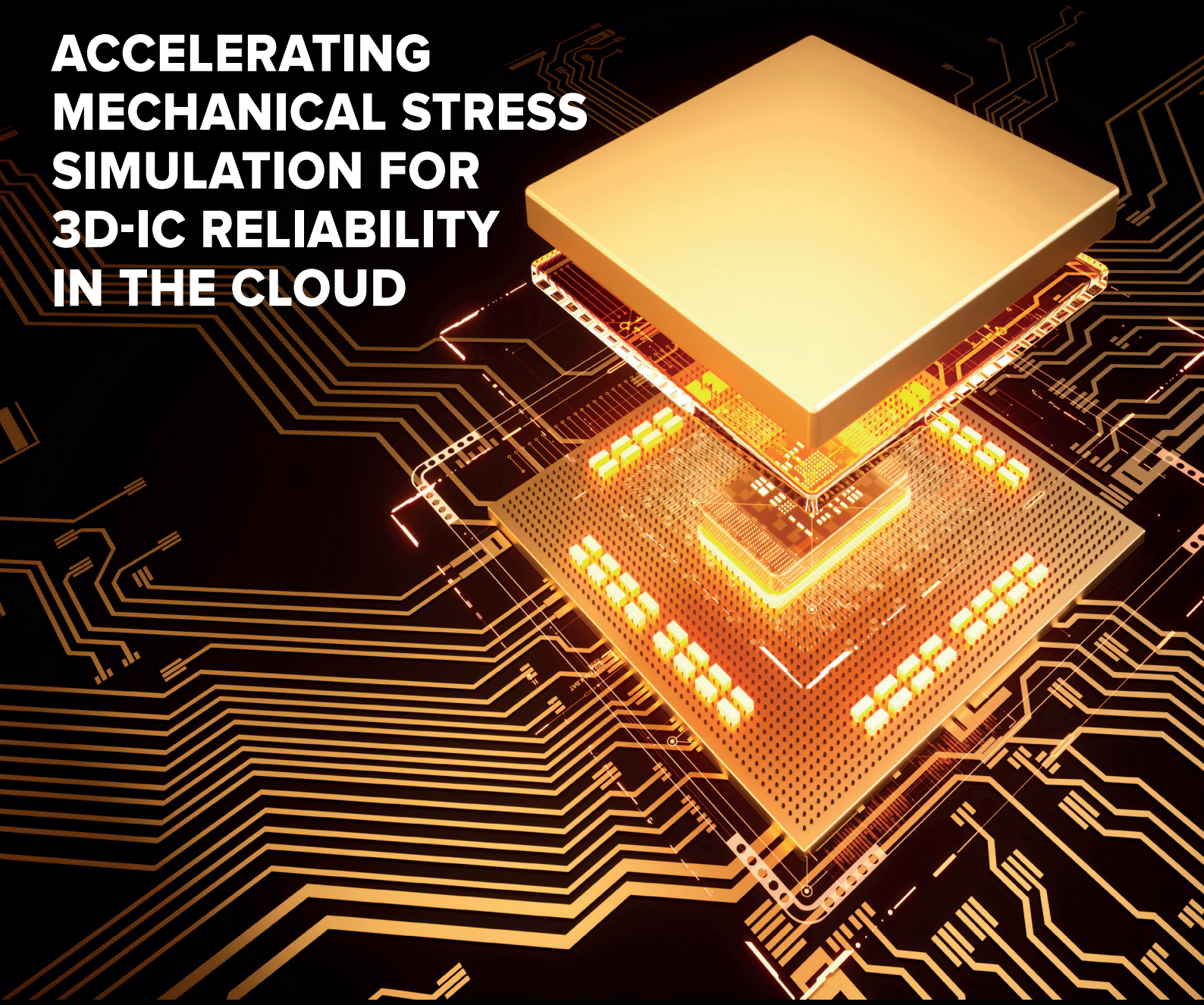
SILICON SEMICONDUCTOR
CONNECTING THE SILICON SEMICONDUCTOR COMMUNITY

ACCELERATING MECHANICAL STRESS SIMULATION FOR 3D-IC RELIABILITY IN THE CLOUD



VOLUME 45 ISSUE I 2024
AN ANGEL BUSINESS COMMUNICATIONS PUBLICATION
SILICONSEMICONDUCTOR.NET

INSIDE

Dry cleaning process for optimum cleaning

Whether energy, mobility or industry - electronic systems and components play a key role towards climate neutrality

Microprocessor architecture choice in the age of Ai

In the age of AI, the microprocessor selection for anchoring an AI solution is especially important

STMicroelectronics boosts chip design speed

Europe's largest semiconductor company optimised performance, cost, and energy consumption by choosing AMD EPYC CPUs

EVG®

invent
innovate
implement

www.EVGroup.com

NANOCLEAVE™ LAYER TRANSFER REVOLUTIONIZES ADVANCED PACKAGING AND LOGIC SCALING

- IR laser cleave technology enables ultra-thin-layer transfer from silicon carriers
- Nanometer-precision release of bonded, deposited or grown layers
- Front-end compatibility through silicon carriers and inorganic release materials
- Supporting future roadmaps in advanced packaging and transistor scaling

GET IN TOUCH to discuss your manufacturing needs
www.EVGroup.com

EVG®850 NanoCleave™

VIEWPOINT

By Phil Alsop, Editor

Promise of a better year ahead, geopolitics permitting

The good news seems to be that, no matter which forecast(s) you subscribe to, the industry is predicted to enjoy some significant growth this year, which might not quite make up for the 2023 downturn numbers, but is nonetheless very welcome. All things AI seem to be fuelling this growth – with the data centre and related high performance computing sectors seeing the lion's share of the resulting increased semiconductor demand. There is some suggestion that AI will also begin to make inroads into the personal device market as well.

2.5/3D packaging is another area highlighted by IDC as enjoying a good increase in numbers over the next few years – something confirmed by Allied Market Research's '3D Semiconductor Packaging Market' research report. As the Ansys interview in this issue highlights, the main challenge for this market right now seems to be that a range of companies are coming at the problem/opportunity from different perspectives and, as yet, there seems to be no agreement as to the development of any kind of standards. So, for now at least, a dynamic space, with plenty of investment and innovation, but also a fair amount of uncertainty as to the exact 2.5D/3D packaging roadmap.

Elsewhere in the news, the ongoing skills shortage and the somewhat linked continuing on-/reshoring trend are front and centre. In no particular order, the Semiconductor Research Corporation (SRC) has unveiled the Microelectronics and Advanced Packaging (MAPT) Roadmap, described as 'the first industry-wide 3D semiconductor roadmap to guide the forthcoming microelectronic revolution' (there's 3D cropping up again!); the European Commission has officially inaugurated the Chips Joint Undertaking (Chips JU), designed to 'reinforce the European semiconductor ecosystem and Europe's technological leadership' – it will do this by bridging the gap between research, innovation and production; and, 'in response to the critical shortage of skilled engineers in semiconductor IC design and layout', a new initiative has been launched by Professor Patrick Reynaert of KU Leuven, working alongside several semiconductor companies in Flanders.

I will resist the urge to bring up the three Ss once more(!), but maybe introduce a fourth for the new year – security. Sadly, no matter where one looks in the world right now, a combination of human-made and natural chaos continues to threaten long-held certainties in so many ways. From the US where it seems, no matter the result of the forthcoming presidential election, some kind of subsequent instability seems unavoidable, through the Middle East's regular disruption, to the ongoing tension in Asia, and not forgetting the war in Europe...security is under scrutiny in so any ways.

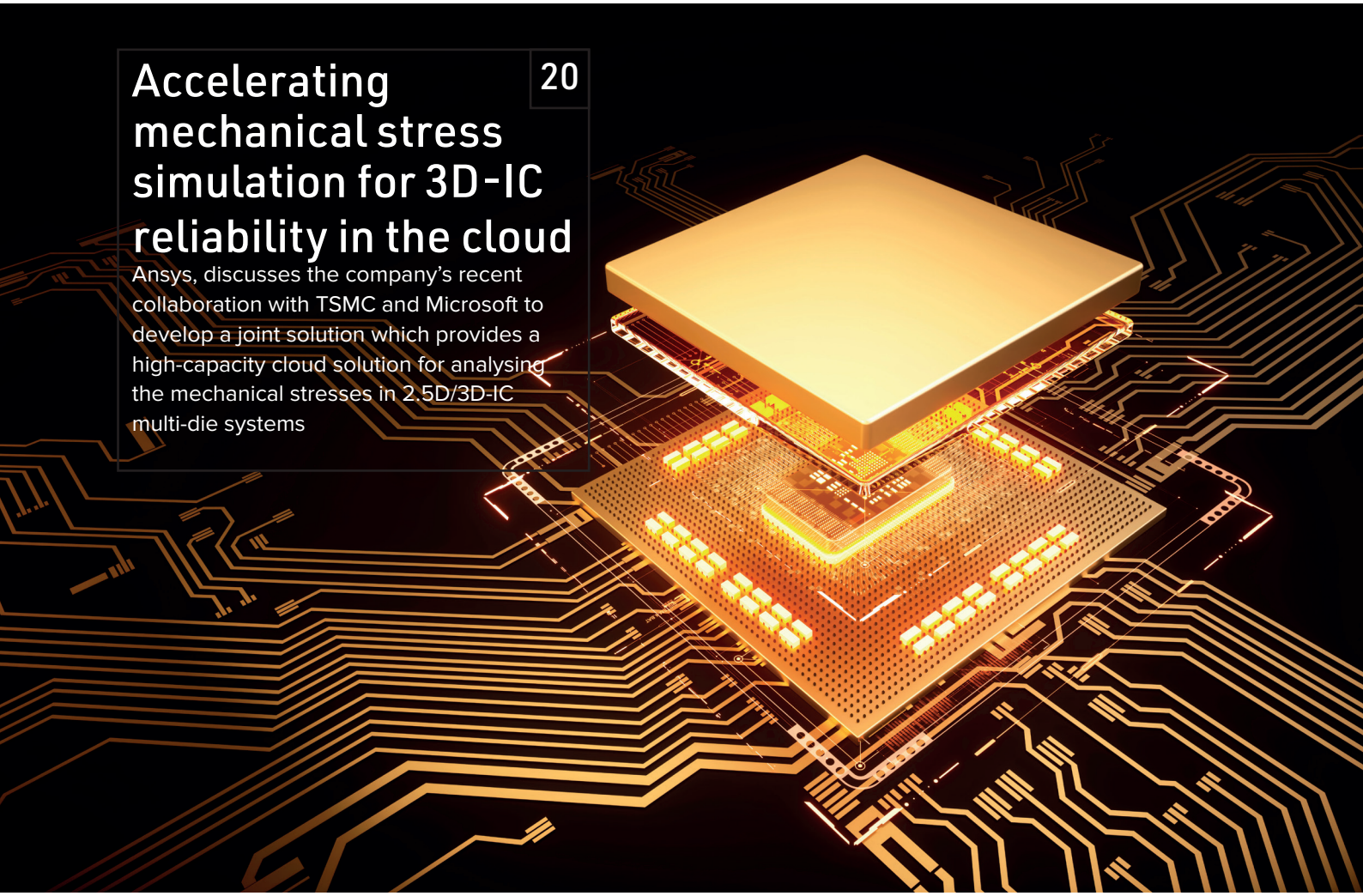
Let us hope that the geopolitical events of 2024 do not derail the semiconductor industry's growth to any great extent. I write more in hope than expectation.

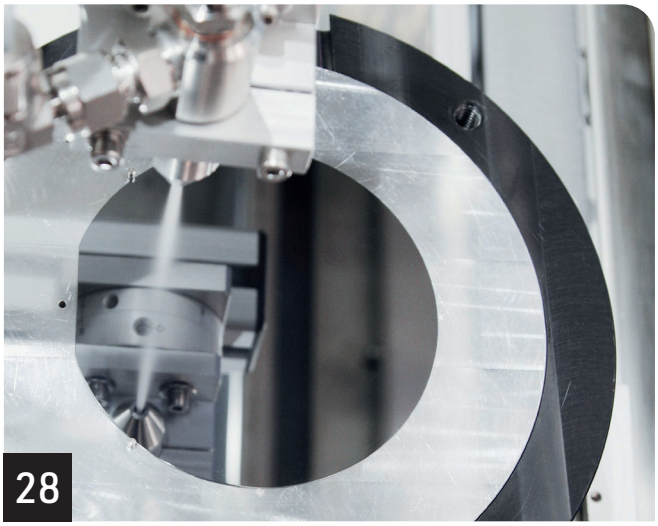
For your free weekly e-news roundup go to: siliconsemiconductor.net

CONTENTS

VOL 45 ISSUE I 2024

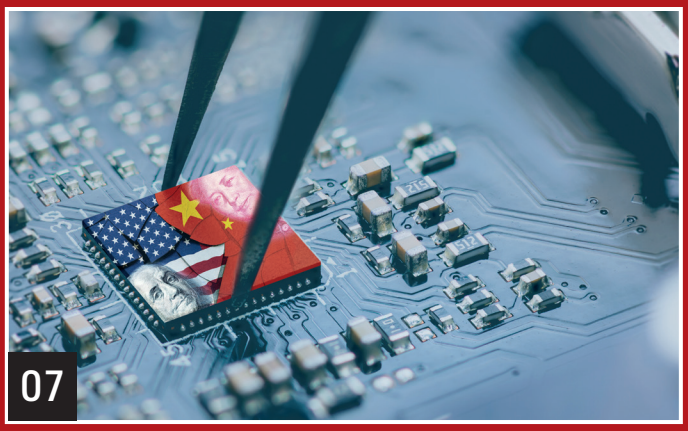
SILICON SEMICONDUCTOR





28

NEWS



Publisher Jackie Cannon jackie.cannon@angelbc.com +44 (0)1923 690205
Editor Phil Alsop phil.alsop@angelbc.com
Sales & Marketing Manager Shehzad Munshi shehzad.munshi@angelbc.com +44 (0)1923 690215
Marketing & Logistics Executive Eve O'Sullivan eve.osullivan@angelbc.com +44 (0)2476 823123
USA Representatives Tom Brun Brun Media tbrun@brunmedia.com +001 724 539-2404
Janice Jenkins jjenkins@brunmedia.com +001 724-929-3550
Director of Logistics Sharon Cowley sharon.cowley@angelbc.com +44 (0)1923 690200
Design & Production Manager Mitch Gaynor mitch.gaynor@angelbc.com +44 (0)1923 690214

Chief Executive Officer Sukhi Bhadal sukhi.bhadal@angelbc.com +44 (0)2476 718970
Chief Technical Officer Scott Adams scott.adams@angelbc.com +44 (0)2476 718970
Directors Jackie Cannon, Sharon Cowley

Published by Angel Business Communications Ltd,
6 Bow Court, Fletchworth Gate, Burnsall Road, Coventry CV5 6SP, UK.
T: +44 (0)2476 718 970 E: info@angelbc.com

Angel BUSINESS COMMUNICATIONS

recycle When you have finished with this magazine please recycle it.

Silicon Semiconductor is published ten times (6x print and 4x video issues) a year on a controlled circulation basis. Non-qualifying individuals can subscribe at: £50.00 per annum (UK) , €60.00 per annum (Europe), $90 per annum (air mail)(USA). Cover price £4.50. All information herein is believed to be correct at time of going to press. The publisher does not accept responsibility for any errors and omissions. The views expressed in this publication are not necessarily those of the publisher. Every effort has been made to obtain copyright permission for the material contained in this publication. Angel Business Communications Ltd will be happy to acknowledge any copyright oversights in a subsequent issue of the publication. Angel Business Communications Ltd © Copyright 2024. All rights reserved. Contents may not be reproduced in whole or
part without the written consent of the publishers. The paper used within this magazine is produced by chain of custody certified manufacturers, guaranteeing sustainable sourcing. US mailing information: Silicon Semiconductor, ISSN 1096-598X, is published 5 times a year by Angel Business Communications Ltd, Unit 6, Bow Court, Fletchworth Gate, Burnsall Rd, Coventry CV5 6SP. UK. Airfreight and mailing in the USA by agent named World Container INC 150-15, 183rd St, Jamaica, NY 11413, USA. Periodicals Postage Paid at Brooklyn, NY 11256. POSTMASTER: Send address changes to Silicon Semiconductor, Air Business Ltd, c/o World Container INC 150-15, 183rd St, Jamaica, NY 11413, USA.

We strive for accuracy in all we publish; readers and contributors are encouraged to contact us if they recognise an error or omission. Once a magazine edition is published [online, in print or both], we do not update previously published articles to align old company names, branding, marketing efforts, taglines, mission statements or other promotional verbiage, images, or logos to newly created or updated names, images, typographic renderings, logos (or similar) when such references/images were accurately stated, rendered or displayed at the time of the original publication. When companies change their names or the images/text used to represent the company, we invite organizations to provide Angel Business Communications with a news release detailing their new business objectives and/or other changes that could impact how customers/prospects might recognise the company, contact the organisation, or engage with them for future commercial enterprise.
Printed by: The Manson Group. ISSN 2050-7798 (Print) ISSN 2050-7801 (Online) © Copyright 2024.

Photonic pursues silicon-based approach to quantum computing

Photonic Inc., a company building one of the world's first scalable, fault-tolerant, and unified quantum computing and networking platforms based on photonically linked silicon spin qubits, has introduced its quantum architecture.

PHOTONIC'S unique approach to quantum computing stands to accelerate innovation across the quantum ecosystem.

"Quantum computing is real, and we believe that—within five years, significantly sooner than the widely accepted timeframe—we will be the first quantum computing company to offer a scalable, distributed, and fault-tolerant solution," said Dr. Stephanie Simmons, Founder and Chief Quantum Officer of Photonic and Co-chair of Canada's National Quantum Strategy Advisory Board. "These are the capabilities that must be delivered for quantum computing to be a relied upon across industries, and we believe that we have correctly identified the silicon T centre as the missing component needed to finally unlock the first credible path to impactful commercial quantum computing."

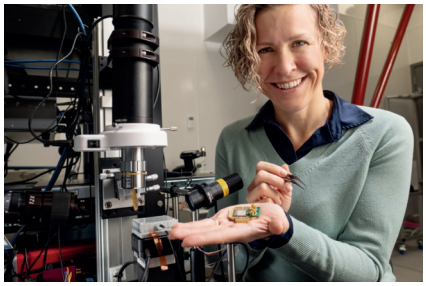
Photonic specializes in spin-photon interfaces in silicon, silicon integrated photonics, and quantum optics. Photonic has committed to silicon—the chemical element serving as the backbone of modern telecommunications and computation – as the dominant enabling design in quantum computing.

Leveraging colour centers and telecom photons, Photonic's patented technology provides computing (with spin qubits), networking (via photons), and memory. Photonic links in silicon deliver quantum entanglement not only between qubits on the same chip but also among multiple quantum chips. Silicon-based qubits enjoy substantially greater microelectronic-style scalability than other types of qubits. Indeed, Photonic's architecture achieves horizontal scaling. Photonic's highly connected qubit architecture also enables use of efficient quantum error correction codes, such as quantum LDPC (Low Density Parity Check) codes. These codes are known for extremely low physical to logical qubit overheads and fast and efficient hardware implementation.

Such attributes will prove increasingly crucial as quantum computing drives more applications. For example, Photonic's technology can deliver unparalleled security to global digital communications, offering a solution to the cybersecurity threat posed by the development of quantum computing. Beyond internet cybersecurity, quantum computing promises transformative capabilities for modelling and simulation of complex systems and processes in areas such as climate modelling, materials development, and creation of life-saving pharmaceuticals.

"Ultimately, the breadth of problems to which quantum computing can offer a solution means it will have a tangible, meaningful impact on people all around the world," said Dr. Paul Terry, Photonic Chief Executive Officer. "We're moving to large-scale, accessible quantum computers networked together to provide access to quantum services that will enable companies and governments to suddenly tackle problems that are, right now, beyond our capabilities because of the inescapable constraints of classical computing. It's incredibly thrilling to be on the cusp of this inflection point in quantum computing and, more broadly, physics history."

Photonic, which has built one of the world's largest teams of quantum silicon experts, has raised a total of $140 million USD in funding to date, and benefits from leveraging the globally mature manufacturing ecosystem for semiconductors and telecoms. The company is led by innovators with deep expertise in delivering disruptive



technologies. Dr. Simmons is a Tier 2 Canada research chair in silicon quantum technologies and a Canadian Institute for Advanced Research fellow in quantum information science. She is an associate professor in the physics department at Simon Fraser University in Burnaby, British Columbia, who has twice been accorded a Physics World Top 10 Breakthrough of the Year (2013 and 2015) for her work in silicon quantum technologies.

Dr. Terry is a seasoned executive, entrepreneur, engineer, and angel investor who has founded or has been one of the founding employees at six successful companies in areas from big data systems to supercomputing, to IP services and telecom networking.

In related news, Photonic also announced today Photonic Collaborating with Microsoft to Power Global Quantum Ecosystem, and Photonic Raises $100 Million USD for Quantum Technology from BCI, Microsoft, and Other Investors.

In addition, a Photonic technical paper, Scalable Fault-Tolerant Quantum Technologies with Silicon Colour Centres, details the company's novel architecture, and a Microsoft blog post, Microsoft and Photonic join forces on the path to quantum at scale, offers more perspective on that company's collaboration with Photonic. offers more perspective on that company's collaboration with Photonic.

China and US bolster semiconductor independence

TrendForce's latest findings reveal that as of 2023, Taiwan holds approximately 46% of global semiconductor foundry capacity, followed by China (26%), South Korea (12%), the US (6%), and Japan (2%).

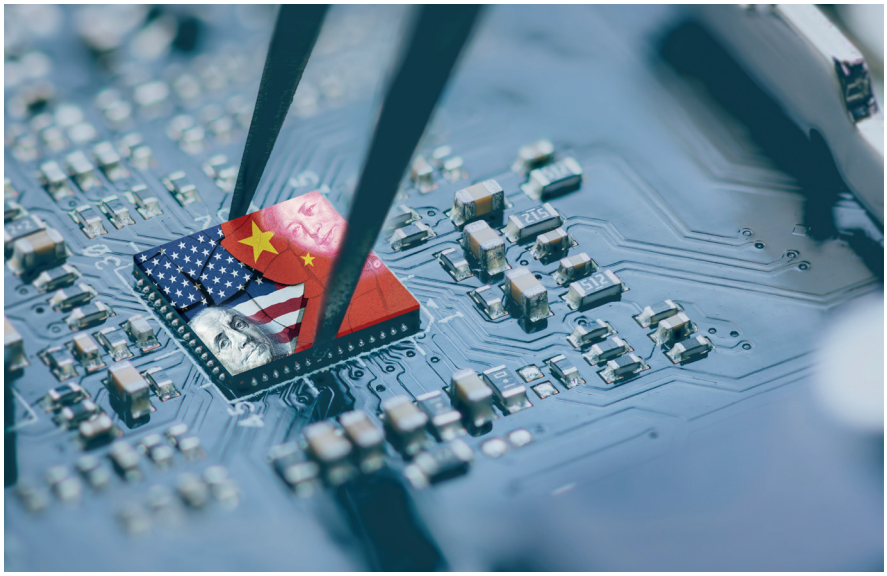
DUE to government incentives and subsidies promoting local production in countries like China and the US, the semiconductor production capacities of Taiwan and South Korea are projected to decrease to 41% and 10%, respectively, by 2027.

Taiwan to concentrate 60% of advanced manufacturing processes by 2027, retaining a hold on key technologies In advanced manufacturing processes (including 16/14nm and more advanced technologies), Taiwan leads with a 68% global capacity share in 2023, followed by the US (12%), South Korea (11%), and China (8%). Meanwhile, Taiwan holds nearly 80% when it comes to EUV generation processes (such as 7nm and beyond).

In response to the concentration of semiconductor manufacturing capacity in Taiwan, the US, which has a high demand for advanced processes, is actively encouraging and supporting major companies such as TSMC, Samsung, and Intel. 2027, the US's share of advanced process capacity is expected to increase to 17%, although TSMC and Samsung will still account for over half of this capacity.

Japan is also planning a return to semiconductor manufacturing, actively supporting local company Rapidus with a goal of reaching the most advanced 2 nm process. They aim to create a semiconductor cluster in Hokkaido and are offering subsidies to foreign companies, including Japan Advanced Semiconductor Manufacturing (JASM) and PSMC's Sendai plant (JSMC). China's mature process capacity is set to grow to 39% as Taiwanese firms actively cultivate unique technological advantages

China is focusing aggressively on mature process technologies (28nm and older), particularly in response to export controls on advanced equipment by the US, Japan, and the Netherlands.



By 2027, China's share in mature process capacity is expected to reach 39%, with room for further growth if equipment procurement proceeds smoothly.

However, as Chinese manufacturers rapidly expand their mature process capacities—backed by government subsidies—this could lead to intense price competition in products like CIS, DDI, PMIC, and power discrete, impacting Taiwan-based foundries like UMC, PSMC, and Vanguard. Vanguard is expected to be most affected due to its product line including LDDI, SDDI, PMIC, and power discrete. Other companies like UMC and PSMC will maintain their advantages in the 28/22nm OLED DDI and memory sectors.

In response to chip shortages and geopolitical influences, fabless customers are diversifying risk by working with multiple foundries, potentially leading to increased IC costs and concerns over duplicate orders. Customers are also requiring global validation of production lines, even with long-term foundry partners, to enable flexible capacity adjustments. Consequently, foundries must navigate larger scale capacity and price competition while needing to maintain profitability, flexibility in capacity adjustments, new capacity depreciation pressures, and technological leadership.

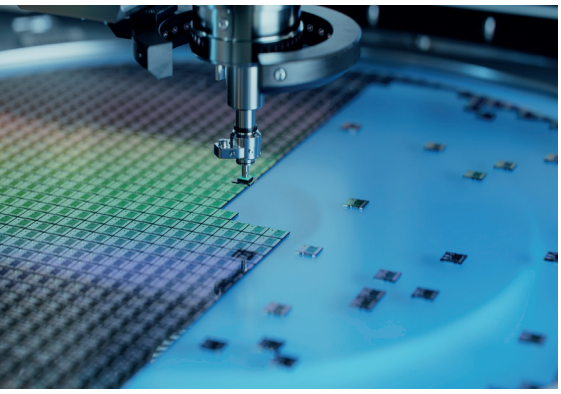
China is focusing aggressively on mature process technologies (28nm and older), particularly in response to export controls on advanced equipment by the US, Japan, and the Netherlands. By 2027, China's share in mature process capacity is expected to reach 39%, with room for further growth

Breakthroughs and opportunities in 3D packaging

Increase in trend of miniaturization in portable electronic industry and rise in dependency on these devices worldwide is shifting device manufacturers toward finding new methods of size reduction.

