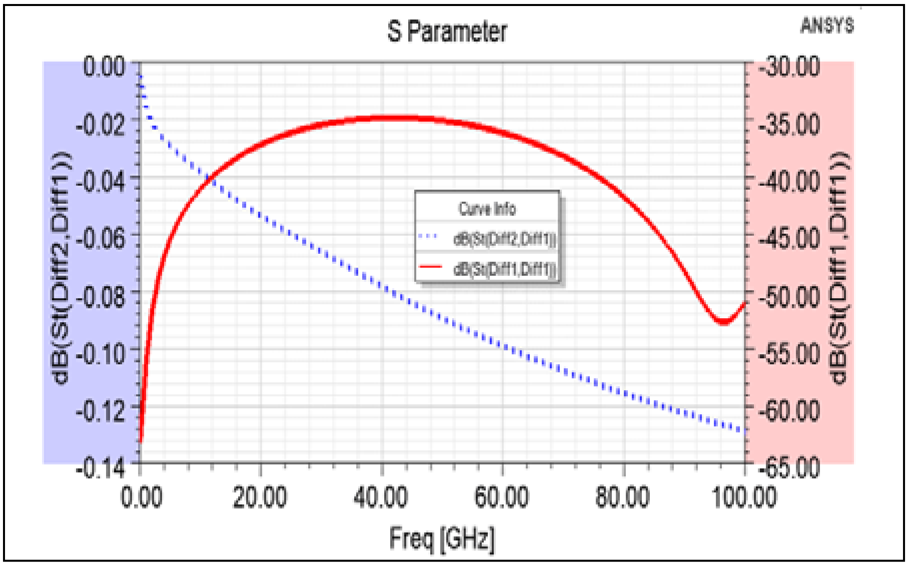


Figure 5: Simulation frequency response of the FOWLP differentiate transmission line up to 100GHz.

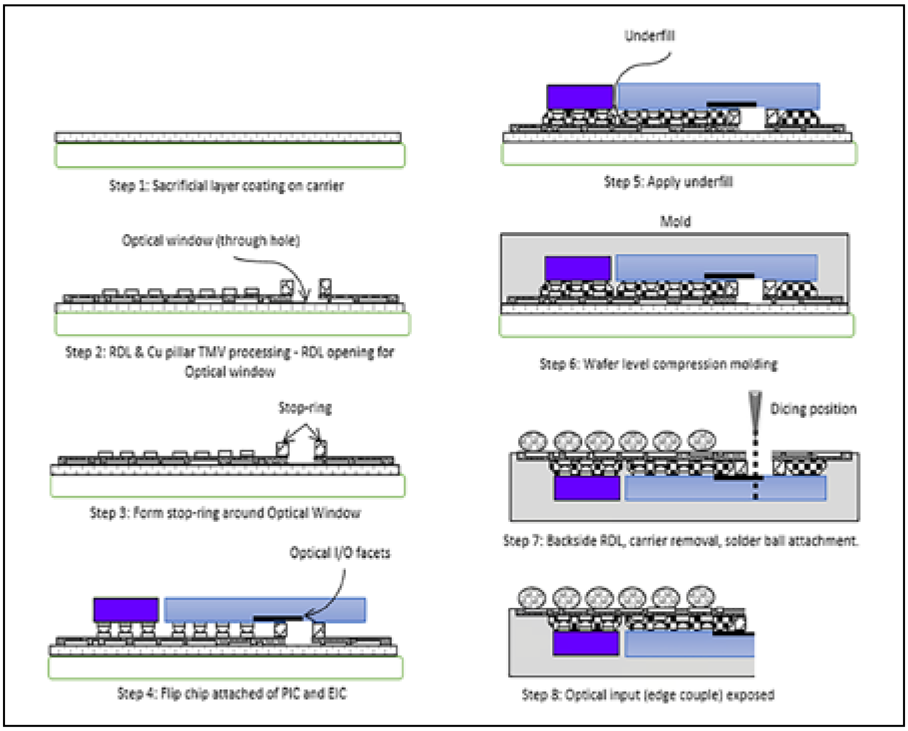

Step 1: Sacrificial layer coating on carrier
Optical window (through hole)
Step 2: RDL & Cu pillar TMV processing - RDL opening for Optical window
Stop-ring
Step 3: Form stop-ring around Optical Window
Optical I/O facets
Step 4: Flip chip attached of PIC and EIC
Underfill
Step 5: Apply underfill
Mold
Step 6: Wafer level compression molding
Dicing position
Step 7: Backside RDL, carrier removal, solder ball attachment.
Step 8: Optical input (edge couple) exposed

