State of the Semiconductor Industry

Trends and drivers shaping the semiconductor landscape

pwc

Table of Contents

Section 1

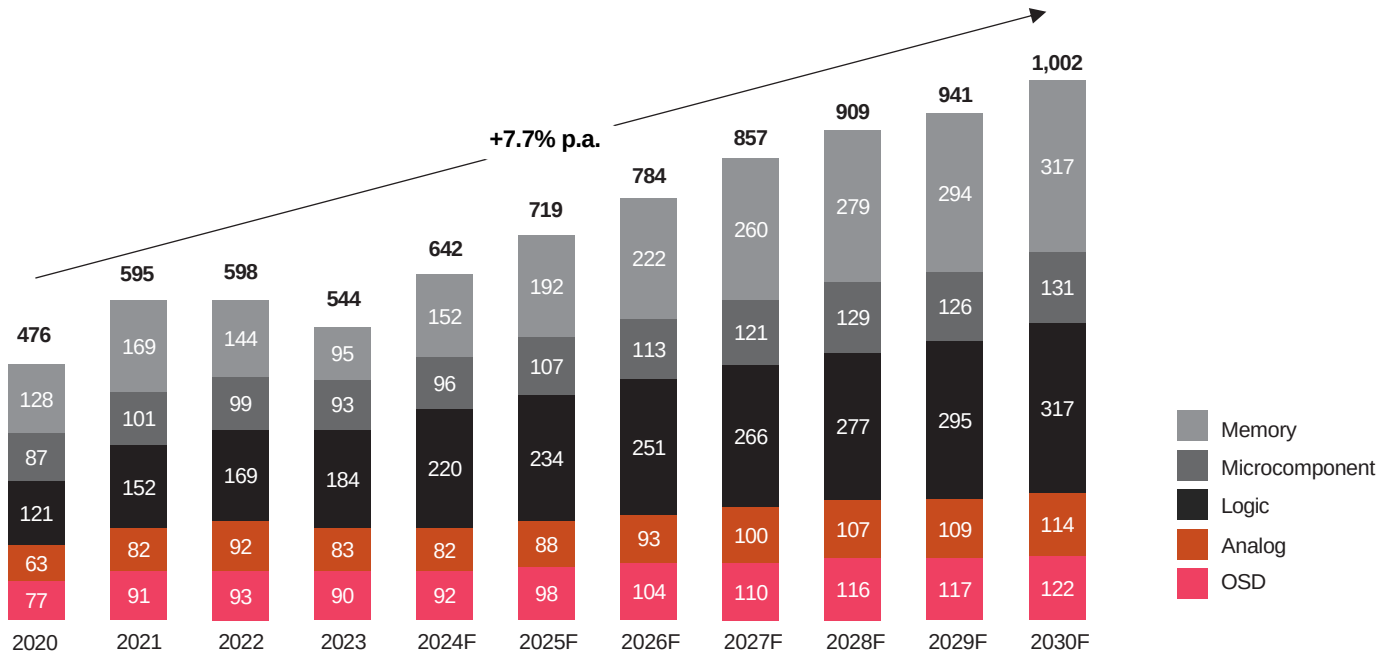
Raising stakes towards a trillion-dollar industry

Semiconductors have been the driving force behind technological evolution for more than 70 years, sparking the fundamental transformation of many industries throughout the world. From personal computing and smartphones, to data centers and cloud computing, semiconductor innovations have shaped the development of vital applications across the economy. As we look ahead, the trajectory of the semiconductor industry is set to be fueled by megatrends such as electrification, digitization, and the accelerating deployment of artificial intelligence (AI) and Internet of Things (IoT) technologies.

These trends place the industry in a position to achieve sustained long-term growth, with annual global revenue projected to reach $642 billion in 2024 and one trillion by the end of the decade (*see Exhibit 1*).

Exhibit 1

Global semiconductor market by component type, 2020–2030 ($bn)



Source: Omdia Q3 2024; OSD – Optoelectronic, sensor and discrete

The semiconductor industry has long operated as a global supply chain, benefiting from efficient manufacturing based on scale, technology, and worldwide distribution. However, over the last five years, the COVID-19 pandemic and escalating trade tensions have highlighted the need to secure supply chain sovereignty through greater investment in local production.

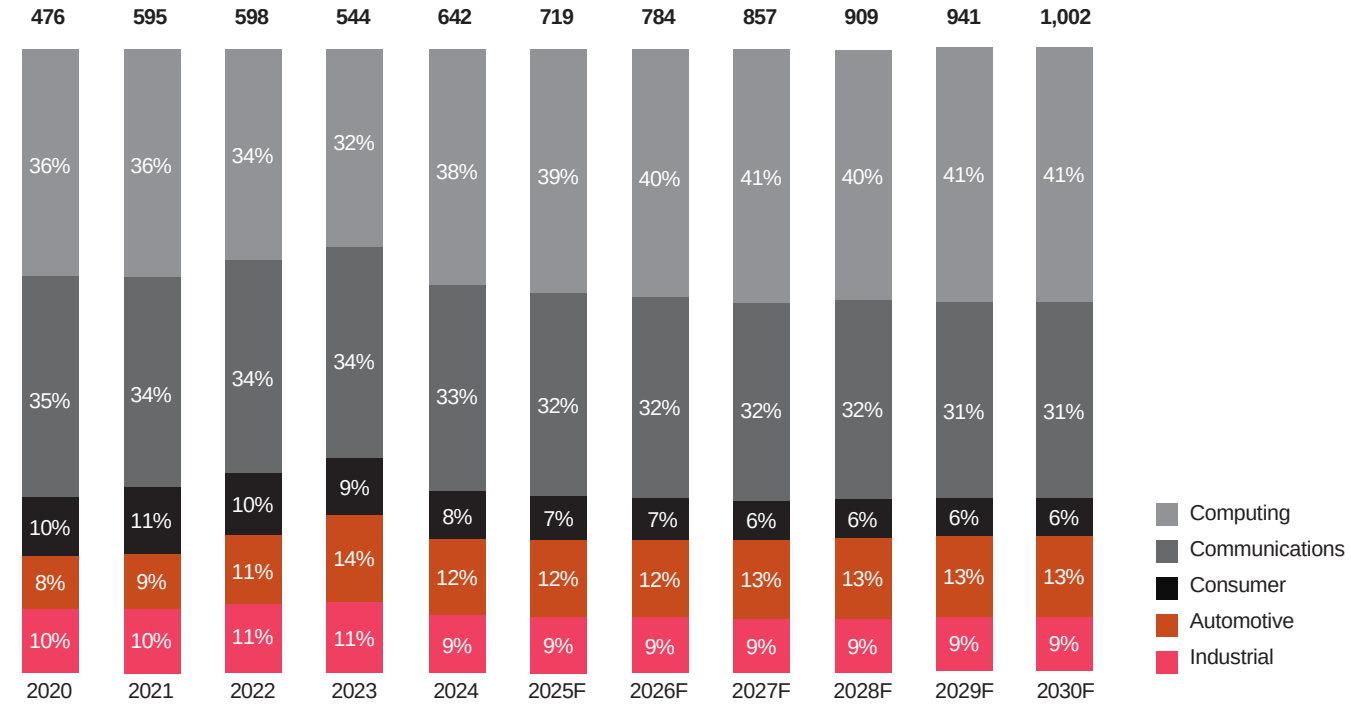
During the pandemic, semiconductor demand surged to record levels as companies equipped employees for remote work and as individuals invested in home electronics. At the same time, the increasing adoption of IoT technologies generated further demand, as industries modernized operations. As some sectors underestimated the scale of the demand, suppliers struggled to secure capacity. This led to a global chip shortage from the second half of 2020 through to the end of 2022. Meanwhile, over supply and less demand resulted in excess inventory, sending the industry into a downturn in 2023. The market is now stabilizing, with semiconductor revenue expected to rebound, surpassing the previous peak from 2022.

AI, IoT and Automotive as key growth drivers for the semiconductor industry

Of the seven component types that make up the semiconductor market (memory, logic, micro-component, analog, optoelectronic, sensor, and discrete), memory and logic products will continue to comprise the largest share of semiconductor revenue. Their dominance can be explained by the essential contribution they make to a wide range of applications, from computing and mobile devices to industrial and automotive products. As the world becomes increasingly data-driven, the demand for faster, more efficient memory solutions will rise, and will be bolstered further by the growing adoption of AI, IoT, and cloud computing across industries. Moreover, the continued proliferation of data centers around the world reinforces the importance of memory and logic in handling vast amounts of real-time data (*see Exhibit 2*).

Exhibit 2

Global semiconductor market by application share, 2020–2030 ($bn)



Source: Omdia Q3 2024

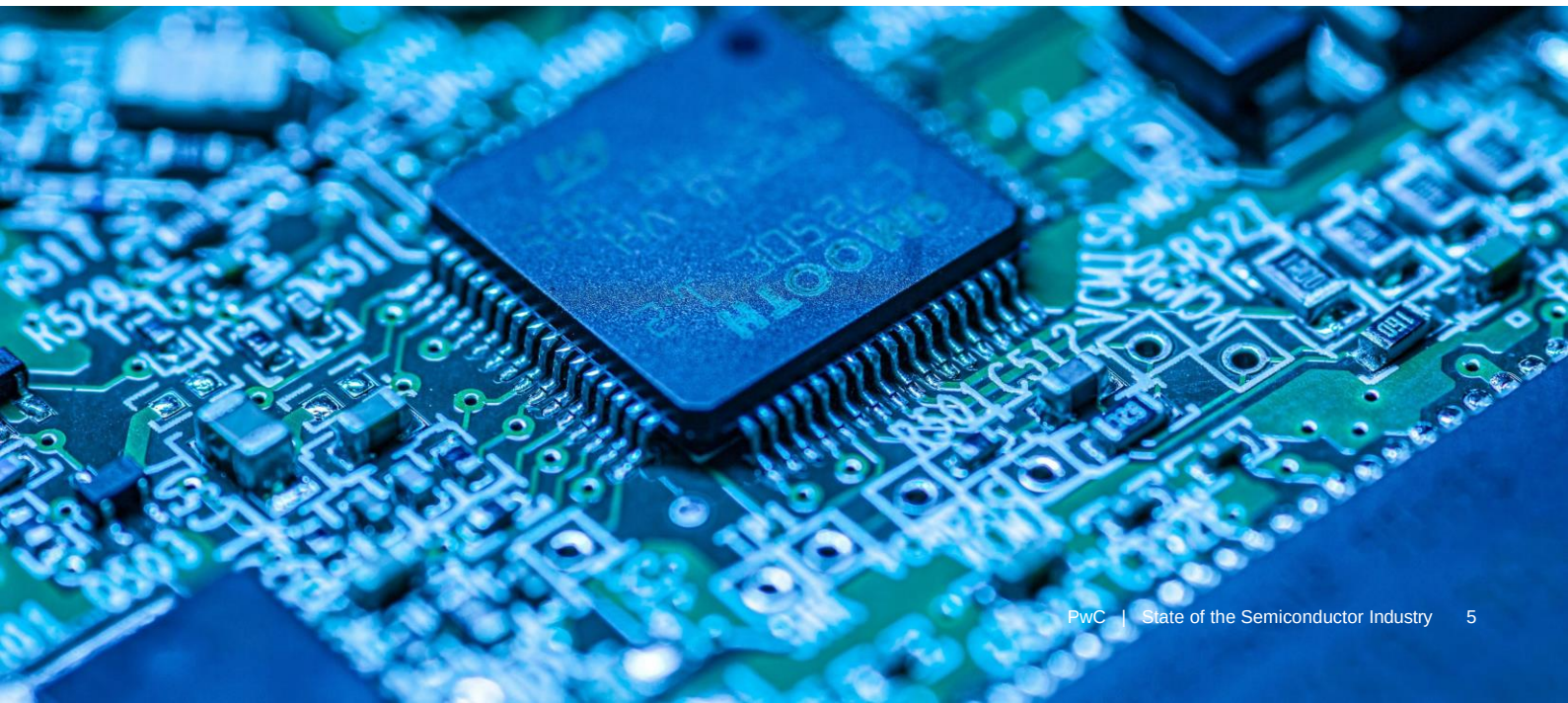
In terms of applications, the computing market is expected to overtake communications to become the largest segment from 2024 onward (*see Exhibit 2, previous page*), with a projected compound annual growth rate (CAGR) of 9% through to 2030. AI adoption, in particular, constitutes a major factor in this growth, as industries increasingly require advanced semiconductor solutions to manage intensive computational tasks. Technologies such as machine learning, neural networks, and data analytics will play a transformative role in shaping the future of the semiconductor industry, with the full impact of these trends still to unfold.

Moreover, the shift toward custom integrated circuits (ICs) represents a significant trend, as companies increasingly favor tailored solutions over standard off-the-shelf components. The reason for this shift is the need for specialized performance, energy efficiency, and enhanced security in applications ranging from data centers to consumer electronics. A notable example is Apple's development of its proprietary M chips, which offer superior integration and performance in comparison with generic processors.

The automotive sector is projected to remain the fastest-growing semiconductor market, with a forecasted CAGR of 10% from 2024 to 2030. A critical element of this projected growth is the ongoing electrification of vehicles, despite some short-term adjustments to the growth projections in the United States and European markets. According to PwC Autofacts, global battery electric vehicle (BEV) penetration within the light vehicle category was estimated to be 13.3% in the second quarter of 2024, and is expected to rise to 42.5% by 2030.[1] BEVs rely on power electronics and power management ICs to drive motors and manage battery systems. Compared to internal combustion engine vehicles (ICEs), BEVs contain more than double the semiconductor content because of their high-voltage systems.[2]

The shift toward software-defined vehicles (SDVs) is also reshaping the automotive sector. SDVs rely on advanced software functions that can be continuously updated, allowing for greater customization and flexibility, and faster innovation cycles. This decoupling of hardware from software leads to swifter advancements. Furthermore, trends such as autonomous driving and enhanced comfort features are boosting demand for high-performance semiconductors. Because of these changing requirements, semiconductor and component manufacturers will need to anticipate future vehicle architectures and stay ahead of product demands. The growing value of the semiconductor content per vehicle, which nearly doubled from $420 in 2019 to $800 in 2023, is expected to reach $1,350 by 2030, thereby increasing threefold over a decade.[2]

$1,350
semiconductor content per vehicle



Section 2

Memory enters a new frontier

Memory plays a pivotal role in the semiconductor market, serving as a fundamental component in data storage and processing across countless applications. As the demand for higher density, capacity, speed, and bandwidth grows, memory finds itself on a long-term trajectory of innovation. This innovation is shaped by the surge in data-intensive applications—such as AI, IoT, cloud computing, and advanced data analytics—which require substantial memory resources to operate efficiently. Over the last two decades, memory ICs have emerged as the fastest-growing segment among major semiconductor devices, with a CAGR of 8.6%. As the digital economy expands, the share of total semiconductor revenue attributed to memory is expected to increase from 18% in 2008 to an estimated 25% by 2024.[2] This rising demand for efficient data storage and processing underscores memory's role as a significant factor in future semiconductor industry growth.

