

Semiconductor Industry Outlook and New Technology Frontiers

Yuh-Jier Mii

Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan. Contact email: yjmii@tsmc.com

Abstract—The semiconductor industry is a dynamic landscape of innovation, where new materials, advanced processing techniques, and cutting-edge design converge to shape the future of technology. Powered by the principle of technology scaling, this field continues to push boundaries, enabling transformative applications in artificial intelligence (AI), high-performance computing (HPC), 5G/6G, autonomous driving, internet of things (IoT), and more. As we progress through time, scaling evolves, unlocking new levels of chip efficiency and performance. The horizon shines bright with the promise of breakthroughs in extreme ultraviolet (EUV) lithography, new device architectures like stacked complementary field-effect transistors (CFETs), novel low-dimensional channel materials, and the strategic synergy of design-technology co-optimization (DTCO), paving the way for exciting new technology eras. Additionally, advanced packaging techniques enhance system-level performance, blending computational power to surpass current limitations. The growth in specialty technology segments such as RF, non-volatile memory, power management, CMOS image sensors (CIS), and silicon photonics expands the range of innovative devices. This keynote paper will explore the latest advancements and emerging trends in the semiconductor industry, offering insights into how these cutting-edge frontiers will drive smart technology integration and create a brighter future for society.

I. SEMICONDUCTOR INDUSTRY & MARKET OUTLOOK

Looking at the growth history of the semiconductor industry, the PC era propelled revenue past the 100-billion-dollar mark in the 1990s. The internet era, which began in the mid-1990s, further expanded the industry, reaching 250 billion dollars by the first decade of the 21st century. The smartphone era, ignited by the launch of the first iPhone in 2007, drove industry growth to nearly half a trillion dollars over the following decade. Cloud computing then emerged in the wake of the smartphone revolution, pushing the industry to new heights. Finally, as shown in Figure 1, AI is poised to play a

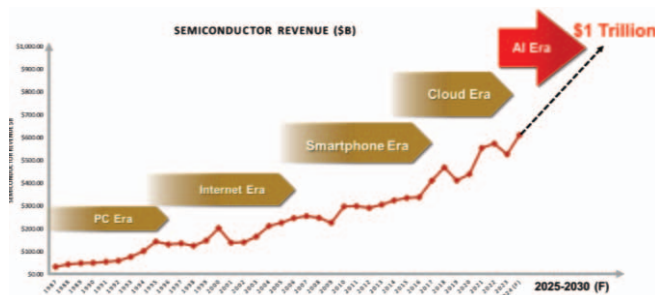


Fig. 1. Semiconductor industry revenue growth trajectory.

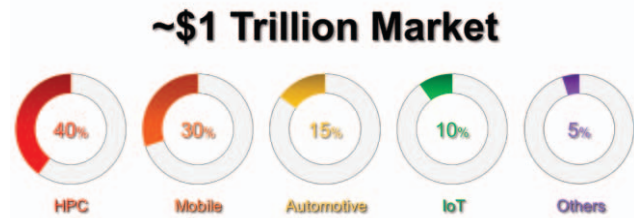


Fig. 2. 2030 semiconductor market by platform.

central role in the industry's pursuit of one trillion dollars in revenue by 2030.

Driven by multi-year megatrends in 5G-, AI-, and HPC-related applications, the structural, long-term growth trajectory of chip demand remains strong, even amidst cyclical macroeconomic fluctuations. By the end of the decade, as depicted in Figure 2, the worldwide semiconductor revenue will approach one trillion dollars, with HPC contributing 40%, mobile 30%, automotive 15%, and IoT 10%. [1]

AI is set to infiltrate all semiconductor products in our daily lives, enabling new features and, in many cases, transforming the user experience with unprecedented intelligence and productivity. As illustrated in Figure 3, AI will begin its rapid growth in data centers and gradually be integrated into smartphones, PCs/tablets, MR/XR gadgets, and IoT devices. Autonomous driving would not be possible without AI, and human-like robots represent the next frontier of AI applications that may one day revolutionize the world as we know it today.

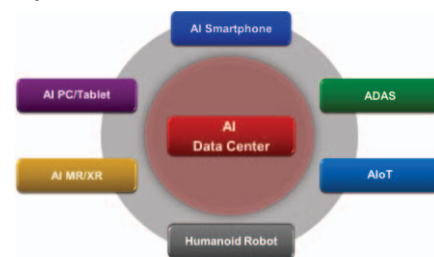


Fig. 3. AI enriches all products and applications.