GLOBAL 3D semiconductor packaging market size is estimated to reach $8.9 billion by 2022, growing at a CAGR of 15.7 % from 2016 to 2022 according to the Allied Market Research Titled “3D Semiconductor Packaging Market”.

The report includes a detailed analysis of the dynamic factors such as drivers, restraints, challenges, and opportunities. The drivers and opportunities help to comprehend the rapidly changing industry trends and how they can impact the growth of the market. Moreover, the challenges and restraints analyzed in the report help recognize profitable market investments. The global 3D Semiconductor Packaging report provides quantitative and qualitative analysis of the market from 2021 to 2030.



The qualitative study focuses on the value chain analysis, key regulations, and pain point analysis. The global 3D Semiconductor Packaging market report includes an overview of the market and highlights market definition and scope along with major factors that shape the 3D Semiconductor Packaging market. The study outlines the major market trends and driving factors that boost the growth of the 3D Semiconductor Packaging market. The report includes an in-depth study of sales, market size, sales analysis, and prime drivers, challenges, and opportunities.

Some of the prime drivers of the 3D Semiconductor Packaging industry are surge in penetration of the aging infrastructure is further anticipated to drive the 3D Semiconductor Packaging market growth. The market for 3D Semiconductor Packaging would be driven by investing in new technology aimed at increasing system life. Another key factor driving the growth of the 3D Semiconductor Packaging market is the increased focus on infrastructure throughout the world.

3D Semiconductor Packaging provides monitoring technology to alert maintenance workers when outdated and overused equipment is about to fail, allowing them to make better decisions by providing real-time data on problems and possibilities for improvement.

Aside from the limits listed above, there are others, such as environmental factors such as temperature and humidity, as well as groundwater seepage, which can have an influence on the operation of switchgear electrical networks, particularly those situated outside. The changing times necessitate changes in the fundamentals as well. In this situation, even small and medium-sized organizations (SMEs) are taking advantage of collocation data hubs’ immense potential and the internet’s enormous capacity.

The market study further promotes a sustainable market scenario on the basis of key product offerings. On the other hand, Porter’s five forces analysis highlights the potency of buyers and suppliers to enable stakeholders make profit-oriented business decisions and strengthen their supplier-buyer network. The report provides an explicit global market breakdown and exemplifies how the opposition will take shape in the new few years to come. Rendering the top ten industry players functional in the market, the study emphasizes on the policies & approaches integrated by them to retain their foothold in the industry.

The analysis highlights the highest revenue generating and fastest growing segments. These insights are helpful in devising strategies and achieving a sustainable growth. The 3D Semiconductor Packaging market is studied on the basis of different segments including type, applications, and region. This makes the study well organized and resourceful along with promoting easy understanding. The report a comprehensive data based on each segment of the 3D Semiconductor Packaging market.

The 3D Semiconductor Packaging market is analyzed on the basis of geographical penetration along with a study of market influence in the various regions such as North America (United States, Canada, and Mexico), Europe (Germany, France, UK, Russia, and Italy), Asia-Pacific (China, Japan, Korea, India, and Southeast Asia), South America (Brazil, Argentina, Colombia), Middle East and Africa (Saudi Arabia, UAE, Egypt, Nigeria, and South Africa).

The global 3D Semiconductor Packaging market offers a detailed overview of the industry based on the main parameters including market extent, probable deals, sales analysis, and essential drivers. The market report is summarized enfolding the operations of an array of different organizations in the sector from different regions. The study is a perfect consolidation of quantitative and qualitative information accentuating on the key industry developments and challenges that the market is facing along with the lucrative opportunities available in the sector.

The 3D Semiconductor Packaging market report also showcases the factual data throughout the forecast period and brings about an estimate till 2031.

Bridging the talent gap in Flanders' Semiconductor Industry

Sofics supports initiative and donates time and resources for the course on analog layout.

IN RESPONSE to the critical shortage of skilled engineers in semiconductor Integrated Circuit (IC) design and layout, a new initiative has been launched by Professor Patrick Reynaert of KU Leuven in collaboration with several leading semiconductor companies in Flanders, Belgium, including Sofics. Flanders, renowned for the prestigious research center imec, boasts a myriad of commercial enterprises at the forefront of cutting-edge technology. These companies play pivotal roles in various electronic applications, spanning automotive, medical, industrial, consumer, data center, optical, and space industries. Despite their ambitious growth plans, the semiconductor industry, also in Flanders, faces a significant hurdle – a shortage of qualified engineers in IC design and layout.

To address this challenge head-on, Professor Reynaert's recent initiative introduces three comprehensive courses designed to coach individuals into proficient 'chip designers'. This program provides students with specialized skills essential for semiconductor industry roles.

Crucially, the courses are not just theoretically driven; they are developed and taught by seasoned experts from Flanders' leading semiconductor companies. Participating firms include BelGan, Caeleste, Cyient, easics, iCana, ICsense, imec, Melexis, Omnivision, and Sofics. This collaboration ensures that students gain practical insights, real-world perspectives, and industry-specific knowledge directly from those shaping the landscape.

"We are excited to be part of this collaborative effort to cultivate the next generation of chip designers," said Koen Verhaege, CEO of Sofics. "It can help to bridge the skills gap, enabling semiconductor companies in Flanders to thrive and achieve their growth aspirations."

Prospective students are encouraged to enroll and embark on a journey towards becoming the future leaders of semiconductor design. For more information about the courses and enrollment details, please visit https://www.esat.kuleuven.be/mc_chips.

Worldwide semiconductor revenue to grow 17% in 2024

Global semiconductor revenue is projected to grow 16.8% in 2024 to total $624 billion, according to the latest forecast from Gartner, Inc. In 2023, the market is forecast to decline 10.9% and reach $534 billion.

"We are at the end of 2023 and strong demand for chips to support artificial intelligence (AI) workloads, such as graphics processing units (GPUs), is not going to be enough to save the semiconductor industry from double-digit decline in 2023," said Alan Priestley, VP Analyst at Gartner. "Reduced demand from smartphones and PC customers coupled with weakness in data centre/hyperscaler spending are influencing the decline in revenue this year."

However, 2024 is forecast to be a bounce-back year where revenue for all chip types will grow (see Figure 1), driven by double-digit growth in the memory market. Memory Revenue to Rebound in 2024 After Double-Digit Decline

The worldwide memory market is forecast to record a 38.8% decline in 2023 and will rebound in 2024 by growing 66.3%.

Anaemic demand and declining pricing due to massive oversupply will lead NAND flash revenue to decline 38.8% and fall to $35.4 billion in revenue in 2023. Over the next 3-6 months, NAND industry pricing will hit bottom, and conditions will improve for vendors. Gartner analysts forecast a robust recovery in 2024, with revenue growing to $53 billion, up 49.6% year-over-year. Due to high oversupply level and lack of demand, DRAM vendors are chasing the market price down to reduce inventory. Through the fourth quarter of 2023, DRAM market's oversupply will continue which will trigger a pricing rebound. However, the full effect of pricing increases will only be seen in 2024, when DRAM revenue is expected to increase 88% to total $87.4 billion. Developments in generative AI (GenAI) and large language models are driving demand for deployment of high-performance GPU-based servers and accelerator cards in data centres. This is creating a need for workload accelerators to be deployed in data centre servers to support both training and inference of AI workloads.

Commission launches chips joint undertaking under the European Chips Act

The European Commission has officially inaugurated the Chips Joint Undertaking (Chips JU), designed to reinforce the European semiconductor ecosystem and Europe's technological leadership.

IT WILL BRIDGE the gap between research, innovation and production thereby facilitating the commercialisation of innovative ideas. The Chips JU will, among others, deploy pilot lines for which the Commission announced today the first call with €1.67 billion of EU funding. This is expected to be matched by funds from Member States to reach €3.3 billion, plus additional private funds.

In addition, the European Semiconductor Board held its first meeting recently. The Board brings together Member States to provide advice to the Commission on the consistent implementation of the European Chips Act and on international collaboration in semiconductors. It will be the key platform for coordination between the Commission, Member States, and stakeholders to address issues relating to the resilience of the supply chain and possible crisis responses.

The Chips Joint Undertaking

The Chips JU is the main implementer of the Chips for Europe Initiative (expected total budget €15.8 billion until 2030). The Chips JU aims at strengthening Europe's semiconductor ecosystem and economic security by managing an expected budget of nearly €11 billion by 2030, provided by the EU and participating states.

The Chips JU will:

- Set up pre-commercial, innovative pilot lines, providing industry state-of-the-art facilities to test, experiment and validate semiconductor technologies and system design concepts;
- Deploy a cloud-based Design Platform for design companies across the EU:
- Support the development of advanced technology and engineering capacities for quantum chips;
- Establish a network of competence centres and promote skills development.

The work of the Chips JU reinforces Europe's technological leadership by facilitating the transfer of knowledge from the lab to the fab, bridging the gap between research, innovation and industrial activities, and by promoting the commercialisation of innovative technologies by European industry including start-ups and SMEs.

First calls for funding Chips pilot lines
To launch its first calls for innovative pilot lines, the Chips JU will make €1.67 billion in EU funding available. The calls are open to organisations that wish to establish pilot lines in Member States, typically research and technology organisations, calling for proposals on:

- **Fully Depleted Silicon on Insulator, towards 7 nm:** This transistor architecture is a European innovation and has distinct advantages for high-speed and energy-efficient applications. A roadmap towards 7 nm will provide a path towards the next generation of high-performance, low-power semiconductor devices.
- **Leading-edge nodes below 2 nm:** This pilot line will focus on developing cutting-edge technology for advanced semiconductors at the size of 2 nanometres and below, which will play essential roles in a variety of applications, from computing to communication devices, transport systems and critical infrastructure.
- **Heterogeneous system integration and assembly:** Heterogeneous integration is an increasingly attractive technology for innovation and increased performance. It refers to the use of advanced packaging technologies and novel techniques to combine semiconductor materials, circuits or components into one compact system.
- **Wide Bandgap semiconductors:** The focus will be on materials that allow electronic devices to operate at much higher voltage, frequency and temperature than standard silicon-based devices. Wide bandgap and ultra-wide bandgap semiconductors are necessary to develop highly efficient power, lighter weight, lower costs and radio-frequency electronics.

The deadline for the calls for these pilot lines is in early March 2024.

AI demand leads semiconductor revenue growth

Led by continued demand for AI, the total semiconductor industry grew 8.4% in 3Q23 from 2Q23 hitting $139bn according to Omdia's latest Competitive Landscape Tracker.

THE INDUSTRY has now grown for a second straight quarter, after previously declining for five consecutive quarters. "The increase in the semiconductor industry wasn't solely due to AI demand, as growth spread throughout other semiconductor segments," said Omdia Principal Analyst Cliff Leimbach. "Fourteen of the top fifteen companies experienced a quarterly increase in semiconductor revenue in 3Q23, and 80 of the 126 companies (63%) tracked grew revenue in the third quarter."
AI continues to drive the market, but improvements observed throughout the industry

Demand for AI continued to be a major theme for the semiconductor industry in 3Q23. Two firms that have benefited from AI, NVIDIA, and SK Hynix, continued to record large semiconductor revenue increases. NVIDIA, which makes GPUs that are used for data-intensive AI, increased semiconductor revenue 18% to $12bn. NVIDIA's semiconductor revenues are up $7.3bn from a year ago when 3Q22 revenues were $4.6bn. SK Hynix, which is dominant in the high bandwidth memory (HBM) used in AI applications, increased semiconductor revenue 26% to $6.7bn.

Since the memory market was at the lowest in 1Q23, SK Hynix's revenue has increased by $2.9bn. The rise of AI has significantly improved revenues for firms with a large presence in this area. However, other areas of the semiconductor market also increased in 3Q23. The wireless segment also progressed along with the release of new smartphone models and better inventory dynamics than previous quarters. Growth was lower in the automotive sector, up 4.3% in 3Q23, but this segment has been reliably steady, having last declined in 3Q20 and currently representing 13.5% of all semiconductor revenue. The consumer segment also improved, up 7.9% from 2Q23, showing the breadth of the increased revenue in the semiconductor market.

Memory market continues improvement
The memory segment was the hardest hit segment during the semiconductor market decline which started in 2022, declining from $43.6bn in 1Q22 to $19.3bn in 1Q23. From the low point in 1Q23, the market bounced back going from $19.3bn to $24.5bn in 3Q23. Four of the top five memory makers increased memory IC revenue by double-digit percentage points in 3Q23 from 2Q23. AI demand is a large part, but other applications are increasing in demand as well. NVIDIA has risen the ranks and has become the second largest semiconductor firm by revenue. This has split up the long-time pair at the top of the semiconductor market shares of Intel and Samsung Electronics. These two firms have traded the position of number one and number two over the years.

A plan to 'revitalise' the semiconductor industry

SEMICONDUCTOR RESEARCH CORPORATION (SRC) has unveiled the Microelectronics and Advanced Packaging (MAPT) Roadmap, crafted through the collective effort of approximately 300 individuals representing 112 organizations from industry, academia, and government.

The MAPT Roadmap defines critical research priorities and technology challenges that must be addressed to support the seismic shifts outlined in the Decadal Plan for Semiconductors released in January of 2021. The MAPT Roadmap, available at https://srcmapt.org/, is the first industry-wide 3D semiconductor roadmap to guide the forthcoming microelectronics revolution.

Funded by the U.S. Department of Commerce's National Institute of Standards and Technology (NIST) in April 2022, SRC was selected to head this effort based on a strong history of thought leadership and forging innovation. SRC Chief Scientist and Director of the MAPT Roadmap, Dr. Victor Zhirnov, commented, "The commitment demonstrated by such a wide array of scientists, engineers, and researchers to the development and production of the MAPT Roadmap indicates the importance of such an effort."

Every day, people across the planet interact with dozens of semiconductor devices without a second thought. There is a crisis at hand, however. The ever-shrinking components are facing fundamental physical limits, and next-gen breakthroughs are unachievable without major advancements. The 2030 Decadal Plan for Semiconductors, released by SRC and Semiconductor Industry Association (SIA) identified five seismic shifts in the industry related to smart sensing, memory and storage, communication, security, and energy efficient computing. The Decadal Plan is, by design, agnostic on specific solutions; it identifies what is needed, rather than how it will be accomplished.

The MAPT Roadmap continues the spirit of the Decadal Plan and discusses how to achieve its system-level goals, outlining the implementation plan for semiconductor industry. The fundamental research that will transform these obstacles is focused on advanced packaging, 3D integration, electronic design automation, nanoscale manufacturing, new materials, and energy-efficient computing. The MAPT Roadmap is framed around fundamental and practical limits of information and communications technology sustainability: energy sustainability, environmental sustainability, and workforce sustainability.

WINNER OF THE PRODUCTRONICA INNOVATION AWARD

AP&S

NEXASTEP

OUR NEW BATCH TOOL FOR HIGH THROUGHPUT WET PROCESSING

- **Compact design**
  for optimal use of the clean room
- **Fast processes**
  through high process temperature of up to 170°C
- **Uninterrupted production flow**
  via warehouse with up to 75 storage locations and automated loading connected to the transport system
- **Reduction of particle generation & contamination**
  by improved access and optimized extraction of the process chamber

WE ARE YOUR PARTNER FOR WET PROCESS EQUIPMENT

AP&S International GmbH . Obere Wiesen 9 . 78166 Donaueschingen . Germany . 0771 8983 0 . sales@ap-s.de . www.ap-s.de

Global capacity to reach record 30 million wafers per month

Global semiconductor capacity is expected to increase 6.4% in 2024 to top the 30 million wafers per month (wpm) mark for the first time after rising 5.5% to 29.6 wpm in 2023, SEMI announced recently in its latest quarterly World Fab Forecast report.

THE 2024 growth will be driven by capacity increases in leading-edge logic and foundry, applications including generative AI and high-performance computing (HPC), and the recovery in end-demand for chips. The capacity expansion slowed in 2023 due to softening semiconductor market demand and the resulting inventory correction.

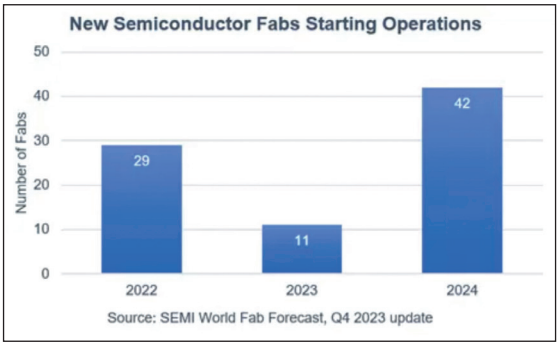
"Resurgent market demand and increased government incentives worldwide are powering an upsurge in fab investments in key chipmaking regions and the projected 6.4% rise in global capacity for 2024," said Ajit Manocha, SEMI President and CEO. "The heightened global attention on the strategic importance of semiconductor manufacturing to national and economic security is a key catalyst of these trends." Covering 2022 to 2024, the World Fab Forecast report shows that the global semiconductor industry plans to begin operation of 82 new volume fabs, including 11 projects in 2023 and 42 projects in 2024 spanning wafer sizes ranging from 300mm to 100mm.

China leads semiconductor industry expansion
Boosted by government funding and other incentives, China is expected to increase its share of global semiconductor production. Chinese chip manufacturers are forecast to start operations of 18 projects in 2024, with 12% YoY capacity growth to 7.6 million wpm in 2023 and 13% YoY capacity growth to 8.6 million wpm in 2024.

Taiwan is projected to remain the second-largest region in semiconductor capacity, increasing capacity 5.6% to 5.4 million wpm in 2023 and posting 4.2% growth to 5.7 million wpm in 2024. The region is poised to begin operations of five fabs in 2024.

Korea ranks third in chip capacity at 4.9 million wpm in 2023 and 5.1 million wpm in 2024, a 5.4% increase as one fab comes online. Japan is expected to place fourth at 4.6 million wpm in 2023 and 4.7 million wpm in 2024, a capacity increase of 2% as it starts operations of four fabs in 2024.

The World Fab Forecast shows the Americas increasing chip capacity by 6% YoY to 3.1 million wpm with six new fabs in 2024. Europe & Mideast is projected to up capacity 3.6% to 2.7 million wpm in 2024 as it launches operations of four new fabs.



Southeast Asia is poised to increase capacity 4% to 1.7 million wpm in 2024 with the start of four new fab projects.

Foundry segment continues strong capacity growth
Foundry suppliers are forecast to rank as the top semiconductor equipment buyers, increasing capacity to 9.3 million wpm in 2023 and a record 10.2 million wpm in 2024.

The memory segment slowed expansion of capacity in 2023 due to weak demand in consumer electronics including PCs and smartphones. The DRAM segment is expected to increase capacity 2% to 3.8 million wpm in 2023 and 5% to 4 million wpm in 2024. Installed capacity for 3D NAND is projected to remain flat at 3.6 million in 2023 and rise 2% to 3.7 million wpm next year.

In the discrete and analog segments, vehicle electrification remains the key driver of capacity expansion. Discrete capacity is forecast to grow 10% to 4.1 million wpm in 2023 and 7% to 4.4 million wpm in 2024, while Analog capacity is projected to grow 11% to 2.1 million wpm in 2023 and 10% to 2.4 million wpm in 2024.

The latest update of the SEMI World Fab Forecast report, published in December, lists 1,500 facilities and lines globally, including 177 volume facilities and lines with various probabilities expected to start operation in 2023 or later.

Global equipment sales to reach $124 billion in 2025

Global sales of total semiconductor manufacturing equipment by original equipment manufacturers are forecast to reach $100 billion in 2023, a contraction of 6.1% from the industry record of $107.4 billion

posted in 2022, SEMI announced in its Year-End Total Semiconductor Equipment Forecast – OEM.

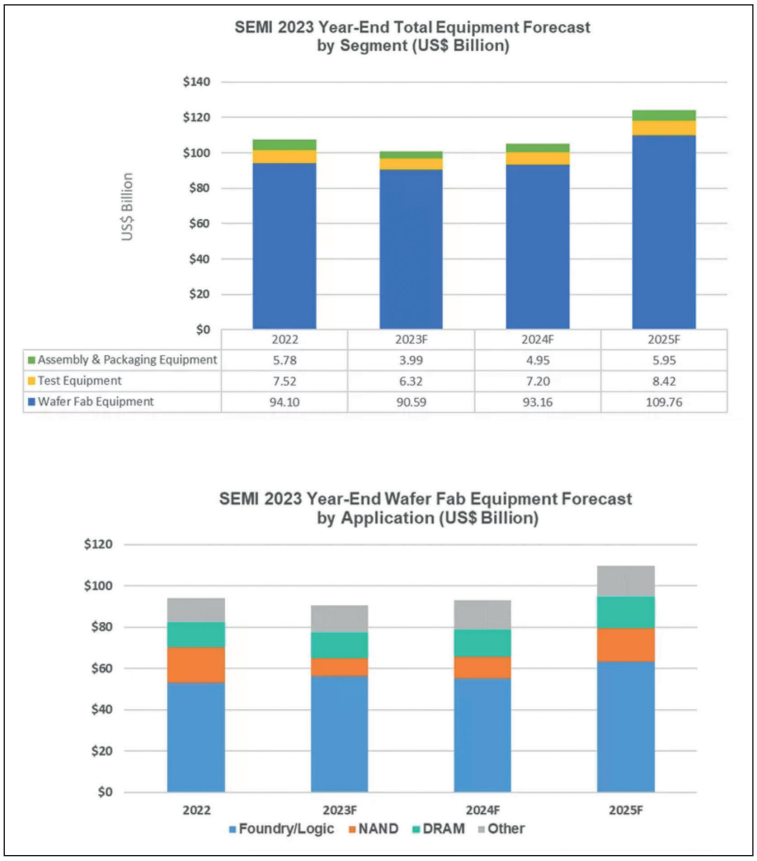
Semiconductor manufacturing equipment growth is expected to resume in 2024, with sales forecast to reach a new high of $124 billion in 2025, supported by both the front-end and back-end segments. "We anticipate a temporary contraction in 2023 due to the cyclical nature of the semiconductor market," said Ajit Manocha, SEMI President and CEO. "2024 will be a transition year. We then expect a strong rebound in 2025, driven by capacity expansion, new fab projects, and high demand for advanced technologies and solutions across the front-end and back-end segments."

Semiconductor equipment sales by segment
After registering a record $94 billion in sales last year, the wafer fab equipment segment, which includes wafer processing, fab facilities and mask/reticle equipment, is projected to slip 3.7% to $90.6 billion in 2023. This contraction marks a significant improvement from the 18.8% decline forecast by SEMI in its Mid-Year Total Semiconductor Equipment Forecast – OEM Perspective. The upward revision is primarily due to strong equipment spending by China. Wafer fab equipment segment sales are projected to grow at a modest 3% in 2024 from the revised 2023 base due to limited memory capacity addition and the pause of mature capacity expansion. A growth path with a further 18% expansion in 2025 is expected as new fab projects, capacity expansion and technology migrations drive investments to nearly $110 billion.

The decline in back-end equipment segment sales started in 2022 and continued in 2023 due to challenging macroeconomic conditions and softening semiconductor demand. Semiconductor test equipment market sales are projected to contract by 15.9% to $6.3 billion in 2023, while assembly and packaging equipment sales are expected to drop by 31% to $4.0 billion in the same year. The test equipment and assembly and packaging equipment segments are forecast to expand by 13.9% and 24.3%, respectively, next year. Back-end segment growth is expected to continue in 2025, with test equipment sales rising 17% and assembly and packaging sales increasing 20%.

Semiconductor equipment sales by application
Equipment sales for foundry and logic applications, accounting for more than half of total wafer fab equipment receipts, are expected to log a 6% increase year-over-year to $56.3 billion in 2023 despite softer end-market conditions. The application segment is forecast to contract 2% in 2024 as mature technology expansion slows and spending on leading-edge technology improves.

Foundry and logic equipment investments are projected to increase 15% in 2025 to $63.3 billion, driven by increased capacity expansion purchases and the introduction of new device architectures.



As anticipated, memory-related capital expenditures will see the sharpest decline in 2023. NAND equipment sales are predicted to drop by 49% to $8.8 billion in 2023 but will surge 21% to $10.7 billion in 2024 and rise another 51% to $16.2 billion in 2025. DRAM equipment sales are expected to remain stable, growing by 1% and 3% in 2023 and 2024, respectively. Supported by continuous technology migration and expanding demand for high-bandwidth memory (HBM), DRAM equipment segment sales are expected to increase an additional 20% to $15.5 billion in 2025.

➤ Source: SEMI December 2023, Equipment Market Data Subscription

Semiconductor equipment sales by region
China, Taiwan and Korea are expected to remain the top three destinations for equipment spending through 2025. China is projected to maintain the top position over the forecast period as the region's equipment billings continue to soar. Equipment shipments to China are projected to surpass a record $30 billion in 2023, widening its lead with other regions. While equipment spending for most tracked regions is expected to fall in 2023 before resuming growth in 2024, China is expected to see a mild contraction in 2024 after heavy investments in 2023.

SEMI recognises Edwards

SEMI Europe and the SEMI European Advisory Council for Diversity and Inclusion have announced Edwards as the recipient of the 2023 SEMI Industry Leader in Diversity and Inclusion Award. Christine Pelissier, General Manager, Customer Center EMEA of Semiconductor at Edwards accepted the award

at SEMICON Europa 2023 during the Future of Work session.

"I'm proud of the work that we've done at Edwards to promote diversity and embed a positive culture of inclusion," Pelissier said. "We started our journey in 2019 and we still have work to do. This award is a recognition of the organization's commitment to not just set targets but to do the work required to deliver on them."

"The semiconductor industry's capacity to innovate and sustain competitiveness in Europe relies significantly on a diverse and skilled workforce," said Laith Altimime, President of SEMI Europe. "Over the past five years, Edwards has continuously focused on improving the diversity and inclusion of its workforce through innovative and progressive programs."

Established in 2019, the SEMI Industry Leader in Diversity and Inclusion Award recognizes European semiconductor companies that exhibit outstanding leadership and make strategic contributions to fostering diversity, equity, and inclusion. The award celebrates their practices that advance both their workforces and the chip industry at large.

Edwards promotes culture-first mindset

Edwards' leaders have placed a strong emphasis on diversity and inclusion in the European semiconductor industry, establishing the company as a pioneer with recent measures to promote worker wellbeing, mental health and inclusivity. Edwards takes a holistic approach to cultivating an inclusive workplace with corporate policies that support inclusive language, trans colleagues, and Fusion, a network for the Edwards LGBTQ+ community. Edwards also offers employee education programs to raise awareness of issues such as unconscious bias and menopause.