> "This rising demand for efficient data storage and processing underscores memory's role as a significant factor in future semiconductor industry growth."
>
> **Glenn Burm**
> PwC Global Semiconductor Sector Leader

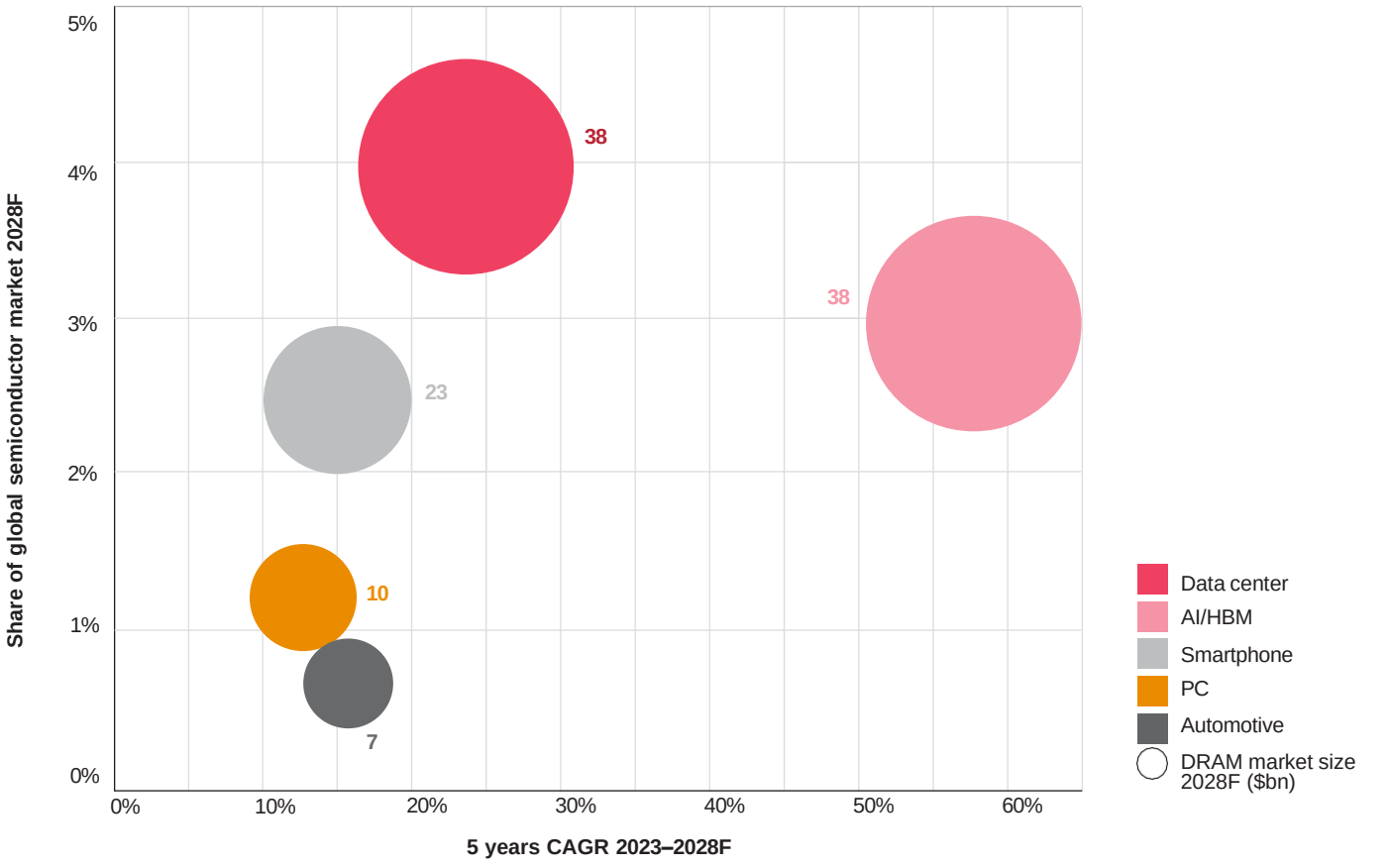
DRAM's evolution: Fueling AI and data-driven growth

Dynamic Random Access Memory (DRAM) is a type of memory used in computers and other electronic devices to store data that is being actively used or processed. It has long been a cornerstone of the semiconductor industry, accounting for an estimated 14% of total revenue in 2024. Over the years, DRAM has evolved significantly, propelled by technological advancements and the increasing demands of data-intensive applications. Initially used in personal computing and data centers, DRAM's role has expanded to serve emerging sectors such as AI, machine learning, and cloud computing, where the need for higher bandwidth, faster speeds, and greater capacity continues to grow.

High Bandwidth Memory (HBM) is projected to rival data center DRAM in market value by 2028. It is expected to account for 4.1% of the global semiconductor market by 2028, growing at a CAGR of 57.5% from 2023 to 2028. In comparison, data center DRAM is forecasted to hold the slightly higher market share of 4.2%, but with a more moderate CAGR of 22.3%. Meanwhile, smartphone DRAM, a more mature segment, will hold a 2.6% market share, with a CAGR of 15.3% during the same period. This highlights the explosive growth of AI-driven memory applications (*see Exhibit 3*).

Exhibit 3

DRAM market by application in share of global semiconductor market (%) and 5-year CAGR 2023–2028 (%)



Source: Omdia Q3 2023

HBM: Accelerating the next era of memory innovation

As AI and high-performance computing continue to develop, traditional DRAM solutions are being pushed to their limits. HBM has emerged as an important innovation that helps to meet these demands. Optimized for parallel computing and AI workloads, HBM features an ultra-wide interface with more than 1,000 input/output channels, enabling significantly higher data transfer rates than conventional DRAM. Graphics processing units (GPUs) from NVIDIA and AMD, which are essential for AI training and inference tasks, rely heavily on HBM to manage large datasets and complex computations in a more efficient way.[3]

With the launch of new platforms and performance enhancements, HBM has been experiencing a 50–100% annual increase in DRAM content. Introduced in 2022, the HBM3 generation used in NVIDIA's H100 featured 80GB of DRAM. By 2024, the B100 was launched with HBM3E, boosting the DRAM capacity to as much as 192GB.[3] NVIDIA's recently announced Rubin platform, based on HBM4, will push these limits even further, offering up to 764GB of DRAM.[2]

This represents a significant increase, marking a six- to tenfold rise in memory capacity over four to five years. Before the advent of HBM3E, memory companies were responsible for manufacturing all parts of the HBM, including the base die. However, with the launch of HBM4, where logic and memory chips start to merge, foundries will begin to produce the base die. Given the need to incorporate custom functions based on each customer's requirements, collaborations between foundries and memory companies will become essential. For instance, TSMC has been collaborating with multiple memory chip makers for more than two years on the HBM used in AI applications.[4]

The broader DRAM market is driven by cost and scale, with products featuring standardized interfaces. In contrast, HBM operates in a closed-loop ecosystem due to its rapidly evolving technological and performance demands, resulting in high barriers to entry. As HBM specifications evolve so quickly, its designs remain more closed, enabling vendors and customers to adapt swiftly to new developments. Consequently, product quality and through-silicon via (TSV) yield become important factors in ensuring performance and reliability.

The HBM market is set to expand rapidly up to and including 2028, with bit growth CAGR of 64% and revenue growth CAGR of 58%. By 2028, HBM will have become a $38 billion segment, accounting for approximately half of the server DRAM market and 27.6% of the $136 billion total DRAM market (*see Exhibit 4, next page*).

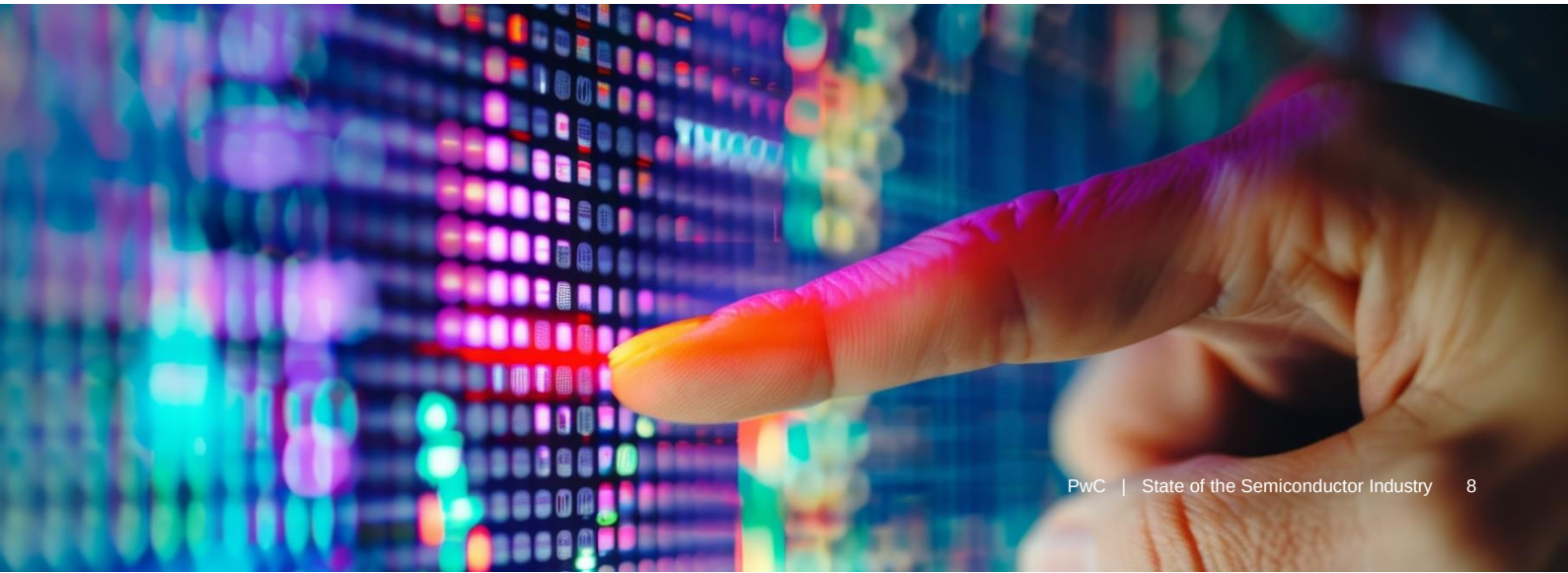
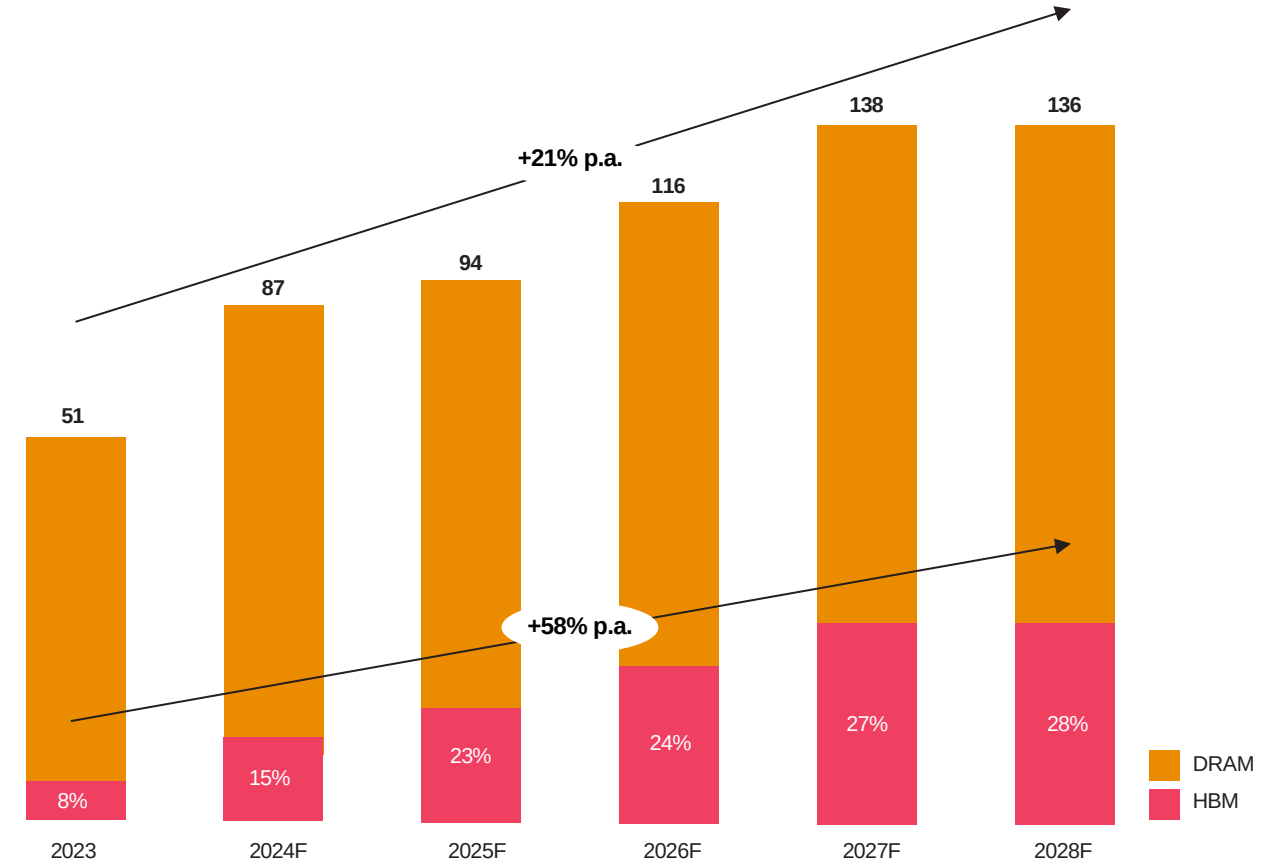


Exhibit 4

HBM bit demand (Gigabit equivalent bn), and HBM & DRAM market, 2023–2028 ($bn)



Source: Omdia Q3 2024

Rising profitability sparks a new wave of DRAM investment

With DRAM supply tightening, prices and profit margins are rising, prompting memory IC vendors to increase their capital expenditures (CAPEX). The DRAM industry saw its operating profit margin (OPM) surpass 20% in early 2024, while projections suggest it could climb to 30–40% by the end of the year. In 2025, this surge in profitability is likely to push CAPEX well beyond the $33 billion spent in 2022, possibly reaching record levels. However, most of this investment will be focused on HBM's back-end manufacturing, so DRAM production capacity in 2025 will remain close to what it was in 2022.