Figure 6: Proposed FOWLP process flow for PIC embedding.

to 100GHz. The simulated results show that the FOWLP interconnect has a large bandwidth and is capable of supporting the high-speed digital signal in the OE. These excellent electrical performances are achieved mainly because of the short interconnect design.

FOWLP photonics process. The key challenge of molding the PIC is to prevent contamination on the photonics I/O facets. The PIC is first singulated or diced at a distance from the photonics I/O so that a buffer area is formed. The buffer area is part of protection for the photonics I/O facet during the subsequent molding process.

One of the possible processes that could be used to achieve the FOWLP is described here. The steps are similar to the RDL-first process developed by IME [5] as shown in **Figure 6**. The additional process step is to form the stop ring (dam). The stop ring is formed using a polymer or other suitable material if necessary. It is used to prevent the underfill in the subsequent step from flowing to the optical I/O area, or protected area. The singulation process is designed so that the PIC buffer area together with its surrounding mold are diced off at Step 7. The FOWLP of the EIC and PIC with optical edge coupling is then formed. The integrated photonics FOWLP module can then be flip-chip assembled to the main optical board or substrate for the main integration. The process described above is compatible with the FOWLP process. IME is currently developing the complete process and evaluating the optical coupling performance.

Figure 7a shows the PIC test vehicle after the FOWLP molding process before the dicing is done. The buffer region in the PIC helps to protect the molding compound from coming in contact with the PIC optical I/O facets. **Figure 7b** shows the PIC test vehicle after dicing. The picture shows that the molding compound adheres well to the PIC sidewall and did not flow in such a manner as to contaminate the PIC; the optical coupler is exposed after dicing off the buffer region.

Photonics through Si-interposer

The other alternative integration platform for the EIC and PIC is based on the through-silicon interposer (TSI), which has also been successfully developed for high-density electrical integration. The schematic of this overall integration design is shown in **Figure 8**. The TSI is used to integrate the EIC and PIC, as well as to support the fiber assembly for the PIC. The main advantage of this Si-interposer is that it can provide the high alignment accuracy needed for multi-channel optical fiber alignment and assembly. For all photonics communication circuits, the optical signal of the PIC must be eventually coupled to the optical fiber for external connection. The fiber assembly needs to achieve a high alignment accuracy in the range of less than 2µm or 1µm (some PIC designs can only accept less than 1µm of misalignment) to achieve good optical coupling efficiency.

In this proposed fiber-to-PIC assembly solution, the Si-interposer serves as a base substrate to support the PIC and a fiber block, which holds the optical fiber array, as shown in

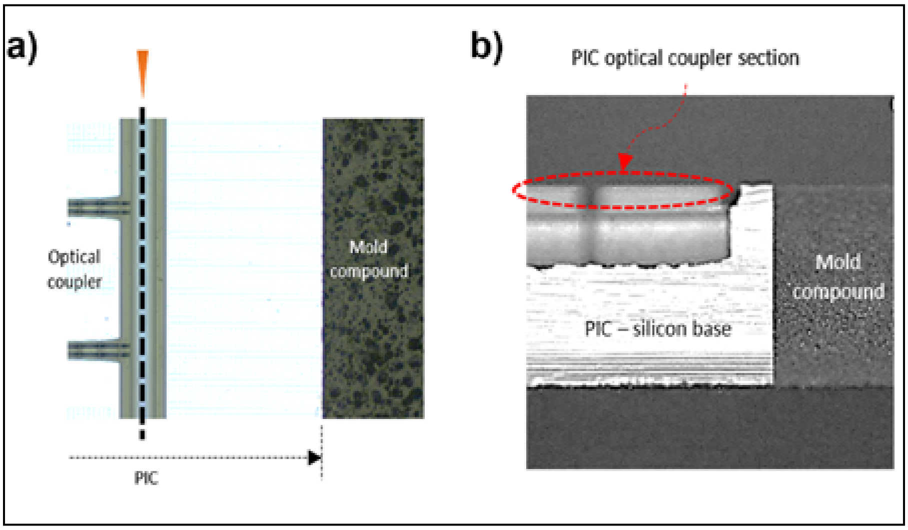


Figure 7: Photonics FOWLP: a) Top view of the test vehicle before dicing; b) Front view of the test vehicle after dicing.