AI is poised to be the next major growth area, and it is undoubtedly the most crucial one. To accommodate the ever-increasing complexity of AI models, as depicted in Figure 4, the demand for training compute power is rapidly expanding [2]. In recent years, the growth rate of training compute power has increased even further, spurred by generative AI and large language models. Taking cues from the data center revenue growth of NVIDIA over the past five years, generative AI has been—and will continue to be—a strong driver of industry growth.

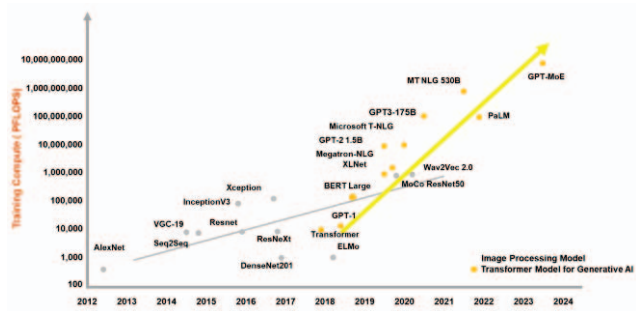


Fig. 4. AI model training computational requirement [2].

While AI technologies have made significant progress in recent years, the substantial energy required to train and operate these models poses limitations to their widespread adoption. Today's supercomputers or large computing PODs already consume megawatts or even tens of megawatts of power. At this rate, AI computing PODs will require gigawatts of power within a few years, as projected in Figure 5. Addressing these challenges requires innovation at all levels, from adopting leading-edge technologies for improved efficiencies and developing innovative architectures to utilizing renewable energy sources for powering these models. Therefore, energy-efficient computing is essential to expand AI applications to new levels.

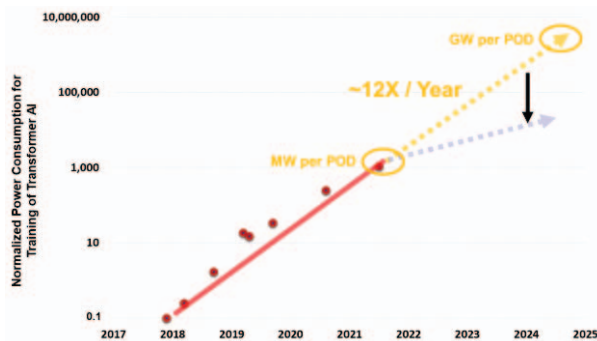


Fig. 5. Energy-efficient compute required for AI deployment.

II. ADVANCED LOGIC TECHNOLOGIES

Logic process technology has evolved significantly over the past decades. As shown in Figure 6, the advancements have primarily focused on geometry shrinkage; however, this alone is not sufficient and requires advancements in both device architecture and lithography. Device architecture transitioned from planar to field-effect-transistor (FinFET) at the 22/16nm node [3-4], substantially improving transistor electrostatics. Today, the industry continues to scale transistor dimensions by transitioning to nanosheet field-effect-transistor (NSFET) devices at the 3/2nm node [5-6]. Concurrently, lithography has advanced from immersion techniques to EUV lithography to maintain pitch scaling [7]. To maximize the benefits of advancements in device architecture and lithography, DTCO is crucial. DTCO not only propels logic technology performance, power, and area (PPA) but is also being extended to optimize performance and power at the system level.

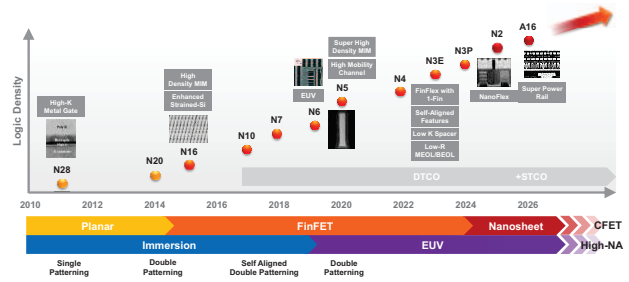


Fig. 6. Evolution of logic process technology over the past decade and future outlook.

To further optimize PPA at the system level, each chiplet's constituent technology can be tailored to best address its respective workload. In today's monolithic SoC designs, technology can be fine-tuned to better serve diverse applications beyond the mobile baseline, such as HPC, AI, and ultra-low power usage. However, the range of speed versus power efficiency trade-offs is limited. In contrast, a system composed of chiplets offers the flexibility to modify the process away from the baseline to best suit the needs of each partition. For example, one can optimize for maximum frequency in HPC, performance per watt at the same speed for AI, and the best performance per watt for ultra-low power applications. This domain-specific technology optimization enables the creation of the most efficient compute systems.