Government incentives are also shaping the landscape. Since 2022, subsidies and tax benefits have been introduced in the US, Japan, and South Korea, with their impact expected to become evident from 2025. Samsung's Pyeongtaek P4 fab is set to begin mass production in 2025–26[5], followed by SK Hynix's Yongin fab in 2027, and Micron's new facilities in Boise, Hiroshima, and New York from 2026 to 2029.

The shift to 3D DRAM: Overcoming the limits of 2D technology

The rate of DRAM cost improvement has slowed significantly since the launch of 10 nm-class nodes in 2017. Prior to that, costs were decreasing by 20 to 30% annually. However, between 2017 and 2023, the pace of the decrease dropped to just 6% per year, and this is expected to slow even further as manufacturers push into even lower production nodes. The diminishing returns from 2D DRAM technology at these fine line widths indicate that further cost reductions below 10nm will be minimal.

3D DRAM is expected to take over where 2D DRAM has left off, keeping the industry on its long-term cost-reduction trajectory. Starting with the second generation of under-10nm-class technology, 3D DRAM is expected to offer significant cost efficiencies. Currently, 8- and 16-layer 3D DRAM products are under development, with mass production of high-stack 3D DRAM anticipated to start around 2030.[6]

NAND flash recovery: From market bottom to AI-driven supercycle

NAND flash memory, a non-volatile storage technology, plays a critical role in devices such as solid-state drives (SSDs) due to its ability to retain data without power. With its high density, scalability, and low cost per bit, NAND has become fundamental to modern storage solutions across sectors and segments such as consumer electronics and data centers.

Over the last decade, NAND flash has experienced substantial bit growth, rising from 37.1 billion 1GB equivalents in 2013 to 744.9 billion in 2023. This surge stems largely from demand for smartphones, PCs, and servers, which rely on NAND-based storage for fast and efficient data handling. Important advancements, such as the transition from 3G to 5G LTE and faster PC processors, have fueled further growth. As AI adoption expands, the demand for high-capacity SSDs is accelerating, particularly when it comes to managing large workloads such as training and inference, which require vast storage capacities. By 2028, NAND market revenue is expected to reach $115 billion, boosted by bit growth and the increasing need for scalable, high-performance storage.[2]

> As AI adoption expands, the demand for high-capacity HBM is accelerating, particularly when it comes to managing large workloads such as training and inference."

Yoo-Shin Chang
Partner, Strategy& Korea

In the third quarter of 2023, the NAND market hit its lowest point before it then started to recover. To tackle oversupply and falling prices, vendors cut wafer production, leading to bit reduction—a deliberate decrease in the number of NAND flash memory bits produced. This strategy helped to stabilize average selling prices (ASPs) by the end of 2023, as reduced supply allowed demand and pricing to rebalance. In 2024, NAND flash bit supply remains constrained due to a 13% decrease in capital expenditure, limiting production expansion.[2]

With the market set to recover by 2025, vendors have restarted idle fab capacity to meet the anticipated demand spurred by AI-capable devices. Indeed, the expected recovery in 2025 could spell the start of a potential supercycle for NAND, with sustained, robust growth expected from that point.

AI-enabled smartphones will continue to stimulate NAND demand, as they require increasingly advanced storage solutions. Smartphones are projected to remain the largest segment of the NAND market, representing 5.4% of the global semiconductor market by 2028, with a CAGR of 30% (*see Exhibit 5, next page*). The data center segment is expected to grow rapidly, reaching 3.5% of the global semiconductor market by 2028, with a higher CAGR of 33.4% due to large-scale AI workloads and expanding storage infrastructure. PCs are forecast to capture 2.1% of the market, as AI applications create the need for more advanced storage capabilities. Although automotive memory currently represents only a fraction of the NAND market, it is expected to experience strong growth, with a 23% CAGR to 2028, as vehicles rely increasingly on advanced memory solutions (*see Exhibit 5, next page*).

The NAND market is expected to experience strong growth, with a CAGR of

23%

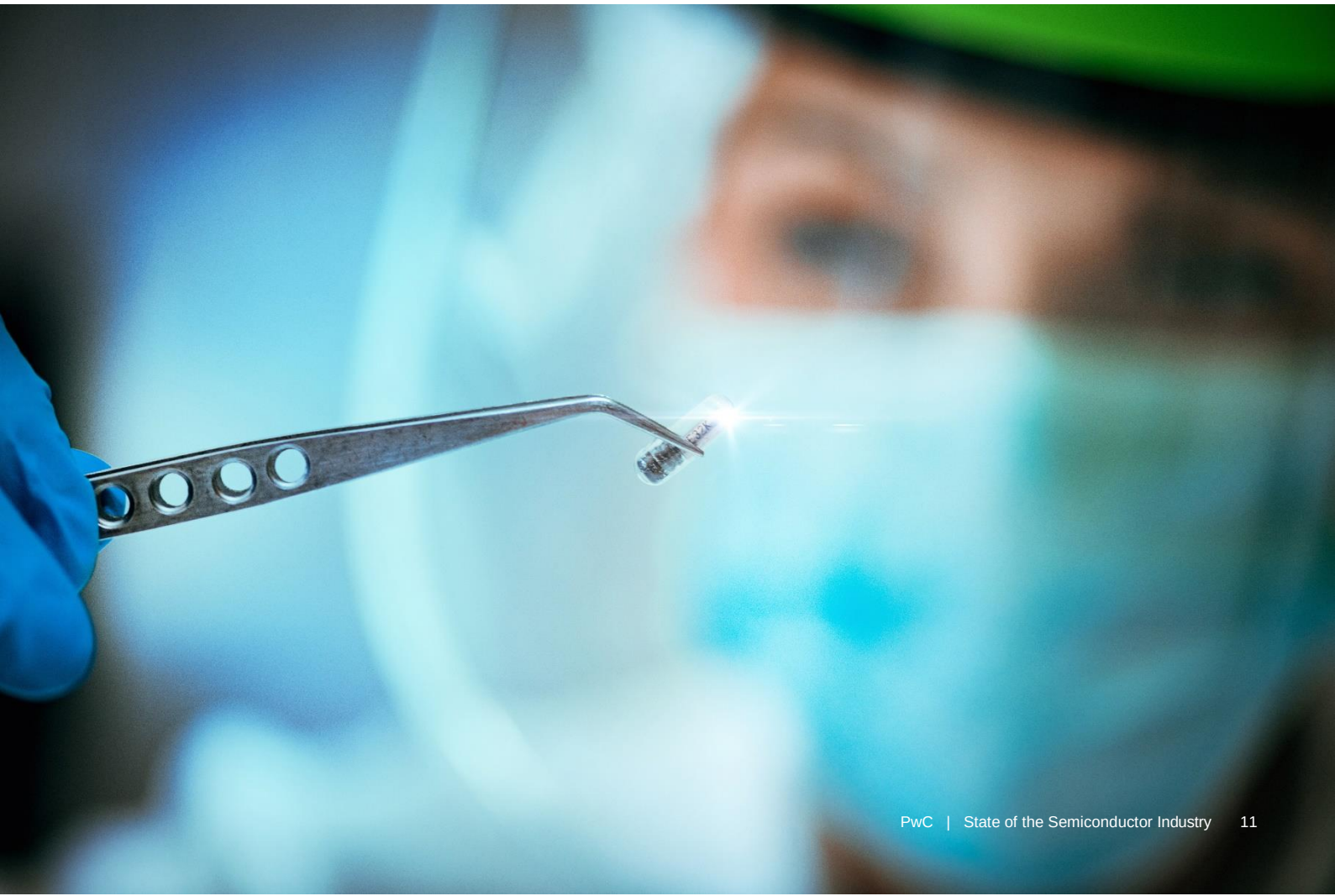
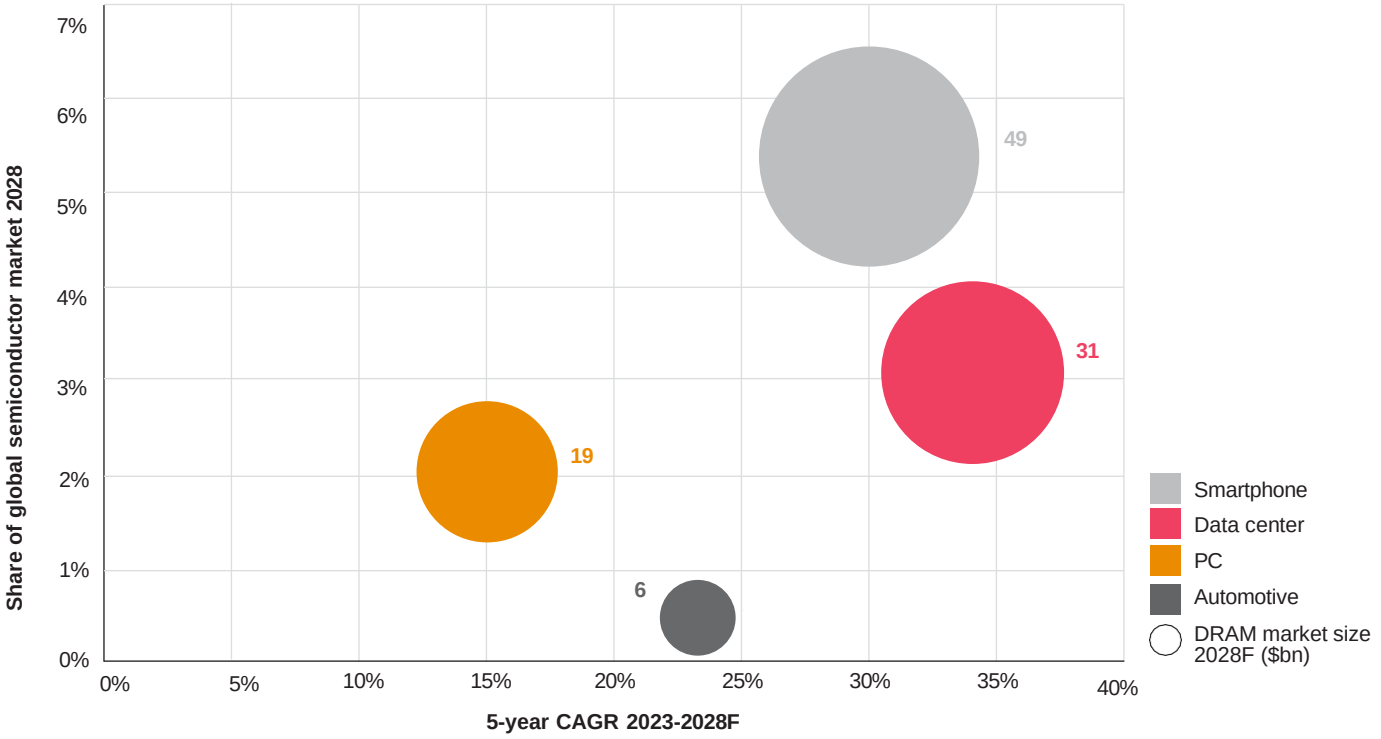


Exhibit 5

NAND market by application in share of global semiconductor market (%), and 5-year CAGR 2023–2028 (%)



Source: Omdia Q3 2024

Scaling NAND storage: The shift to QLC and 1,000-layer technology

NAND memory cells are stacked in layers within a chip, with more layers enabling higher storage capacity. In 2024, most NAND vendors are set to surpass 200 layers, although as the number of layers increases, technical challenges arise. New technologies will be introduced to facilitate further advances, with developments underway for 1,000-layer NAND, which could shape the industry for the next decade.[2]

To meet the growing demand for high-capacity storage, driven in particular by AI workloads, NAND vendors are shifting from triple-level cell (TLC) technology, which stores three bits per memory cell, to quad-level cell (QLC) technology, which stores four bits per cell, providing greater storage density at a lower cost. By 2029, QLC is expected to account for more than 50% of the NAND market.[2] SK Hynix/Solidigm leads the way in QLC-based enterprise SSDs for data centers[7], while major vendors such as Micron are adopting charge trap flash (CTF) technology, which reduces interference between memory cells and improves scalability for client SSDs. In comparison with TLC, QLC faces trade-offs in terms of performance and endurance. However, these challenges are being confronted through techniques such as over-provisioning or hybrid solutions tailored to specific workloads.