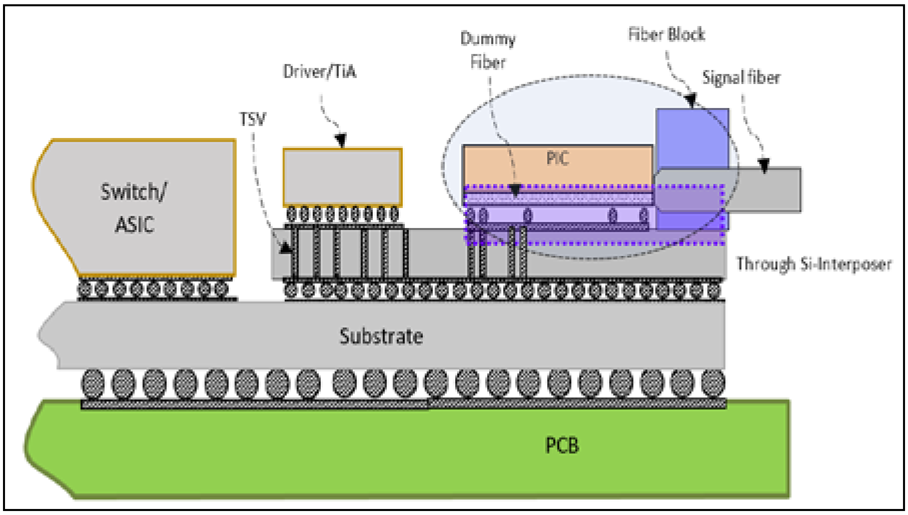


Figure 8: Proposed solution using a Si interposer to provide the fiber coupling and integration for PIC and EIC.

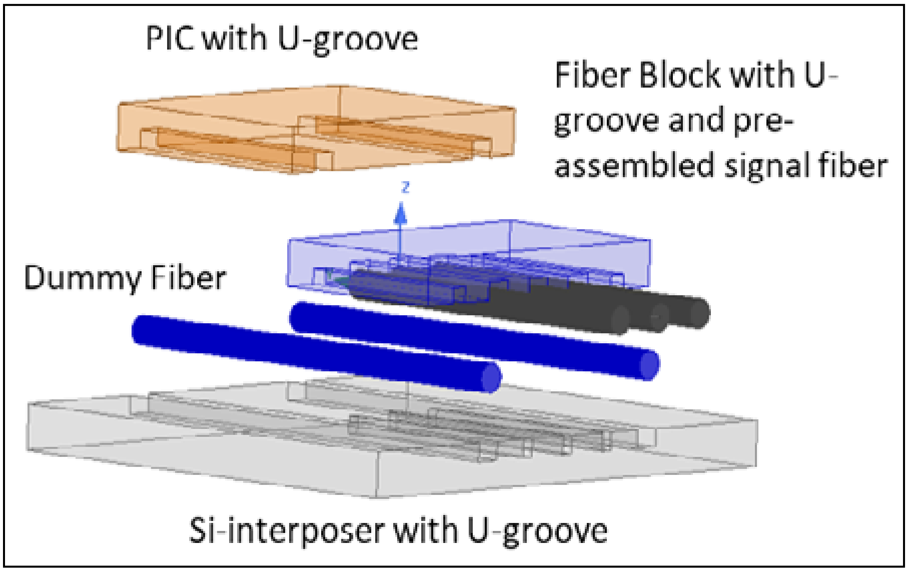

PIC with U-groove
Fiber Block with U-groove and pre-assembled signal fiber
z
Dummy Fiber
Si-interposer with U-groove

Figure 9: Side view of the Si electrical optical interposer.