Figure 7 illustrates the cross-section of a state-of-the-art 2nm technology featuring nanosheet devices and backside power delivery [6]. On the front side of the wafer, nanosheet devices offer excellent power efficiency from structural and DTCO innovations. Additionally, BEOL process and material innovations contribute up to a 10% reduction in RC delay and offer design rules that enable a 3% to 4% gain in logic density. On the backside of the wafer, direct contact with the device preserves gate density and maintains flexibility in device width. Metallization on the backside enhances power delivery, reduces IR drop, and improves chip density and performance by dedicating front-side routing exclusively to signal paths.

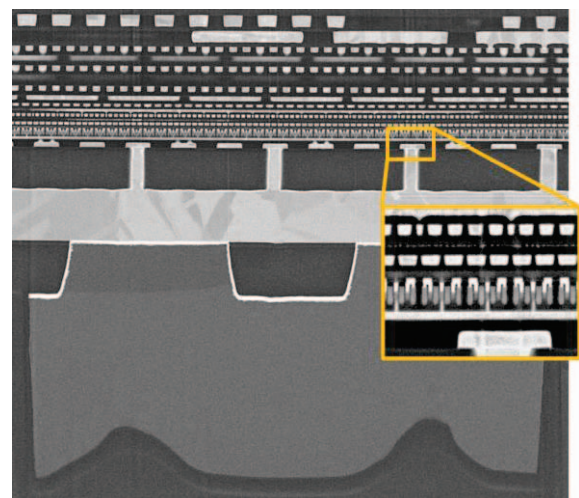


Fig. 7. State-of-the-art nanosheet-based 2nm logic technology.

III. LOGIC TECHNOLOGY FRONTIERS

The scaling of CMOS logic transistors will continue to be the backbone of advances in future semiconductor computing technology. Following the introduction of FinFET technologies and nanosheet architecture, as illustrated in Figure 8, the complementary field-effect transistor (CFET) architecture has emerged as a leading contender for future logic scaling [8-11]. Although the vertically stacked nFET and pFET configuration in CFET is expected to increase process complexity and fabrication costs, it offers a significant density advantage—approximately 1.5 to 2 times higher—compared to conventional CMOS architecture with n/p FETs placed side by side at the same gate pitch.

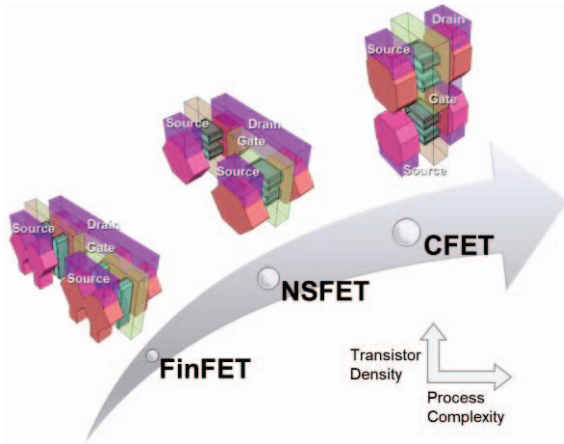


Fig. 8. Device architecture evolution from FinFETs, Nanosheet FETs (NSFETs), to vertically stacked CFET architecture.

Figure 9 showcases the most advanced, fully operational CFET inverter, featuring an industry-leading 48nm gate pitch. This inverter includes a backside contact as V_{DD} , a front-side source contact as V_{SS} , a common gate for input V_{in} , and a vertical metallized drain local interconnect for the common drain output V_{out} . It demonstrates robust voltage transfer characteristics (VTC) in response to V_{DD} up to 1.2 V. This marks a pioneering breakthrough in monolithic CFET technology, laying the groundwork for a process architecture poised to drive the scaling of future logic technologies.