SEMI Europe welcomes nominations for the 2024 award for companies that have inspired the industry through their exemplary leadership and execution, volunteerism, sponsorship, and innovation in diversity, equity, and inclusion. Prior award recipients hailed from companies including Soitec, Melexis, and SPTS Technologies, a KLA company.

"The semiconductor industry's capacity to innovate and sustain competitiveness in Europe relies significantly on a diverse and skilled workforce," said Laith Altimime, President of SEMI Europe. "Over the past five years, Edwards has continuously focused on improving the diversity and inclusion of its workforce through innovative and progressive programs."

SEMI honours sustainability leaders

SEMI has announced recipients of the 2023 SEMI Sustainability Excellence Award. SEMI President and CEO Ajit Manocha selected the award recipients based on their inspirational leadership and innovative approaches to expanding sustainability practices at their companies and in the semiconductor industry in order to meet aggressive net zero emission goals.

The leaders were also recognized for their contributions to the Path to Net Zero session during the SEMICON West 2023 CEO Summit keynote program. The honorees will be commemorated on the SEMI Wall of Fame at the association's headquarters in Milpitas.

2023 SEMI Sustainability Excellence Award honorees:

- Gary Dickerson, President and CEO, Applied Materials
- Keyvan Esfarjani, Executive VP, Chief Global Operations Officer, GM of Manufacturing, Supply Chain and Operations, Intel Corporation
- Mukesh Khare, General Manager of IBM Semiconductors and VP of Hybrid Cloud Research, IBM
- Aamir Paul, President, North America, Schneider Electric

"The drive toward net zero is a challenge that requires intense collaboration across international borders and supply chains," said Manocha. "SEMI is pleased to honor Gary, Keyvan, Mukesh and Aamir for their tremendous inspiration as the semiconductor industry works to achieve its emissions goals. Their leadership and passion are strong drivers for their companies adopting leading-edge sustainability practices, and their actions are a powerful motivator for the entire industry supply chain to follow suit."

The recipients were selected based on their alignment with principles of the SEMI Sustainability Initiative and its focus on Environment, Social and Governance (ESG) issues.

Recipient accomplishments

Gary Dickerson is a longtime sustainability leader and a major contributor to the semiconductor industry's net zero pathway. His passion for making a positive impact is reflected in Applied Materials' vision to Make Possible a Better Future. Dickerson champions Applied's ESG leadership initiatives, including the company's "3x30" program, which is aimed at reducing the environmental impact of its products, and Applied's Net Zero 2040 Playbook, a bold framework for reducing the industry's carbon emissions through close collaboration with Applied's customers and suppliers.

Keyvan Esfarjani leads Intel's ambitious sustainability efforts, which include achieving 100% renewable energy, net positive water and zero waste to landfill by 2030, reaching net zero Scope 1 and 2 GHG emissions by 2040, and collaborating with suppliers to meaningfully reduce Scope 3 emissions and achieve net-zero upstream GHG emissions by 2050. He is a consistent force in reducing Intel's environmental footprint while driving global capacity expansion plans and calling for the semiconductor industry to accelerate its timeline in meeting climate goals.

Dr. Mukesh Khare leads a global IBM team that is redefining the future of computing for products such as generative artificial intelligence (AI) and high-performance computing, two energy-intensive technologies. He leads teams of researchers devising new methods to process data more efficiently and accelerate the pace of progress towards sustainable quantum, AI and hybrid cloud infrastructure while advancing the industry's drive to net zero emissions.

Aamir Paul led Schneider Electric's partnership with Applied Materials and Intel to launch the pioneering Catalyze program. Unveiled at SEMICON West 2023, the Catalyze program promotes the adoption of a renewable and low-carbon energy roadmap for the chip sector and provides global partners in the semiconductor value chain with better access to renewable energy solutions and comprehensive decarbonization assistance. By collaboratively digitalizing energy for efficiency and sustainability, the Catalyze program's partners allow energy transition to be advantageous for businesses and environmentally beneficial for all.

Leadership accelerator launch

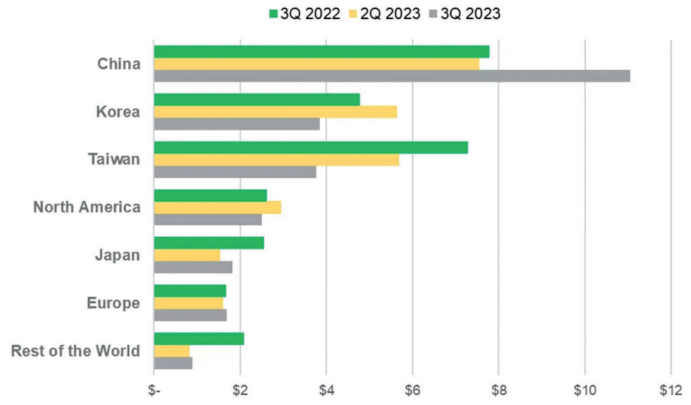
SEMI has introduced the SEMI Leadership Accelerator to foster industry growth by assisting organizations in cultivating the next generation of leaders. This initiative will support semiconductor supply chain organizations in evaluating high-potential employees, offering personalized leadership coaching, and delivering tailored training for individual and team development. Developed in a collaboration between SEMI Europe and Mercuri Urval, the SEMI Leadership Accelerator provides science-based, ISO-certified programs, empowering senior leaders, executives, and board members to navigate global challenges, drive business transformation, and achieve sustained growth in the semiconductor sector.

With Europe projected to need 350,000 new workers in the semiconductor industry by 2030, Laith Altimime, President of SEMI Europe, emphasized the imperative for cultivating leaders capable of navigating the complexities of the semiconductor industry. "The semiconductor industry must further develop diverse leaders who can reshape businesses in a complex and fast-changing world of uncertainty. The growing number of job vacancies in Europe's semiconductor sector risks undercutting its growth potential as companies expand manufacturing capacities. It is essential for leaders in the industry to unlearn outdated approaches and prepare for unforeseen challenges. The SEMI Leadership Accelerator will facilitate this transformative process."

Mercuri Urval, is a global leadership advisory firm with over 50 years of experience, an extensive global presence, and specialized knowledge across industry sectors. Methodically crafted, semiconductor industry-specific programs tailored by Mercuri Urval and leveraging SEMI's extensive industry insights and network, will enable the SEMI Leadership Accelerator to address leadership development needs, empowering companies to identify and nurture leaders internally.

Equipment billings drop 11%

Global semiconductor equipment billings contracted 11% year-over-year to US$25.6 billion in the third quarter of 2023, while quarter-over-quarter billings slipped 1% during the same period, SEMI announced today in its Worldwide Semiconductor Equipment Market Statistics (WWSEMS) Report.
"The Q3 2023 dip in equipment billings was due to softening chip demand," said Ajit Manocha, SEMI President and CEO. "However, China has shown strong demand and spending power for mature-node technologies, a sign of the industry's resilience and growth potential in the long run."



Source: SEMI (www.semi.org) and SEAJ (www.seaj.or.jp), November 2023

Region	3Q 2023	2Q 2023	3Q 2022	3Q (QoQ)	3Q (YoY)
China	$11.06	$7.55	$7.78	46%	42%
Korea	$3.85	$5.65	$4.78	-32%	-19%
Taiwan	$3.77	$5.69	$7.28	-34%	-48%
North America	$2.50	$2.95	$2.61	-15%	-5%
Japan	$1.82	$1.54	$2.55	19%	-29%
Europe	$1.70	$1.61	$1.67	5%	2%
Rest of the World	$0.89	$0.83	$2.08	8%	-57%
Total	**$25.59**	**$25.81**	**$28.75**	**-1%**	**-11%**

Sources: SEMI (www.semi.org) and SEAJ (www.seaj.or.jp), November 2023
Note: Summed subtotals may not equal the total due to rounding.

Compiled from data submitted by members of SEMI and the Semiconductor Equipment Association of Japan (SEAJ), the WWSEMS Report is a summary of the monthly billings figures for the global semiconductor equipment industry.
Following are quarterly billings data in billions of U.S. dollars with quarter-over-quarter and year-over-year changes by region:

New energy collaborative

Aiming to reduce global semiconductor ecosystem carbon emissions, SEMI and the Semiconductor Climate Consortium (SCC) have created the Energy Collaborative (EC) to understand and clear roadblocks to the installation of low-carbon energy sources in the Asia-Pacific region. The EC, a collective of industry leaders, will provide a consolidated view of priorities for low-carbon energy in the region.

"Sharing resources to start this foundational semiconductor industry sustainability work now is important to enable wider access to low-carbon energy in the next five to ten years," stated Young Bae, SCC Governing Council member and sponsor for the SCC Scope 2 Working Group, and Global Business Director, Advanced Cleans Technologies at DuPont. "One of the key action areas the SCC has identified is the lack of low-carbon energy plans and actions in the Asia-Pacific region. The EC will help the SCC by accelerating investments to broaden access."

The EC sponsoring companies will anchor the collective's work to engage in roundtables and fact-finding sessions.

McKinsey & Company is a Knowledge Partner to the initiative, providing fact-based analysis and support.

"The semiconductor value chain and its downstream partners and customers have a pivotal role to play in the acceleration of low-carbon energy installations, due to their scale in high-priority markets and the extent to which they will drive growth in future energy demands," stated Ajit Manocha, President and CEO of SEMI. "To reach the emissions reductions goals of the sector, a step function change in ambition and action is required. The EC is focused on that goal – increasing the pace and scaling of access to low-carbon energy."

Sharing resources to start this foundational semiconductor industry sustainability work now is important to enable wider access to low-carbon energy in the next five to ten years

A recent SCC report found that the semiconductor value chain is a significant consumer of energy in almost all key Asian markets. Additionally, a recent analysis by McKinsey & Company shows that even with major semiconductor companies' latest commitments, which are more stringent than past measures, the industry is not on track to limit emissions to the extent required under the 2015 Paris Agreement. The analysis finds that both individual and collective actions by semiconductor players can help the entire industry increase its sustainability effort and meet the 1.5°C challenge.

Two new international board members

SEMI has announced the election of two new members to the SEMI International Board: Kai Beckmann, Member of the Executive Board and CEO of the Electronics business sector of Merck KGaA, Darmstadt, Germany, and Benjamin Loh, Chair of the Management Board and President and CEO of ASM. The tenure of the new board members begins immediately.

"We congratulate the newly elected International Board members, Kai Beckmann and Benjamin Loh, and thank them for their commitment to guide SEMI in representing the global electronics manufacturing and design supply chain," said SEMI President and CEO Ajit Manocha. "Kai and Benjamin add to the breadth and depth of knowledge on our Board of Directors and will help us to ensure that SEMI programs continue to advance semiconductor industry growth, address top concerns, and deliver exceptional member value."

SEMI's 18 independent voting directors and 11 emeritus directors represent companies from Europe, China, Japan, Korea, North America, and Taiwan, reflecting the global scope of the association's activities. SEMI directors are elected by the general membership as voting members of the Board and can serve a total of five three-year terms. Kai Beckmann joined the Executive Board of Merck KGaA, Darmstadt, Germany in April 2011. He has served as CEO of the Electronics business sector since September 2017. Under his leadership the business successfully transformed into a leading player in the global electronics materials market. Beyond Merck, he has several external mandates such as President of the German Federation of Chemical Employers' Associations and Chairman of the Advisory Board of Fraunhofer Institute for Computer Graphics Research.

Benjamin Loh was appointed Chair of the Management Board and President and CEO of ASM in May 2020. He is a non-executive director of ASMPT and in the past held positions as non-executive director at Schneeberger, Schweiter Technologies AG, and Liteq BV. Loh was a SEMI China regional advisory board member while employed with Unaxis Corporation and living in China from 2002 to 2005.

siconnex

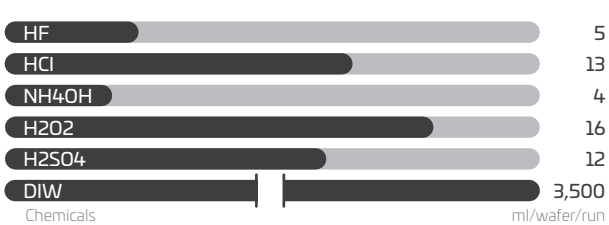
SOLUTION FOR CHEMICALS SHORTAGE

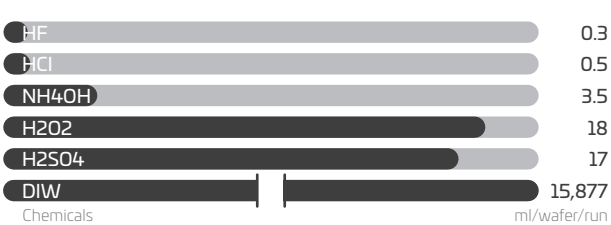
ELIMINATE H_2SO_4 AND H_2O_2 WITH OUR TECHNOLOGY

sicOzone by siconnex

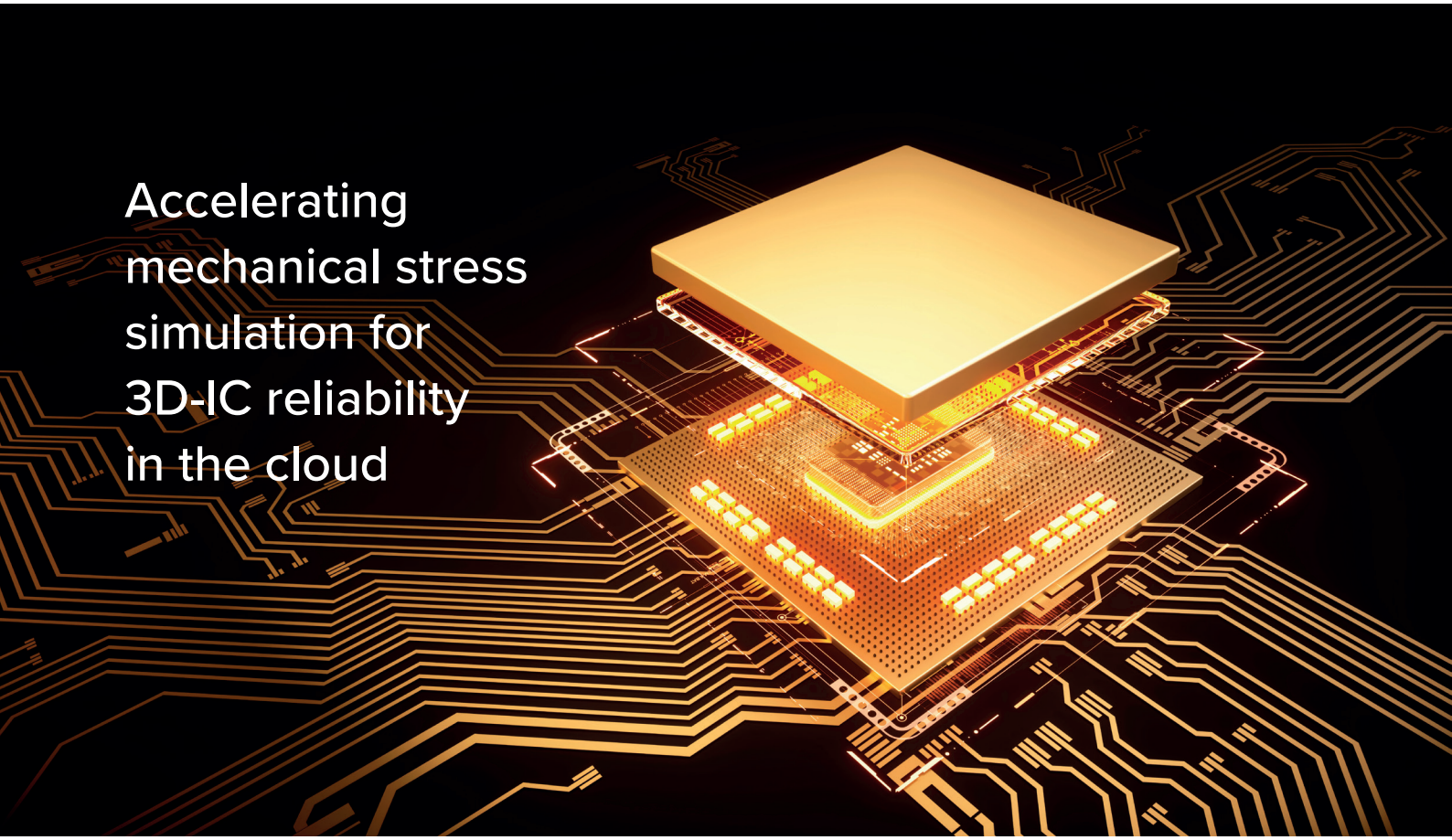
Ongoing supply chain disruptions impact the industry, leading to anticipated shortages and rising prices in cleaning chemicals, as forecasted by researchers like TECHCET. Smart utilization of ozone, along with the usage of minimal amounts of ammonia and acids, result in significant cost savings. This approach can replicate traditional cleaning steps and enhance organic cleaning efficiency in SC1, SC2, and DHF processes, with precise control by using inline spiking technology.

Chemical consumption: SicOzone vs. conventional cleaning applications





Accelerating mechanical stress simulation for 3D-IC reliability in the cloud



Marc Swinnen, Product Marketing Director at Ansys, discusses recent collaboration with TSMC and Microsoft to develop a joint solution which provides a high-capacity cloud solution for analysing the mechanical stresses in 2.5D/3D-IC multi-die systems, which lets joint customers avoid field failures and extends product lifetime and reliability.

PHILIP ALSOP, EDITOR asks the questions

PA: *Ansys has been collaborating with TSMC and Microsoft, focusing on analysing mechanical stresses in multi-die 3D integrated circuit systems. The obvious place to start would be to understand how did the collaboration come about?*

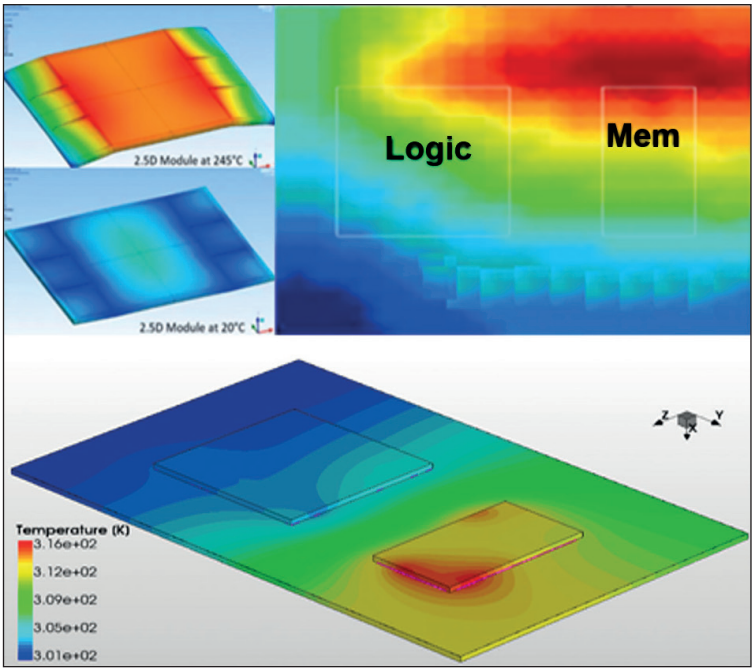
MS: Our collaboration with TSMC has been going on for decades. That's based largely on the fact that Ansys sells and produces the RedHawk-SC™ product, which is an electronic design automation (EDA) software tool used by chip designers to verify the power integrity of their chip. Basically, every chip has a power and ground network on it. Every single transistor has to be connected to power and has to be connected to ground, like any electronic device. If you have 50 billion transistors on your chip, that means you must design two electrical networks, each with 50 billion endpoints. So, these are incredibly large and complicated on-chip networks that are vital to the proper functioning of the chip. They need to be checked because there is always voltage drop on the power lines. And these days, to save power, the voltage is so low that you really can't afford to lose even 100 millivolts going from the package pin to the actual transistor. Hence everything has to be very carefully analysed to make sure your power integrity, or voltage drop, is properly accounted for and will meet your spec.

And that's a very big, tough problem. And that's what RedHawk-SC does. It does the final sign off for manufacturing, saying, yes, this will work. Of course, this all relies critically on the manufacturing rules. We work very closely with all of the major foundries, including TSMC. TSMC and Ansys have a longstanding collaborative association to get this golden sign off tool out to the industry - the large majority of all the chips in the world are signed off for their power integrity using Ansys Redhawk-SC. That is the foundation of our deep and ongoing collaboration with TSMC.

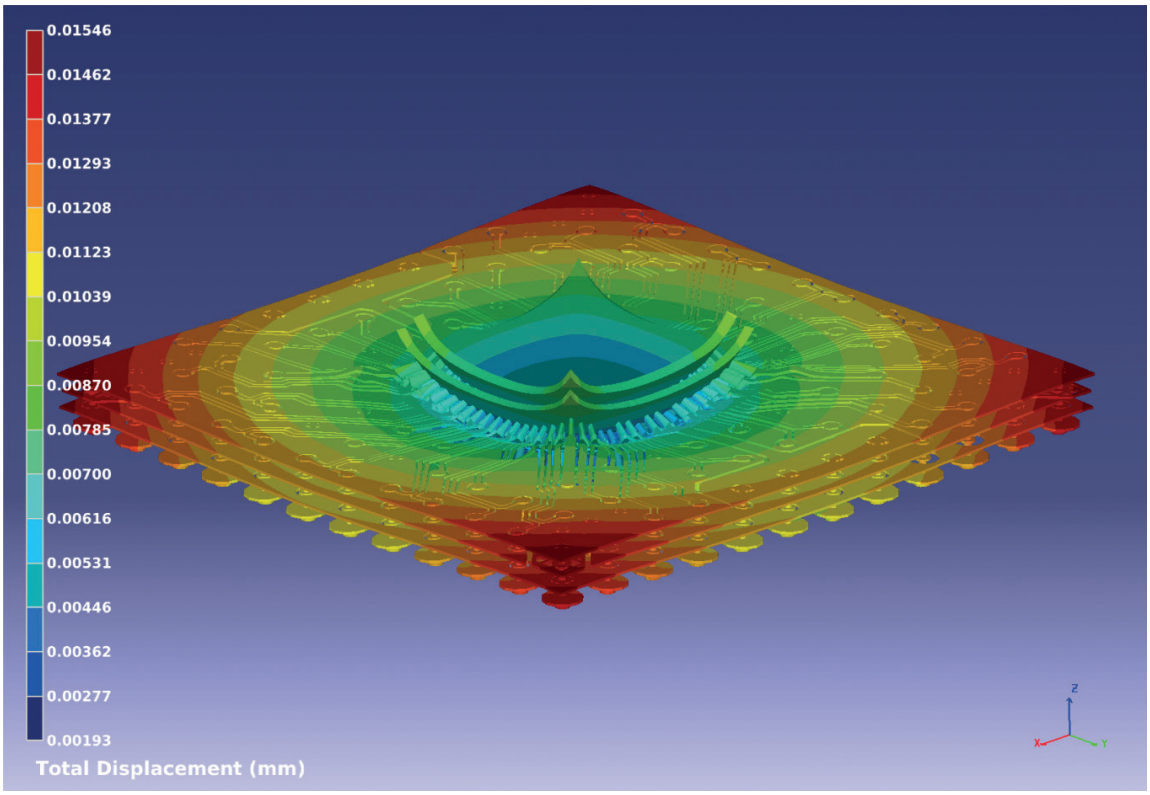
Which brings us to the topic under discussion. Traditionally, a chip, or integrated circuit (IC), is a monolithic piece of silicon - it's all one thing. You cut it out of the wafer and it's one little chip of silicon which gets embedded in a package. But now, for multiple reasons, it is no longer possible to build the big systems you want today on just a single

chip. So, they've started making multiple chips and putting them together into a system we call 3D IC - where you stack several chips on top of each other or, more commonly, you put them right next to each other, which we call two-and-a-half D. I'll just call all of these configurations 3DIC - all these different ways of stacking or putting them right next to each other. The idea of a 3DIC is that it contains multiple dice. So bare dice, not packaged dice, that are placed right next to each other. Usually they're placed on top of another chip, called an interposer, which connects them all together. All high-performance computing is going there today.

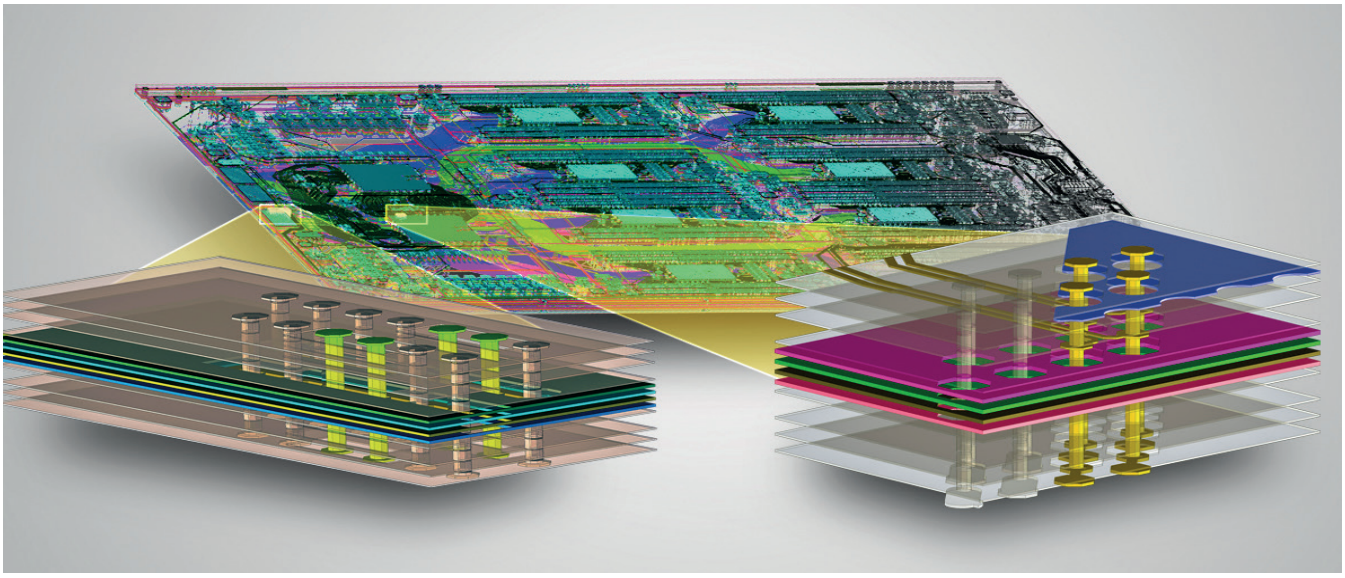
Now, some of these dice get hot and some of them get less hot and so you have differential thermal expansion. The dice are connected to each other with micro bumps. These are tiny, tiny bumps - up to a thousand per square millimetre - and they can't stand very much shear stress. If your assembly starts expanding and contracting differentially and cycling through these thermal cycles, you're going to get mechanical deformation, warping and stresses in this 3D assembly. And that is something radically new for chip designers. I mean, someone always had to worry about thermal expansion at some point. Usually a system or package designer way down the line, after the chip was assembled on a board and the board was put in the system and the system was in the heatsink - at that point, somebody did some mechanical analysis. But now it's come crashing down onto the chip, designers who now have to worry right out of the gate, how is this thing



➤ Display of thermal and warpage results from RedHawk-SC Electrothermal™ for a multi-die 2.5D assembly of a logic chip and memory chip on an interposer.