Figure 8. The alignment of the PIC and the fiber block is achieved by a pair of alignment or dummy fibers placed on the deep trench or U-groove of the Si interposer. The Si-interposer is designed with a set of U-grooves along its length. The U-groove is formed using a dry etch process that is compatible with the PIC and TSI.

The PIC and the fiber block will also have a set of matching U-grooves designed along the length of the device. When the PIC and the fiber block are assembled face down on the Si-interposer, with the U-grooves of the PIC and the fiber block aligned on the dummy fiber of the Si-interposer, only the edges of the U-grooves will be physically in contact with the dummy fiber. **Figure 9** shows the concept of using the U-groove edges to support the fiber. In this way, the optical axis of the fiber and the PIC waveguide will be self-aligned with a vertical height offset determined by the width of the deep trench and the diameter of the dummy fiber. The dummy fiber, which is a standard bare optical fiber, is used as a cylindrical mechanical alignment and support feature because of its inherent high precision diameter, yet low cost. The standard optical bare fiber has a very precise diameter of 125µm diameter (Ø) ±0.7µm. This fiber-to-PIC alignment solution relies only on the edges of the U-grooves, which are used to support the alignment fiber. The depth and profile of the U-grooves are not critical as they do not have any physical contact with the alignment fiber (dummy fiber). Because epoxy will be applied in the U-grooves to assemble the alignment fiber, the depth of each U-groove just has to be deep enough to support the epoxy volume without pushing up the fiber, or affecting the fiber position, as shown in **Figure 10**. More details on the Si-interposer design can be found in [6].

It should be noted that the coupling efficiency also depends on the performance of the optical coupler of the PIC. An optical coupler is usually used to mitigate the coupling loss because of the large mode size mismatch between the fiber and the optical waveguide. Besides having low coupling loss, the optical coupler should have a large misalignment

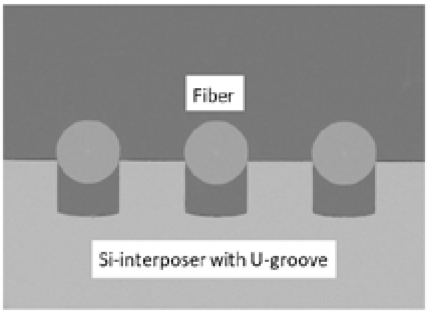


Figure 10: Cross-section of the test vehicle assembly showing the U-groove edge supporting the fiber.

tolerance to relax the alignment accuracy requirement. For these reasons, the suspended coupler design is the preferred design over the typical inverse tapered coupler. The reported suspended coupler [7] has a 1dB excess loss of more than 2µm in the cross-section misalignment, and more than 40µm along the fiber direction. The performance of the suspended coupler closely matches the requirement of this Si-interposer integration platform.

Both the PIC and the EIC are flip-chip assembled on the Si-interposer.

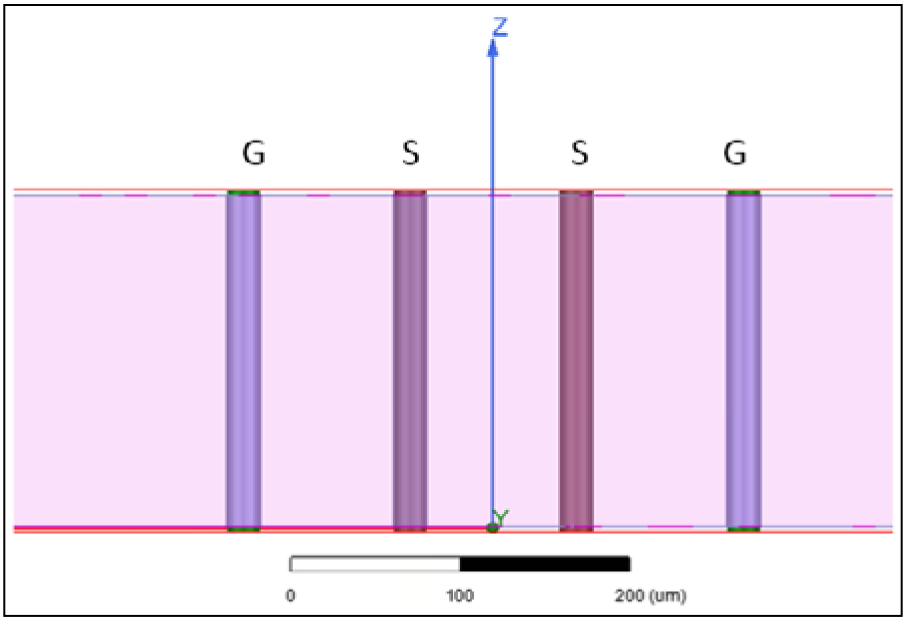


Figure 11: GSSG TSV model.