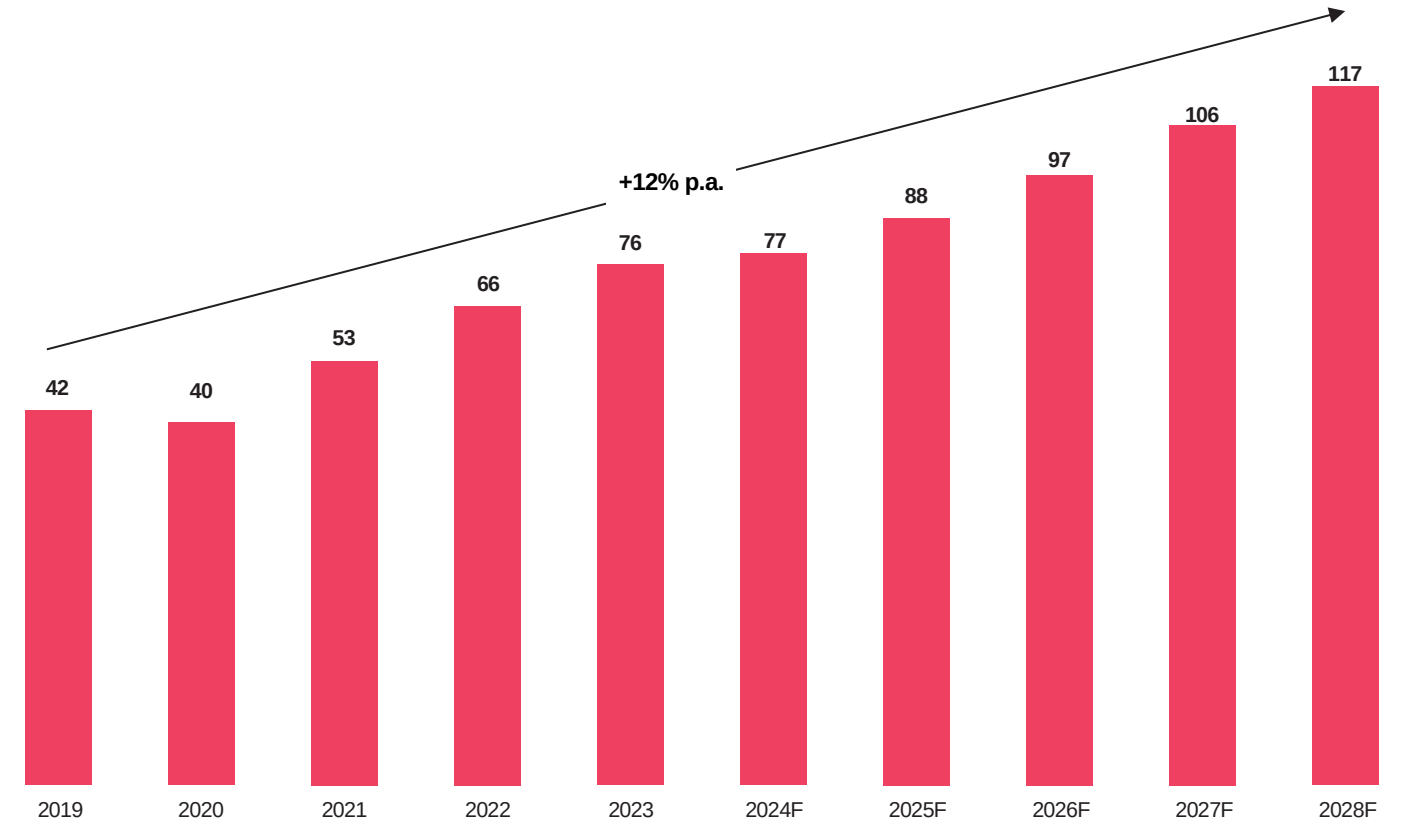
Section 3

Automotive semis shift into higher gear

The automotive semiconductor market is undergoing a substantial transformation due to the adoption of EVs and the emerging trend toward SDVs. In 2023, global vehicle production surpassed pre-pandemic levels for the first time since the onset of the COVID-19 crisis, suggesting that the automotive market has recovered and returned to normalcy. Over the next five years, we expect the market to continue to grow at 8.9% CAGR, reaching revenue of $117 billion by the end of 2028 (*see Exhibit 6*).

Exhibit 6

Automotive semiconductor market, 2019–2028 ($bn)



Source: Omdia Q3 2024

Electrification boost: Power semiconductors driving the eV revolution

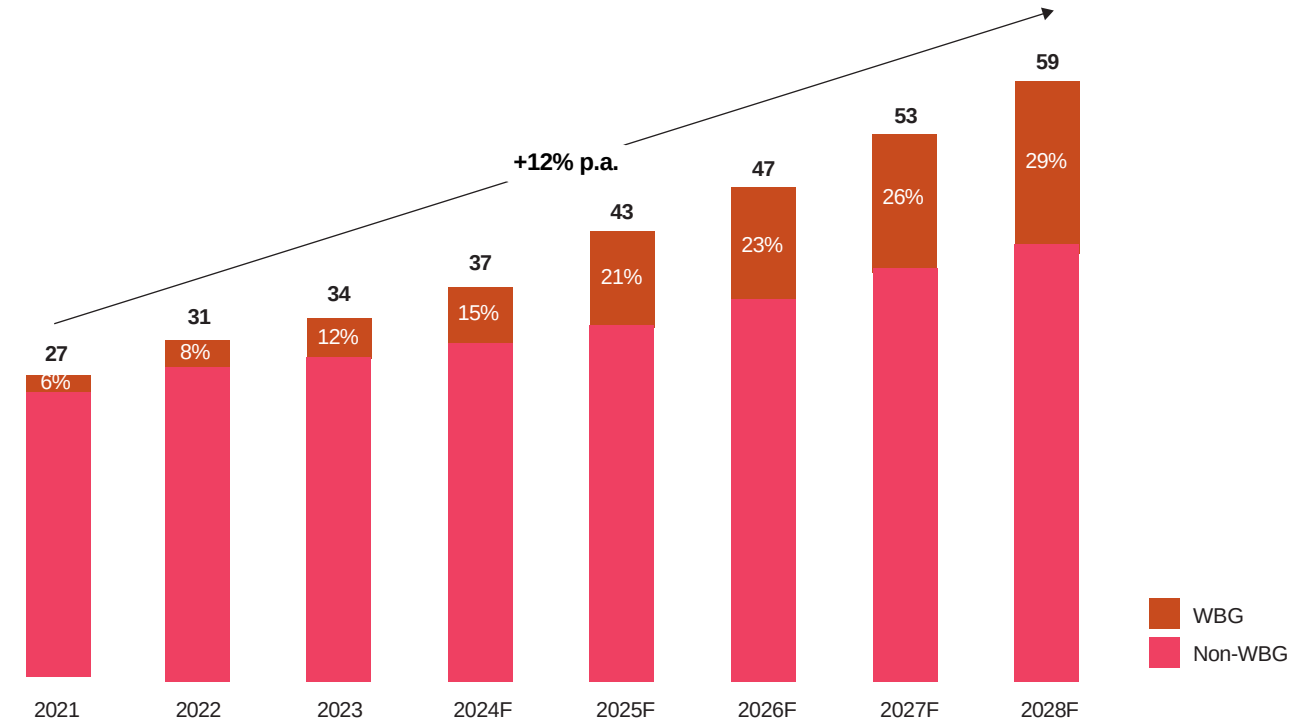
The shift toward vehicle electrification has created significant market opportunities for power semiconductors. These components are essential across various vehicle applications, and play a particularly prominent role in EV systems such as main inverters, DC-DC converters, onboard chargers (OBC), and battery management systems (BMS). According to the technology research and advisory group Omdia, the value of power semiconductors per vehicle is six times higher for EVs than it is for ICEs. As a result, sales in the automotive market reached $21 billion in 2023, reflecting a 30.2% year-on-year growth, with double-digit growth projected to continue.

The widespread adoption of EVs and advancements in battery technology are stimulating innovation in this area. Wide-bandgap (WBG) semiconductors—such as silicon carbide (SiC), gallium nitride (GaN), and emerging materials such as gallium oxide (Ga2O3)—are transforming the automotive industry. Compared to traditional silicon (Si), WBG devices offer more energy efficiency, greater power density, and the ability to operate at higher temperatures, making them ideal for key systems such as inverters, chargers, and DC-DC converters. SiC and GaN, for example, enable faster switching speeds and better heat management, resulting in smaller, lighter components—critical for improving vehicle range and performance.

Leading power semiconductor suppliers, including Infineon, STMicroelectronics, Onsemi, ROHM, and Nexperia, are increasingly focusing on WBG technology. Their share within power electronics is therefore growing significantly. While WBG accounted for 12% of the automotive power market in 2023, Omdia is forecasting annual revenue to grow to $16.8 billion by 2028, denoting an increased share of 29% (*see Exhibit 7*).

Exhibit 7

Power discrete and modules market, 2021–2028 ($bn)



Source: Omdia Q3 2024

Silicon and WBG: Competition versus coexistence?

While silicon remains dominant in lower-power, cost-sensitive applications such as compact city EVs and hybrids, wide-bandgap materials such as SiC are becoming essential for high-performance electric vehicles. SiC is particularly suited for main inverters in long-range and performance models, where power density and energy efficiency are critical. However, the higher cost of SiC remains a challenge, driven by its energy-intensive and complex crystal growth process. Moreover, SiC's hardness and brittleness increase processing difficulty, resulting in lower yield rates.

On the other hand, as major players invest heavily in new production capacities, SiC prices are expected to fall over the coming years. This trend will be reinforced by the entry of more than 50 suppliers in China alone who have moved into the SiC market in recent years to challenge established players. Intensifying competition, combined with the high capital investment required for SiC manufacturing, is likely to foster this consolidation, reshaping the competitive landscape as the industry evolves.

While the influx of new capacity raises concerns about potential oversupply and sharp price declines, the transition to 200mm SiC wafers still poses considerable challenges. The complexity of the process and the limited availability of equipment are causing long lead times and delays in capacity ramp-ups. These delays reduce the immediate risk of oversupply, although future market dynamics will depend heavily on EV demand growth, particularly in Europe and the US over the next few years.

In contrast, GaN, which is compatible with larger silicon wafers, offers significant potential for cost reduction through utilizing existing manufacturing infrastructure. This makes GaN particularly attractive for high-speed chargers and power converters. Its ability to reduce component size and weight also makes it ideal for onboard charging systems. As technology advances, GaN is poised to challenge SiC in higher-voltage applications, with ongoing research aimed at expanding its voltage capabilities and enhancing its reliability. Recent breakthroughs, such as Infineon's development of the world's first 300-millimeter GaN wafer technology, mark a major step toward lowering production costs, positioning GaN as a strong contender for the rest of the decade. The GaN market is also expected to undergo consolidation, as demonstrated by significant M&A activity, including Infineon's acquisition of GaN Systems in 2023 and Renesas's planned acquisition of Transphorm in the second half of 2024.

As the automotive industry continues to evolve, we can expect a coexistence of technologies in power electronics, based on the specific application needs of different vehicle types. In summary, the choice of semiconductor technology will depend heavily on the specific requirements in terms of performance, power, and cost. As these technologies evolve, they will complement each other in a range of applications, pushing the boundaries of what is possible in electric vehicle design and efficiency.

The changing landscape within the power semiconductor market

The power semiconductor value chain comprises several critical stages, each contributing to the performance and cost structure of the final product. The value chain begins with substrate creation and crystal growth, where raw materials such as Si or SiC are processed into high-quality wafers, creating the foundation for semiconductor devices. Next, during wafer processing, the semiconductor structures are fabricated on these wafers, imparting the essential electrical properties. The wafers are then diced into individual dies, which can be used as discrete components or assembled into power modules. The final stages involve packaging, where the devices are enclosed in protective structures to ensure optimal thermal management, durability, and performance. Packaging is especially important for power modules, which combine multiple dies to handle higher levels of power.

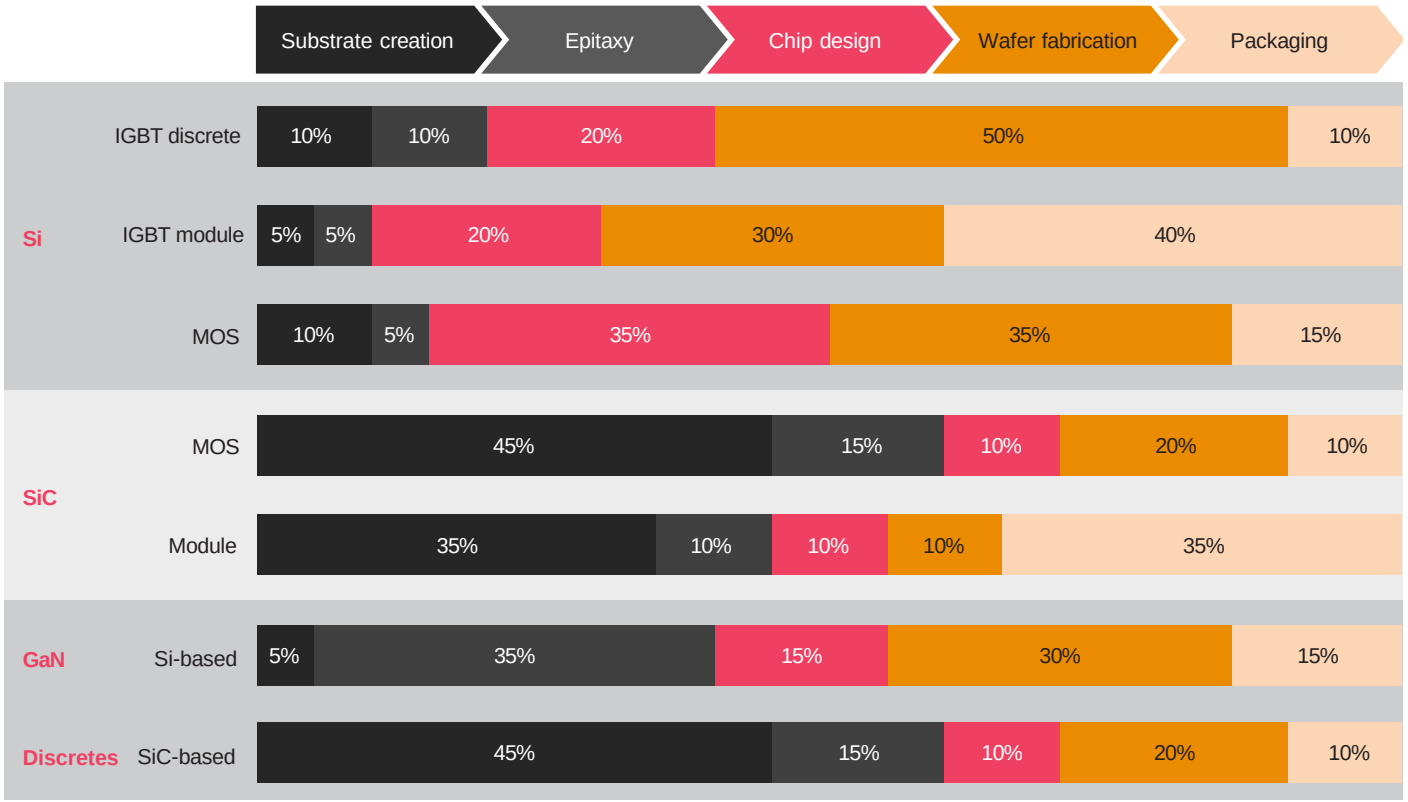
In EV power electronics, manufacturers utilize both discrete and power module solutions for the main inverters, each designed for specific power and efficiency requirements. For SiC-based power modules and discrete solutions, the substrate creation and crystal growth stages are particularly critical, contributing 35–45% of the total added value to the power package (*see Exhibit 8, next page*). This reflects the complexity of producing high-quality SiC substrates, which are vital in producing the superior efficiency and performance that SiC offers in high-voltage automotive applications. As SiC technology advances and production scales up, the value attributed to these early stages is expected to decrease. Enhanced manufacturing processes and economies of scale are likely to reduce costs, making SiC-based solutions more competitive and encouraging broader adoption in electric vehicles.