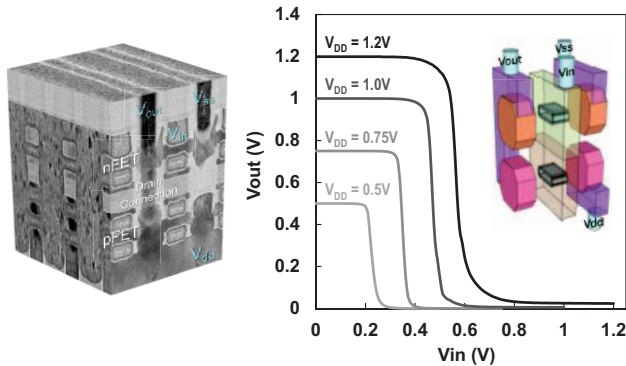


Fig. 9. (Left) Cross-sectional TEM visualization of a monolithic CFET inverter with an industry-leading 48nm gate pitch. (Right) Voltage transfer characteristic (VTC) of monolithic CFET inverter measured up to $V_{DD} = 1.2$ V [12].

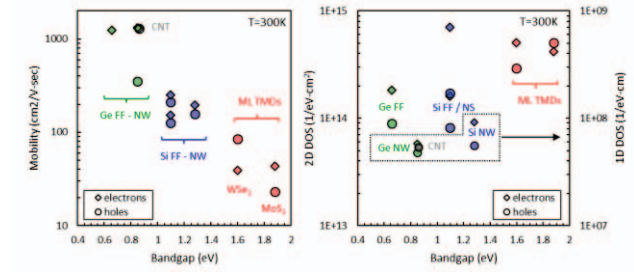


Fig. 10. Comparison of semiconductor properties among potential channel materials for advanced logic scaling and advancements.

Beyond CFET, the ongoing quest for higher performance and more energy-efficient logic technologies necessitates an accelerated search for channel materials that go beyond those based on silicon. Figure 10 provides a theoretical overview of semiconductor properties from some potential transistor channel candidates beyond Si [13]. Germanium, a high-mobility bulk semiconductor [14-15], has long been considered a potential candidate for low supply voltage applications. Additionally, low-dimensional materials such as carbon nanotubes (CNTs) and transition metal dichalcogenides (TMDs) have garnered significant interest due to both their physical and electronic properties.

The ability to precisely control the diameter of CNTs [16] and the thickness of TMDs [17] opens up opportunities for innovative advancements in device scaling. While challenges exist, recent advancements continues to fuel the research interest. For CNTs, achieving NMOS performance on par with PMOS has been demonstrated through specific doping techniques, enhancing their high current density capabilities [18]. For TMDs, the most notable advantages are their high density of states and the ability to scale gate lengths below 10 nm. Although very short-dimension devices have been reported, full device scaling requires further reduction in EOT and gate dielectric thickness, better interface defect control, and CMOS integration with low-resistance contact solutions [19-21]. While alternative channel materials have interesting properties and have shown performance improvement over time, significant progress is still needed before they can surpass silicon for industry consideration.

Interconnect innovation is another critical field for technology advancement. Figure 11 highlights several new developments the industry has been working on. As technology scales, middle-of-the-line (MOL) resistance becomes critical for system performance. By using new low-resistance materials

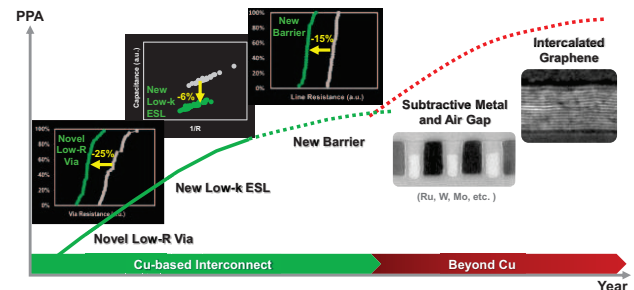


Fig. 11. Interconnect material and architecture Outlook.

and advanced interface engineering, a 40% reduction in MOL resistance has been achieved. Additionally, beyond traditional Cu damascene interconnects, a novel metal reactive ion etching (RIE) process with an air-gap approach is being developed to reduce capacitance and enhance performance, potentially lowering line capacitance by 20% to 30%. Furthermore, a new 2D material is being explored as a superior alternative to Cu for interconnects. This material shows lower thin film resistivity than Cu at reduced thicknesses, helping to mitigate line resistance increases in scaled geometries and enhance overall performance.

IV. SYSTEM INTEGRATION TECHNOLOGIES

In addition to pushing 2D technology scaling to enable better transistors and higher packing density in monolithically integrated SoCs, innovations beyond the chip level have also become a must to extend integration into heterogeneous domain. To unlock the power of heterogeneous integration and enhance system-level performance by more than tenfold, 3D stacking and 2.5D advanced packaging technologies have been introduced in tandem, as illustrated in Figure 12.