➤ 3DIC stack showing thermal induced warping and stresses from Ansys RedHawk-SC Electrothermal™. Colors indicate displacement (exaggerated in Z-axis for easy visibility).



going to deform and warp? And if I use the wrong materials or use the wrong floor plan, my design will have much lower reliability than if it's done properly. Hence, they need to do mechanical simulations early on, and predict thermo-mechanical behaviour.

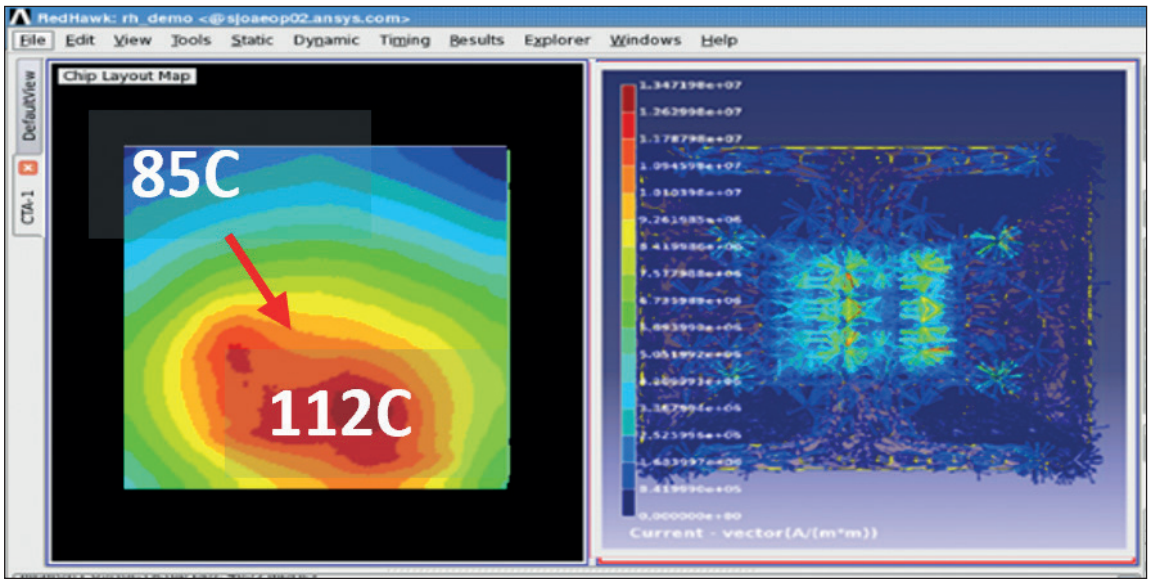
Ansys has a rich history in this area - beyond the semiconductor division we have many other simulation tools. We have computational fluid dynamics, we have mechanical, we have safety, we have optics, we have electromagnetics - many, many fields. But mechanical is one of our specialties where we are industry leading. It was natural for us to take those algorithms and apply them to our semiconductor problem.

TSMC worked with us to solve some of the issues they've seen in their own production and design side of things. They saw this as a problem which they needed solving. It's a tough computational problem so they pulled in Microsoft Azure to give the cloud computing capabilities required to really solve this in the required timeframe.

With cloud computing from Microsoft, the mechanical/thermal simulation from Ansys, and the manufacturing capability from TSMC, together we came up with a solution flow that worked and has been proven to work.

PA: *And the objective of the project, I believe it's to provide added confidence to address novel multi-physics requirements that improve the functional reliability and increase product lifetimes of advanced 3D fabric designs?*

MS: So, there's two points to that. One is novel, and the other one's reliability. So why is this novel? Mechanical simulation is not novel in itself, but for semiconductor designers, it is. As I already mentioned, this was something monolithic designers never had to worry about. But 3D assemblies - and I use 3D as a catch-all name for all those different architectures that the foundries supply with chip-on-chip and chip-next-to-chip - there's lots of ways of arranging these chips. I'll just call them all 3D-IC.



➤ Temperature gradient on a chip based on current distribution analysis from Ansys RedHawk-SC™.

Chips have very specific rules about tolerance to bending. For example, a chip, is allowed to bend to a certain amount in the concave way - only so much. And by the way, it's not only heating that causes it to bend, but also during assembly, when they're actually mounting these chips on top of each other, you press the chips down and that bends them a little bit

So that is novel for chip designers that they now have to think at the floor planning stage: Okay, which of these chips is going to get hot, which is going to stay cooler? If I put two hot chips right next to each other, is that going to be a thermal problem that I can't fix? Or especially if two chips get hot in the same activity mode, like in streaming mode while you're playing video, both these chips get really hot in this corner. That could doom my project right from the get go.

Thermal is the number-one limitation on achievable integration density today. You can very easily stack chips several layers deep. You can design that. You can manufacture that. All that is not a problem. The problem is you can't cool it! It'll get too hot, and it'll melt. So how close and how compact you can make a system is determined, number one, far and away, by power dissipation. How do I control and manage my heat dissipation? So, when you're assembling these large systems, multiple chips together - and we're talking up to a dozen chips - how do I manage my power?

With heat comes thermal expansion, temperature cycling, and differential expansion of multiple materials. Monolithic chips are made out of one thing, silicon. Now suddenly you have systems that may include silicon chips together with silicon carbide chips and organic substrates. The individual chips are usually placed on an interconnection substrate called an interposer. The interposer is often made of silicon, like a giant chip made with an old process, like 35 nanometre or something.

But interposers can be made of organic resin as well. It can also be glass. There are foundries working with glass substrates. So, there's multiple materials in this stack-up and they all expand differentially. This mechanical concern is novel to most chip designers today. They may have expertise in the company somewhere that deals with thermal issues in electronic systems, but it's not in the chip design group. As a result, they may have to reorganise how the company is set up so that this mechanical expert is involved right from the get go. So that's novel to them.

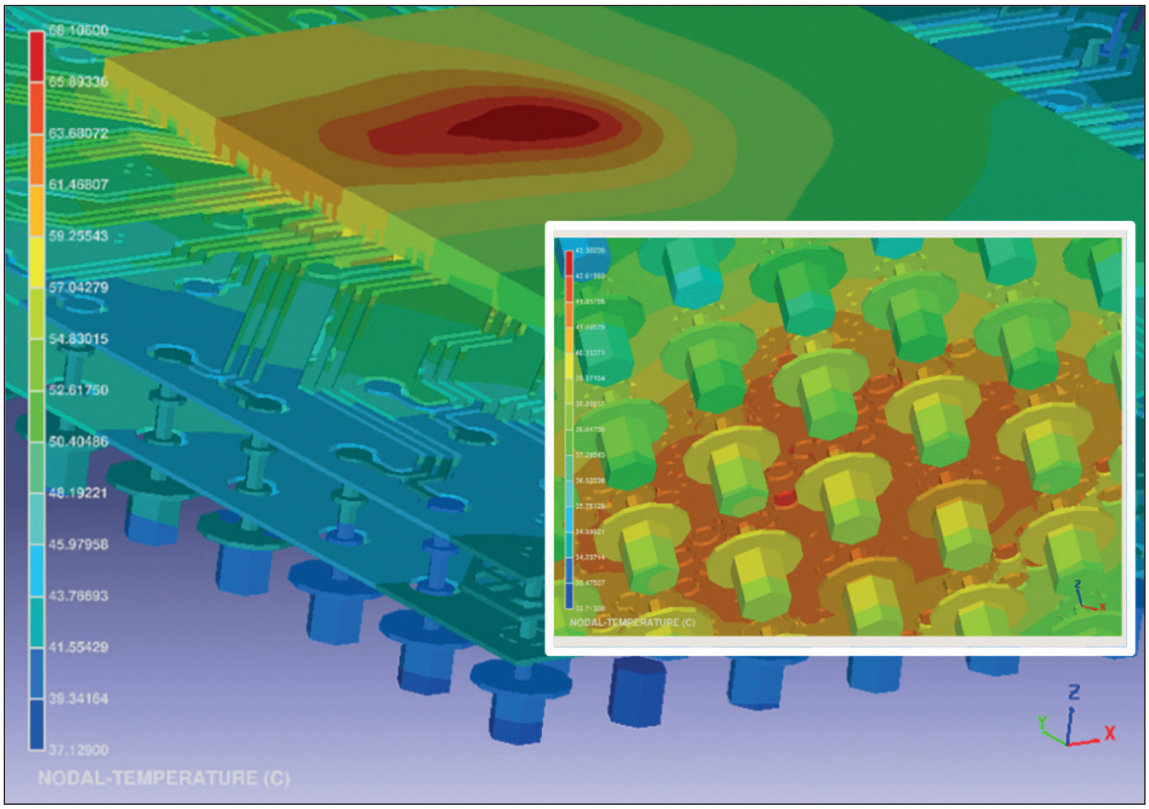
Ansys is taking our multiphysics simulation capabilities and making them available in a way that works with semiconductor data formats in a semiconductor flow. That is what we believe to be the answer. We have a tool called RedHawk-SC Electrothermal™, which takes care of integrating these. On the reliability side, the fundamental benefit is that thermal expansion is something that will, over time, degrade the chip. As I said, these chips are connected through little solder bumps that connect when you press the chips together.

They're about 10 micrometres apart, which gives you a very high density of interconnect. But if these chips start bending, warping or twisting and you get shear stresses between them, it's very easy for these micro bumps to shear, to break or to deform slightly so that their contact point becomes diminished. Which means that the power density or current density through that bump becomes higher and the chance of them melting becomes higher.

We did a 3D-IC assembly recently, with 400,000 micro bumps in the system, and you can have millions of micro bumps. If any one of them fails, you could have a system failure. It really behoves you to very carefully consider how your multi-die system warps as it heats up and cools down over time. That will give you a reliability problem if not properly managed.

Chips have very specific rules about tolerance to bending. For example, a chip, is allowed to bend to a certain amount in the concave way - only so much. And by the way, it's not only heating that causes it to bend, but also during assembly, when they're actually mounting these chips on top of each other, you press the chips down and that bends them a little bit. So that needs to be taken into account, too. Concave bending is allowed to a certain degree, but convex bending is absolutely not allowed. Apparently, there's very low tolerances on that. One way of bending is allowed, the other one isn't. These are some of the reliability issues that come up and there's trade-offs to be made there.

PA: *If I understand you correctly, the issue is, because this is a completely novel process, there needs to be a lot of learning to take place. Once they have understood how all you've described happens, they will then find out the optimum way to make these 3D systems. But until they do, they have to carry out multiple tests under all different loads and conditions. And if they do things in exactly the same way each time, they will get the same results? Or are there variations, what we might call random variations, even in identical conditions?*



➤ 3D thermal analysis results from Ansys RedHawk-SC Electrothermal™ for a chip on a package substrate, with pin-by-pin thermal resolution.

MS: Yes, so that is the beauty of simulation. People might say, well, I prefer working on a real system that doesn't predict what it's going to do, but actually you can measure on the test bench. Yes, there's a benefit to measuring reality, but the problem is you're measuring a specific instance of reality. You're measuring this particular device with this particular set of parameters. As an example, imagine measuring the parameters of something like steel bolts. How strong is a steel bolt? Well, some will be a little stronger than others. Batches of steel vary. There's a range there. You can test a particular bolt and that might be a strong one or it might be a weak one, but it doesn't tell you what the range of possibilities are when I'm going to build thousands of these things into a bridge. So too, when you take a chip under test or 3D assembly, you can put it on the test bench and test it, but you're testing a particular device. But there's so many different parameters in thickness variation, material property variation. Also temperature or thermal is due to activity, right? If the chip isn't doing anything, it doesn't get hot, it's only when it's actually active that it starts to get hot.

So what activity are you going to give it? And that's a problem because there's many, many thousands, millions of possible activity combinations. Simulation allows you to test across a variety of parameters, a variety of conditions, a variety of environmental, ambient conditions, too, and do all this with a variety of material combinations and verify that the system will work across this entire envelope of possibilities. So that's the power of simulation in general. That is what RedHawk-SC Electrothermal allows you to do as well. You can identify problems and then decide what you're going to do about fixing these or how to prevent them. For example, one of the trade-offs you have to consider is this silicon interposer. It's much bigger than a standard chip, we're talking three inches by three inches or so. This interposer, if you make it thick, then it will not warp as much. It'll be stiffer, it won't deform as much, which is good, but that means that you get much higher thermal stresses. The differential expansion wants to be there but it can't bend, so the stresses will be very high, but it won't deform as much. But if you make the interposer much thinner, it'll deform much more, but the stresses will be lower. Like everything in engineering, it's a trade-off.

Also, there are holes drilled through that punch through the silicon interposers to connect from one side to the other. These holes are called through-silicon vias (TSVs) and they're typically copper pillars. They're tiny, but for a chip size they're actually quite big. Even though they're electrical conductors meant to connect signal wires, they also serve a good purpose of dissipating the heat down through the silicon. There are different ways you can arrange these, not just for the electrical functionality, but also for the heat functionality. This is another way in which you can even-out the heat to reduce the amount of differential expansion, which will

reduce the stresses. Getting an optimal pattern of TSVs is another way you could optimise your design. There are a lot of angles to look at to see if the design can be improved.

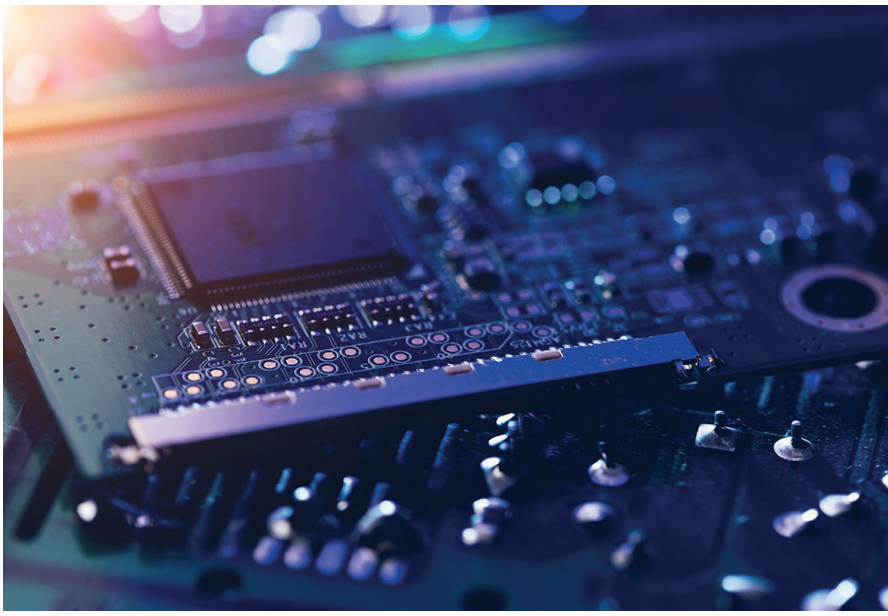
PA: *What you're describing, the main challenge is scaling up what you might call computationally demanding stress simulations, but crucially, at the same time, maintaining the predictive accuracy alongside. How big is that challenge?*

MS: These thermal simulations can be very accurate and rely on meshing. You break the design down into a finite element mesh, which means you break down the whole pattern into millions of tiny little triangles that model the design geometry and can be analysed as a small local problem. A fine mesh gives you high accuracy but takes a long time to solve. And we're talking days to weeks of computation on hundreds of CPUs, and so that can become prohibitively expensive.

There are two ways to address the huge computational demands of these simulations. One is to improve your algorithms, and that's Ansys's job. We have things like adaptive meshing, where you check if the mesh really needs to be that fine everywhere. Some places have a lot of thermal gradient, and yes, I have to model that with a fine mesh to capture those curves. But a lot of the chip, the cooler parts, have a pretty even temperature across them. I don't need a fine mesh to model that. I can do a much faster, coarser mesh and still get accurate results. So, you have adaptive meshing which is fine, where it needs to be, and it can be coarser where it doesn't need to be, and that really brings the computation time down. That's the algorithmic side.

But the other side is, no matter what you do, it's still a big tough problem. So that's where we pull in people like Microsoft and cloud providers to say, okay, we need to routinely allow for big compute times for big 3DICs. Routinely, these systems are simulated and analysed on the cloud. And that can be a mixed on-prem or off-prem cloud, or completely on the commercial cloud. But we work with the cloud vendors like Microsoft to ensure that the Ansys product works easily on the cloud, that these algorithms are distributed efficiently, but also things like resiliency. What if one of the CPUs goes down? Can the whole job recover from a single failure and not lose two days of simulation time because one CPU went down? Also, can we use the cloud spot market? It's a lot cheaper than using on-demand resources.

There are a lot of issues to deal with in making sure this all works in the cloud. And that was the reason why Microsoft was involved in this particular collaboration with TSMC. Microsoft makes its own chips too, so it's sort of mutual relationship. We work with them as a customer, we work with them as a cloud vendor. It all works together. So yes,



they make sure that these computationally very demanding jobs can be done within eight hours or so - if you throw several hundred or several thousand CPUs at it. We've run RedHawk-SC at one customer on up to 4000 CPUs to get a detailed, full-system simulation result.

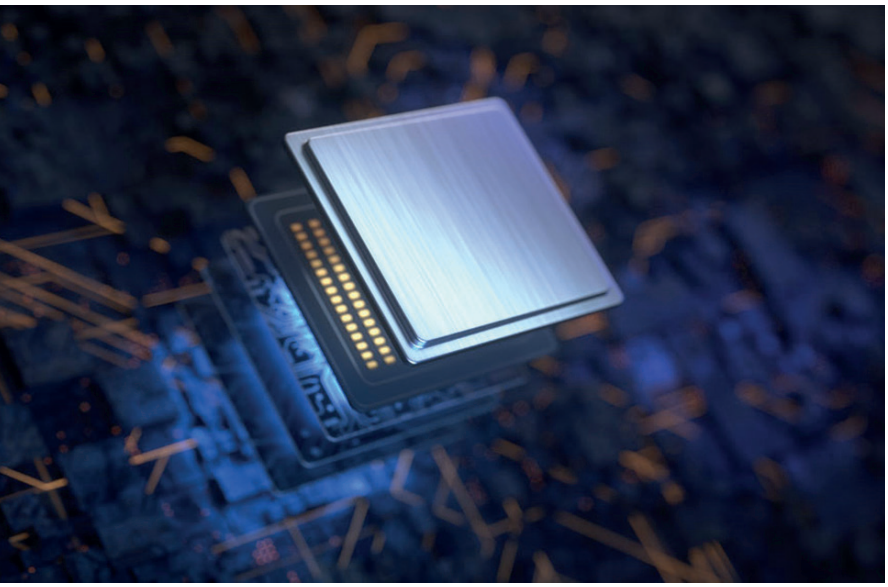
PA: *The 3D chip as a concept is relatively new and requires a new manufacturing process. Hence TSMC's need to understand much of what you have outlined in terms of 3D chip design and subsequent performance?*

MS: Yes, multi-die assembly is not something TSMC, being a chip manufacturing operation, had to worry too much about this in the past, especially since they're not system designers and don't do board assemblies. They just worry about the monolithic chip. Mechanics wasn't really an issue. However, now they have developed 3D architectures and technologies. TSMC has a reputation of making sure that their manufacturing is not just accurate manufacturing, but also that flows are available for their customers to be able to design these things. So, their concern goes beyond just manufacturing, but also can this fit in a flow that customers can take advantage of? They're very proactive in that sense. And they work with developing these reference flows for customers so that customers know what tools to use and how to use them to come to a good result that TSMC has validated against silicon and in the lab. So, for TSMC, mechanical simulation was a novel sort of thing to worry about. Most customer design teams still haven't grappled with mechanical deformation,. It is only the very leading edge that has dealt with this.

If you look at who has adopted 3D-IC today, it's the really the big HPC providers, the high end of the semiconductor design community. Nvidia, AMD, Intel, IBM, ST, those are the companies that have the volume and the density system requirements

to go for full 3D. But it's starting to trickle down more to the mainstream. Since there are so many options for building a 3DIC, the foundries have stepped up and said, okay, we will propose a number of architectures that you could use, and that are supported by our manufacturing and that we've shown work reliably. This is something new. Typically, the foundries have stayed a little away from packaging. It was done by a different set of industries called OSAT (outsourced system assembly and test). And the foundries focused on making the die, which then went off for packaging.

With 3D, it's not clear where the die ends and the package starts and so the foundries have stepped up and said, okay, we've proposed these architectures, and you can go on their websites and find these. There are multiple architectures with reference flows to validate them. So that's pulled foundries into something like mechanical simulation, because that's what their customers will have to do and it's on their architecture, with their manufacturing. They feel a responsibility to verify this all works.



PA: *The importance of what you're providing to them, from what you described, has ramped up several notches because they're now underpinning these architectures, as opposed to potentially handing off some of the problem in a different direction as they might have done previously?*

MS: Yes, exactly. 3D-IC is an emerging market that's very much still in flux. Like I say, there's many, many architectures which indicates that there's no settled agreement on which is the best way of doing it. There's still lots of technologies being tried. When you look at a 3D chip, you can look at it from two perspectives. You can say this is just a PCB but sort of shrunk down like a very high-density PCB. Or you can look at it from the other side and say no, it's like a chip but only bigger and more expanded. And customers have approached it from both sides.

With 3D, it's not clear where the die ends and the package starts and so the foundries have stepped up and said, okay, we've proposed these architectures, and you can go on their websites and find these

Some have a more PCB background, some have a more a chip background. Is it as a small PCB or big chip? It's somewhere in the middle and it sort of collapses all those problems into a single ball of wax, which is the seat of the problem. What I've seen is that, while both approaches are still used by customers, the foundries have stepped up their efforts to the point where I think this is becoming much more a silicon-oriented problem and I think in the future it'll be designed more like a giant chip.

Both perspectives have issues: If you look at it from the chip side, these interposer layers and the connections between the chips are very PCB-like. They have river routing, there are redistribution layers that are not Manhattan. You have through-silicon vias which are very large. Chip tools are not comfortable with all this PCB variability. PCB tools can handle all that.

The problem is PCB tools can't handle the scale and the capacity of these systems. So, they're used to running a few thousand signals, but on an interposer you can have millions of signals and the PCB tools just don't have the capacity to handle these millions of signals. So, both sides have issues and have to up their game and so new tools are coming on the market to handle that. For example, Synopsys has a product called 3DIC Compiler™, which is derived from their chip tools, but intended specifically for 3DIC.

PA: *In terms of where you are with TSMC, I believe the solution has been validated?*

MS: TSMC does have reference flows for their customers. So they can say, hey, here's a validated flow that we've been able to run through, and we know this works. But they also have their own backroom operations where they design their own IPs and some chips. This particular collaboration originated more from TSMC's own internal needs. It was more their own production people saying, "Hey, we see this as a problem". They collaborated with us on this flow and have also published an application note on how this all works. It's available on the TSMC design portal for all TSMC customers to access. So, it is very much a practical user perspective as opposed to simply a reference flow for the foundry.

semi

Connecting Semiconductors and Electronics

About SEMI:
SEMI connects more than 2,500 member companies and 1.3 million professionals worldwide to advance the technology and business of electronics design and manufacturing. The breadth and depth of our events, programs and services help members accelerate innovation and collaboration on the toughest challenges to growth.

ESDA Electronic System Design Alliance

FlexTech

FOA Fab Owners Alliance

MEMS & SENSORS INDUSTRY GROUP®

SOI Industry Consortium

SMART MANUFACTURING

SMART MOBILITY

SMART MEDTECH

SMART DATA-AI

WORKFORCE DEVELOPMENT

Global ADVOCACY

Sustainability

MIT MARKET INTELLIGENCE TEAM

1000+ STANDARDS

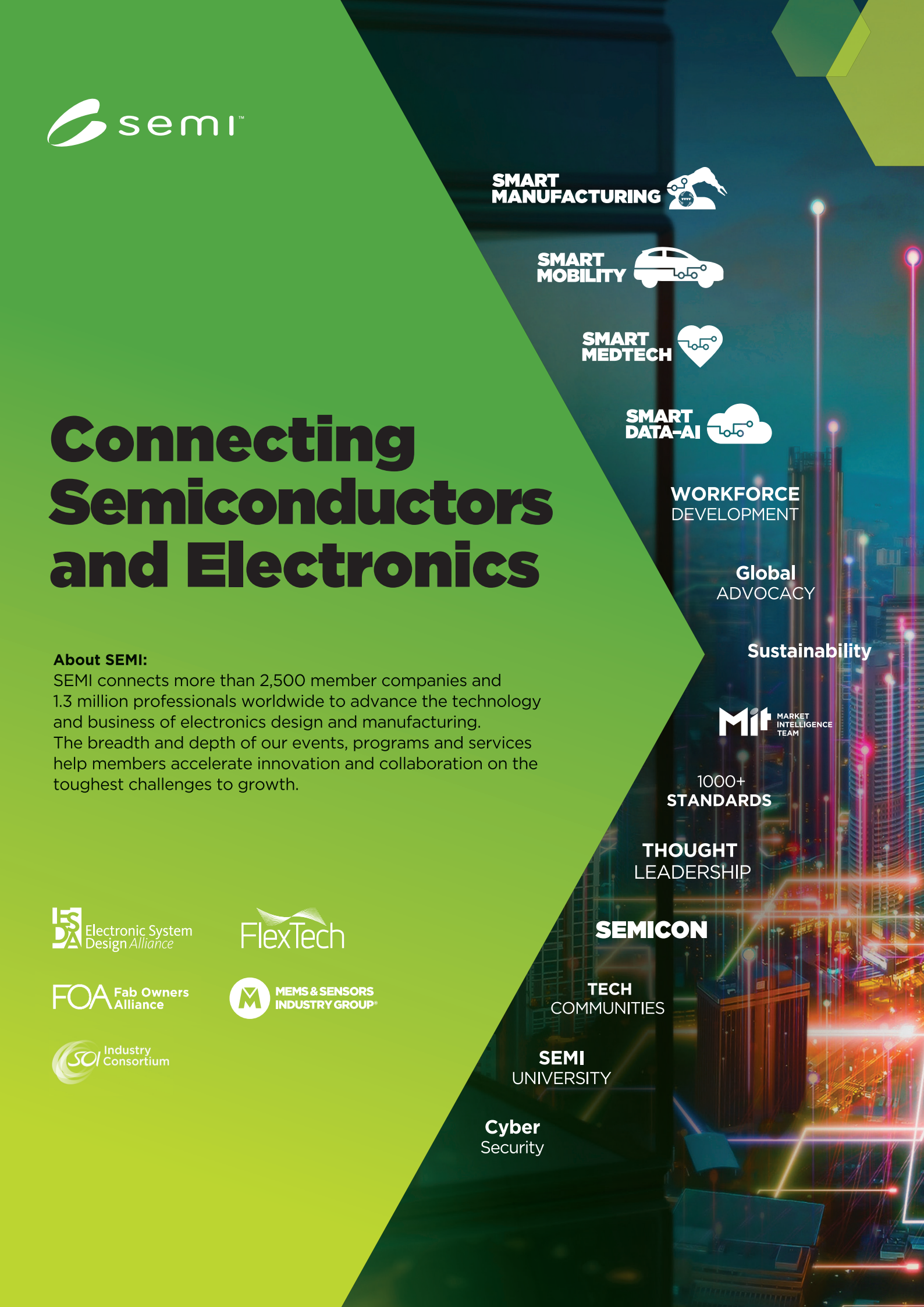
THOUGHT LEADERSHIP

SEMICON

TECH COMMUNITIES

SEMI UNIVERSITY

Cyber Security



Dry cleaning process for optimum cleaning results in electronics production

Whether energy, mobility or industry - electronic components and systems play a key role in the transformation towards climate neutrality. Due to higher demands on power density and reliability as well as new solutions, the requirements for technical cleanliness during the production of these components have increased enormously. The scalable, cleanroom-compatible and inline-capable quattroClean snow jet technology from **ACP SYSTEMS** is a dry, reliable and sustainable means of meeting these requirements for a wide range of cleaning applications.