In power module solutions, the packaging stage—which involves module design and assembly—accounts for 35–40% of the total added value (*see Exhibit 8, next page*). This step is essential for ensuring the module's ability to withstand high thermal and mechanical stresses, thus enhancing both reliability and performance. Having recognized the increasing importance of packaging, semiconductor manufacturers are expanding their capabilities to develop their own power packages. In this way, they can extend control over the value chain beyond chip design and fabrication. This vertical integration allows them to improve the differentiation of their products, optimize performance, and boost competitiveness. As players along the value chain pursue these strategies, the competitive landscape will continue to evolve.

> "The trend toward SDVs reflects the increasingly widespread implementation of vehicle functions through software that can be continuously updated and improved. As hardware and software are largely decoupled, this development enables more customization, greater flexibility for consumers, and faster innovation cycles."
>
> **Tanjeff Schadt**
> Partner Strategy& Germany

Exhibit 8

Typical value contribution per value chain step for different power semiconductor types



Source: PwC analysis based on research from the beginning of 2023

From hardware to software: Software-defined vehicles (SDV) and semiconductor demand

The trend toward SDVs reflects the increasingly widespread implementation of vehicle functions through software that can be continuously updated and improved. As hardware and software are largely decoupled, this development enables more customization, greater flexibility for consumers, and faster innovation cycles. Modern electrical/electronic (E/E) architectures, heavily defined by semiconductors, create the foundation for realizing SDVs. As the industry shifts toward zone and central compute architectures, the demand for high-performance processors will rise. This transition will, in turn, reduce the need for traditional computing and power management devices that currently manage the distributed ECU systems, and will thus streamline vehicle electronics.

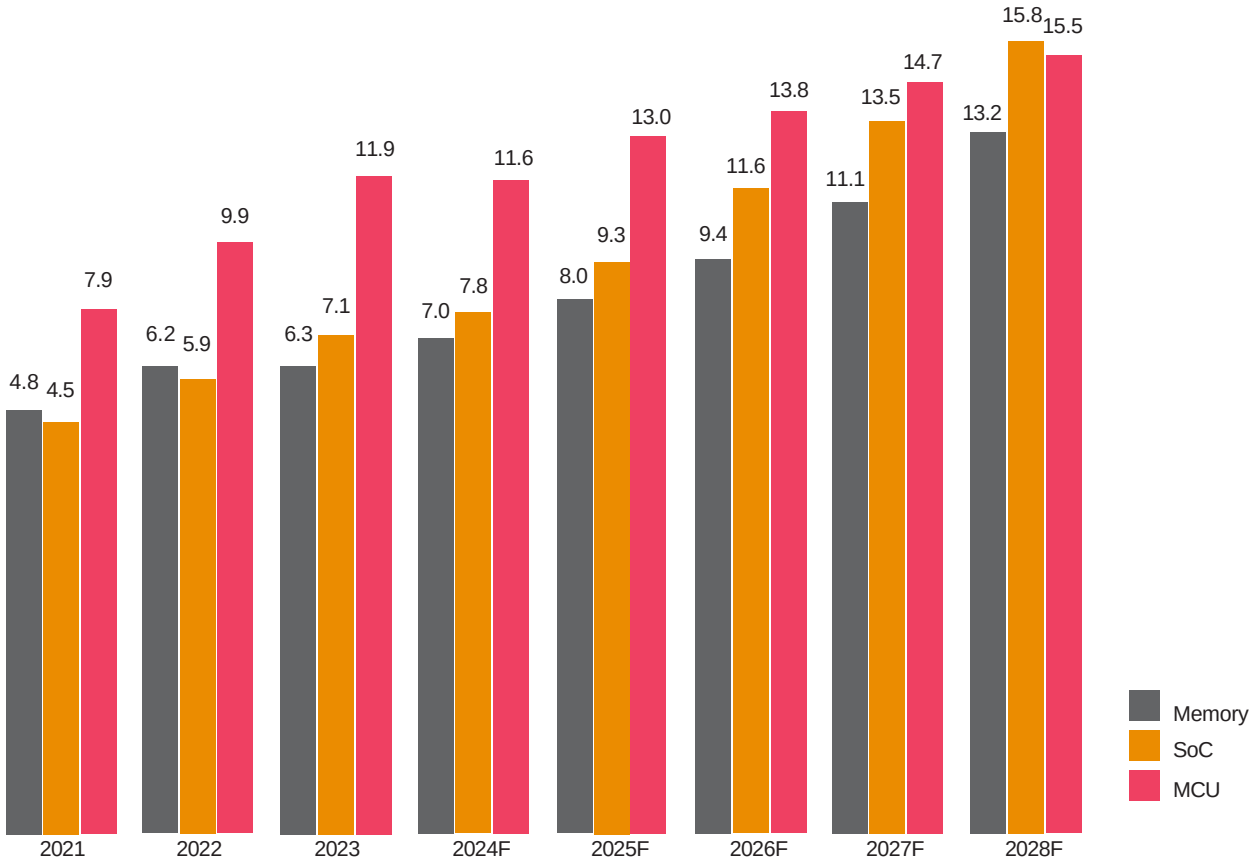
At the core of SDVs are advanced processors, commonly referred to as systems-on-chip (SoCs). These SoCs integrate various functions—such as central processing units (CPUs), memory, and peripherals—into a single chip architecture. Such integration is essential for managing the complex software tasks required by SDVs, including real-time data processing, ADAS control, security modules, and infotainment systems. Moreover, the growing need for GPUs and neural processing units (NPUs) reflects their role in powering the machine learning algorithms central to autonomous driving and other advanced functionalities. The automotive SoC market is expected to reach $16 billion in 2028, with a projected CAGR in the meantime of 17% (*see Exhibit 9, next page*).

While SoCs are critical for central processing, microcontrollers (MCUs) remain vital for handling specific control tasks and peripheral interfaces. MCUs specialize in real-time operations and are often deployed in systems where precise timing, power efficiency, and reliability are crucial, such as engine control, embedded sensors, and battery management systems. Modern MCUs support advanced connectivity options, including Ethernet, Wi-Fi, Bluetooth, and vehicle-to-everything (V2X) communications that rely on real-time data exchange between vehicles and external networks.

Memory solutions are critical for storing and accessing the vast amounts of data generated by SDVs. These components support the high-speed data access and storage requirements of complex vehicle software systems. High-capacity, high-speed memory allows vehicles to store large volumes of sensor data and software applications, enabling real-time processing and decision making. In 2023, memory chips accounted for 8% of the overall automotive semiconductor market, a share set to increase to 11% in 2028 when the relevant revenue is projected to reach $13 billion (*see Exhibit 9*).

Exhibit 9

Automotive SoC, MCU, and memory forecast by component type ($bn)



Source: PwC analysis based on research from the beginning of 2023

Section 4

Adapting to decoupling: Strategies for resilience

As global political landscapes shift, the semiconductor industry has become a focal point of geopolitical tensions. Export control restrictions and local content requirements are challenging traditional global supply chains. For tech companies, building resilience against these disruptions has become increasingly essential. Ensuring a stable supply chain, mitigating risks, and maintaining competitiveness in this volatile environment are now critical factors for success.

Technological and economical decoupling of US and China

In recent years, the US and China have been gradually moving towards a de facto technological and economic decoupling. Such a trend has profound implications for the semiconductor industry, which relies heavily on both nations for various stages of its supply chain. As a result, we are seeing the emergence of both US-centric and China-centric technospheres that focus on digital and connectivity technologies. In the Global South, Chinese tech corporations are pursuing many large-scale digital infrastructure projects.

Chinese carmakers are already being urged by the government to increase their local chip content by the upcoming year to

25%

Relevant legislative measures by the US include the Export Control Reform Act (2018), the Holding Foreign Companies Accountable Act (2021), the Secure Equipment Act (2021), and the US CHIPS Act (2022). Currently, a proposed rulemaking for Section 5949 of the National Defense Authorization Act (NDAA) is being discussed. This would prohibit US government entities from acquiring electronics and devices that contain certain Chinese semiconductors. If passed, these prohibitions will become effective on December 23, 2027. In the past, restrictions initially imposed on the government sector have often expanded within a few years to the broader civilian market, suggesting that these semiconductor restrictions could eventually affect a wider range of industries.

At the same time, mainland China is itself building up similar regulations—in particular “Made in China 2025” (since 2015), the IT Application Innovation (ITAI) Program (since 2016), the Export Control Law (2020), the Data Security Law (2021), the Anti-Foreign Sanctions Law (2021), or the Foreign Sovereign Immunity Law (2023). Chinese carmakers are already being urged by the government to increase their local chip content to 25% in the upcoming year. Moreover, China is rapidly building up its semiconductor capabilities to reduce its dependency on the outside world.

Considering the critical role played by certain regions in global semiconductor manufacturing, heightened geopolitical tensions present a significant risk to major supply chain routes and partnerships. If tensions escalate, economic and trade pressures could escalate, potentially damaging business operations and the stability of supply chains. The extent of the impact would certainly depend on the intensity and speed of these developments, but nevertheless there would be wide-ranging implications for the semiconductor industry’s ability to maintain consistent production and distribution.

Strategies for enhancing geopolitical resilience

Geopolitical resilience in connection with semiconductors is becoming a critical success factor for tech companies serving the global market. By recognizing the geopolitical risks and adopting effective strategies, companies can enhance their resilience while still maintaining innovation and competitiveness. As the global landscape continues to evolve, staying ahead of these challenges will be key to long-term success.

Multi-source strategy

Leading players are adopting multi-fab and multi-sourcing strategies to minimize supply chain disruptions. By diversifying their manufacturing locations and suppliers, companies can reduce the risk that geopolitical events will affect their operations.

Risk-based approach

A risk-based approach can help companies to identify, assess, and mitigate risks in semiconductor components throughout their lifecycle. This approach, enriched with enhanced risk criteria for geopolitical disruptions, enables companies to confront potential issues proactively at the semiconductor, module, sub-component, or product level.

Product localization archetypes

As regulations evolve, governments could mandate local or hybrid solutions, tailored to specific markets. To navigate these changes, tech companies can adopt strategies that enhance compliance and strengthen the resilience of their technology stacks. Proactive approaches can involve regionally differentiated products, or using drop-in replacements for individual semiconductors to meet local requirements. Alternatively, companies may take a reactive stance by maintaining global products and making adjustments only when required.

Talent strategy

Successful localization efforts will demand significant talent resources. As companies strive to build regional capacities, the demand for skilled professionals will increase. A robust talent pipeline is needed to support these initiatives. According to the 2023 Strategy& study Bridging the talent gap[8], Europe's semiconductor landscape alone needs around 350,000 more professionals by 2030 to achieve the regional target set by the European Union (EU) of a 20% global market share.[9]

Section 5

The renaissance of purpose-built silicon

The market for purpose-built, application-specific ICs is set for a resurgence over the next decade, driven by increasing demands for performance, efficiency, and security. This surge in demand has expanded a value chain that includes design services, foundries, and electric design automation (EDA) tools. Companies such as GUC, Alchip, and ADTechnology offer specialized design services, while initiatives such as the Open Compute Project's chiplet marketplace are democratizing access to pre-designed processor components. Small-volume foundry services, such as those from the Fraunhofer Institute for Integrated Circuits (IIS), now allow for the production of custom ICs at a scale of tens of thousands rather than millions, rendering this technology accessible to enterprises without hyperscale capabilities. With the increased availability of intellectual property rights, the development of software tools, and the reduction in design costs, the custom IC market is set to continue growing across various industries (*see Exhibit 10*).

However, scalability remains a key challenge, especially in relation to advanced semiconductors where design costs are high. According to research by IBS, a consulting firm for the electronics industry, designing a 10nm chip costs more than $170 million, and a 5nm chip more than $500 million. The evolving landscape of the data center industry provides fertile ground for custom IC development, with the four largest data center computing users—Amazon, Meta, Microsoft, and Google—accounting for 30–40% of data center IT capital expenditure.[2] These companies operate applications with the scale and performance needs that make custom IC development a strategic priority.