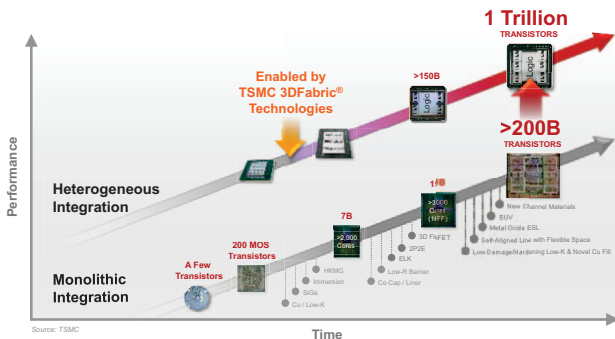


Fig. 12. 2.5D and 3D integration at system level.

To achieve future system scaling and performance, improving the 3D chip-to-chip interconnect density is critical. Over the past decades, interconnect density among chips within a package has advanced rapidly. Advanced silicon stacking and packaging technologies, including SoIC, InFO, and CoWoS®, continue to aggressively scale down the chip-to-chip interconnect pitch, offering the potential to improve 3D interconnect density by another six orders of magnitude, as depicted in Figure 13. These advanced integration capabilities enhance data transfer rates, reduce latency, optimize power

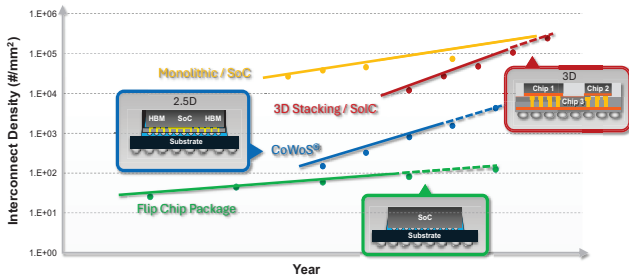


Fig. 13. Trends in interconnect density scaling enabled by advanced packaging technologies.

consumption, and elevate the overall performance of computing systems.

The insatiable demand for computing power has driven rapid growth in system-level heterogeneous integration, as illustrated in Figure 14. The Chip-on-Wafer-on-Substrate (CoWoS) technology has expanded from 3.3 reticles in 2023 to 5.5 reticles soon, with projections exceeding 8 reticles in the next couple of years. Meanwhile, the System-on-Wafer (SoW) technology leverages CoWoS to elevate compute power to new heights. By integrating high-bandwidth memory (HBM) and vertically stacked compute chiplets, SoW is expected to deliver unprecedented compute performance upon its introduction in the coming years.

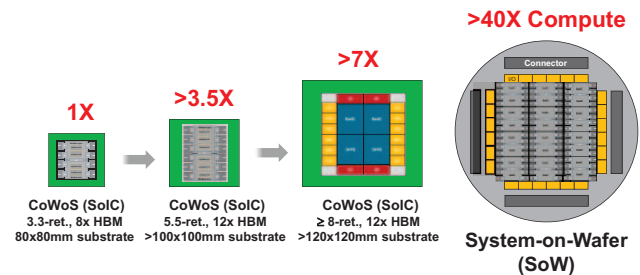


Fig. 14. System-level scaling demands rapid growth in heterogeneous integration.

Optical transceivers are crucial for future AI systems, enabling high-speed, low-energy, and reliable data transmission between chips. Our Compact Universal Photonic Engine (COUPE) technology uses an innovative SoIC-X process to seamlessly stack electrical and photonic dies, significantly reducing power consumption and latency. Integrating optical engines at the board, package, and interposer levels offers substantial benefits in form factor and power reduction, as shown in Figure 15.

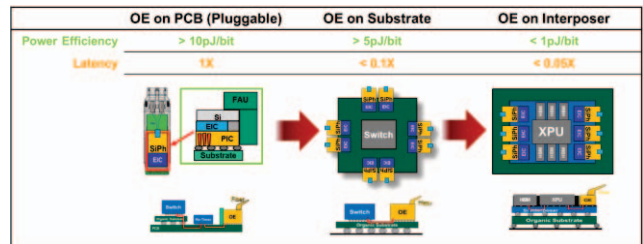


Fig. 15. Silicon photonics revamps data transmission.

Customers encounter significantly more complexities in 3DIC design compared to SoC design. Figure 16 illustrates the necessary collaboration among partners from various domains to address these complexities, which include 3DIC-specific EDA tools, die-to-die interface IPs and connections, high-bandwidth memory, substrates, and testing for the integration of multiple chips or chiplets. To streamline 3DIC design across various packaging structures and configurations, seamless interaction between EDA tools at different design stages is crucial. This requires an industry-standard language that can support the entire design process, from architecture and prototyping to design implementation and signoff.