BY DORIS SCHULZ, FREELANCE JOURNALIST

TECHNICAL CLEANLINESS has a decisive impact on the performance, functionality, reliability and service life of electronic components and systems, from semiconductor & microsystem products, PCB joining & assembly technologies to the manufacture of battery cells & modules and the production of photovoltaics & laser-structured organic electronics.

Photo credit: ACP Systems

The benchmark is the level of cleanliness defined as necessary for the downstream processes or end application, which must be achieved in a consistent and reproducible manner. To ensure compliance with the sometimes extremely high limits regarding particulate, filmic-organic and inorganic residual contamination, cleaning is usually carried out in a cleanroom. In a design adapted to the respective cleanroom class, the quattroClean snow jet technology from acp systems has established itself as a solution for such high-purity tasks.

SCC 1 and highest filmic cleanliness levels demonstrably achieved

The technology concerns a dry process which is used for full-surface or selective parts cleaning. The cleaning medium is liquid carbon dioxide recycled from chemical production processes

➤ The compact, stainless steel JetCell cleaning cell with rotating nozzles was developed for automated cleaning applications and is capable of cleaning large areas in a short time. It can be easily integrated into an interlinked manufacturing environment or operated as a stand-alone device.

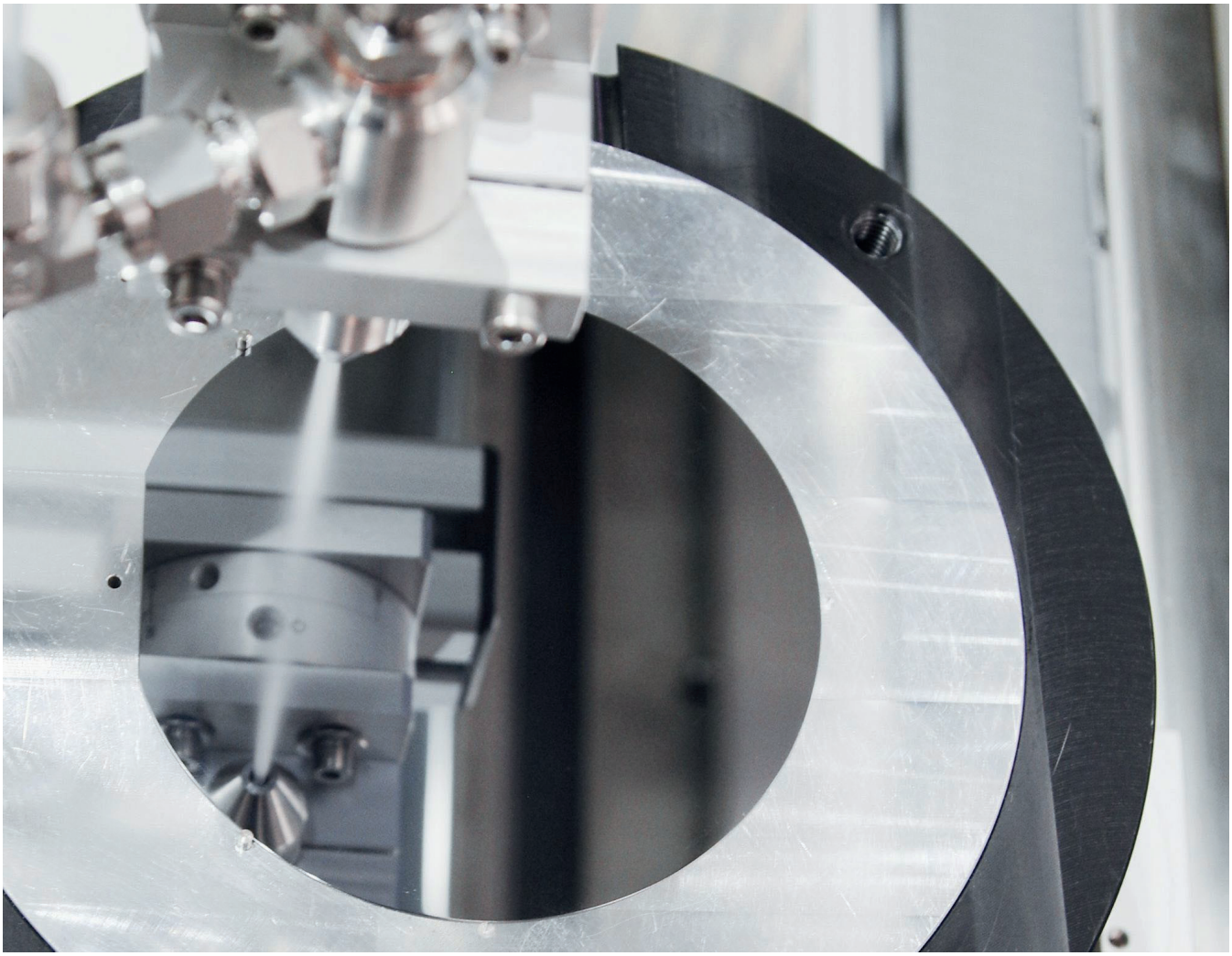


Photo credit: ACP Systems

and the generation of energy from biomass. The carbon dioxide is guided through a wear-free, two-substance ring nozzle and expands on exiting to form fine snow crystals. These are bundled by a separate jacket jet of compressed air and accelerated to supersonic speed. When the easily-focused jet of snow and compressed air impacts on the surface to be cleaned, a combination of thermal, mechanical, solvent and sublimation effects occur, which is the basis of the cleaning action. The crystalline carbon dioxide sublimates completely during the process, leaving the treated surfaces completely dry. The detached contaminants are then extracted together with the process gas, thus preventing re-contamination of the parts or contamination of the surroundings. Time-consuming and energy-intensive rinsing and drying processes are no longer necessary.

Extensive tests have shown that the quattroClean snow jet technology consistently meets the requirements of surface cleanliness class (SCC) 1 according to VDI 2083, Sheet 9.1:2006 with regard to residual particulate contamination. In addition, particulate cleanliness corresponding to SCC 0.1 can be achieved with repeat accuracy for components that have been cleaned beforehand in an ultra-fine wet-chemical process. In the case of film residues, cleaning results are comparable to those of other ultra-fine cleaning processes such as wet chemical and plasma cleaning or vacuum bake-out.

Sustainable cleaning with Green Screen certified medium

In combination with process gas from Linde, the effective quattroClean process has been awarded the Green Screen Certificate for Cleaners & Degreasers from Clean Production Action. The independent, non-profit organization has thus created a benchmark specifically for cleaning media that are used as process chemicals in manufacturing, especially in the electronics industry. The certificate confirms that the process gas does not contain any chemical substances which are harmful to humans or the environment.

Cleanroom-compatible cleaning solutions for every application

To optimally tailor the cleaning solution to the respective requirements and production situation, acp systems offers a range of machines based on standardized modules as well as individually planned systems. The design and features of the solutions, which are made entirely of stainless steel, are adapted to the cleanroom class concerned. This ensures a cleanliness of 99.995 percent for

➤ The dry quattroClean snow jet technology ensures that particulate cleanliness in accordance with SCC 1 is consistently achieved. In the case of filmy residues, cleaning results correspond to those attained with other ultra-fine cleaning processes.

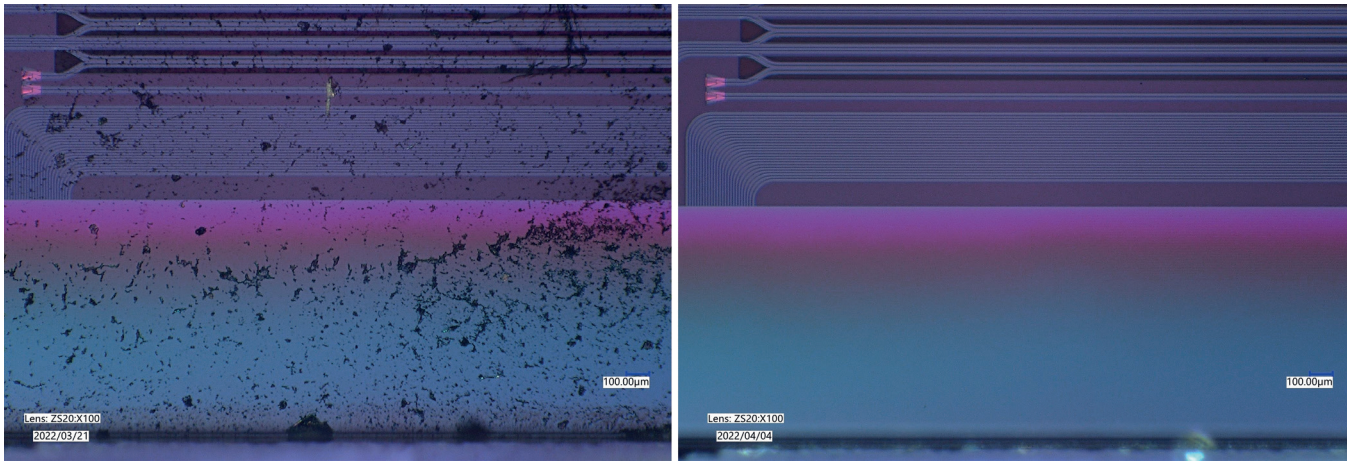


Photo credit: ACP Systems

➤ Effective, dry and cleanroom-compatible cleaning solutions such as the quattroClean snow jet technology are a must in order to reliably meet the sometimes enormously increased requirements for technical cleanliness in the electronics industry or to clean pre-assembled components.

the liquid carbon dioxide when it is prepared and a quality of 1.2.1. according to ISO 8573-1:2010 for the compressed air. If necessary, a gas scrubber can be integrated into the ultra-clean compressed air supply (XCDA) to filter out any traces of organic substances. The quality of the air generated meets semiconductor production requirements, for instance.

One of the solutions offered by acp systems is the compact JetCell with rotating nozzles. The flexible plug & play cleaning cell was developed for automated cleaning applications and is easily integrated into an interlinked manufacturing environment, e.g. immediately downstream of a punching or injection molding machine. Alternatively, it can be operated as a stand-alone solution.
It boasts a high cleaning performance with an extremely low media consumption. During surface cleaning, which can be performed simultaneously from above and below, seamless process monitoring & control ensure that each part is treated with the previously validated process parameters. Integrated interfaces enable the digitally-controllable cleaning system to be connected to a higher-level host computer.

Photo credit: ACP Systems

Cleanroom test facility for process design and validation

acp systems has its own cleanroom test facility for process design. This allows cleaning tests to be carried out in a validated ISO Class 7 cleanroom with zones up to Class 5 according to ISO 14644-1. All process parameters, such as the volume flows for compressed air and carbon dioxide, the number of nozzles, the area to be cleaned and the duration of the jet, are precisely tailored to the respective application, while also taking the material properties, type of contamination and required level of cleanliness into account. The values determined can be stored as part-specific cleaning programs in the system's control unit.

Thanks to the dry quattroClean snow jet technology, a cleaning solution is available that can be used to perform a wide range of cleaning tasks in the electronics industry reliably and efficiently. The process not only ensures the required cleanliness of electronic, mechanical, optical and sensor components but also saves energy and resources at the same time.

➤ The process is designed in the company's own cleanroom test facility. Optimum process parameters are also defined here. The process media - carbon dioxide and compressed air - are prepared in such a way so as to ensure the air quality required by the semiconductor industry, for example.

ANGELTECH

CS INTERNATIONAL CONFERENCE

PIC INTERNATIONAL CONFERENCE

PE INTERNATIONAL CONFERENCE

SAVE THE DATE

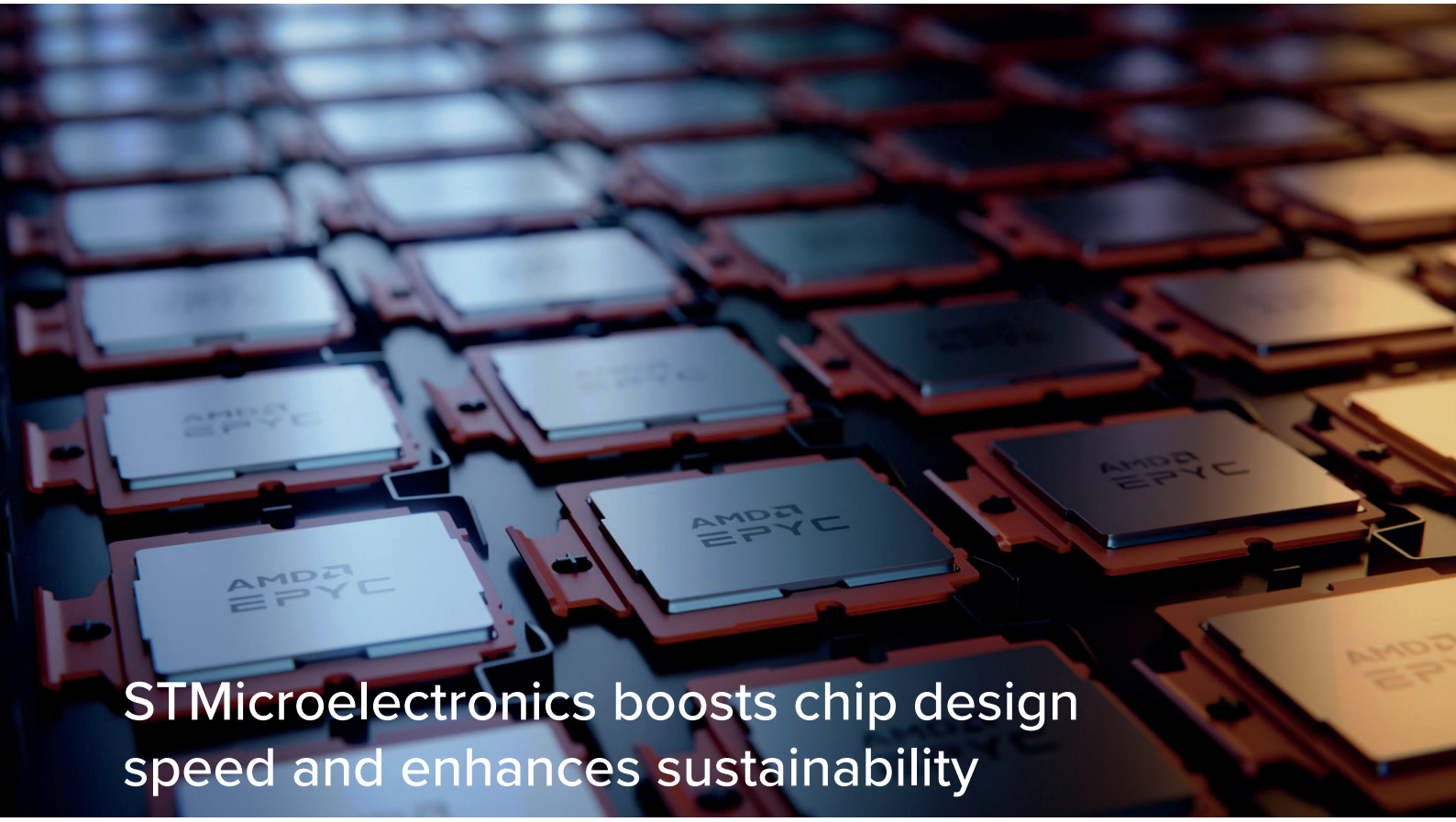
16 – 17 APRIL 2024

Sheraton Brussels Airport Hotel

To find out more about our sponsor and speaker opportunities
contact us today on: +44 (0) 2476 718970
or email:
info@csinternational.net
info@picinternational.net
info@pe-international.net

ANGELTECH
EXECUTIVE STRATEGY SUMMIT

info@angel-tech.net

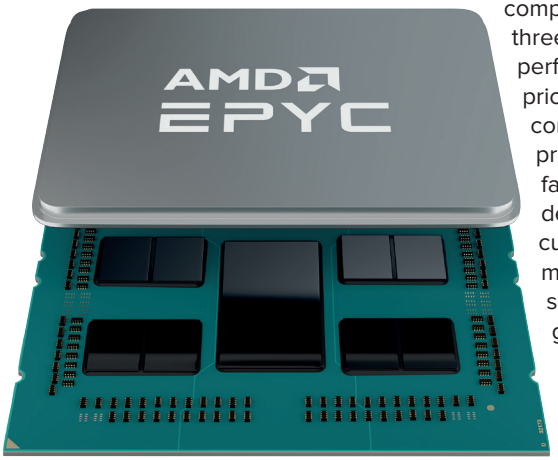


STMicroelectronics boosts chip design speed and enhances sustainability

Europe's largest semiconductor company optimised the performance, cost, and energy consumption in its R&D data centre by choosing AMD EPYC CPUs.

BY AMD

THE DESIGN and manufacturing of computer chips gets more intensive with every generation. As transistor sizes reduce, complexity increases exponentially. This creates an insatiable need for greater computational performance for production workflows. Europe's largest semiconductor company STMicroelectronics has a constant battle to keep the pace of its data centers up with the demands of its design processes. AMD EPYC™ processors enabled the company to fulfill all three of its criteria—performance, price, and power consumption—providing much faster chip design delivery to its customers while maintaining sustainability goals.



R&D performance and energy use challenges "We want to decrease our energy consumption by 150 gigawatt hours per year and become carbonneutral by 2027," says Olivier Joubert, DTIT Senior Infrastructure Architect, STMicroelectronics. "It's a big challenge for a company with doubledigit growth for the past few years as we add new factories, data centers, and increased capacity, so it's very important that we choose the right data center server technology."

Since STMicroelectronics both designs and manufactures semiconductors, it needs data center computation in three main areas. Each factory has at least two data centers, while the general business IT infrastructure is centralized in both onpremises data centers and the cloud. But its largest requirement comes from its research and development data centers. "When we halve the size of our semiconductor transistors, the density increases, and we need four times the data storage and compute," says Joubert.

The largest STMicroelectronics R&D data capacity is in France. Data centers in Italy and India provide the remainder of the processing capability. Improving R&D data center computational capacity has a

real impact on STMicroelectronics' ability to build newer, more complex chip designs. However, this poses further challenges for the company's environmental goals. "Sustainability is in the DNA of STMicroelectronics," says Joubert. "We've been in business for 36 years, and we've been publishing a sustainability report for 25 of them. We strive to consume less."

These challenges mean that STMicroelectronics is constantly evaluating new server technologies as they arrive. "We start with the most important thing for us in a server, which is the CPU," says Joubert. "We have benchmarked all the CPUs for the past 10 years. We always need the fastest processor with the most cores. For a while, there was no competition. But when AMD launched the EPYC processor in 2017, alternatives returned. When a processor has a good result in SPEC CPU benchmarks, we shortlist it for our own testing." When the 2nd Gen AMD EPYC processor arrived, STMicroelectronics decided to benchmark it. The performance figures far exceeded the company's expectations.

Better performance, cost, and power efficiency
"We were really impressed by the result of the benchmark,"says Joubert. "But we had to wait for the 3rd Gen AMD EPYC processor to enable support for our operating system and application stack." STMicroelectronics had already begun to use AMD EPYC processors through the Microsoft Azure Cloud, which the company bursts its R&D workloads into heavily when required. "We work closely with Microsoft to ensure they use the same CPU as we do in our data center.

HPE provided servers and CPUs for STMicroelectronics to test in its own data center. "The benchmark we're running for R&D is Electronic Design Automation (EDA) and Computer-Aided Design (CAD)," says Joubert. "With 2nd Gen AMD EPYC processors, we saw 6 percent more performance than any other CPU on the market.

With 3rd Gen AMD EPYC processors, that rose to 12 percent. We got 25 percent better value per core for the single-socket CPUs and 30 percent for the dual socket ones. We also introduced power efficiency tests. The 3rd Gen AMD EPYC reduced consumption per core by 30 percent. This was the first time we met all three criteria—performance, price, and efficiency—so we started to purchase a lot of 3rd Gen AMD EPYC CPUs."

Dependable roadmap to 4th Gen AMD EPYC
STMicroelectronics began rolling AMD EPYC processors out in its R&D data centers. "We started with R&D because this is the most demanding workload," says Joubert. "Better performance means faster time to market. Now we can complete a design in much less time. We're able to design a chip faster and run more designs in parallel than before."

This is particularly valuable for industries where semiconductor designs are evolving fast, such as automotive Advanced Driver Assistance Systems (ADAS) or communications networking. STMicroelectronics replaces its R&D data center servers every 27 months on a rolling basis, so is continually increasing its AMD EPYC fleet.

By the end of May, it had deployed 550 single-socket HPE servers powered by 64-core 3rd Gen AMD EPYC CPUs in its data center, and since then the total has exceeded 1,000 CPUs, with many thousands in use via Microsoft Azure as well.

The company has also been able to fulfil its sustainability goals. The increasing energy consumption of core-dense server CPUs is offset by vastly greater performance. The AMD EPYC CPUs and HPE servers are also rated to operate at higher temperatures, enabling the data center air conditioning to be run more sparingly. "We achieved a 33 percent reduction in electricty consumption using 3rd Gen AMD EPYC processors," says Joubert. STMicroelectronics also licenses its software on a time basis. The greater performance of AMD EPYC processors means more jobs can be run within that time, saving money.

CHALLENGES
Enabling R&D data center performance to keep pace with increasing chip design complexity while keeping within sustainability goals.

SOLUTION
Deploy HPE servers and Microsoft Azure cloud instances powered by 3rd Gen AMD EPYC™ processors.

RESULTS
12 percent better performance, up to 30 percent lower cost per core, 30 percent lower power consumption per core, 33 percent lower data center electricity consumption, faster chip design delivery.

STMicroelectronics is now testing 4th Gen AMD EPYC processors, with promising results. “We can get at least 25 percent more performance than the previous generation with 4th Gen AMD EPYC CPUs,” says Joubert. “That’s far better than any competitor available in the market. In our experience, even with 3rd Gen AMD EPYC processors, competitors need 50 percent more power per core, and the CPUs cost 75% more for lower performance.

So, 4th Gen AMD EPYC will be even further ahead. There are four things we like about AMD EPYC processors. First, AMD’s ability to execute the roadmap on time. Secondly, supply delivery without any semiconductor shortage. Third, AMD works with the ecosystem, supporting our stack. And finally, there is alignment between preferred partners, so HPE and Microsoft Azure have the same technology.” Encouraged by these benefits, STMicroelectronics is now also considering the addition of AMD GPUs to its data center fleet. “We are incredibly happy to have an alternative to the other CPUs on the market and a partner like AMD.”

Olivier Joubert, senior infrastructure architect at STMicroelectronics comments:

“We have tested all processors over the last 10 years. We always needed the fastest ones with the most cores. For a while, there was no competition, but since AMD introduced the first EPYC processors in 2017, an alternative has appeared again.

“After the premiere of the 2nd generation of AMD EPYC processors, we also tested them, and their performance significantly exceeded the company’s expectations. “With the 3rd generation of AMD EPYC processors, the performance advantage over the competition increased to 12%. Moreover, we achieved a 25% better price ratio to the number of cores in 1-processor machines and 30% in 2-processor machines. We also started efficiency tests and with EPYC 3 generation, energy consumption per core dropped by 30%. This was the first time we could meet all the criteria - performance, price, and efficiency, so we started buying a lot of 3rd generation AMD EPYC processors

“With the 4th generation of AMD EPYC we can achieve at least 25% more performance than the previous one. This is much more than any other competitive processor available on the market. Our experience shows that even compared to the 3rd generation of AMD EPYC, the competition needs 50% more energy per core, and processors cost 75% more for lower performance. The 4th generation of AMD EPYC will therefore be even further ahead.”

SIS ROUNDTABLE

Connecting Leaders & Experts in the Silicon Semiconductor Field

Not every discussion is a **heated debate...**

- Based around a hot topic for your company, this 60-minute recorded, moderated ZOOM roundtable would be a platform for debate and discussion
- Moderated by an editor, Phil Alsop, this can include 3 speakers
- Questions prepared and shared in advance
- There would be an opportunity to view and edit out any unflattering bloopers

This event would be publicised for 4 weeks through all our mediums including:

- A banner on the Silicon Semiconductor homepage for 8 weeks
- 4x weekly dedicated HTMLs
- 4x news pieces which would also appear on the weekly e-newsletters
- Promoted through our social media platforms for 8 weeks (pre and post event)
- Available as an on-demand asset through all mediums
- All registered attendees' details would be made available to you

Cost: €5995

Contact: Jackie Cannon at jackie.cannon@angelbc.com

ANGEL EVENTS

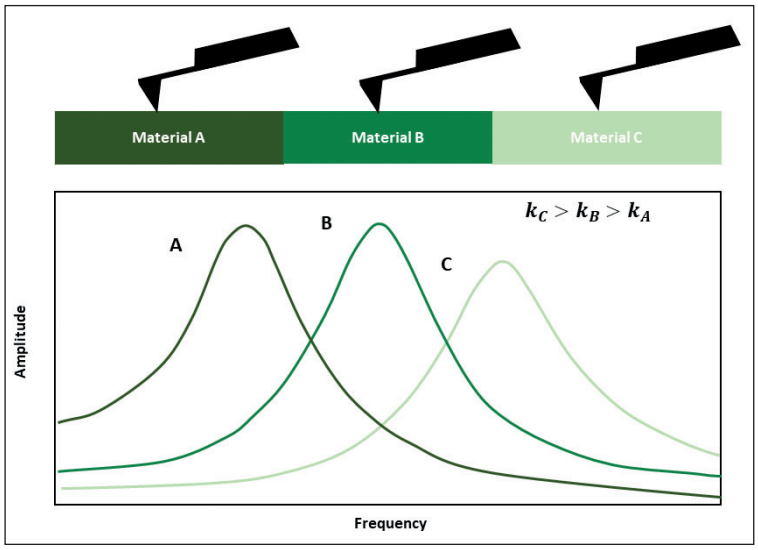
Filling the metrology void

Nearfield Instruments B.V. recently launched AUDIRA – said to be the industry's first and only in-line, non-destructive subsurface metrology system for advanced semiconductor manufacturing.