Exhibit 10

Custom IC development value chain (selected players)

Design	IP & EDA	Design optimization	Foundry	Packaging and test
In-house Google Amazon Microsoft Meta **Fabless** Qualcomm MediaTek Nvidia Broadcom Marvell	**IP vendor** Arm Synopsys Imagination SST Cadence **Electronic design automation (EDA)** Synopsys Cadence Siemens	GUC Alchip Faraday Asicland Gaonchips ADTechnology CoAsia SemiFive	Intel TSMC Samsung Global Foundries SMIC	TSMC Intel Samsung ASE Amkor JCET

Source: Public available information

Custom ICs for video processing

Video processing became the first target for custom IC development due to the massive scale of video streaming demand. Google estimates that 60% of global internet traffic can be attributed to video streaming.[8] Over the last five years, the company has developed custom video encoding ICs, known as video coding units (VCUs). These ICs have allowed Google to achieve a significant reduction in the number of servers needed for YouTube. Using the VP9 video coding format, Google reported that a server equipped with 20 VCUs replaced several racks of Skylake-based servers. Despite the initial investment in custom IC design and production, Google estimates that the VCU project reduced YouTube's costs by anything between 20 and 33 times over a three-year period.[8]

Meta and Tencent have also developed their own video processors, reporting significant performance improvements. Other video streaming companies, such as NETINT, deploy ASICs to maximize performance per server, watt, and dollar (*see Exhibit 11, next page*).

Custom ICs for network and security applications

Network and security applications, which are compute-intensive and handle large-scale workloads, are another prime area for custom IC development. In a conversation with Omdia, Amazon indicated that roughly 20% of their infrastructure is dedicated to network and security processing. This state of affairs led to the development of a custom IC integrated with an Ethernet controller on a DPU, colloquially known as the Nitro card. The in-house DPU handles virtual PC data plane processing (such as encapsulation or routing), encryption, and other network functions. With each iteration, Amazon expanded the Nitro card's capabilities, adding storage control, security monitoring, system control, and analytics functions. By offloading these tasks onto custom ICs, Amazon freed up CPU cores, which they then sold as Infrastructure as a Service (IaaS) to enterprises. The success of this project led to the development of custom security chips for hardware-based root of trust and the Nitro hypervisor.

Custom ICs for AI processing

AI is among the most performance-intensive and commercially significant applications, encouraging all major cloud service providers to advance the development of custom ICs. These custom chips are designed to give providers a competitive edge by accelerating AI applications or improving efficiency. Google was the first cloud provider to develop a custom IC for AI with its tensor processing unit (TPU), initially deployed in a specialized version for AI inference. By 2024, Google is expected to deploy more than a million TPUs, primarily for AI inference tasks.[2]

Forging ahead: custom AI chips in China's silicon race

In China, custom IC development for AI has become a necessity due to ongoing GPU export sanctions. Chinese cloud providers are building custom ICs optimized for both AI inference and training to maintain competitiveness.

In 2023, Tencent began expanding the deployment of its custom ASIC – Zixiao v1 – positioning it as an alternative to the NVIDIA A10 GPU for inferencing tasks. Tencent is also deploying the Zixiao v2Pro for AI training, claiming that its performance is comparable to the NVIDIA L40 GPU, doubtless in response to US sanctions.

Huawei has also developed custom silicon for AI workloads, starting with the Ascend 910 in 2019. However, US sanctions have disrupted the supply chain, prompting Huawei to collaborate with SMIC to develop new variants. In August 2023, Huawei and iFLYTEK introduced the StarDesk AI Workstation, powered by the Ascend 910B, reportedly manufactured using SMIC's N+2 7nm process. Huawei has since deployed more than 100,000 Ascend 910Bs in cluster sizes ranging from a few hundred to 20,000 units.[2]

Exhibit 11

Exemplary players with purpose-built Ics

	Company	Chip type	Chip name	Task
Video processing	Google	Video processing unit	Argos	Video processing and encoding
	Meta	Video processing unit	Meta scalable video processor	Video processing and encoding
	Tencent	Video processing unit	Canghai	Video processing and encoding
	Netint	Video processing unit	G4/5; T400	Video processing and encoding
Network/ security	aws	Data processing unit	Nitro	Data plane processing Encryption and other network functions
	Cisco	Networking	UADP and Silicon One	Networking, switching and security
AI processing	Google	AI accelerator	TPU v4/5	AI processing (training and interference)
	Aws	AI accelerator	Tranium; Inferentia	AI processing
	Microsoft	AI accelerator	Maia	AI processing
	Meta	AI accelerator	MTIA	AI processing
	Tencent	AI accelerator	Zixiao V1/V2	Image and speech recognition
	Huawei	AI accelerator	Ascend 910B	AI workloads
	Baidu	AI accelerator	Kunlun	AI computing
	Alibaba	AI accelerator	Hanguang 800	AI inference

Source: Public available information

The next waves: Security, web services, databases, and analytics

The next wave of custom IC development is expected to target workloads in security, web services, databases, and analytics. As these applications continue to grow in scale, custom ICs will be key to optimizing computational efficiency and maximizing performance. For example, custom chips designed for database processing can reduce query response times and increase the number of users that a single server can handle. Early experimentation in this area is already underway, with companies such as Microsoft deploying field-programmable gate arrays. Similarly, the growing need for enhanced cybersecurity in a rapidly digitizing global economy will stimulate the development of custom ICs for security applications. The sovereign cloud movement has already established security requirements that will shape the design of these chips.

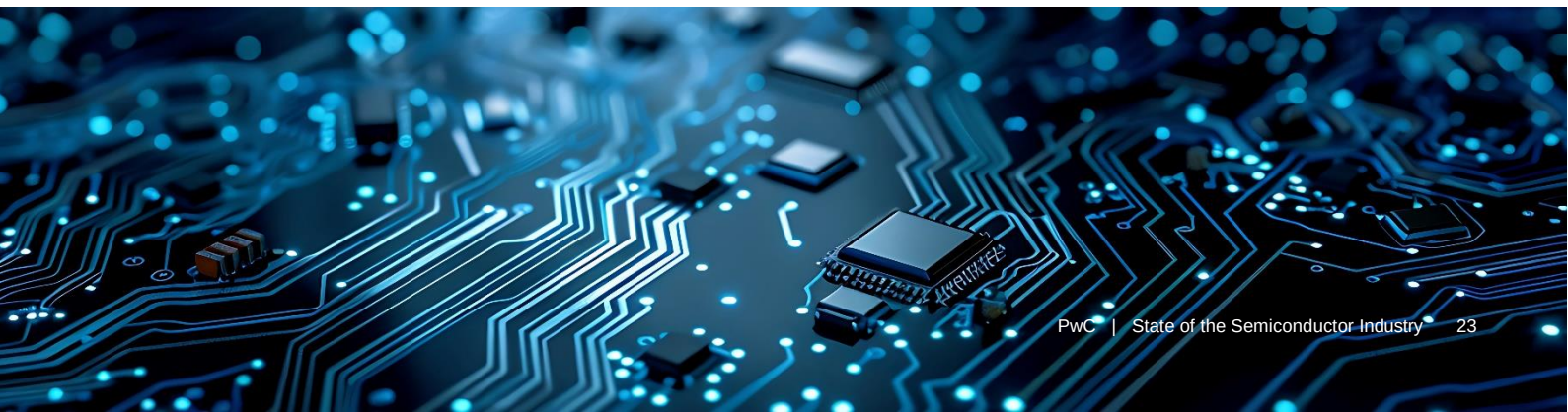
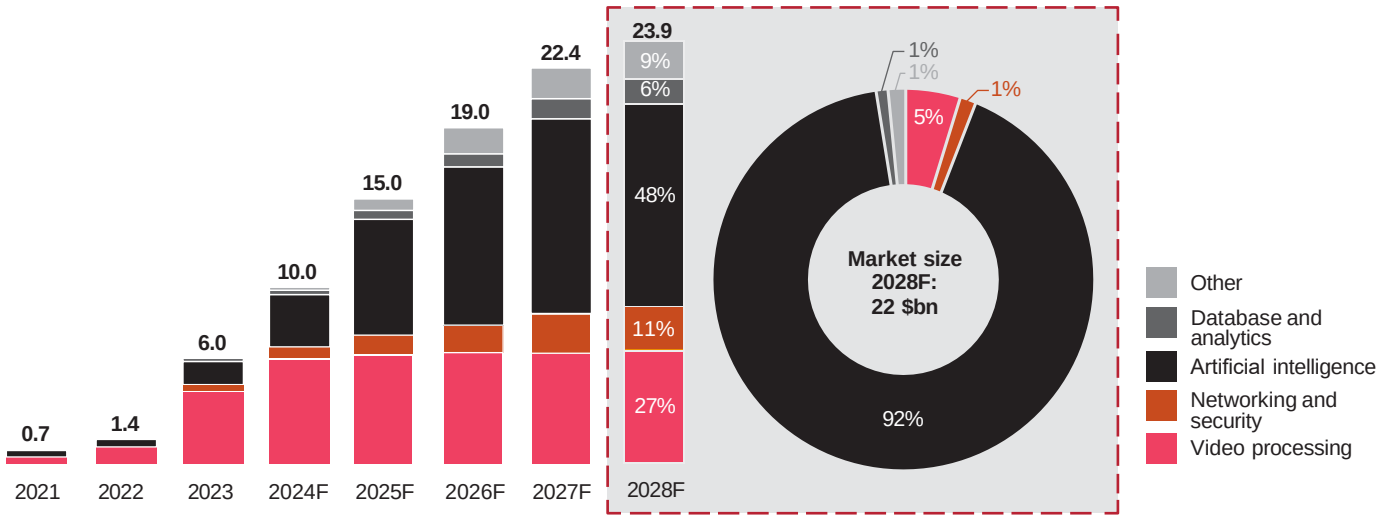


Exhibit 12

Custom ICs in the data center industry 2021–2028 (million units) and market size in 2028 ($bn)



Source: Omdia Q3 2024

Impact on the computing market landscape

Omdia predicts that these trends will create annual demand for nearly 25 million custom units by 2028. On average, servers designed to harness the processing power of these chips will be configured with more than ten per system. As a result, annual deployments of servers accelerated by these specialized processors are expected to surpass two million units by 2028. In 2023 and 2024, video processing chips will be the most widely deployed, with AI investments only beginning to garner significant results. Video processing chips are also significantly less expensive than those optimized for AI, which are packed with high-cost HBM memory and numerous computing cores (*see Exhibit 12*).

In dollar terms, the annual custom IC opportunity in the data center market is expected to reach nearly $24 billion by 2028 (*see Exhibit 12*), offering significant potential for semiconductor designers and manufacturers who were unable to compete with NVIDIA in the GPU space. Companies such as Broadcom, Marvell, and Intel have aligned their strategies to capitalize on the growing demand for custom ICs in the data center sector.

In dollar terms, the annual custom IC opportunity in the data center market is expected to reach nearly

$24bn

While the largest data center operators were the first to develop purpose-built, application-specific ICs in-house, this trend is broadening out to industries such as automotive and healthcare. In the automotive industry, many players have optimized off-the-shelf processors through partnerships with vendors such as NVIDIA, Intel, and Huawei. Tesla, for example, designs its own processors for neural network computations in autonomous driving, incorporating redundancy and safety features.[10] Denso creates custom processors for automotive applications and supplies them to automakers throughout the world. BYD has also integrated semiconductor manufacturing internally, focusing on MCUs for battery management systems (BMS), powertrain control, and real-time sensor data processing.

As other players move from optimizing off-the-shelf processors to developing own designs, the value of the automotive custom IC market will grow exponentially. We expect a significant share of the $15.8 billion automotive SoC revenue in 2028 to consist of purpose-built, application-specific ICs.

How to decide: Customization or build-up of own capabilities?

Deciding whether to pursue custom IC development depends on a company's strategic needs, operational scale, and desired level of control over technology. Companies must then assess their internal capabilities, including expertise and resources, to determine if they can manage the complexities of in-house development or whether collaborating with external partners is the more appropriate route. For businesses seeking to reduce risk while still benefiting from customization, partnerships with fabless semiconductor companies or integrated device manufacturers (IDMs) provide access to tailored solutions. To meet precise technical and performance specifications, these collaborations can vary from fully customized services to a co-design.

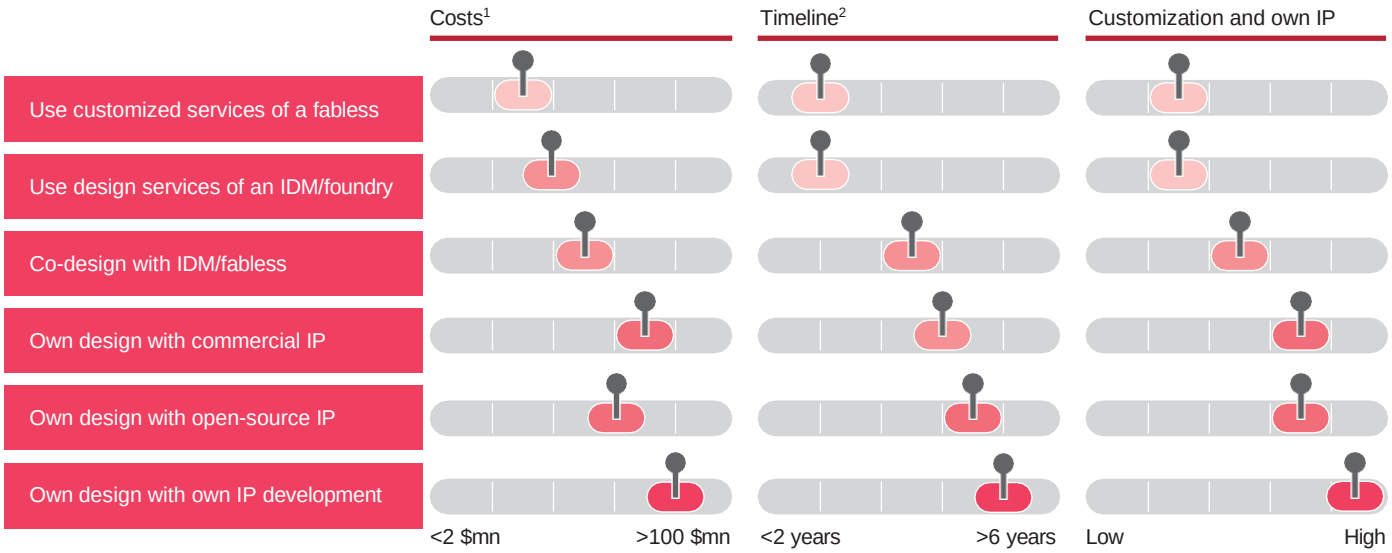
For companies with greater in-house capabilities, developing custom ICs in-house offers significant benefits. The approach enables optimization for specific objectives, such as reducing power consumption or increasing performance, allowing for deeper customization to meet specific use cases. Owning the design process also provides enhanced control over IP security, making it easier to safeguard proprietary technology. Moreover, in-house development fosters innovation and generates new business opportunities, such as licensing IP or developing niche products.