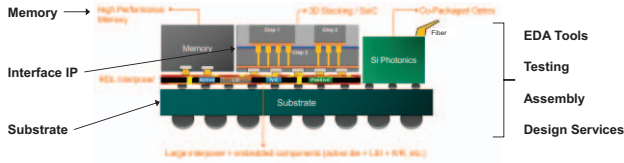


Fig. 16. Industry-wide collaboration is required for design and manufacturing of 3DIC.

Creating the AI systems of today and tomorrow demands a platform that offers comprehensive coverage of technologies required for the heterogeneous integration. This includes advanced logic technologies that go into chiplets, SoIC technology that stacks the chiplets, RDL interposers that incorporate embedded components and bridging interconnects connecting to the compute die stack and high-performance memory, and silicon photonics engines mounted on the same substrate or carried by the RDL interposer to provide sufficient I/O bandwidth to meet compute requirement.

V. SPECIAL TECHNOLOGIES

The expansion of specialty technology segments like RF, non-volatile memory, power management, CMOS image sensors, and Si photonics is broadening the spectrum of innovative devices. In this section, we will delve into the latest advancements and emerging trends in the semiconductor industry, providing insights into how these cutting-edge developments will drive the integration of smart technologies.

A. RF Technology

Given the growing importance of Wi-Fi wireless connectivity in our daily lives, as highlighted by Cisco's analysis showing 51% edge IP traffic coverage [22], the Wi-Fi standard continues to evolve to meet increasing data demands and support emerging applications. Wi-Fi 6, for instance, offers a theoretical peak data throughput of 2.4 Gbps using a popular 2x2 multi-input multi-output (MIMO) configuration in mobile devices [23]. Its successor, Wi-Fi 7, enhances this throughput by 2.4 times, achieving 5.8 Gbps [24]. Looking ahead, Wi-Fi 8 is expected to further boost peak throughput to 10 Gbps, although the standard is still under discussion.

This escalating data rate opens up new applications and market opportunities for Wi-Fi, but it also brings about significant challenges. Achieving these higher throughputs necessitates additional features such as the 6 GHz band, wider channel bandwidth, more complex modulation, and multi-link operation, all of which contribute to increased chip area and

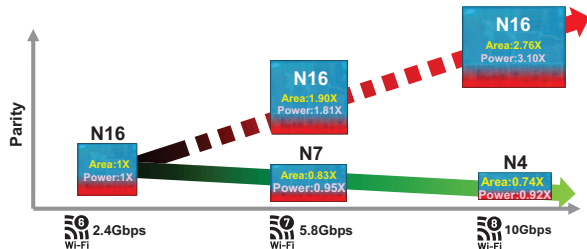


Fig. 17. Wi-Fi area and power parity trend with system evolutions and silicon technologies.

power consumption. As illustrated in the upper part of Figure 18, a Wi-Fi 7 RFSoc could result in a 90% increase in die area and an 81% increase in power consumption compared to its Wi-Fi 6 counterpart using the same N16 silicon technology. A Wi-Fi 8 RFSoc is projected to see even more significant increases, with die area and power consumption growing by 176% and 210%, respectively.

These increases in die area and power consumption can severely impact the user experience of battery-powered mobile devices, thereby undermining their competitiveness. To address these challenges, technology scaling becomes essential for solution providers to achieve PPA scaling benefits. As illustrated in the lower part of Figure 17, migrating the Wi-Fi 7 RFSoc from N16 to N7 and the Wi-Fi 8 RFSoc from N7 to N4 enables the achievement of smaller die sizes and lower power consumption compared to their Wi-Fi 6 counterparts, despite the superior performance of the newer standards.

B. eNVM technology

Conventional eFlash technology is expected to halt at the 28nm node due to the complexity of process integration. In contrast, non-volatile memory technologies such as MRAM [25] and RRAM [26] have successfully scaled down to 16nm and 12nm. These technologies are projected to further scale down to 5nm and 4nm, as illustrated in Figure 18.