Usually, the frontside electrical connection of the interposer is formed using back-end-of-line (BEOL) processing, which supports the bare die assembly using micro solder bumps. Typically, up to three metal layers can be formed on the top side of the Si-interposer to support the front-side routing. The minimum metal line width is around 0.8µm, which is more than sufficient to support all the electrical routing between the PIC and the EIC. The electrical connection on the backside of the interposer is formed using RDL, which has a much larger line width resolution. As most likely the top side BEOL has provided all the electrical routing, the backside RDL is used mainly to support the solder bump for the assembly of the Si-interposer to the substrate or printed circuit board (PCB). The electrical connection between the micro solder bump and the RDL is accomplished by using the TSV. The TSV helps to overcome the bandwidth-limiting inductance of the wire bond and increases the I/O density of the interconnect. For high-speed electrical connection, a high-resistivity Si wafer is required to reduce the electrical loss.

For the TSI integration, the length of the TSV has to be much greater than the U-groove depth. The dummy fiber is designed to be supported by the edge of the U-groove, therefore, the depth of the U-groove needs to be deep enough so that it can accommodate the epoxy without overflow or pushing up the dummy fiber. Based on the preliminary design, the depth of the U-groove needs to be more than 50µm, but to prevent the Si-interposer breakage during handling, it is recommended to increase the height of the Si-interposer to about 200µm.

The EM-simulated response of a differential ground-signal-signal-ground (GSSG) TSV with a height of 200µm is shown in Figure 11. The typical aspect ratio of the TSV is 1:10, so for a 200µm high TSV, the diameter will be around 20µm. The pitch of the TSV is set to 100µm, and a high-resistivity wafer of 700Ω.cm is used.

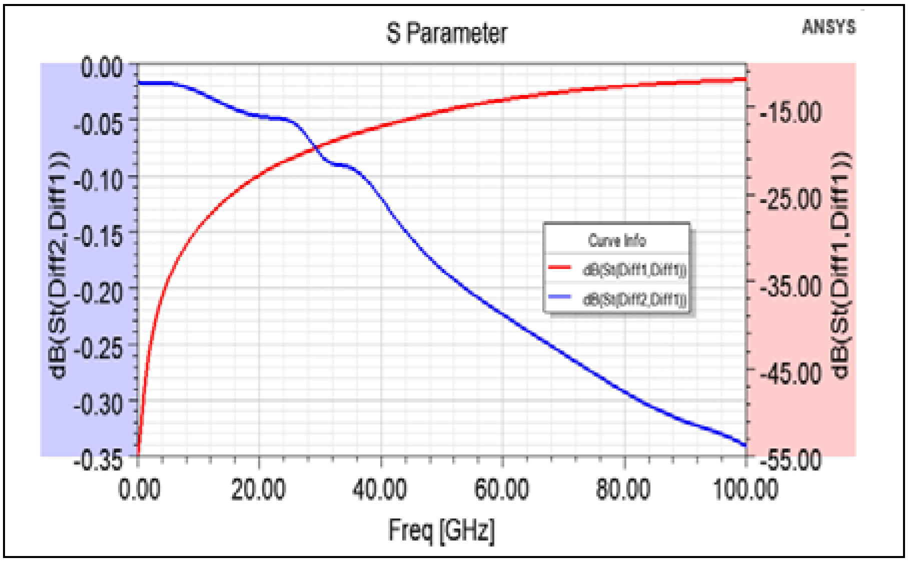


Figure 12: Simulated frequency response of the differential (or GSSG) TSV.