Increased customization also enables companies to tightly integrate hardware with software, making products more adaptable to specific use cases. This synergy between hardware and software is leading to a fundamental shift, where traditional hardware manufacturers are increasingly becoming software-driven companies. Such a transition not only enhances the product offering but also creates new revenue models and builds competitive advantage in the market.

> "This synergy between hardware and software is leading to a fundamental shift, where traditional hardware manufacturers are increasingly becoming software-driven companies."
>
> **Tom Archer**
> Global Technology Consulting and Alliances Leader, PwC US

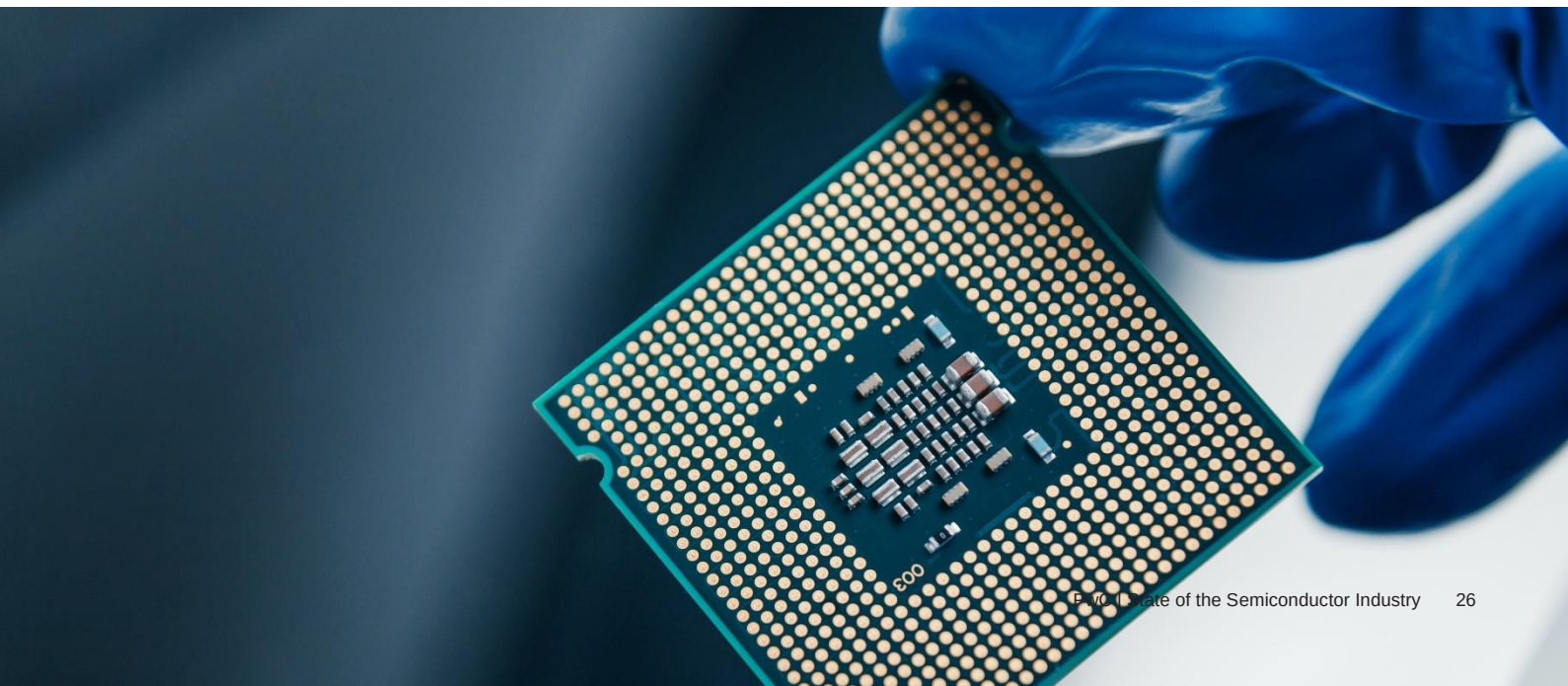
Exhibit 13

Assessment of ways to play for custom IC development



1 Detailed costs depending on semiconductor technology (node size) as well as product type and suppliers
2 Including build-up phase of required capabilities
Source: PwC analysis

Ultimately, the decision about which avenue to pursue depends on a company's objectives, application complexity, and the balance between customization and speed to market. Using commercial or open-source IP may shorten development timelines while still retaining the requisite flexibility. For smaller or mid-sized companies, partnering with established players offers an efficient way to enter the custom IC market. On the other hand, larger firms may benefit from full ownership of the IP, enabling them to innovate and optimize solutions for their unique use cases (*see Exhibit 13*).



Artificial Intelligence – from scale to diversity

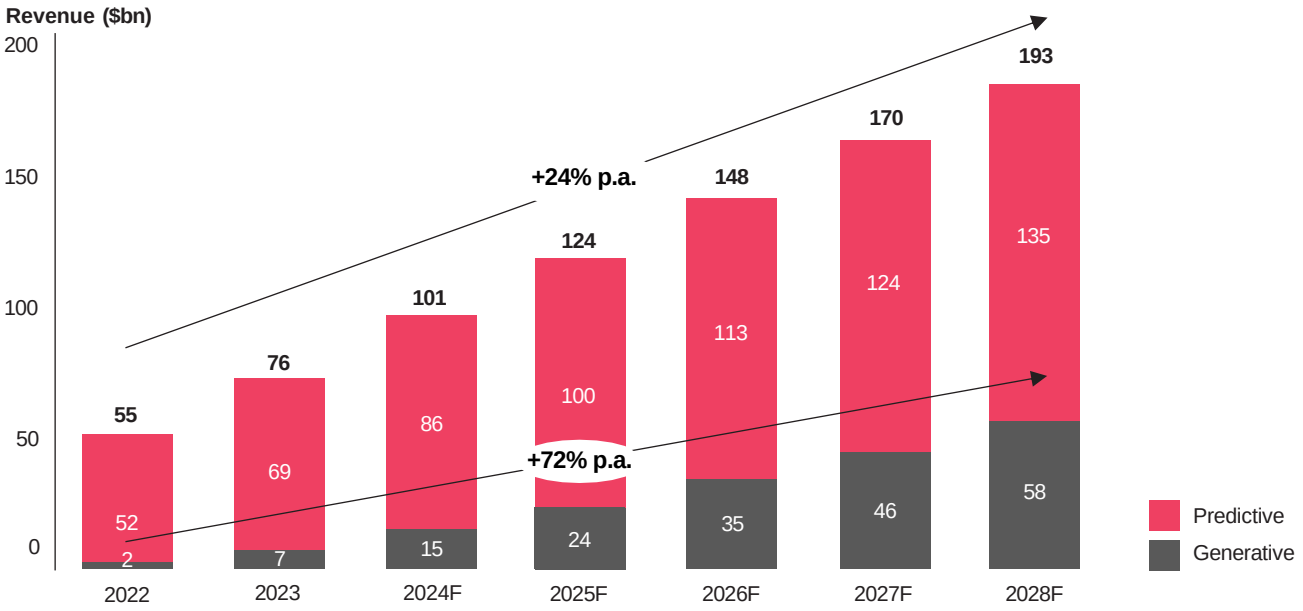
AI is enjoying rapid adoption and having a profound impact on industries and economies worldwide. For the semiconductor industry specifically, AI presents two major opportunities. First, AI-assisted chip design and optimization of manufacturing processes promise greater efficiency, fewer errors, and a faster time to market. Second, the explosive growth of AI applications, particularly predictive and generative AI, is generating a significant surge in demand for advanced semiconductor components, representing a key growth driver for the years ahead.

Predictive AI applications, such as automated quality inspection and supply chain optimization, are projected to grow significantly, reaching annual revenue of $135 billion by 2028 (*see Exhibit 14*). These systems make use of existing data to forecast outcomes and enhance business processes.

In contrast, generative AI is designed to create new content based on learned data patterns. It powers applications that generate text, images, music, or other outputs by learning from existing datasets. These types of AI systems are progressing from early adoption to mass market uptake and are expected to account for revenue of $58 billion in 2028, growing at a CAGR of 54% from 2023 onwards (*see Exhibit 14*).

Exhibit 14

Predictive and generative AI software market, 2022–2028 ($bn)



Source: Omdia Q3 2024

The race for scale in AI, which peaked between 2019 and 2021, has slowed, with Google's 1.6 trillion parameter Switch-C setting the record back in 2021.[11] Although GPT-4 and similar models have not surpassed this record, their typical model size is increasing. Since Meta's LLaMa model leak in February 2023[12], the proliferation of smaller AI models (five to 70 billion parameters) has surged, particularly in the open-source community. This "missing middle" category is now a hotbed for innovation, as these models are small enough for researchers and individuals to explore with high-end personal hardware.

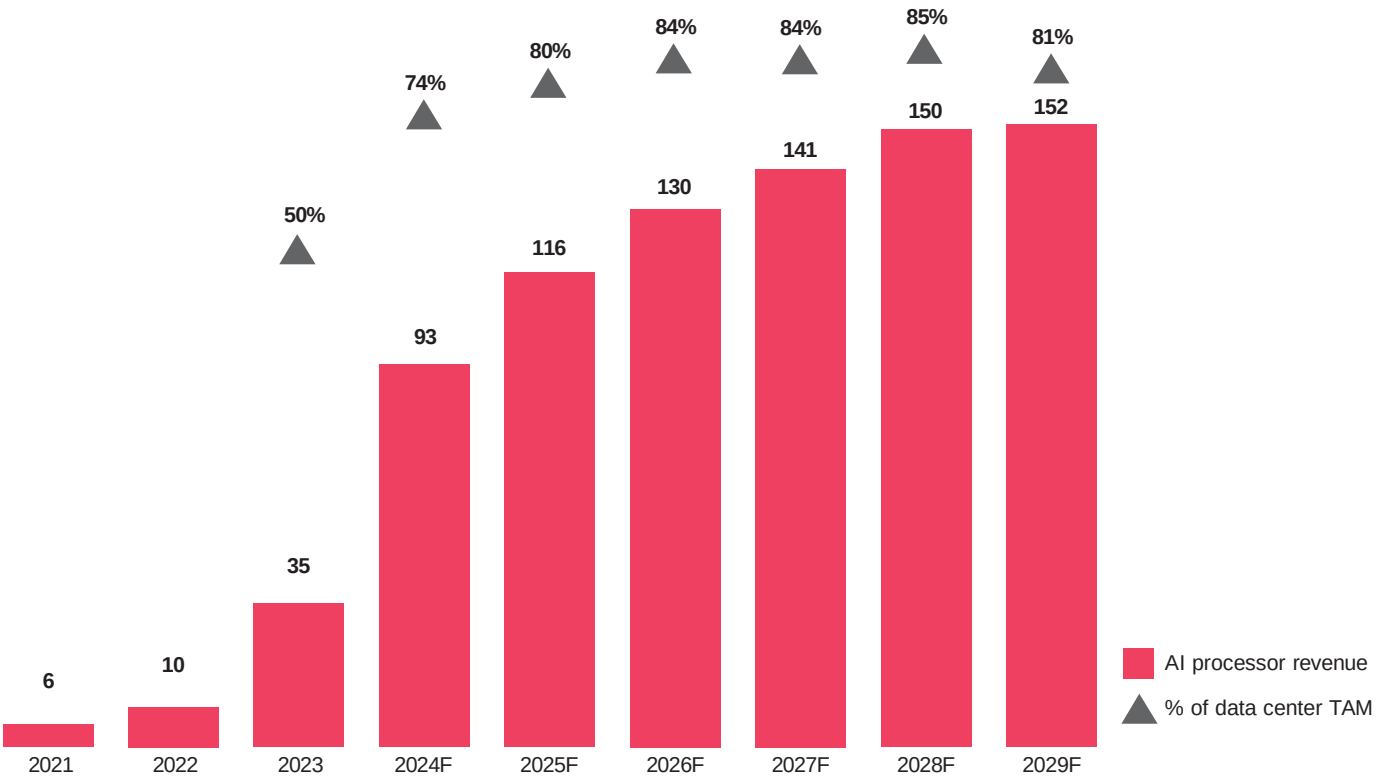
Transforming data centers with AI accelerators

AI adoption, particularly the shift from predictive to generative AI, is leading to burgeoning demand for semiconductors, specifically AI accelerators and high-bandwidth memory (HBM). Traditional machine learning models and smaller neural networks in the sub-500M weight class, such as YOLO-v5, are being replaced by larger models in the 5B–70B weight class, such as Phi, Gemma, and Mistral-7B. As AI computing demand surges, NVIDIA's data center business is expected to generate more than $80 billion this year, despite not all revenue being strictly related to compute or AI. This projection is particularly striking given that, prior to its standout 2Q23 results, the company's total revenue was less than half of that figure.

The rapid growth in demand for AI silicon, especially in data centers, has shifted the traditional relationship between CPUs and co-processors such as GPUs or domain-specific ASICs. Increasingly, CPUs are becoming the business logic co-processors for GPUs, or are acting as orchestrators for a diverse range of specialized accelerators (*see Exhibit 15*).

Exhibit 15

AI processors market for cloud and data center, 2021–2029 ($bn)



Source: Omdia Q3 2024

We estimate that by 2028/29, more than 80% of data center processors by value will either be AI accelerators or feature AI capabilities, contributing to a total addressable market of approximately $150 billion. The primary driver of this growth is not tied to a specific use case, but rather to the adoption of the Transformer model architecture. Originally designed for machine translation, Transformer models have since become the most widely used AI architecture across nearly every application. However, a critical limitation of this design is that its memory requirements scale quadratically with the size of the context window, while its inference throughput is dependent on memory bandwidth. Bandwidth, latency, and particularly predictable latency are crucial, as current AI methods require all-to-all communication at some point in the training process, slowing the operation to the speed of the slowest machine or network link.