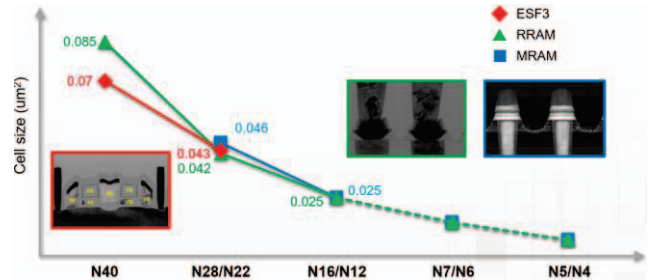


Fig. 18. eNVM cell size scaling will continue with RRAM and MRAM into FinFET Era.

RRAM is emerging as an ideal replacement for eFlash due to its reliability performance and significantly lower mask requirements—only two to three masks compared to the 10 to 18 masks needed for eFlash. Production of 40/28/22nm RRAM has been ongoing since 2021. We are also making significant progress on 12nm RRAM, which is slated for release in the second half of 2024 for consumer applications, offering read access times of less than 10 nanoseconds for high-performance uses. Early silicon assessments indicate that RRAM can scale down to 6nm. For MRAM, it offers superior performance in terms of reliability and write throughput, making it suitable for industrial and automotive applications. The 22nm MRAM technologies are now ready to support customers, with 22nm MRAM production having commenced in 2020. The 16nm MRAM macros offer read speeds of less than 10 nanoseconds, along with superior reliability characterized by over 1 million cycles of endurance and 20 years of data retention at 150°C. These features meet the most demanding application requirements.

C. CIS technology

Advancements in image sensor technology have significantly transformed how people communicate and share information. The integration of digital cameras into new devices has revolutionized user interaction with AI-era products. However, the continuous evolution of image sensor technologies is essential for realizing these innovative products and applications. Recently, there have been two critical breakthroughs in image sensor technology that are poised to drive new product developments in the coming years.

The first breakthrough is the three-wafer stacked technology, as shown in Figure 19, which improves upon the two-wafer stacked CIS technology (one pixel wafer and one ISP wafer). Adding a third wafer allows for design optimization, such as moving pixel circuits to the middle wafer to increase the effective photodiode area, enhancing optical performance. It also enables the addition of circuits to support new CIS features, like a three-wafer stacked backside-illuminated structure, which significantly enhances the voltage-domain global shutter sensor's footprint. This improvement is achieved by better integrating pixel, storage, readout, and processing circuits. This compact footprint CIS is crucial for augmented reality (AR) and virtual reality (VR) applications. Another example of three-wafer stacking is the fusion of an event-based vision sensor (EVS) with a conventional RGB sensor. EVS offers advantages such as low latency and low power capability compared to traditional shutter- or frame-based sensors, making it ideal for applications like deblurring, ultra-low power operation, and high-speed tracking. By re-integrating the two-sensor solution with three-wafer architecture, a fusion sensor with two different types of pixels and two distinct readout circuits can be achieved. [27-28]

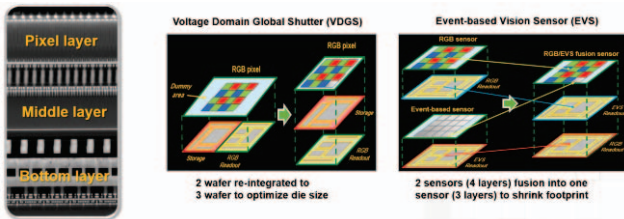


Fig. 19. Three-wafer stack technology, voltage domain global shutter (VDGS) and event-based vision sensor (EVS).

The second breakthrough is the development of high-density in-pixel capacitors. By integrating these capacitors with the LOFIC pixel design, the dynamic range of image sensors can be significantly enhanced, even with limited pixel sizes. These high-dynamic-range (HDR) image sensors are crucial for advanced driver assistance systems (ADAS), which will make cars safer. Additionally, this technology has the potential to improve smartphone camera performance in the near future.

D. Silicon Photonics

Our silicon photonics technology integrates both passive and active photonic devices into a single chip, as illustrated in Figure 20. This integration includes components such as grating couplers, modulators, waveguides, and Germanium (Ge) photodiodes, with the exception of the laser source, which can

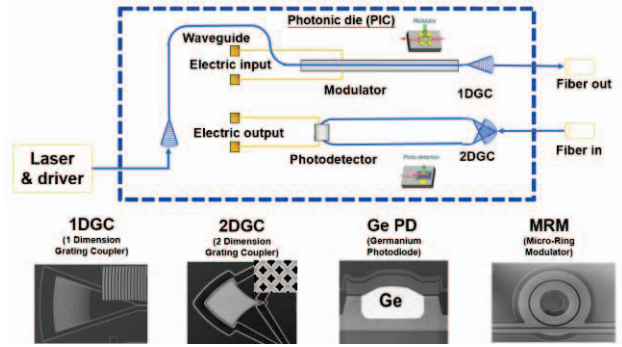


Fig. 20. Silicon photonics integrated chip and key components.

be coupled through fibers. Notably, TSMC's silicon photonics leverages advanced 12-inch process technologies, offering excellent process capability and controllability. These factors enable customers to design a modern single-chip optical engine (OE).