As shown in Figure 12, the simulated ideal insertion loss is less than 0.5dB and the return loss can go from 10dB up to 100GHz. Again, the simulated result shows that the TSI can support the high-speed high-density EIC and PIC integration for the data center OE.

Summary

Two EIC and PIC integration platforms are proposed here. The FOWLP and the TSI platforms allow the PIC and the EIC to be integrated close together with high-density routing. This helps to improve the electrical performance and reduce the form factor, both of which are required to meet the current and future OE for the data center. For the FOWLP platform, the main solution to enable the embedding of the edge-coupled PIC has been illustrated. An extra section of the silicon substrate is designed on the PIC to protect the optical I/O. This additional section of silicon substrate is diced off after the FOWLP process to expose the PIC optical I/Os. This design is simple and is compatible with the current FOWLP process. This will enable the high-speed photonics circuit to be integrated using the cost-effective FOWLP platform.

The TSI integration platform, besides providing high-speed PIC and EIC integration, also provides the alignment feature for the fiber to the PIC assembly. Together with the PIC optical coupler, the fiber can be aligned using passive alignment, which is one of the main photonics packaging challenges.

Acknowledgments

We would like to thank our fellow co-authors: Chockanathan Vinoth Kanna, Wu Jiaqi, Chong Ser Choong, and Surya Bhattacharya. This article is based on materials originally presented at the 2021 IEEE 71st Electronic Components and Technology Conference (ECTC), June 2021. Additionally, this work is partially funded by A*ccelerate of A*STAR (ETPL/18-GAP056-R20A).

References

1. G. Denoyer, C. Cole, A. Santipo, R. Russo, C. Robinson, L. Li, et al., “Hybrid silicon photonic circuits and transceiver for 50Gb/s NRZ transmission over single-mode fiber,” Jour. of Lightwave Tech., 2015 | Vol. 33, Issue: 6.
2. Y. Yang, M. Yu, Q. Fang, J. Song, X. Tu, P. G-Q. Lo, et al., “3D silicon photonics packaging based on TSV interposer for high-density on-board optics module,“ 2016 IEEE 66th Elec. Comp. and Tech. Conf. (ECTC).
3. T G. Lim, D. S. W. Ho, E. W. L. Ching, Z. Chen, S. Bhattacharya, “FOWLP design for digital and RF circuits,” 2019 IEEE 69th ECTC.
4. H. Uemura, K. Warabi, K. Ohira, Y. Kurita, H. Yoshida, H. Furuyama, et al., “Backside optical I/O module for Si photonics integrated with electrical ICs using fan-out wafer-level packaging technology,” 2018 IEEE 68th ECTC.
5. V. S. Rao, C. T. Chong, D. Ho, D. M. Zhi, C. Ser Choong, L. P. S. Sharon, et al., “Development of high-density fan-out wafer-level package (HD FOWLP) with multi-layer fine-pitch RDL for mobile applications,” 2016 IEEE 66th ECTC.
6. L. Teck Guan, L. Hong Yu, J. Ming Chinq, E. Wai Leong Ching, C. Ser Choong, L. Soon Thor, et al., “Silicon Optical Electrical Interposer - Fiber to the Chip,” 2019 IEEE 21st Elec. Packaging Tech. Conf. (EPTC).
7. J. Lianxi, L. T-Yang, L. Chao, L. Xianshu, T. Xiaoguang, H. Ying, et al., “High efficient suspended coupler based on IME’s MPW platform with 193nm lithography,” 2017 Optical Fiber Comm. Conf. and Exhibition (OFC).

Biographies

Lim Teck Guan is Senior Scientist at the Institute of Microelectronics, A*STAR, Singapore. He received his PhD in Microwave Photonics from U. of Surrey, UK. His current research focus is on the development of FOWLP and Si-interposer integration solutions for mmWave and photonics circuits. Email: limtg@ime.a-star.edu.sg