By Niranjan Saikumar; Co-authors: Nelda Antonovaité, Irene Battisti, Taras Piskunov, Mehdi Soozande

THE AUDIRA SYSTEM is aimed at providing highly accurate and precise nanometer-level measurements of buried features and defects, such as voids, in advanced memory and logic devices. With AUDIRA, Nearfield Instruments targets the introduction of complementary subsurface process control measurements to transmission electron microscopy (TEM) and critical dimension-scanning electron microscopy (CD-SEM).

Nearfield Instruments' proven in-line surface metrology solution, QUADRA, features an innovative architecture based on multi miniaturized atomic force microscopes (termed MAFM), enabling high-speed, non-destructive, on-device topography measurements. The very high speed MAFM head technology combined with the multi MAFM approach that allows for parallel imaging across the wafer results in throughputs more than 100x higher than existing state-of-the-art automated AFM systems, enabling the use of QUADRA as in-line tool in HVM fabs. QUADRA's unique architecture also ensures excellent precision and dynamic stability during measurements of CMP application layers as well as high-aspect ratio (HAR) structures and high-NA EUV (Numerical aperture Extreme Ultra Violet) applications. The proprietary advanced imaging mode uniquely places QUADRA as the only AFM tool capable of 3D measurements of critical HAR structures. Coupled with custom CD extraction algorithms and tip wear detection mechanisms for automatic probe exchange, QUADRA provides excellent in-line process monitoring capabilities for CMP process as needed with hybrid bonding, HAR structures after etching and EUV resist CD metrology.



➤ Figure 1: Contact Resonance shift from material to material. A stiffer material has a higher contact resonance frequency. Figure adapted from B.D. Huey et al, Annu. Rev. Mater. Res. 37 (2007).

The AUDIRA subsurface metrology tool will be based on the QUADRA architecture retaining the benefits of a unique multi-MAFM system architecture and the proprietary atomic force microscopy (AFM) technology aimed at high-throughput for in-line process control. In AUDIRA, Nearfield Instruments introduces two new acoustic microscopy techniques addressing the need for subsurface metrology at two different depth ranges. With both technologies, the AFM probe is used as a waveguide for sending and then "listening" to sound waves from the wafer layers.

Applications

AUDIRA is targeting both the metrology and inspection markets specifically for measuring nanometer-range subsurface structures through opaque layers.

For metrology, the principal application is overlay accuracy, which is the misalignment of patterns on different layers. Currently, many traditional optical overlay metrology techniques are limited to measuring large optical targets that are ten times larger than the dimensions of the device itself. For the 5 nm node, the overlay tolerance is in the order of a few nm, with most leading memory and logic manufacturers achieving overlay in the order of

2 nm, and expected to shrink further. With shrinking dimensions, direct on-device overlay measurements will be required to remove the possible error from metrology to the device (MTD) calibration and guarantee the performance of the device.
The need for an on-device overlay metrology solution is also driven by the increasing stochastics effects in the EUV lithography process due to interaction between EUV light and resist. Stochastic effects lead to increased CD non-uniformity (line-edge and line-width roughness) which eats up a significant part of edge placement error (EPE), which is below 5 nm for the 5 nm node, leaving even less budget for overlay error.

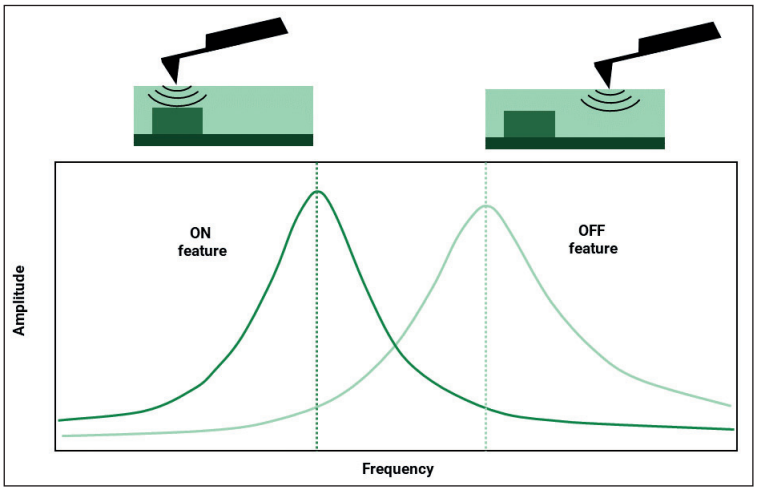
This strict overlay control places a challenge to the accuracy and precision of metrology tools into sub-nanometer range. While e-beam imaging can detect very local on-product alignment errors, including stochastics, it has limited penetration depth and is damaging to sensitive photoresists, limiting its applications. Furthermore, the memory market shows an increase in overlay metrology after etching while logic customers choose after development inspection due to large variation of feature dimensions, which points to the need for diverse overlay metrology capabilities. For example, multi-layer measurements are required both for transparent as well as opaque layers.

The inspection and review market application target for AUDIRA is primarily, but not limited to voids. Void formation in layers can be tracked at the device level to at least one critical process where the sacrificial silicon nitride is removed and replaced with conductive metal resulting in shallow voids buried 10s to 100s of nm deep.

While this issue is particularly exacerbated in 3D NAND structures with the trend towards densely packed large holes with reduced hole-spacing, this is also seen with interconnects due to the scaling down of metal line cross-sections. The increased current density due to scaling is further stressed in the presence of voids and can lead to device failure.

At the wafer/die level during advanced packaging step using hybrid bonding, surface defects on the Cu pad can create a void as large as 10 times the defect size after the bonding step with these large voids buried microns deep below opaque layers. With the size of the Cu pads shrinking to well below a micron and dense Cu pads for bonding moving towards the norm, defect sensitivity with resolution of 100s of nm or better is needed.

Defects in bonding material/process can also lead to a different type of defect which is delamination. While this is seen with hybrid bonding at the wafer level, with the device fabrication roadmap leading to 2D FET, delamination of the channel layer is a major concern for yield at the device level. In both cases, these defects result in electrical or mechanical failures.



➤ Figure 2: Cartoon picture showing the difference in the AFM probe response on and off a buried structure. Adapted from M.H. van Es et al, Ultramicroscopy 184 (2018).

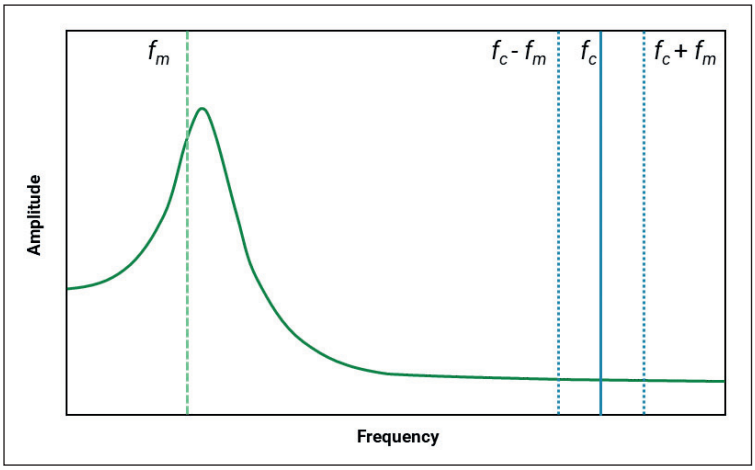
Measurement technology

AUDIRA brings forth two different complementary subsurface metrology technologies, where comparatively the trade-off is on the side of penetration depth of imaging and lateral resolution capability. From a technical standpoint, the two technologies differ in terms of the acoustic frequency range of operation and the underlying physics used in the measurement with frequencies in the MHz range used to measure shallow structures going up to 10um in depth based on the difference in the material properties, and GHz range used to measure deeper structures or structures with comparable material properties buried shallow below the surface.

Sensing viscoelastic variations

The first technology is based on sensing the viscoelastic variations of the materials, and it is suited for measuring shallow and fine features or relatively large features buried deep going down to a few ums. Lateral resolution of nm or sub-nm level can be obtained depending on the application. Every material shows a characteristic resonance frequency when sensed by a scanning probe, referred to as contact resonance (CR) (See Figure 1). The change in contact resonance can also be caused by material changes buried under the surface, like in the case of voids/delaminations or subsurface alignment markers (See Figure 2).

In literature, CR-AFM for both surface characterization as well as subsurface imaging is typically being done with the source of the acoustic wave and the sensing of the CR being separated. The acoustic wave is generally applied at the bottom of the wafer transferring acoustic energy to the full wafer, while the scanning probe on the top senses the CR at specific locations. In AUDIRA, a shift from the traditional way of working is modified by combining source and detection at the probe.



➤ Figure 3: SSURFM technique customized for direct probe-based actuation and sensing in AUDIRA. Adapted from M.H. van Es et al, Ultramicroscopy 184 (2018)

The probe is used as a waveguide to introduce acoustic waves at the carrier frequency fc in MHz range which is then modulated at a frequency fm. The nonlinear tip-wafer interaction causes a downmixed signal at fm which is amplified by the cantilever response. This is used to detect the CR changes with this measurement mode referred to as Subsurface ultrasonic resonance force microscopy (SSURFM).

The penetration depth in terms of sensitivity is determined by the feature size as well as the mechanical stiffness variation of the different materials involved. AUDIRA innovates this technique further by enhancing the resolution and the sensitivity to buried material through custom cantilever and tip designs aimed at subsurface measurements.

Wave propagation scheme

The sensitivity of the viscoelastic variation sensing drops with depth: at depths larger than 10um, it can only detect large features with significant material property variation. To address the needs at depths larger than 10um, the second technology employed by AUDIRA utilizes injection of acoustic waves in the order of GHz where the target resolution is in the range of a few nm or tens of nm depending on the target depth.

The sample is scanned using an ultrasound piezoelectric transducer (in PZT technique) or using an ultrashort laser pulse (in pump-probe PTAI method). Ultrasound waves generated at the surface of the sample propagate through its depth, reflect on the buried structures, and create mechanical deformation of the surface which is registered as an ultrasound echo

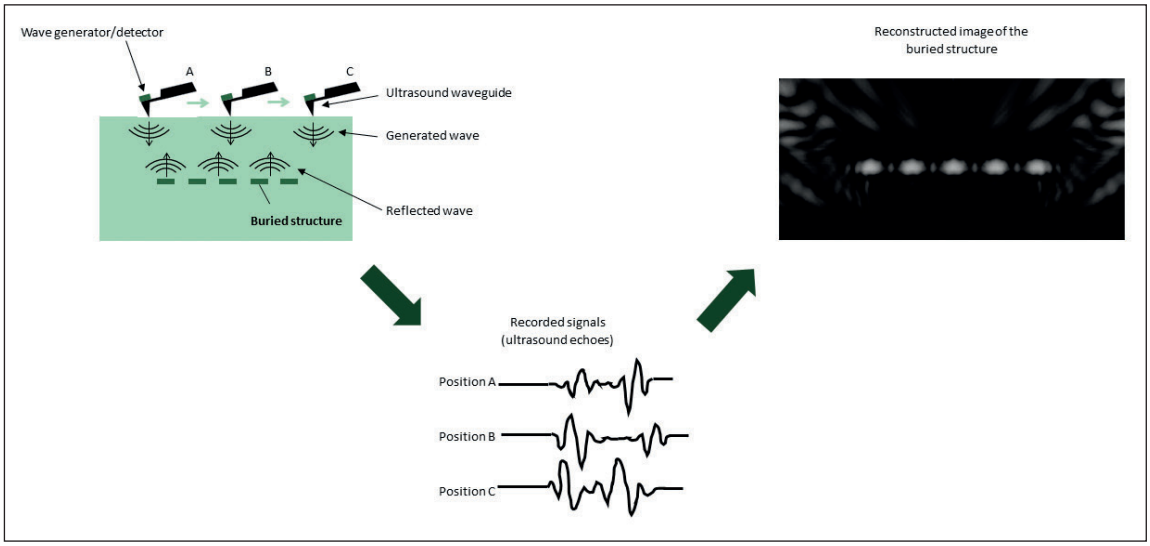
The physics utilized with this technology shifts to a wave propagation scheme wherein the transmitted wave interacts with the wafer at all the surface transitions within the device and reflects back. This reflected acoustic wave contains information linked to the strength and time of arrival which is then used to reconstruct the 3D subsurface structure image.

Two primary technology directions are considered for imaging using this scheme: the Piezoelectric Transducer (PZT) approach and Photothermal Acoustic Imaging (PTAI) with AUDIRA capable of accommodating both within the same architecture. In the PZT technique, advanced electronics are utilized to efficiently launch and detect GHz acoustic waves. The rapid switching between emission and detection enabled by these electronics allows precise control of the acoustic wavelength. This level of control plays a pivotal role in determining both the depth resolution and the signal-to-noise ratio of the measurements.

In the PTAI method ultrashort laser pulses are used to initiate acoustic waves, detect the reflections, and reconstruct acoustic echoes, operating within a pump-probe scheme. Unlike the PZT technique, PTAI places less stringent demands on the electronics component. However, while PTAI offers the advantage of simplified electronics, the echoes it reconstructs are inherently indirect measurements when compared to the direct measurements achievable through the PZT method.

The sample is scanned using an ultrasound piezoelectric transducer (in PZT technique) or using an ultrashort laser pulse (in pump-probe PTAI method). Ultrasound waves generated at the surface of the sample propagate through its depth, reflect on the buried structures, and create mechanical deformation of the surface which is registered as an ultrasound echo. The array of echoes is processed to reconstruct the buried structure. 2D scan of the surface allows to generate a full 3D image of the subsurface, which can be used to generate different cross-sections, for example the B-scan (side-view of the sample) shown on the top-right image.

From an application perspective, the two technologies provide a trade-off in the penetration depth due to the acoustic frequency that can be employed as the depth is directly dependent on the wavelength of the acoustic excitation. The PZT technique is limited by the thickness of the piezo material which can be consistently fabricated translating to lower frequencies in the range of a few GHz going up to 10 GHz, while PTAI can act



➤ Figure 4: Ultrasonic beamforming principle.

as a high frequency source going all the way up to 100 GHz.

Signal processing plays a vital role to obtain a comprehensive reconstruction of the full 3D structures for both the techniques. AUDIRA deploys a customized ultrasound beamforming technique (See figure 4). Such technique is a well-known post-processing approach used in non-destructive inspections. This method involves processing the array of echoes collected by the scanning probe during the imaging, mimicking acoustic lenses used for focusing ultrasonic beams at a point of a solid object or structure. This processing forms the backbone of the method's ability to produce high-fidelity representations of subsurface features and enables the extraction of detailed structural information.

Image Processing in AUDIRA

Nearfield Instruments recognizes the critical role played by data processing algorithms linked either to overlay extraction or defect identification. For overlay extraction, while existing algorithms for optical techniques are applicable to some extent, the unique per pixel capture of information requires special attention. To this end, Nearfield Instruments with AUDIRA also introduces custom algorithms aimed at enhancing the contrast of subsurface images and importantly improving the accuracy of the extracted overlay. Additionally, since overlay extraction through multiple opaque layers with their own respective alignment markers is needed, algorithms for resolving interfering structures and simultaneous overlay extraction are employed. Nearfield Instruments is also working with customers for alignment markers customized for AUDIRA based measurement with a view towards improving throughput and precision.

On the side of defect identification and classification, Nearfield Instruments is developing traditional algorithms, but is also in addition to traditional methods has been developing machine learning based classification algorithms which can be deployed in the fab. The machine learning route is especially needed for defect detection after hybrid bonding since the back-end process with wafer-edge trimming, wafer grinding and dicing creates artifacts related to the surface roughness and also renders low contrast images.

Platform and link to M&I roadmap

Nearfield Instruments works with customers to develop a recipe and choosing for which applications local elasticity measurements are best, and when to switch to wave propagation schemes. With both MHz viscoelastic sensing and GHz ultrasound imaging technologies, the AFM head scans across the die by capturing information from the wafer with as small as a tenth of a nanometre step. The data captured at each step is then combined and translated into a comprehensive pattern of the underlying layers.

The elasticity and wave propagation schemes have been developed so far mainly through the transmission of the acoustic waves from the bottom of the sample for the former and from the top for latter. This is severely restricting in capabilities and does not allow a single tool to handle both modalities. The bottom actuated approach for elasticity is also possible for small samples and practically impossible for large samples like a 300mm wafer. Nearfield Instruments has been successfully able to combine these technologies to allow for the scan head to use the AFM probe as the waveguide for the acoustic waves.

As a result, AUDIRA can utilize the proven multi-head architecture of QUADRA enabling faster adaptation in the fab. The multi-head architecture allows for AUDIRA to be simultaneously used for measuring at different depths on a wafer and additionally work as a defect inspection and review tool with different scan heads performing different functions. Enabling such a functionality in the scan head allows AUDIRA to function as a hybrid metrology tool in the fab and is a critical breakthrough.

Microprocessor architecture choice in the age of AI: exploring ARM and RISC-V

In a world where the everyday devices we use are based on programmable silicon, choosing the right microprocessor architecture is key to delivering a successful product. In the age of AI, the microprocessor selection for anchoring an AI solution is especially important.

BY GOPAL HEGDE, SENIOR VICE PRESIDENT OF ENGINEERING AND OPERATIONS AT SIMA.AI.

INNOVATORS building their own silicon are familiar with the process of selecting an Instruction Set Architecture (ISA) that dictates how the instructions and data types a CPU uses to perform calculations, manage data and interact with memory and other components in a given product. It is also important to understand the ecosystem of design implementations, tools and extended software support, as well as the flexibility of the licensing options.

For AI/ML products, other considerations include support for data types used in AI/ML and native instructions to accelerate AI/ML applications. While x86 ISA has dominated the computing market, x86 CPU core licenses are widely available outside of the Intel foundry ecosystem. There are two primary players in the ISA space with licensable ISA and strong ecosystems: ARM and RISC-V. ARM boasts a legacy of domination in the embedded device market, while open source RISC-V claims to be the architecture of choice for flexibility desired by emerging AI companies. These factors are largely determining their adoption and how they are used in AI systems.

ARM and RISC-V side-by-side

ARM emerged in the 1990s and has become ubiquitous due to its energy efficiency, broad ecosystem support, flexible licensing terms and integrated design implementations. The general RISC design philosophy prioritizes a reduced

number of instruction classes, parallel pipeline units and a large general-purpose register set, though ARM has further evolved this with extensions. The ARM processor has been specially designed to reduce power consumption and extend battery operation, and contains features for multi-threading, co-processors and higher code density, along with comprehensive software compilation and hardware debug technology.

ARM is a licensable IP, allowing many companies to implement custom ARM-based designs into their products. These license fees fund ongoing ARM developments and allow the company to continue improving its technology, such as the development of new extensions and optimizations for modern workloads like AI and hardened implementations targeted at deep submicron processes. However, ARM has utilized its own view of how to support an AI/ML workload as it defines its roadmap for the ARM ISA. This has not always been met with acceptance by companies, as they feel there are better approaches to support AI/ML algorithms on a programmable processor than ARM is promoting.

ARM's licensing model and vast ecosystem support have made it the dominant architecture for mobile, IoT and embedded use cases. In fact, chips containing ARM IP power most of today's devices, used by Apple, Nvidia, Qualcomm, Mediatek, Google and many more vendors in mobile, consumer and embedded silicon products. ARM has not been adopted widely for AI/ML workloads (except in microcontrollers for tinyml use cases), but to host stacks, due to the higher computational performance and power efficiency needed to support these emerging algorithms.

Companies flock to ARM for a simple reason: Its solutions work with their software. In addition, Arm closely controls the ISA and provides Arm validation suites (AVS) to ensure that software implemented for any Arm ISA implementation is binary compatible with other Arm ISA implementations.

Meanwhile, RISC-V provides an open alternative to expensive ARM licensing models. Its flexibility and modular design allows for tailored implementations for AI, IoT and many other applications, as well as the freedom to customize these extensions for specific use cases, such as AI/ML.

RISC-V was introduced in 2010 at UC Berkeley, built on a reduced instruction set architecture. As an open ISA, RISC-V's design emphasizes simplicity and efficiency by using a small set of simple and general-purpose instructions. Though often misconstrued as "open source," RISC-V is an open ISA standard, meaning RISC-V International defines and manages standards that chip developers are free to implement as they choose. Therefore, the RISC-V ISA is open, and anyone can download the documentation to use for free without requesting permission or paying a fee. In turn, these developers sometimes open source their implementations of RISC-V ISA for others to utilize. This open nature encourages third party validation, as RISC-V architecture is scrutinized closely in the public domain. There is a cottage industry of IP vendors who license RTL implementations of RISC-V, leading vendors include Si-Five, Akeana, Andes and Codasip. Some RISC-V silicon vendors with RISC-V that include AI/ML extensions also offer licensable versions of their designs.

RISC-V users benefit from an active open source ecosystem providing software, tools, a community for developers using the technology and more. Unencumbered by ISA licensing fees, RISC-V offers an attractive alternative model for designing customized chips, an offer taken up by Nvidia, QCOM, Western Digital, Huawei and more. Even more telling, many AI/ML start-ups have adopted RISC-V as the core for their AI/ML compute engine. The ability to define and customize the ISA to support new sets of AI/ML algorithms is attractive, along with the freedom to innovate without being compliant to the ISA specification, as in an ARM license.

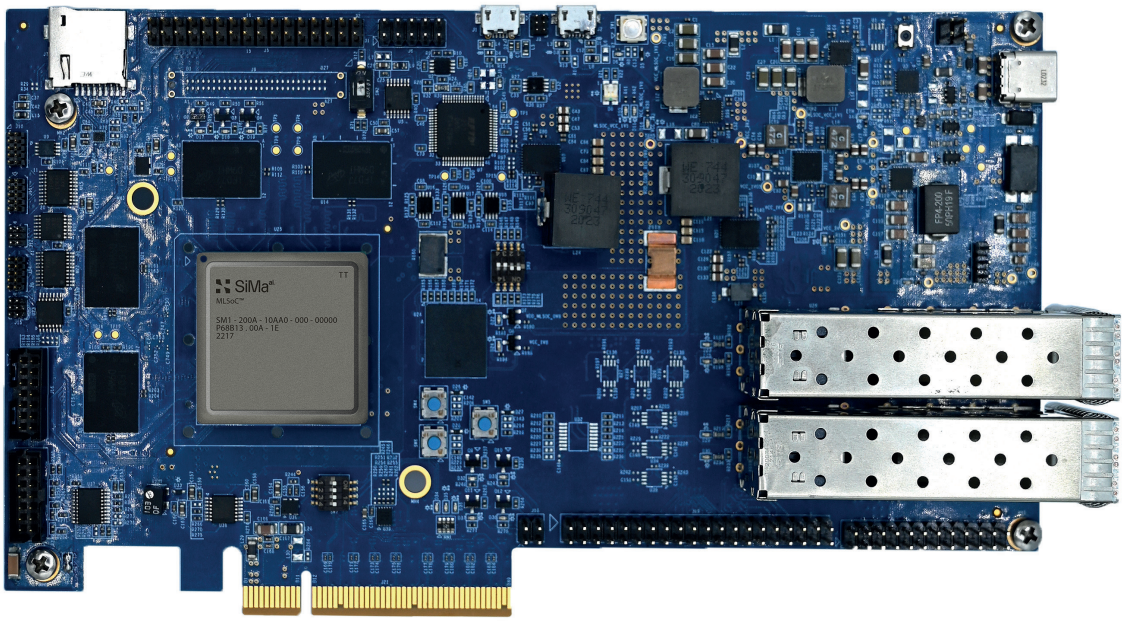
Despite the open source nature of RISC-V, lack of standardization is a recurring issue for companies implementing the architecture. As various chip developers using RISC-V create and implement a mix of open and proprietary extensions for their CPU cores, the end result are silicon chips that can run one company's software, but fail to support another.

Using a RISC-V ISA for AI/ML designs

This backdrop shows the contrasting benefits of the ARM versus RISC-V architectures and it is increasingly clear that many AI/ML start-ups have selected RISC-V over ARM. It is worth understanding why this is the case and what is the expected outcome of this preference from RISC-V.

If we go back to the restrictive ARM license, and of course the fee structures, it is clear that RISC-V can enable more innovation in the customization and extension of the architecture to address AI/ML data types and use cases. It has been a trend that short integer, floating point and block floating point representations are being utilized to reduce compute complexity, reduce area and power dissipation while still retaining accuracy. For reference, a 32 bit floating point multiply produces a 64 bit result

➤ SiMa.ai's purpose-built MLSoC™



➤ SiMa.ai's MLSoC™ Evaluation Board

whereas an 8 bit integer calculation produces a 16 bit result. This means at least four-fold less memory and even less gates to produce a result, as long as the accuracy is sufficient, this is a key trade-off. As algorithms get more complex, the data types deployed must have more dynamic range, more precision and accumulation to retain the accuracy of the computation. This has driven 16 bit integer and 16 bit floating point data type representations. Many of these designs are based around RISC-V with the incorporation of custom instructions with the new data types. However enhanced, the RISC-V single instruction scalar core is not efficient or performant on AI/ML algorithms.

To address this, companies are deploying vector coprocessors to the RISC-V core and utilizing the scalar RISC-V operations for control and data management. In the case of RISC-V, many of the extensions are customized by the vendor to extract more performance outside of the datapath as loosely coupled cores. Even an enhanced vector RISC-V core in of itself is not powerful enough for leading AI/ML use cases, so the parallelism is further expanded by scaling the number of RISC-V cores utilized in a design. This scaling of RISC-V cores into a network of processors now requires scheduling across a large number of independently executing resources. This ability to customize and extend RISC-V core architecture-has been a key driver of adoption of RISC-V for AI/ML markets.