Furthermore, the highly integrated single-chip OE has a compact size, allowing for placement onto a substrate or even onto an interposer. With "OE on Substrate" or "OE on Interposer" configurations, it is possible to achieve the shortest path to connect the ASIC and OE, significantly reducing transmission power and latency, which are critical factors for AI networking. Due to its high integration, small size, and short links, our silicon photonics presents a viable option for AI networking applications.

VI. CONCLUSION

Semiconductor innovations, encompassing advancements in device technology, system-level scaling, and customer-specific design ecosystems, will remain pivotal in driving rapid technological progress in the era of AI. TSMC is actively exploring a new array of innovations for future generations of technology, system integration platforms, and design ecosystems. These efforts will be crucial in meeting the increasing societal demands for energy-efficient, data-intensive computing in the coming decades.

REFERENCES

- [1] K. Zhang, *ISSCC* 2024, pp. 10-15
- [2] R. Merritt, *NVIDIA Blogs*, March 2022, <https://blogs.nvidia.com/blog/what-is-a-transformer-model/>
- [3] C. Auth *et al.*, *VLSI* 2012, pp. 131-132
- [4] S.-Y. Wu *et al.*, *IEDM* 2013, pp. 9.1.1-9.1.4
- [5] J. Jeong *et al.*, *VLSI* 2023, pp. T1.2.1-T1.2.2
- [6] G. Yeap, *IEDM* 2024, pp. 2.1.1-2.1.4
- [7] Y.-J. Mii, *VLSI* 2022, pp. T276-T281
- [8] S. Liao *et al.*, *IEDM* 2023, pp. 29.6.1-29.6.4
- [9] M. Radosavljević *et al.*, *IEDM* 2023, pp. 29.2.1-29.2.4
- [10] J. Park *et al.*, *VLSI* 2024, pp. T1.2.1-T1.2.2
- [11] S. Demuyneck *et al.*, *VLSI* 2024, pp. T5.2.1-T5.2.2
- [12] S. Liao *et al.*, *IEDM* 2024, pp. 2.5.1-2.5.4
- [13] C. Diaz, *IEDM* 2024, pp. 28.2.1-28.2.4
- [14] M.J.H. van Dal *et al.*, *IEDM* 2018, pp. 21.1.1-21.1.4
- [15] C.-T. Tu *et al.*, *IEDM* 2023, pp. 29.5.1-29.5.4
- [16] N. Safron *et al.*, *VLSI* 2024, pp. T3.3.1-T3.3.2
- [17] Y. Y. Chung *et al.*, *IEDM* 2024, 12.5.1-12.5.4
- [18] S. Li *et al.*, *IEDM* 2024, 10.2.1-10.2.4
- [19] W.-C. Wu *et al.*, *VLSI* 2024, pp. T1.4.1-T1.4.2
- [20] A. S. Chou *et al.*, *IEDM* 2024, 24.8.1-24.8.4
- [21] A. Azizi *et al.*, *IEDM* 2024, 24.4.1-24.4.4
- [22] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022
- [23] Broadcom's BCM56960 Series, <https://www.broadcom.com/products/wireless/wireless-lan/bluetooth/bcm4389>
- [24] Intel "What Is Wi-Fi 7?", September 2023, <https://www.intel.com/content/www/us/en/products/docs/wireless/wi-fi-7.html>
- [25] P.-H. Lee *et al.*, *ISSCC* 2023, pp. 494-496
- [26] A. Grossi *et al.*, *IMV* 2023, pp. 1-4
- [27] K. Zaitus *et al.*, *VLSI* 2022, pp. T1.3.1-T1.3.2
- [28] S. Fukuoka *et al.*, *SIST* 2023, doi: 10.2352/EI.2023.35.6.ISS-182
- [29] C. Loi *et al.*, *ISSCC* 2019, pp. 120-121
- [30] A. Talkhooncheh *et al.*, *ISSCC* 2022, pp. 284-285
- [31] M. Raj *et al.*, *ISSCC* 2023, pp. 204-205