ADVERTISER INDEX

Advertiser	Website	Page
Amkor Technology	www.amkor.com	1
Brewer Science	www.brewerscience.com	47
CyberOptics	www.cyberoptics.com	29
DL Technology	www.dltechnology.com	10
E-tec Interconnect	www.e-tec.com	48
ECTC	www.ectc.net	56
EV Group	www.evgroup.com	2
IMAPS Device Packaging	www.devicepackaging.org	53
INTEKPLUS CO., LTD.	www.intekplus.com	23
Ironwood Electronics	www.ironwoodelectronics.com	42
ISC CO., LTD.	www.isc21.kr	36
ITW EAE	www.itweae.com	13
Johnstech International	www.johnstech.com/future	IBC
Leeno Industrial	www.leeno.com	IFC, 32
Plasma Etch	www.plasmaetch.com	54
Sonix	www.sonix.com	44
SÜSS MicroTec	www.suss.com	7
Technic	www.technic.com	18
Test Tooling Solutions Group	www.tts-grp.com	14, 15
TSE Co. Ltd	www.tse21.com	4,5
Universal Instruments	www.uic.com	OBC
WinWay Technology	www.winwayglobal.com	27
Yield Engineering Systems	www.yieldengineering.com	35

March April 2022 • Space Close March 1st • Ad Materials Close March 4th • For Advertising Inquiries • ads@chipscalereview.com

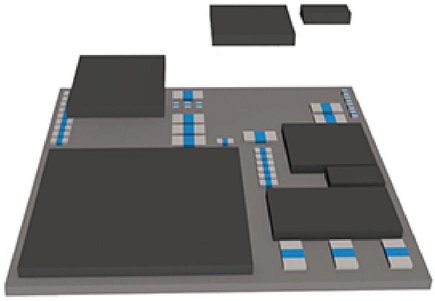
THE FUTURE STARTED WITH
Johnstech™
SINCE 1991
Johnstech
CERTIFIED
ISO 9001:2015
CERTIFIED

Universal Instruments

MAXIMIZE YIELDS for Heterogeneous Integration

Heterogenous Integration is enabling higher-bandwidth, lower power consumption, and increased functionality in virtually all of the newest high-tech products – all within a smaller form factor.

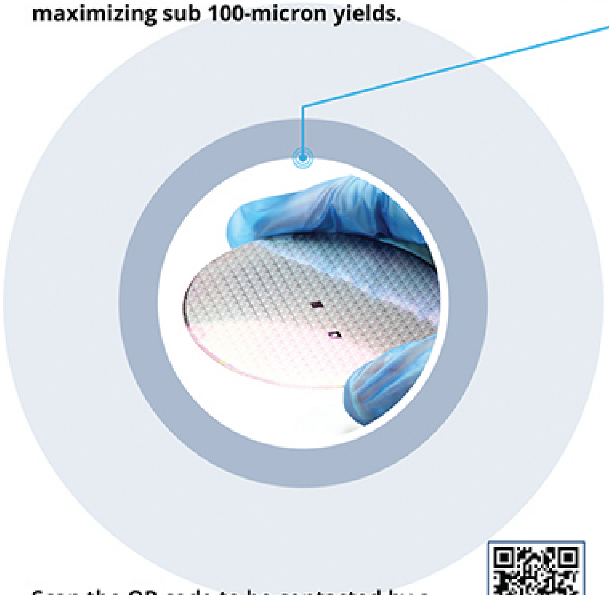
However, building the HI modules that power these devices brings a host of new challenges, demanding a comprehensive solution that breaks traditional boundaries for efficient multi-die assembly.



THIN DIE HANDLING

One such challenge is precision die processing for multi-die packages requiring die stacking with a silicon interposer and chiplets.

Universal's FuzionSC + HSWF solution has the thin die handling, precision ejection capability and system accuracy to transform this challenge into a significant competitive advantage by **maximizing sub 100-micron yields.**



Scan the QR code to be contacted by a Universal Representative and learn more:

FUZIONSC & HIGH-SPEED WAFER FEEDER

The High-Speed Wafer Feeder (HSWF) is the world's fastest rapid-exchange multi-die feeder. Combined with Universal's FuzionSC™ Platform, it is the **ultimate multi-die solution for heterogeneous integration.**

- **WAFER CAPACITY** – to minimize replenishment rate
- **SPEED** – to meet volume requirements
- **MULTIPLE DIE TYPES** – to maximize utilization
- **LARGE SUBSTRATE** – to reduce manufacturing costs
- **THIN DIE** – to maximize sub 100-micron yields

DISCOVER THE HSWF

uic.com