The algorithm scheduling is the achilles heel of these multi-core processor RISC-V architecture designs today, since new compilers that require extensive knowledge of the interconnect, the state of each processor node and the mapping of non-contiguous memory spaces need to be developed. This is a significant investment with a very specialized team several years to produce good results. Supporting this non-standard programming environment requires hand coding, extensive support and longer turnaround time. Since AI/ML is a rapidly evolving field, the ability to compile, not hand code an ML model, is crucial. The hardware has to be future-proof and compatible with upcoming and future ML model network architectures.

For an entire AI/ML system solution, there are additional considerations. Customer applications include video decoding/encoding, image pre/post processing and visualization. These have different sets of compute requirements, with a different coding approach and require additional twists to the architecture and new set of software libraries to support. Using RISC-V as a system host requires support for a wide variety of system software packages including embedded OS, real time OS, drivers, user stacks, security stacks, real time scheduler, I/O, video processing libraries and more. While the RISC-V ecosystem is growing rapidly, the current state of the ecosystem is not sufficient to support end-to-end customer applications.

One could design an AI/ML chip with an ARM host processor and its broad ecosystem support, controlling a RISC-V based AI/ML accelerator, but we have not seen this. It almost seems antithetical to those in the RISC-V community and would be interesting in how ARM would address this licensee. To mitigate using a RISC-V system host processor, the x86 platform becomes the host, as in the server marketplace today, providing complete system software support and the RISC-V AI/ML chip as an accelerator card. However, x86 is not as power efficient as ARM and additional power, latency and board interconnect incurs delays in processing the pipeline due to an external compute subsystem, lowering overall application performance.

The flexibility RISC-V touts is not only difficult to program, but also has significant costs in power

and performance for AI/ML versus alternatives. The most efficient execution of AI/ML algorithms has concentrated on matrix multiplier arrays sometimes called tensor units or tensor processing units (TPUs). The basic premise is that the basic AI/ML kernel algorithms can be run on these compute blocks and minimize data movement and storage, which are the two largest power consumption factors. The ability to load/store is very expensive if users simply want to multiply-add. The use of TPUs with local memory and in-memory compute architectures, including analog techniques, seeks to address this by focusing on reducing data movement and storage. These in-memory AI/ML compute chips can leverage an x86 host as well, making it difficult for RISC-V based accelerator chip designs to compete against these in-memory accelerators in performance and power efficiency.

An alternative approach

It is clear that ease of software development and lowest power consumption are the two most important metrics for an AI/ML chip, especially at the edge. To achieve this, SiMa.ai designed and built a very low power highly programmable compute array that delivers industry leading fps/watt performance, with a flexible compiler that generates highly optimized machine learning code for a wide variety of ML networks combined with an ARM CPU with its broad ecosystem of tools and SW.

In addition, SiMa.ai's MLSoC integrates an industry standard computer vision processor for image pre/post processing, an industry standard video encoder/decoder processor with all the necessary IOs to support a wide variety of sensors. All these building blocks are interconnected on a single die by a very high performance network on a chip fabric to ensure highly efficient processing of computer vision applications.

ARM is the clear choice from a host system perspective, since building a complete embedded system-on-chip where deployment requirements for system software security, reliability and maintainability for AI/ML designs can leverage a mature ecosystem for support. This is why the industry leaders are using ARM today and combining it with an ML power efficient compute architecture for the future.

The ARM advantage

Chip developers who want to optimize their products for customer accessibility will continue to implement ARM's technology into their silicon ecosystems.

ARM continues to dominate mobile and embedded devices due to its maturity and software support, as demonstrated by its recent IPO. It is used by system level companies that are integrating AI and ML into their devices, as the core of its product architecture ensures reduced power consumption and extended battery operation.

RISC-V adoption continues to surge for embedded microcontrollers and is seeing more targeting for data center accelerators, fully vertically integrated systems where the software is provided by the platform vendor, HPC, as well as emerging applications in networking where customization and privacy is a priority. RISC-V is embraced by academic institutions who utilize open source implementations of the RISC-V ISA for research projects and training programs. The open architecture also gives developers the opportunity to build a platform outside of the control of a single company with potential backdoors and auditable code.

RISC-V specifically falls short in validation, regression and verification of cores when compared to ARM. As an established institution with internal structures to support continued evolution, ARM's progression leads to subsequent generations of processor models, placing microprocessors designed with RISC-V architecture less robust than ARM's processor models by some market observers.

ARM's proprietary architecture and licensing fees somewhat inhibit customization opportunities and wider adoption, but with its proven track record, stability and established standards, it is still the frontrunner in ISA adoption.

ARM and RISC-V represent divergent models and philosophies, to be sure. Though RISC-V's platform has the potential to make it the ubiquitous architecture from wearables to supercomputers, it must overcome the issues in its standardization to create a larger software ecosystem before it can gain true leeway over ARM.

ARM's maturity, software support and licensing ecosystem give it the advantage for the foreseeable future as the dominant force behind mobile and Internet of Things devices operating at the edge.

ARM and RISC-V represent divergent models and philosophies, to be sure. Though RISC-V's platform has the potential to make it the ubiquitous architecture from wearables to supercomputers, it must overcome the issues in its standardization to create a larger software ecosystem before it can gain true leeway over ARM

Semiconductor market to recover in 2024

IDC reveals eight key trends for the global semiconductor market in 2024.

ACCORDING TO IDC's latest research, with the global demand for artificial intelligence (AI) and high-performance computing (HPC) exploding, coupled with the stabilizing demand for smartphones, personal computers, infrastructure, and resilient growth in automotive; the semiconductor industry is expected to usher in a new wave of growth. Semiconductor products cover logic integrated circuits (IC), analog IC, microprocessor and microcontroller IC, and memories.

"Memory manufacturers' strict control of supply and output has led to increasing prices from the start of November, and the demand for AI across all major applications will drive the overall semiconductor sales market to recover in 2024. The semiconductor supply chain, including design, manufacturing, packaging, and testing, will bid farewell to the downturn in 2023," says Galen Zeng, Senior Research Manager, Semiconductor Research, IDC Asia/Pacific.

IDC forecasts eight trends for the semiconductor market in 2024:

#1: The semiconductor sales market will recover in 2024 with an annual growth rate of 20%

The supply chain inventory depletion process continues due to weak market demand. Although there are some sporadic short orders and rush orders in the second half of 2023, it is still difficult to reverse the first half annual decline of 20%, thus, the semiconductor sales market is expected to still decline by 12% in 2023. After the memory market recession of over 40% in 2023, in 2024 the effect of the reduction in production to push up the price of the product, coupled with the increase in the penetration of high-priced HBM is expected to become a driving force for market growth.

With the gradual recovery of smartphone demand and the strong demand for AI chips, IDC expects the semiconductor market to return to a growth trend in 2024, with an annual growth rate above 20%.

#2: ADAS (Advanced Driver Assistance System) & infotainment drive automotive semiconductor market development

Although the growth of the automotive market has remained resilient, the trend of automotive intelligence and electrification is clear and an important driver for the future semiconductor market. ADAS accounts for the largest share of the automotive semiconductor market, with a compound annual growth rate (CAGR) of 19.8% by 2027, accounting for 30% of the automotive semiconductor market in that year.

Infotainment accounts for the second largest share of the automotive semiconductor market, with a CAGR of 14.6% by 2027, accounting for 20% of the market in that year, driven by automotive intelligence and connectivity. Overall, more and more automotive electronics will rely on chips, which means the demand for semiconductors will be long-term and steady.

#3: Semiconductor AI applications spread from data centers to personal devices
AI is making a big splash because data centers require higher computing power, data processing, complex large language models, and big data analytics. With the advancement of semiconductor technology, it is expected that more AI functions will be integrated into personal devices starting in 2024. AI smartphones, AI PCs, and AI wearable devices will be gradually launched into the market. It is expected that there will be more innovative applications for personal devices after the introduction of AI, which will positively stimulate the increase in demand for semiconductors and advanced packaging.

#4: IC design inventory depletion gradually ends, Asia-Pacific Market Expected to Grow 14% by 2024
Although the performance of IC designers in the Asia Pacific was relatively sluggish in 2023 due to the long inventory rationalization, most vendors remained resilient despite the pressure of the market. Each vendor was active in investing and innovating in order to remain relevant in the supply chain. . In addition, IC design companies continue to cultivate technologies by leveraging the adoption of AI in client devices and automotive. With the gradual recovery of the global personal device market, there will be new growth opportunities, and it is estimated that the overall market will grow by 14% annually in 2024.

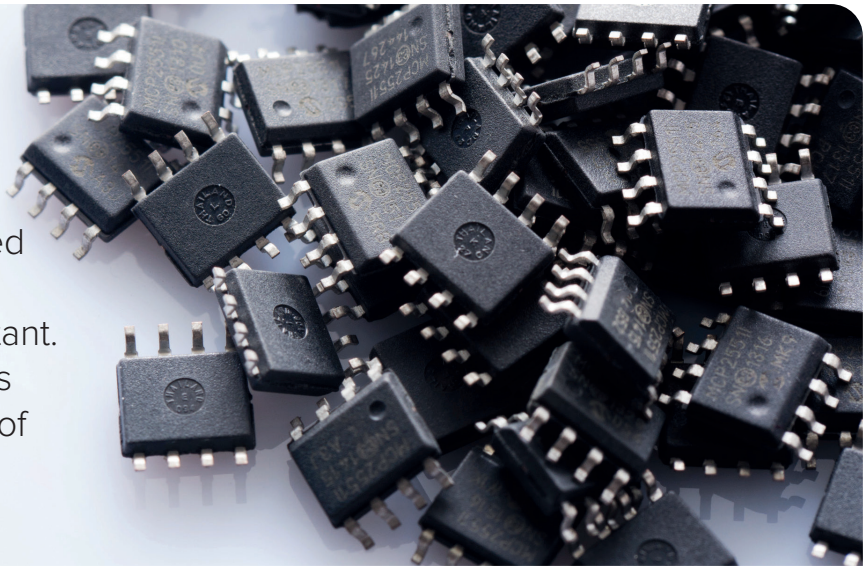
#5: Demand for advanced processes in foundry industry soars
The foundry industry has been affected by the inventory correction and weak demand environment, as capacity utilization rates dropped significantly in 2023, especially for mature processes technologies above 28nm. However, due to the rebound in demand for some consumer electronics and the demand for AI, 12-inch fabs have been recovering slowly in the second half of 2023, with the recovery of advanced processes being the most obvious. Looking forward to 2024, with TSMC, Samsung, and Intel's efforts, and the gradual stabilization of end-user demand, the market will continue to rise and it is expected that the global semiconductor foundry industry will grow in double-digits next year.

#6: Growth of China's production capacity and intensified price competition for mature processes
Under the influence of the U.S. ban, China has been actively expanding its production capacity. To maintain its capacity utilization rate, the Chinese industry has continued to offer preferential pricing, which is expected to put pressure on "non-Chinese" foundries. In addition, the inventory of industrial control and automotive IC in the second half of 2023 to the first half of 2024 will have to be de-stocked in the short term, as wafer production is primarily concentrated on mature processes, which will continue to put pressure on suppliers and their ability to regain bargaining power.

#7: The CAGR of 2.5/3D packaging market is expected to be 22% from 2023 to 2028
As the functionality and performance requirements of semiconductor chips continue to improve, advanced packaging technologies are becoming increasingly important. The 2.5/3D package market is expected to grow at a CAGR of 22% from 2023 to 2028, making it an area of high interest in the semiconductor package testing market.

#8: CoWoS supply chain capacity expands twice, boosting AI chip supply
The wave of AI has led to a surge in server demand, which relies on TSMC's advanced packaging technology CoWoS. Currently, there is still a 20% gap between supply and demand for CoWoS. In addition to NVIDIA, international IC design houses are also increasing their orders. It is estimated that the capacity of CoWoS will increase by 130% by the second half of 2024, and more vendors will actively enter the CoWoS supply chain, which is expected to make the supply of AI chips even more robust in 2024, and will be an important growth booster for the development of AI adoption.

As the functionality and performance requirements of semiconductor chips continue to improve, advanced packaging technologies are becoming increasingly important. The 2.5/3D package market is expected to grow at a CAGR of 22% from 2023 to 2028



Leveraging AI to efficiently test AI chips

The semiconductor devices that bring artificial intelligence (AI) and machine learning (ML) to the cloud as well as to the edge present significant test challenges. The semiconductor test industry is well positioned to address these challenges by leveraging AI and ML techniques to analyze and correlate data across the entire semiconductor value chain.

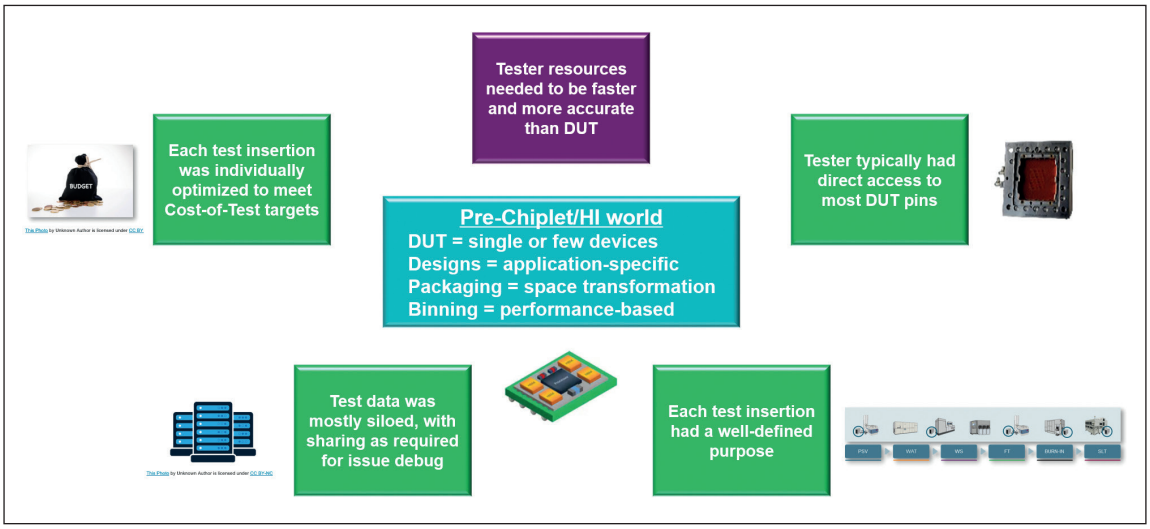
BY IRA LEVENTHAL, VICE PRESIDENT, APPLIED RESEARCH & TECHNOLOGY, ADVANTEST AMERICA, INC.

EMERGING ARTIFICIAL-INTELLIGENCE (AI) and machine-learning (ML) applications are driving a revolution in the semiconductor industry. Traditional central processing units (CPUs) have yielded ground to graphical-processing units (GPUs) to handle the massively parallel computations common in AI and ML programming, and the GPUs themselves are facing competition from a new generation of AI processors with architectures that are optimized for AI and ML inferencing and training. Many of these AI processors are being developed by new entrants into the semiconductor space such as the hyperscaler companies and startups.

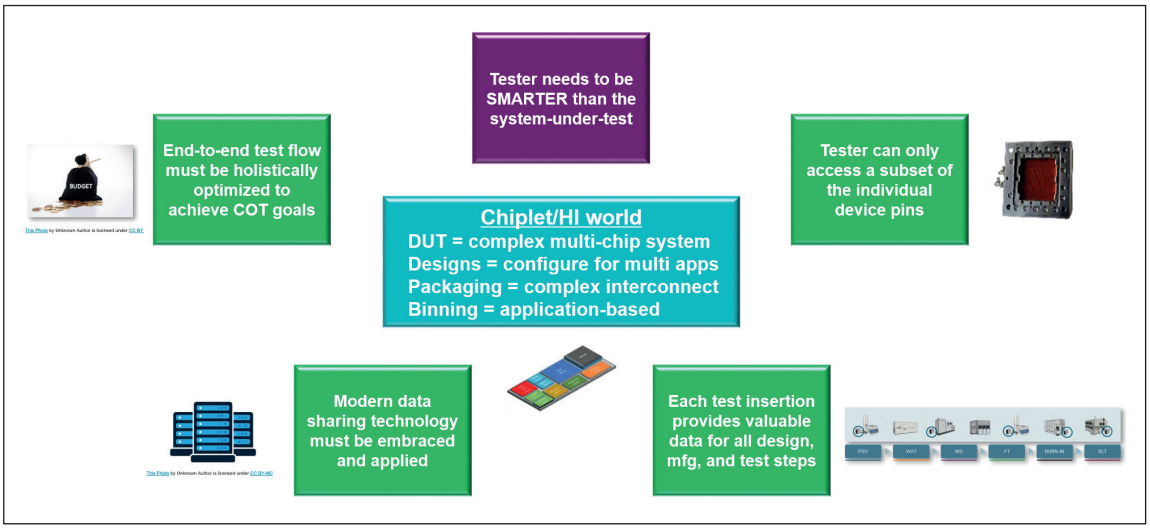
The semiconductor test industry is playing a critically important role in helping its customers solve the problems brought on by the move to dedicated AI processors and other high-performance compute devices. The test industry is also addressing the accompanying challenges of shrinking process nodes, the proliferation of heterogeneous-integration (HI) packaging, the deployment of advanced communications architectures, and chipmakers pushing the limits of power consumption. Initiatives include real-time adaptive testing, predictive maintenance, DC parametric test improvements, in situ thermal test and optimization, and efficient workflow balancing. The key to addressing the challenges is to leverage AI and ML in the production and test processes. Advantest is deploying multiple capabilities to enable AI device test, including advances in hardware, interfacing, handling, thermal management, and data analytics, including AI and ML.

Pre- and post-chiplet worlds

The traditional CPUs of the pre-chiplet world can be described as “purposeful,” or focused on specific well-understood applications. In contrast, the new chiplet-based AI devices may ultimately serve use cases that had yet to be identified when the devices were designed. When a new application arises, the devices cannot be redesigned or even retested in a cost-effective manner. What is needed is a test strategy that generates sufficient data to map these devices to new applications after test is completed.



➤ To test traditional devices, the tester must stay ahead of the DUT.

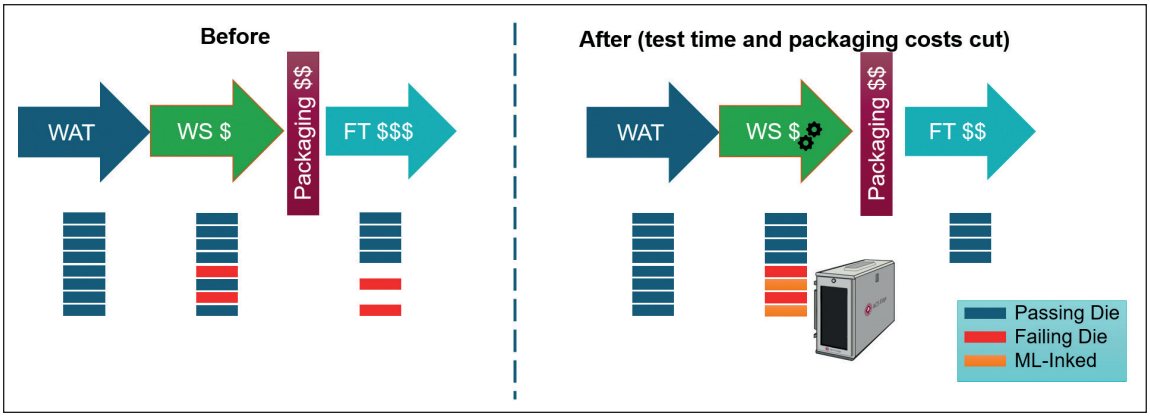


➤ In the chiplet world, the tester must stay ahead of the system.

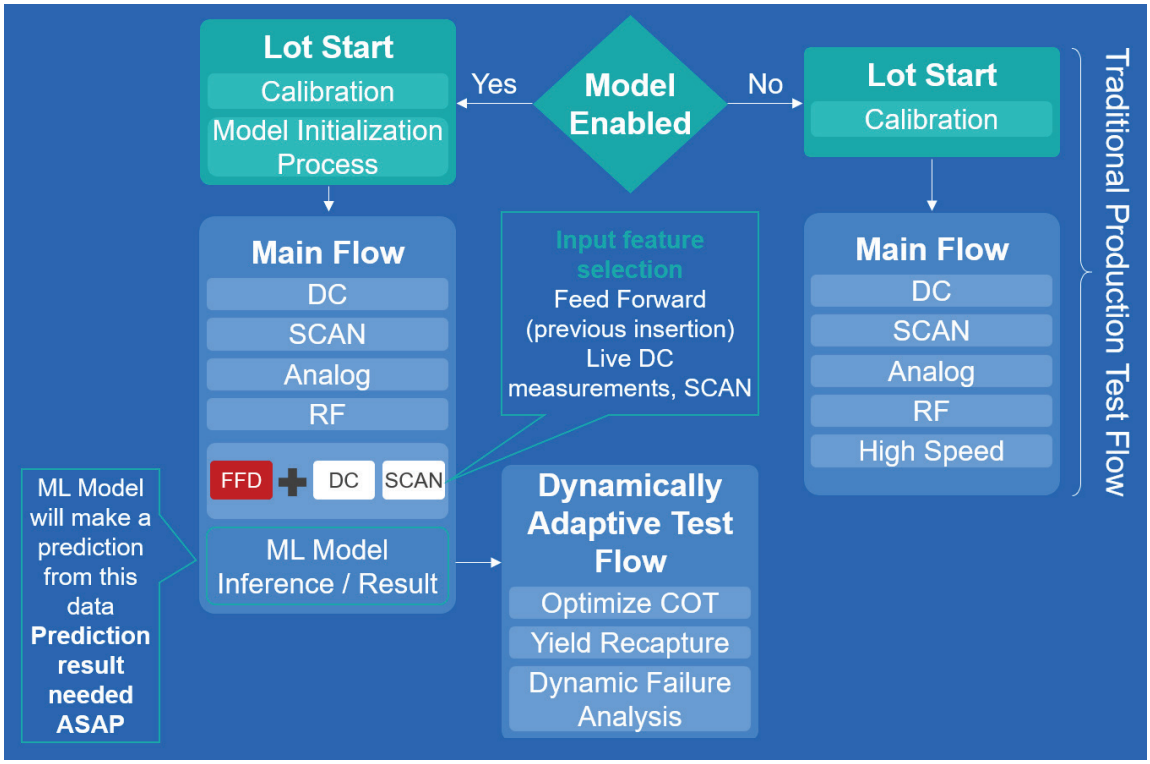
With traditional test in the pre-chiplet world, the device under test (DUT) consists of a package containing one or a few devices. In such applications, the tester performance must stay ahead of the DUT. Binning is performance-based, requiring that tester resources be faster and more accurate than the DUT. In this world, the tester typically has direct access to most DUT pins, and each test insertion – from post-silicon validation (PSV) to system-level test (SLT) – has well-defined purposes, including the collection of test data. However, the data has typically been stored without context, and has been siloed and shared only as required to address major issues such as significant yield loss.

The industry is now moving to chiplet-based modules, using a "Lego-like" approach to integrate CPU, GPU, cache, I/O, high-bandwidth memory (HBM), and other functions. In the new world of chiplets[1] and HI, the DUT is a complex multichip system with the integration of many devices in a single 2.5D or 3D package. Consequently, the tester can only access a subset of individual device pins. Nevertheless, at each test insertion, the tester must be able to extract valuable data that is used to optimize the current test insertion as well as other design, manufacturing, and test steps. With limited pin access, the tester must infer what is happening on unobservable nodes. To best achieve this goal, it is imperative to extract the most value possible out of the data that can be directly collected across all manufacturing and test steps, including data from on-chip sensors. The test flow in the chiplet world already includes PSV, wafer acceptance test (WAT), wafer sort (WS), final test (FT), burn-in, and SLT, and additional test insertions to account for the increased complexity of a package with multiple chiplets are not feasible from a cost perspective. Adding to the challenge, binning goes from performance-based to application-based. In this world, the tester must stay ahead of the system – the tester must be smarter than the complex system-under-test.

One example of application-based real-time binning is smart pairing, also called die matching. Smart pairing considers factors such as voltage, speed, and power consumption to ensure that all dies destined for a particular package have compatible electrical and thermal characteristics, minimizing differences related to factors such as reliability, signal delays, and power imbalances. As another example, a multi-core device might be designed for an application that requires the cores be closely matched. A subsequent application might arise that requires fewer cores or that can tolerate more relaxed matching requirements. Real-time binning can select the highest performance, highest quality devices for the first application while reserving other



➤ An ML-optimized shift left strategy helps ensure that only KGD are packaged.



➤ An ML-enabled dynamically adaptive test flow will minimize COT, maximize yield, and facilitate dynamic failure analysis.

devices for the second less stringent application. The rules-based decision models that have worked well in the pre-chiplet world will continue to serve certain purposes. But for chiplet-era production the ML model is the essential path forward to meet the complex real-time decision-making requirements.

Detection of subtle failure mechanisms in AI devices requires a wide range of input data. In addition, with limited test access in the final package, on-chip sensor data, equipment settings and sensor data, and upstream data from design and manufacturing will all become increasingly important in identifying defects. ML can leverage all this data to fully enable powerful predictive and adaptive test approaches. In general, the goal is to enable the customer to "shift left" in terms of debug and disposition – catching defects at WS, for example, before assembling the wafers into packaged devices.

Packaging cost and complexity are aggressively increasing in the chiplet/HI world, so it is critical that only known-good die (KGD) make it into a package. An ML model running on a WS system can make real-time decisions about KGD based on data from multiple insertions. Prior research has shown that applying a gaussian process-based outlier detection model at WAT can prevent many test escapes with minimal overkill (good dies incorrectly marked as defective), with FT yield increases resulting in significant cost savings.[2]

Keep in mind, however, that there are exceptions to the shift-left rule. For example, it may be beneficial to shift a lengthy test from FT to the right onto an SLT system, where test time is less costly. In either case, an ML strategy can help optimize test flow.

Applying AI and ML in a secure environment

Throughout the test process, the key is to leverage AI and ML to test the AI devices) – to analyze test data, to correlate data across multiple wafers, and to track exactly how each die made its way through the assembly process. A vast array of advanced tools and algorithms applied to design, manufacturing, test, and equipment data will provide correlation across the full manufacturing flow to provide insights that will enable flow optimization.