>80%

of data center processors by value will either be AI accelerators or feature AI capabilities

Providing sufficient memory and I/O bandwidth is, counterintuitively, more energy-intensive than providing computational power (FLOPs). Memory is roughly 10 times more expensive in terms of energy than FLOPs, and I/O adds another factor of ten. As both memory capacity and fast access are essential, simply adding more DRAM is not enough; high-bandwidth memory (HBM) must be integrated directly with the accelerator. This need has given rise to flagship GPUs such as NVIDIA's B200 and AMD's MI300, which draw significant power—more than a kilowatt in the case of the B200.[2] Achieving substantial efficiency improvements, particularly at the semiconductor level, will be critical in reducing the operational costs of AI applications.

Custom accelerators are disrupting GPU dominance

The increasing power demand in data centers, coupled with the capital cost of GPUs, has spurred a wave of custom silicon projects and AI accelerator startups. This reflects the electronics industry's cyclical shift, as described by Sony CTO Tsugio Makimoto in 1991, between phases of customization (when operating at the technology frontier) and standardization (when demand is stable). AI has sparked a new phase of this "Makimoto's wave," with hyperscale cloud providers, IBM, Tesla, Huawei, and Apple introducing their own custom accelerator ASICs. Marvell's unnamed "Customer C" is also expected to ramp up its custom AI ASIC production by 2026.[13]

Thus far, custom ASICs, particularly Google's Cloud TPU, are the only ones making significant inroads into NVIDIA's dominance in the GPU market. Broadcom, which serves as the ASIC outsourcing partner for both Google's TPU and Meta's Training and Inference Accelerator (MTIA), has seen its AI-related revenue increase threefold in the past year—growing even faster than NVIDIA's data center business.[11]

> "Achieving substantial efficiency improvements, particularly at the semiconductor level, will be critical in reducing the operational costs of AI applications."
>
> **Kimihiko Uchimura**
> Partner, PwC Japan

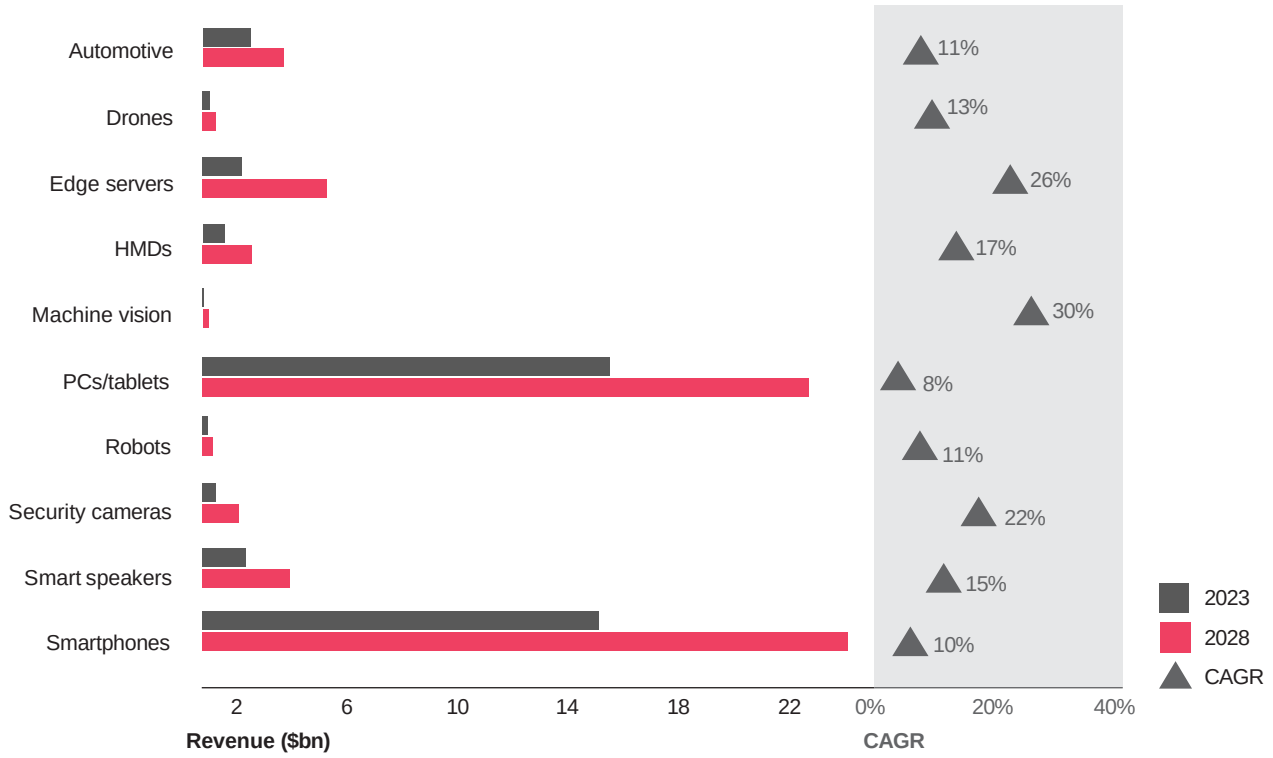
AI chip startups are struggling to make an impact, largely due to their inability to attract the software developers needed to build out the product software development kit (SDK), tools, services, and vertical solutions. A significant opportunity thus arises for foundries and ASIC partners to offer integrated services that lower the barriers to entry for customers. These services must cover not only packaging and lithography but also adjacent IP and, crucially, should extend into the software domain. Providing a robust developer toolchain is an important factor in overcoming adoption challenges.

AI at the edge: The evolution of accelerator silicon

The growth of AI applications is also being fueled by the spread of AI accelerator silicon into edge and client computing. Smartphones were early adopters, with Apple and Qualcomm integrating dedicated AI accelerator cores into their system-on-chips as early as 2017. By 2023, approximately 66% of smartphones featured some form of AI acceleration, with penetration now so widespread that it is even appearing in devices priced below $120. Companies such as Qualcomm are now capable of running 7–10 billion parameter models on mobile devices.[14] PCs, however, have lagged behind. Prior to the launch of Intel's Meteor Lake CPUs, AI acceleration on PCs was limited to power-hungry, GPU-based gaming rigs or Apple Silicon Macs, which possessed built-in AI capabilities from the start. Since then, Intel, AMD, and Qualcomm have all introduced PC CPUs with AI acceleration. These processors have increasingly resembled the system-on-chip designs found in smartphones (*see Exhibit 16*).

Exhibit 16

AI processors market for the edge data centers by device type, 2023–2028 ($bn)

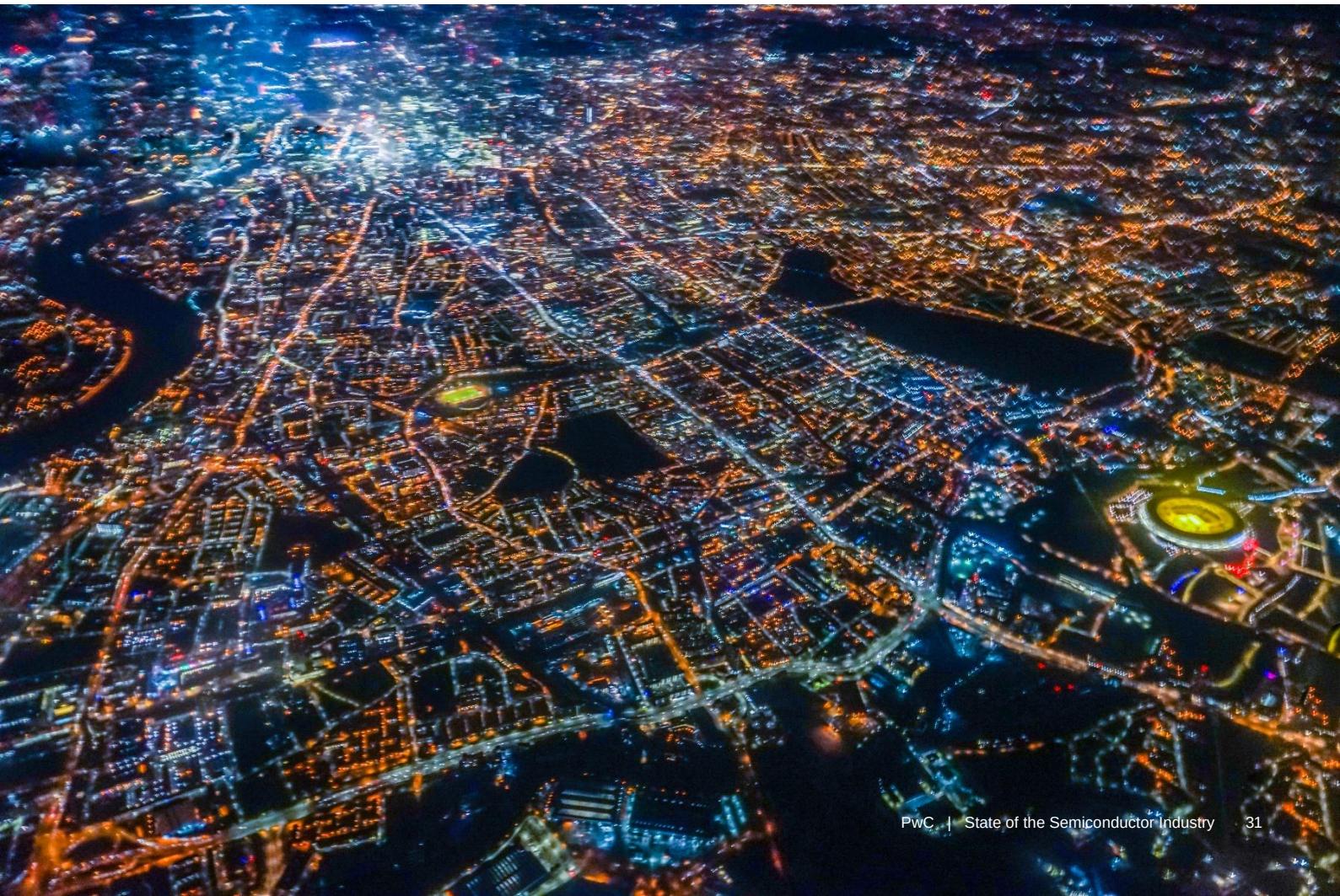


Source: Omdia Q3 2024

What is next: overcoming barriers and rethinking AI post-transformer?

A key question for the coming years is whether more players can adopt custom silicon and what the minimum investment threshold will be. While a figure of $50 million in non-recurring expenses to reach tapeout in around twelve months is commonly cited, the true challenge may lie in developing and maintaining the requisite software tools to support the chip. This has certainly proven to be a major barrier for AI chip startups.

Another pressing question is what comes after the Transformer model. Several alternatives are being explored, such as Striped Hyena and Mamba, which aim to replace the memory-intensive self-attention mechanism in Transformers with a state machine—a concept borrowed from older recurrent neural networks. Furthermore, moving to a ternary, 1.5-bit number representation could extend the use of Transformers. If a breakthrough renders AI compute-bound again, the extensive global investment in flagship GPUs could have less impact, although their massive parallel processing capabilities would still offer value. To stay competitive, industry players must remain deeply engaged with ongoing AI research and be prepared for future technological shifts.



Endnotes

1. PwC 2024: Electric Vehicle Sales Review Q2-2024
2. Omdia analysis and research Q3 2024
3. AnandTech March 2024: NVIDIA Blackwell Architecture and B200/B100 Accelerators Announced: Going Bigger With Smaller Data
4. Taipeitimes September 2024: Samsung, TSMC collaborating in HBM solutions
5. Trendforce August 2024
6. TrendForce June 2024: SK Hynix’s 5-layer 3D DRAM Yield Reportedly Hits 56.1%
7. SK Hynix press release May 2019; SOLIDIGM press release July 2023
8. PwC 2023: Bridging the talent gap
9. European Commission (2022). European Chips Act, retrieved 16th August 2023.
10. Tesla FSD chip: company website
11. William Fedus, Barret Zoph, Noam Shazeer: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity
12. Center for Security and Emerging Technology 2023: Meta’s Language Models Leak
13. TrendForce April 2024: Marvell’s AI Business Reportedly Accelerates, Potentially Benefiting TSMC
14. Forbes: Qualcomm On-Device AI Powers Future Products From Phones To PCs March 2024

About the authors

Glenn Burm is the global leader of the PwC Semiconductor sector and Strategy& Korea. He is an industry advisor in the technology, media and telecommunications and platform sectors and oversees Strategy&/PwC's global relationship with the Samsung Group. He supports clients on growth strategies, investment strategies, value creation, and digital strategies.

PwC Semiconductor Center of Excellence (CoE) is a specialized global team of semiconductor experts within the PwC Global Network, covering key regions such as South Korea, Germany, the US, Japan, and beyond. The PwC Semiconductor CoE is dedicated to delivering innovative solutions that address the challenges faced by clients across the semiconductor ecosystem

Many thanks to Omdia Semiconductor Research for their invaluable support and insightful contributions to this report

Contacts

Korea
Yoo-Shin Chang
Partner,
Strategy& Korea
yoo-shin.chang@pwc.com

Tommy Lee
Partner,
Strategy& Korea
tommy.lee@pwc.com

Tae-Young Kim
Partner,
Strategy& Korea
ty.kim@pwc.com

EMEA
Tanjeff Schadt
Partner,
Strategy& Germany
t.schadt@pwc.com

US
Tom Archer
Partner,
PwC US
thomas.archer@pwc.com

Amit Dhir
Partner,
PwC US
amit.dhir@pwc.com

Arup Chatterji
Partner,
PwC US
arup.chatterji@pwc.com

Japan
Kimihiko Uchimura
Partner,
PwC Japan
kimihiko.uchimura@pwc.com

Toshihiro Murata
Partner,
PwC Japan
toshihiro.murata@pwc.com