Increasing test requirements will drive the need for AI- and ML-driven predictive methods, with real-time execution of test code enabling automated corrective actions during production to control costs. One test customer has noted that ML-based parametric prediction can reduce parametric test time by 50% or more while successfully identifying performance-matched devices with no impact on quality.[3]

Aggressive adoption of modern data security technologies will be required to protect sensitive IP in outsourced manufacturing environments. The semiconductor supply chain often includes contract assembly and test services, and test equipment in one facility may process parts from multiple IC providers – presenting challenges for providers who want to secure all aspects of sensitive test IP. A facility can be expected to provide perimeter security, but with complex supply chains it may become difficult to define the perimeter.

One approach is the zero-trust security model. "Zero trust" implies that products and services by default must not trust other products or services. A zero-

<---------- Input Data Streams ---------->									Test Point
Material Descriptions In Hierarchy	Meta Data	MES/ WIP Equipment History	Fault Detection and Control (FDC)	Defect & Metrology	Equipment and Non-Lot	Assembly / System	ERP / Business / Financial	Data Sources & Connectors	Product
Technology Family Process Product Source Lot Lot Wafer # Die Layout	Equipment Operator Program Recipe Date/Time Process Flow Stages Steps SEMI E142 Traceability	Equipment TrackIn/Out Recipe Operator Chamber QueueTime CycleTime Wfr Counts Rework Reticle	Indicators Summaries Trace Charts Model prediction & Management Real time data collection Sub-fab FDC	Parametric Categorical Lot / Wafer Summaries Defect Summary Kill Ratio Defect Images Bump Msmts	Equipment SPC Fab chemical delivery Equipment counter data Equipment Event / PM Consumables	Die traceability Location of reel/tube Solder paste batch, vendor Equipment parameters Operator logs	Production Order Product Costing Consumables Parts Alarms / Events Split / Merge Scrap / Unscrap Financial EBOM	Other 3rd party Apps & databases ... Data Lake connectivity BDAPI ... Sapience Manufacturing HUB	Test chip (CV) PCM / WAT Wafer Sort Bin Map Multi-Bin Final Test Module Data MCM Chiplets WLA

trust inference environment used in conjunction with an ATE system protects against threats both inside and outside of the traditional network boundaries, reducing the "zone of trust" from the network level to individual nodes or applications. Interactions between zones of trust require continual authentication.

Running AI/ML workloads

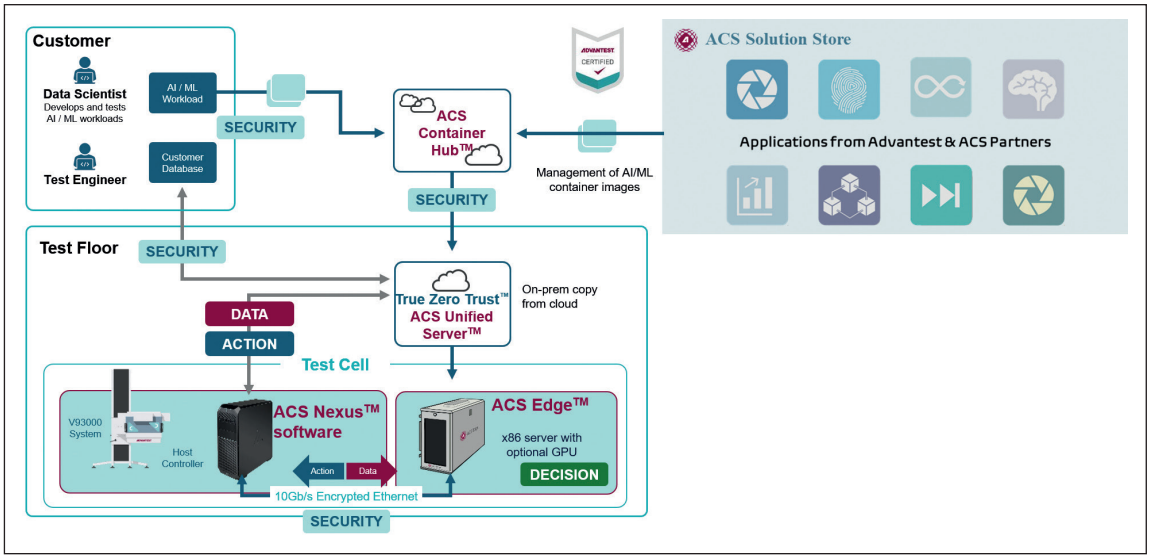
A key consideration moving forward involves how to run AI and ML workloads in a test environment. The ATE host controller is not a good candidate. It is busy with standard ATE control tasks such as managing HW resources, running scan tests, and logging data. In addition, any changes to the host controller OS or SW configuration trigger costly recorrelation exercises, and hardware-refresh opportunities are limited. It is unlikely that a host controller could be upgraded with an ML-optimized processor, and even were that possible, running ML and sharing the resulting data with OSATs and other third parties could make sensitive test IP vulnerable.

Edge computing provides an effective and efficient way of performing real-time, ML-based decision-making in a test environment. Edge computing offers several benefits for applications in general: it provides low and predictable latency, improves bandwidth efficiency, is scalable, and enables enhanced security and privacy. In the test environment, it can handle chores spanning multiple time domains, providing processing in the milliseconds required for a single touchdown at wafer probe; in near real time, after five or 10 touchdowns, for example; and periodically, after a certain number of wafers or devices have been tested. Finally, at lot end, it can transmit the relevant data to the cloud for post processing and offline analytics.

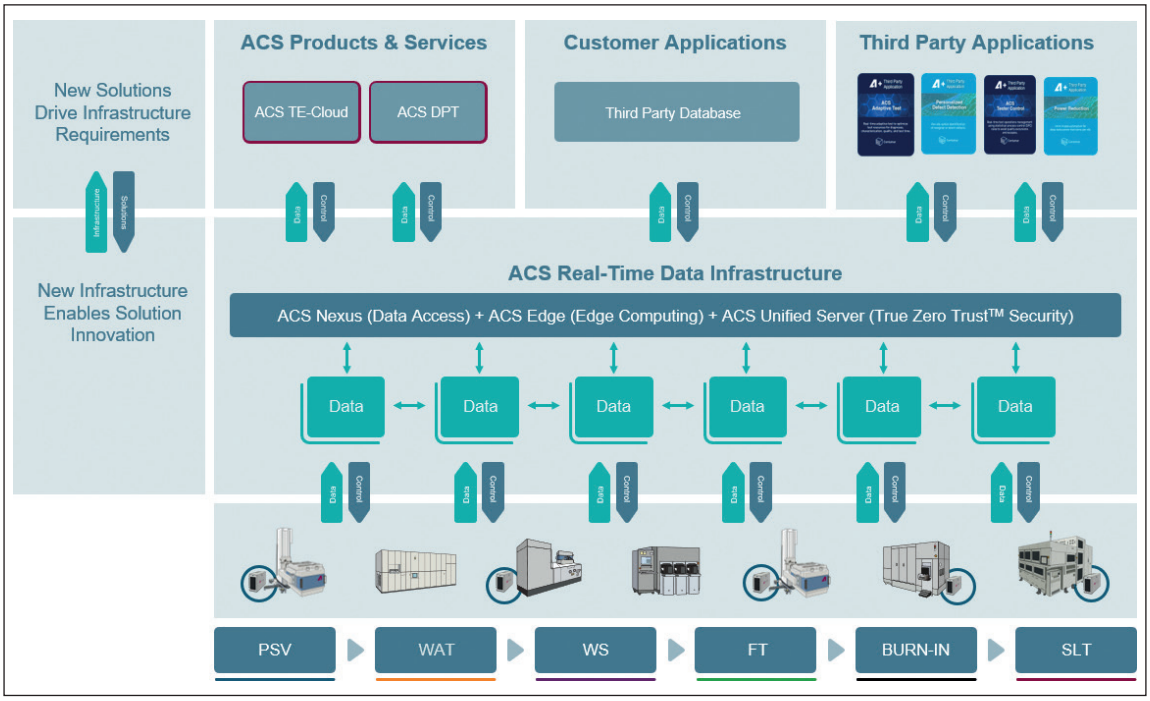
An effective edge solution applies analytic models to enable real-time dynamically adaptive test flows.[4] The goal is to minimize COT, to maximize yield, and—when needed – perform dynamic failure analysis to find root-cause correlations with other test insertions to facilitate continuous process improvements.

The solution should be compatible with a test vendor's entire lineup of hardware and software platforms, and it should also have an open architecture to support customer-owned analytics models. It should be reliable, secure, and easy to update. In addition, it should be able to derive insights by integrating all data sources across the entire manufacturing supply chain, including the IDM, foundry, or fabless semiconductor company.

➤ An effective edge solution should support the full range of input data streams available in semiconductor manufacturing.



➤ The ACS RTDI platform accelerates data analytics and AI/ML decision-making.



➤ Advantest's Open Solutions Ecosystem provides a reliable, secure real-time data structure that integrates data sources across the IC manufacturing supply chain.

The adoption of edge compute for ML test applications is well underway. Advantest's recently launched ACS Real-Time Data Infrastructure (ACS RTDI™) platform accelerates data analytics and AI/ML decision-making within a single integrated platform. It collects, analyzes, stores, and monitors semiconductor test data as well as data sources across the IC manufacturing supply chain while employing low-latency edge computing and analytics in a secure zero-trust environment. ACS RTDI minimizes the need for human intervention, streamlining overall data utilization across multiple insertions to boost quality, yield, and operational efficiencies. It includes Advantest's ACS Edge™ HPC server, which works in conjunction with its V93000 and other ATE systems to handle computationally intensive workloads adjacent to the tester's host controller.

In this configuration, the ACS Edge provides low, consistent, and predictable latency compared with a datacenter-hosted alternative. It supports a user execution environment independent of the tester host controller to ease development and deployment. It also provides a reliable and secure real-time data infrastructure that integrates all data sources across the entire IC manufacturing supply chain, applying analytics models that enable real-time decision-making during production test.

For its part, the V93000 brings flexibility and performance to the test of HI and other devices. Its wide range of available instruments make it suitable for the test of a variety of AI, ML, HI, and other chiplet-based devices across a range of test insertions. With instrument cards such as the PS5000, XPS256, and Link Scale, the V93000 enables customers to shift left—for example, using Link Scale to run mission-mode-like tests at WS.[5]

COT reduction through ML

Initial work by Advantest and our customers has involved COT reduction through the use of ML to identify time-consuming portions of the test flow that can be shortened or eliminated through correlation with previous insertion data. Work to date confirms that adopting ML at the edge provides significant ROI. Additional work over the next one to three years will connect more data across the value chain to develop advanced models that can

The adoption of edge compute for ML test applications is well underway. Advantest's recently launched ACS Real-Time Data Infrastructure (ACS RTDI™) platform accelerates data analytics and AI/ML decision-making within a single integrated platform. It collects, analyzes, stores, and monitors semiconductor test data as well as data sources across the IC manufacturing supply chain while employing low-latency edge computing and analytics in a secure zero-trust environment.

identify more complex correlations and accurately predict yield, quality, and reliability. Yet another initiative, achievable within five years, is bidirectional optimization across the semiconductor value chain with feed-forward and feed-backward data flow to optimize front-end and back-end processes, with edge-compute resources enabling real-time decision-making and optimization.

Probe cleaning and yield enhancement

In addition to ACS Edge and the V93000, Advantest also offers products that leverage AI and ML to enhance the test of AI-enabling devices. For example, Advantest ACS Adaptive Probe Cleaning (ACS APC) applies AI/ML to schedule probe needle cleaning on an as needed basis. In addition, ACS Engineering AI Studio for Yield Improvement (ACS EASY) is an AI-powered, user-friendly software solution that automates the analysis of yield issues, reducing engineers' workloads and speeding correction turnaround time.

Conclusion

Chiplets and HI in 2.5D/3D packages are bringing new challenges in achieving yield and reliability targets at acceptable COT levels for the emerging class of AI processors and HPC devices. With rules-based point solutions inadequate for the task, AI and ML techniques have emerged to process test data as well as data from the entire semiconductor value chain to ensure success. Edge computing is a key driver for processing data from chip sensor IP, equipment sensors, parametric test results, and upstream data sources to fully analyze and appropriately react to results. Advantest's customers have achieved multiple successes by embracing ML and edge computing to address multifaceted challenges with a comprehensive and cohesive set of solutions.

Additional progress will require an "all in" mentality across the semiconductor supply chain to leverage data sharing to enable even more complex models and accelerate the financial benefits that will accrue to all industry participants.

FURTHER READING / REFERENCE

- 1. Hioki, Shinji, "Data Analytics for the Chiplet Era," GO SEMI & BEYOND, November 2023.
- 2. Shimozato, Kyohei, et al., "Adaptive Outlier Detection for Power MOSFETs Based on Gaussian Process Regression," IEEE Applied Power Electronics Conference, March 2022.
- 3. Mier, Sergio, "ML in Semiconductor Test (A 'Balanced Approach')," Road to Chiplets: Data & Test, MEPTEC, November 2021.
- 4. Butler, Ken, and Cortez, Guy, "Deploying Cutting-Edge Adaptive Test Analytics Apps: Innovation Based on a Closed-Loop Real-Time Edge Analytics and Control Process Flow into the Test Cell," GO SEMI & BEYOND, November 2023.
- 5. Armstrong, Dave, "Device Validation: The Ultimate Test Frontier," GO SEMI & BEYOND, March 2023.

How to get two-digit savings on the cost of semiconductor test, with device-oriented test architecture

The adoption of testing solutions that help increase throughput and lower the cost of testing is a key factor to keep competitiveness up in the semiconductor industry. While, for many years, the way to reduce the cost of test was exclusively based on increasing the multi-site factor, a new test architecture approach demonstrates to boost test performances and efficiency.

There is no argument that testing represents a large share of the overall manufacturing costs for semiconductor devices. This cost percentage is experiencing a rising trend, due to the reliability concerns in mission-critical applications like automotive and security, as well as the longer lifetime expectations for many devices.

For decades, all the efforts to reduce the cost of test have focused on increasing the multi-site capabilities of testing equipment, called to perform multiple tests on higher numbers of devices at the same time. That, in turn, paved the way for highly sophisticated and expensive test equipment, housing all the specialized instrumentation to address the requirements of the whole variety of increasingly complex technologies. The company SPEA has been on the track for more than 45 years in developing cutting-edge test solutions, being a partner to semiconductor manufacturers and continuously innovating to meet the needs of this fast-moving industry.

The company has recently inaugurated a ground-breaking approach in the test architecture design, introducing an entirely new test platform called DOT (Device-Oriented Tester).

The innovative architecture DOT testers are based on has proven on the field to bring savings on the cost of test per device that often exceed 60% (going up to 95% savings in the best cases), on several challenging applications such as BMS, SerDes, OpAmp, PMIC, MEMS and sensors, Microcontrollers, Converters.

All the test resources are enclosed in a modular and composable board

The innovation at the basis of DOT architecture resides in the tester instrumentation, which is designed around the device requirements, instead of being based on the specific function to perform. In other words, instead of having a big-iron cabinet containing specialized instruments for every specific function (analog instruments, digital instruments, power instruments, digitizers, waveform generators, and so on), a device-oriented architecture allows for a tester configuration based on modular boards that incorporate, on a single instrument, all the analog, digital and signal-processing resources needed to test the devices.

The number of channels per board is the highest on the ATE market, and it contributes to lower the cost of test, offering more channels available per chip. This high-density board can be replicated inside the machine to reach the required multi-site capability, without losing on multi-site efficiency: as every board incorporates its own CPU and memory resources, the tester's memory size and the strategy to access it efficiently do not represent a limitation. The advantages of this type of test solution are tangible at different levels.

➤ Left: The innovative DOT test platform addresses the test requirements of applications like BMS, SerDes, OpAmp, PMIC, MEMS and sensors, Microcontrollers.

Savings over industrial costs
First of all, the overall tester cost and footprint are greatly reduced, as the total number of boards required is significantly lower than that of a traditional tester. The entire tester can be docked on a wafer prober, while also the power consumption is minimized.

Simple tester use, programming, maintenance
In second place, the possibility to populate a tester with all boards of the same type makes the tester easy to use and program, speeds up the learning curve for test engineers, simplifies maintenance operations and spare parts management.

Superior performance and test efficiency
The modular channel instrument is a high-density, multi-processor, multi-function board, composed of an on-board controller, four matrix cards and up to four different channel cards that can be selected to build the required performance mix, for a total of up to 256 channels in a single instrument slot. Multi-site management is transparent, and test efficiency is practically 100% when scaling from 1 to multiple sites.

Ensure best accuracy through time, with tester self-calibration

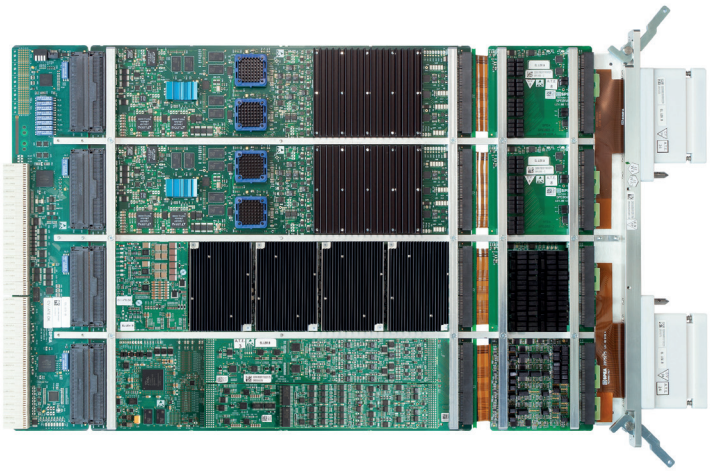
On every test floor, test equipment requires periodic calibration procedures to ensure that the test instrumentation meets the expected measurement accuracy, that is subject to deterioration through time.

Verifying the operation and functional stability of the semiconductor tester is mandatory to ensure accurate results: a tester that is out of calibration would contaminate the product yield with unreliable performance.

However, the tester calibration procedure normally implies a downtime scheduling, the stop of the tester operation for several hours, and the intervention of a qualified operator to undock the production load board, connect the tester to specific diagnostics and calibration hardware, and dock the tester again to the prober or handler at the end of the procedure.

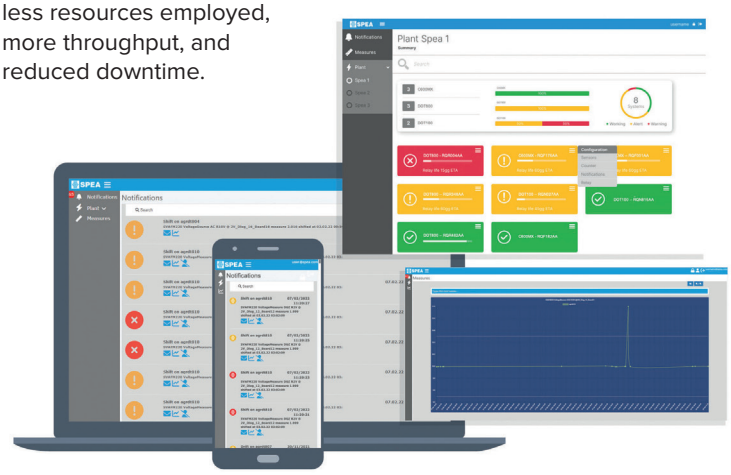
DOT test platform introduces predictive maintenance and self-calibration capabilities as strategic elements to minimize downtime and cost of ownership, while improving quality and control. The incorporation of artificial intelligence and smart sensors inside the test equipment enables the performance of automatic diagnostic procedures, self-corrections and comprehensive data analysis, without needing any scheduling, operator's intervention or specific hardware.

The tester is capable of autonomously verifying the instruments' health status, assessing the instrument measurement accuracy and detecting any trend towards out-of-specs performance, then independently adjusting the involved parameters. When a deviation from the standard occurs, the instrument calibration procedure is automatically launched and the accuracy is restored. In addition to that, key indicators - like the number of relay commutations - give information about the instrument wear, allowing for the schedule of predictive maintenance interventions. This prevents unexpected stops of the tester operations, and permits to reduce the number of interventions required. Dedicated sensors constantly monitor the operating temperature and humidity, to detect any dangerous condition. The data are stored both locally, on the instrument memory, and made available for further analysis and correlation to any anomaly in the machine performance.



➤ The channel instrument of DOT testers is a modular board, incorporating all the analog, digital, signal processing resources to test the devices.

Thanks to these features, testing equipment can become an interactive part of the production floor, in an Industry 4.0 perspective. The adoption of this solution on a fab-wide scale helps semiconductor manufacturers save on scheduled and unscheduled downtime, keeping the process quality under absolute control. Testing semiconductor devices more intelligently can bring better product quality, less resources employed, more throughput, and reduced downtime.



➤ The tester is able to self-assess its measurement accuracy, recalibrating the instrumentation when needed. Key parameters about the instrument wear status, and the operating conditions, are constantly monitored, notified, and analyzed, allowing for an effective predictive maintenance schedule.

To promote your products and services contact: **Shehzad Munshi**
T: +44 (0)1923 690 215 **E:** shehzad.munshi@angelbc.com

CLEAN ROOMS

GAS DELIVERY SYSTEMS

INSPECTION SYSTEMS

VACUUM EQUIPMENT

ETCH

GAS PURIFIERS

PLASMA TECHNOLOGY

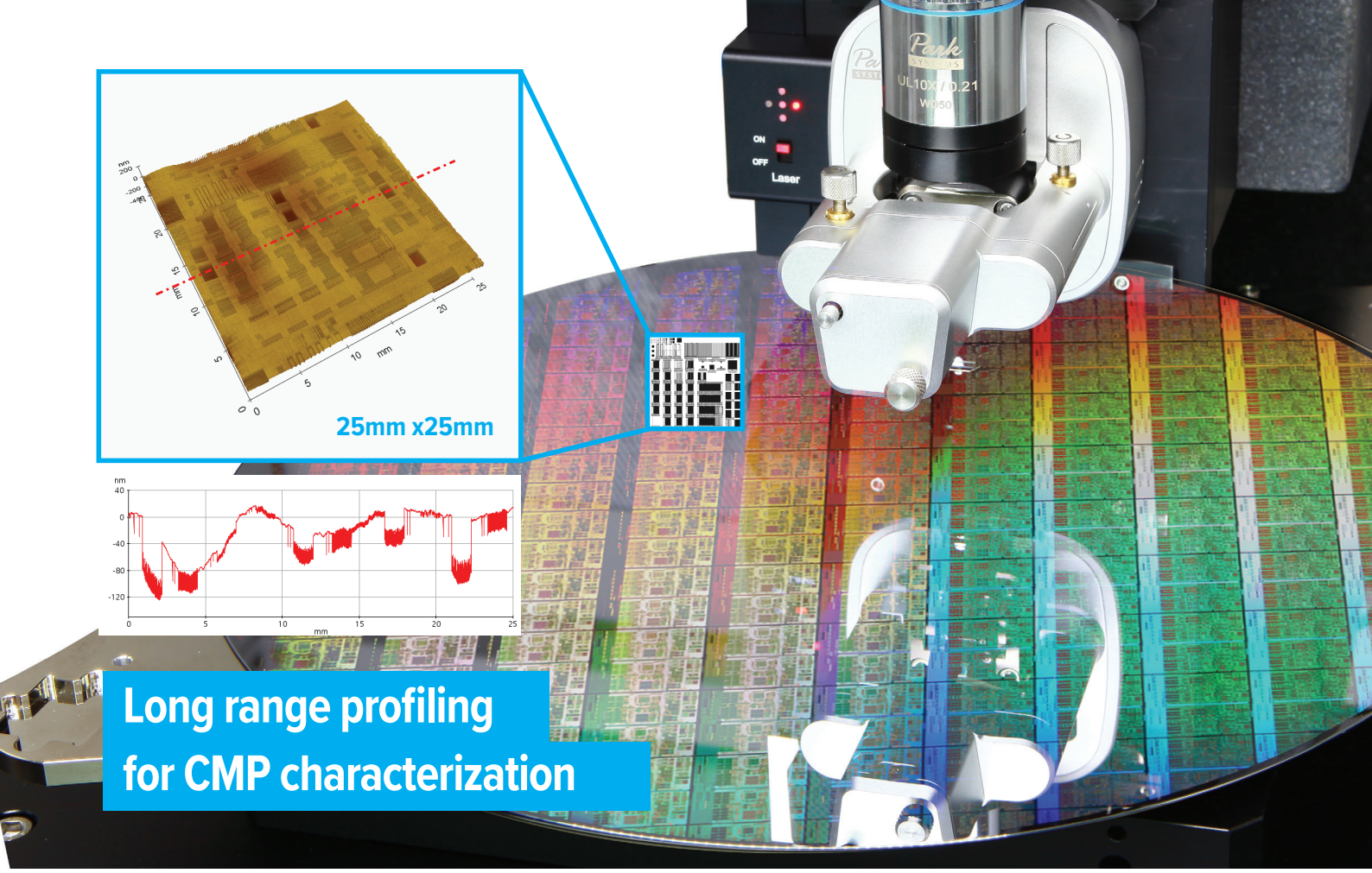
COMPONENTS

FLOW MEASUREMENT

HYDROGEN GENERATORS

TEST

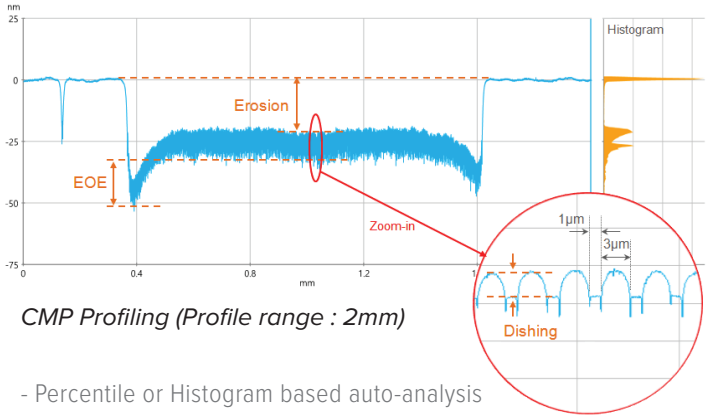
WAFER SERVICES



Park NX-Wafer

Low Noise, High Throughput Atomic Force Microscope with Automatic Defect Review & Atomic Force Profiler

- Low noise atomic force profiler for more accurate CMP profile measurements
- Sub-angstrom surface roughness measurements
- Fully automated AFM solution for defect imaging and analysis
- Capable of scanning 300 mm wafers
- Can improve defect review productivity by up to 1000%



CMP Profiling (Profile range : 2mm)

- Percentile or Histogram based auto-analysis
- Erosion, EOE (Edge-Over-Erosion) and Dishing

Please visit parksystems.com/nx-wafer to learn more about Park NX-Wafer or email inquiry@parksystems.com.

Park SYSTEMS